Identification and correction of systematic error in high-throughput sequence data

Frazer Meacham [1], Dario Boffelli[2], Joseph Dhahbi[2], David I.K. Martin[2],
Meromit Singer[3] ,[*] and Lior Pachter [1,3,4] [*]

Abstract

A feature common to all DNA sequencing technologies is the presence of base-call errors in the sequenced reads. The implications of such errors are application specific, ranging from minor informatics nuisances to major problems affecting biological inferences. Recently developed "next-gen" sequencing technologies have greatly reduced the cost of sequencing, but have been shown to be more error prone than previous technologies. Both position specific (depending on the location in the read) and sequence specific (depending on the sequence in the read) errors have been identified in Illumina and Life Technology sequencing platforms. We describe a new type of *systematic* error that manifests as statistically unlikely accumulations of errors at specific genome (or transcriptome) locations. We characterize and describe systematic errors using overlapping paired reads form high-coverage data. We show that such errors occur in approximately 1 in 1000 base pairs, and that quality scores at systematic error sites do not account for the extent of errors. We identify motifs that are frequent at systematic error sites, and describe a classifier that distinguishes heterozygous sites from systematic error. Our classifier is designed to accommodate data from experiments in which the allele frequencies at heterozygous sites are not necessarily 0.5 (such as in the case of RNA-Seq). Systematic errors can easily be mistaken for heterozygous sites in individuals, or for SNPs in population analyses. Systematic errors are particularly problematic in low coverage experiments, or in estimates of allele-specific expression from RNA-Seq data. Our characterization of systematic error has allowed us to develop a program, called SysCall, for identifying and correcting such errors. We conclude that correction of systematic errors is important to consider in the design and interpretation of high-throughput sequencing experiments.

Availability: Software for correcting systematic errors is freely available from
`http://www.cs.berkeley.edu/~meromit/SysCall`

Introduction

The technological advances that have produced "the third phase of human genomics": sequencing of individual genomes and the determination of rare variants across populations by enabling whole genome sequencing at low cost [11], are accompanied by higher error rates. Improved statistical methods that accommodate these high error rates are needed in the calling of heterozygous sites

[1] Department of Mathematics, University of California, Berkeley, CA.

[2] Children's Hospital Oakland Research Institute, Oakland, CA

[3] Department of Computer Science, University of California, Berkeley, CA.

[4] Department of Molecular and Cellular Biology, University of California, Berkeley, CA.

[*] corresponding authors. contact: meromit@cs.berkeley.edu, lpachter@math.berkeley.edu .

from low coverage data [11]. The design of effective statistical methods requires precise characterization of error in high-throughput sequence data. Previous work has examined the behavior of individual base-call errors in sequence reads [2, 10, 15]. In this paper we discuss a previously undescribed phenomenon in sequence data where these base-call errors aggregate at specific genomic locations across multiple sequence reads. We focus on Illumina technology, although we have observed systematic error on other platforms and return to this in the Discussion.

We begin by describing the types of sequencing error that have been previously characterized, and their relationship to the systematic error we have discovered. The likelihood of a base-call error occurring at any particular location in a sequence read is influenced by several parameters. It is known that base-call errors are more likely towards the ends of reads and that surrounding sequence motifs influence error frequencies [2,10,15]. For example, errors are more likely at positions preceded by GG or following a number of GGC motifs [10], but regardless of the preceding motif, errors become more likely towards the end of reads [2]. However, we have found that errors at some *genomic* positions appear with greater frequency than can be explained by these effects, and we refer to this as *systematic error*. Systematic error manifests as many individual base-call errors from separate sequence reads occurring at the same genomic position (Figure 1). Thus, a systematic error comprises many individual base-call errors (from different reads) that fall at the same genomic location.

These errors have the potential to be especially troublesome because they can confound methods that identify errors based on their sparsity among reads. For example, systematic errors affect current SNP (Single-Nucleotide Polymorphism) calling methods, where the first step involves computing the posterior probability for a SNP at every site based on relative nucleotide counts. Although filters based on the depth of reads are frequently applied (mostly to screen for indels, copy number variants, or other structural variation), most existing approaches will not identify systematic errors, or be able to distinguish them from true SNPs. Similarly, the detection of RNA editing sites in RNA-Seq data is complicated by systematic error, because an accumulation of errors at a transcriptome site can appear to be an edit event when compared with a reference genome that may have been sequenced using another technology [8].

In this paper we present a thorough characterization of systematic errors using Illumina short-read sequencing data that is optimized for the detection of errors because of high coverage. We show that systematic errors must be accounted for when annotating heterozygous alleles, and that although improved base calling software can correct a small number of systematic errors, it is not sufficient by itself. We present an efficient statistical algorithm for the detection of systematic error and use it to show that systematic errors are present in other datasets, including an RNA-Seq dataset, a viral reference genome and new Illumina HiSeq 2000 data from the 1000 genomes project.

Results and Discussion

To investigate the types of errors present in whole-genome Illumina high throughput sequencing data, we conducted a paired-end methyl-Seq experiment on a male human individual with read length of 76 bp (Methods). A methyl-Seq experiment is ideal for investigating systematic error because the experiment results in high average coverage due to the fact that only sites cut by the restriction enzyme are assayed. The reads were mapped with Bowtie [6] allowing up to two mismatches. Our experiment spanned 29,827,077 genomic locations at an average coverage of 35.4. Due to the small fragment size in methyl-Seq experiments many of the mate-pair reads overlapped, providing for each such location two base calls sequenced from the same DNA molecule (Figure 1) albeit from different directions. We made use of this to distinguish between base-call errors and true heterozygosity calls in the following manner: each pair of bases originating from a single mate-pair and sequencing the same position was denoted a *reference-pair* if both calls agreed with the reference genome, a *SNP-pair* if both calls disagreed with the reference genome and agreed among themselves, and an *error-pair* if one of the calls agreed with the reference genome but the other did not. A *SNP-pair* could consist of two base-call errors, in the case that both of the paired reads had an error at the same location, but the probability of such an event was low and we ignored such cases in this study.

Because we focused on overlapping mate-pairs, we report all results in terms of pairs. For example, when stating coverage we state the number of pairs overlapping a site (the coverage of the systematic error location in Figure 1 is 11), and when we state a location has 40% errors it means that of the pairs overlapping the location 40% were *error-pairs*. In our experiment 3,985,926 genomic locations were covered by both reads of some mate-pair but we restricted our analysis to the 2,226,445 of these locations with a coverage depth of at least 10. These 2,226,445 genomic locations where covered by a total of 85,782,923 base-call pairs, 223,957 of which were error-pairs.

Extent of Systematic Error

We found many locations at which there seemed to be an accumulation of errors. To test the extent of this phenomenon we computed the expected number of locations with each possible proportion of error. Let $c_{10}, \ldots, c_j, \ldots, c_{565}$ be the number of locations with j coverage in our data ($\sum c_j = 2,226,445$), and $p := \frac{\#error-pairs}{\#pairs} = 0.002611$ be the probability of sequencing error. Let B_i be a random variable for the number of locations from $c_{10}, \ldots, c_j, \ldots, c_{565}$ with proportion of errors i, and let B_{ij} be a random variable for the number of locations with coverage j and proportion of error i. We computed the expected number of locations to have each proportion of errors i as

$$E[B_i] = \sum_j E[B_{ij}] = \sum_j c_j \binom{j}{k_{ij}} p^{k_{ij}} (1-p)^{(j-k_{ij})},$$

where k_{ij} is the number of errors for coverage j that results in proportion of error i. Figure 2 shows a log-scale histogram of the expected and observed counts for these different error-proportions. The observed counts in the higher frequencies of errors are larger than the expected counts, indicating that there are more locations than expected that have a high frequency of base-call errors. We called such locations systematic errors, and set out to determine the characteristics of these locations, with the goal of lowering the false-positive rates in calling heterozygous sites.

For further characterization, we annotated a set of locations in which the number of *error-pairs* was significantly higher than expected, given the observed frequency of error. Setting $p = 0.002611$ as in the previous section, we compute a p-value for a given location with i errors and n coverage as $p(K \geq i|n) = \sum_{k=i}^{n} \binom{n}{k} p^k (1-p)^{(n-k)}$, where K is a random variable indicating the number of errors at a location. Of the 2,226,445 locations with coverage of at least 10, 2,116 locations were annotated as systematic errors, using a Bonferroni correction for a 0.05 significance level. We used a Bonferroni correction because it ensures that the probability of even one false-positive is ≤ 0.05, resulting in a set that is low in false-positives, and therefore suitable for characterizing the nature of systematic error. We note that this calculation yielded a lower bound on the frequency of systematic errors in our dataset of approximately 1 in 1000 bp.

Characterizing Systematic Errors

Having annotated the set of 2,116 systematic errors, we looked for characteristic features that could be identified in any high throughput sequencing experiment. Of the 2,116 sites we have determined as systematic errors, 953 had all base-call errors on the forward read and 1,062 had all base-call errors on the reverse read (an example is seen in Figure 1). We conclude from this that in systematic errors the base-call errors tend to appear on just one of the sequencing directions (forward or reverse). This tendency was noticed in [3], where the directionality on which errors occurred was used to filter false-positives from the set of heterozygous sites annotated. A possible explanation for this phenomenon is that the sequencing of some motifs, which are different on the opposite strands, have higher probability than others for base-call errors, resulting in systematic errors. This is consistent with the known overlap in absorption spectra of the G and T channels identified by a single laser in Illumina sequencing.

We therefore tested whether there are significant motifs surrounding systematic errors by generating a sequence logo [1, 14] for the reference sequences around the systematic errors (Figure 3). Interestingly, we found that the first base upstream of the systematic error has greater information regarding the presence of a systematic error than the base at which the error is present. We found that the large majority of systematic errors are preceded by a G, and that two G bases followed by a T at the error site is by far the most common and characteristic sequence at systematic error locations. We computed the error rate in our dataset for such GGT sites: $p_{GGT} := \frac{\#error-pairs\ at\ GGT}{\#all\ pairs\ at\ GGT} = 0.0194$. Observing that this error rate is much higher than that for the whole dataset, we tested whether there are GGT locations at which there is a significant excess of errors by conducting the same test for significance as described before but considering only GGT locations and using $p_{GGT} = 0.0194$ rather than the previously used $p = 0.002611$. The number of significant locations (at GGT sites) remained substantial at 660 (out of 61,779 GGT sites considered). This shows that although the motif GGT is a strong characteristic of systematic errors, systematic errors are not accounted for by this motif alone.

To gain insight into the types of sequencing errors present at systematic errors we computed the frequencies of the different base substitutions in both systematic errors and throughout the entire dataset (Figure 4). We witnessed an extremely strong tendency for the $T > G$ error compared to all others. Our results show that there is a higher substitution rate to Gs than to the other nucleotides and that the substitution rate to A or T is considerably lower than the substitution rate to C. With respect to the reference bases at which systematic errors occur, there is a stronger tendency of error at A or T than at C or G. We divided the systematic error locations based on the reference base at which the error occurred, and tested for motifs in each of the fours sets (Figure 3.b). We concluded that the strongest motif at systematic errors is that of GGT where the error is at the T, resulting in an incorrect base call of G.

To test whether the quality scores at the locations of systematic errors account for the extent of base-call errors observed, it would be ideal to compute a p-value for each location given its specific quality scores. However, doing so is not computationally feasible, and we therefore calculated for each location an upper bound on its p-value under the specified model (Methods). Of the 2,226,445 positions with read count of at least 10, 119 had an upper bound on the p-value that was significant under a Bonferroni correction for a significance level of 0.05. This is particularly interesting due to our use of a crude upper bound, and to the fact that in general throughout the experiment the quality scores tend to predict a higher error rate than that observed ($\frac{\#error-pairs}{\#pairs} = 0.002611$ while the quality scores predict an error-pair frequency of 0.00416).

The characteristics of systematic errors, occurring mostly at GGT motifs where the error that occurs is a $T > G$ substitution implies that the errors could be a result of the sequencing technology which makes it hard to distinguish between a GGG and a GGT instance. It is the base-calling algorithm that makes such distinctions, given the images output from the Illumina machine. We asked whether systematic errors could be accounted for by base-callers that utilize sophisticated statistical techniques to reduce error. To test this we compared the systematic errors present in a dataset base-called by Bustard (Illumina's base-caller) to those present in the same dataset when base-called by NaiveBayesCall [5], to our knowledge the most accurate base-calling algorithm available. We used for this the dataset that was used in [5]: 74,686 non-paired reads sequenced from the phiX174 virus. We mapped the sequenced reads from each method to the virus genome, obtaining 382.2x coverage for the Bustard called reads and 394.2x coverage for the NaiveBayesCall called reads. Since phiX174 is only 5,386 bp long and has been thoroughly studied for heterozygous sites due to its use as a sequencing control, we excluded the five known SNP sites from our analysis, and at the remaining sites called all base-calls that were different from the reference as base-call errors. We computed the probability of a base-call error for each dataset of mapped reads by $p = \frac{\#\ base-call\ errors}{\#\ base\ calls}$, and identified locations with a significant accumulation of errors by computing a p-value for each location as described previously and using a Bonferroni correction for a 0.05 significance level. We used the frequency of base-call errors in the Bustard called reads of 0.0029 for the error probability for both data sets, since this was the higher of the two frequencies. We found 59 systematic errors in the Bustard called data set and 40 systematic errors in the NaiveBayesCall data set, amounting to a systematic error rate of 1 in 91 bp and 1 in 135 bp respectively. We believe the higher frequency of systematic errors is due to the phiX174 genome being richer than human in GGX motifs and to the high sequencing coverage (see "Conclusions" section). When restricting to cases in which more than 10% of the base-calls had errors we found 15 systematic errors for Bustard and 10 systematic errors for NaiveBayesCall, 7 of which were at the same sites. Our results show that while systematic error can be reduced with more sophisticated base calling, it is a persistent problem at a significant level even when using state of the art methods.

To test replicability of the locations at which systematic errors occur, we conducted a second methyl-Seq experiment on the same individual (Methods). The error frequency in this second experiment was determined as $p = \frac{\#error-pairs}{\#pairs} = 0.00162$ and of the 2,419,666 locations with coverage of at least 10 pair-calls, 3,272 locations were annotated as systematic errors using a Bonferroni correction of 0.05. From the 2,160,736 positions with at least 10 pair-calls in both of the experiments, 1,916 and 2,519 were annotated as systematic errors in the first and second experiments, respectively, and of those 1,279 locations were annotated as systematic errors in both experiments. This shows that while there is some variability in the locations determined as systematic errors, locations at which systematic errors occur are highly replicable (the expected number of systematic errors to be called at the same locations under random distribution is 2). We tested whether the significant overlap of the locations at which systematic errors were detected was due to GGT motifs being more prone for systematic errors than other motifs. Of the 61,779 GGT sites that were overlapped by at least 10 pair-calls in each experiment, 1,596 and 2,080 locations were annotated as systematic errors in the first and second experiments, respectively, and of these 1,095 locations were annotated as systematic errors in both experiments (the expected number of systematic errors to be called at the same locations under random distribution and restriction to GGT positions is 54). The lists of systematic errors for both experiments is available at: `http://www.cs.berkeley.edu/~meromit/systematic/`.

Identification and Correction of Systematic Errors

The main concern regarding systematic errors is that they may be incorrectly annotated as heterozygous sites in an individual or as rare variants in a population. Fortunately, in systematic error the extent of error at a location usually does not result in an equal ratio of reference to non-matching reference calls, making it easier for methods that expect such a ratio to identify these sites as non-SNPs. Nonetheless, samtools [7] identified 12 of the 2,116 systematic errors in our methyl-Seq dataset as SNPs (three of these are annotated as SNPs in dbSNP130), and in the SNP-calling procedure for the 1000 genomes project a filtering step based on directionality of sequencing was used to account for systematic errors (supplementary material of [3]). Systematic error may pose an even greater difficulty in population studies, where allele ratios are not expected to be 1:1. This difficulty also arises in RNA-Seq experiments in which variants are annotated alongside expression levels [17]. Systematic error may also affect RNA-Seq experiments in the bias it can introduce in coverage at systematic error sites. Such bias can in turn affect expression level estimates [16].

To account for this we have designed SysCall - a classifier which given a list of potential heterozygous sites and the reads overlapping these sites classifies each location as a systematic error or a heterozygous site (Figure 5). Our classifier uses logistic regression to combine the different characteristics of systematic errors to make predictions, and does not assume that the expected frequency of variant observations is half or that the experiment preformed is paired-end, making it applicable to the different types of high throughput experiments discussed.

SysCall annotates a vector of features for a given location, l, as follows: First a sequencing direction is chosen (forward or reverse) as the direction with the larger proportion of base-calls that differ from the reference. Let q_{l1} and q_{l2} be that proportion for the chosen and not chosen directions respectively. For example, for the location annotated as a SNP in Figure 1, we would choose the forward direction and have $q_1 = 1$ and $q_2 = 0$. Let b_i be the nucleotide that is i places from l in the chosen direction and let w_i be the vector of quality scores i places from l, attained from the reads overlapping that location. A feature vector is then annotated as:

$$v_l = (b_{-2}, b_{-1}, b_0, q_{l1} - q_{l2}, q_{l1}, \mathrm{PT}(w_0, w_1)),$$

where $\mathrm{PT}(w_0, w_1)$ is the paired t-test result on the two vectors w_0 and w_1. For example, for the location annotated as a SNP in Figure 1 we would have $v = (G, G, T, 1, 1, -5.56)$. SysCall then runs logistic regression on the vector of features and classifies location l as either a heterozygous site or a systematic error. The paired t-test feature is computed due to our observation that the quality scores at systematic error locations tend to be lower relative to the quality scores at their neighboring sites (Figure 6), and this can help distinguish them from true heterozygous sites.

In order to train our classifier and test its performance we annotated a set of locations in our dataset that would be candidates for heterozygous sites (where a significant amount of the base-calls differ from the reference) and for which using the overlap between paired reads we could call as systematic errors or heterozygous sites with high certainty. Of the 905 locations in our dataset with coverage of at least 40 (paired-calls) and at which 10-90% of the base-calls on the forward strand differed from the reference we annotated two sets: (1) “SNPs” - the 491 locations at which all differences from the reference were *SNP-pairs*. (2) “Systematic errors” - the 338 locations at which all differences from the reference were *error-pairs*. From each mate-pair one of the reads was chosen at random to simulate a non-overlapping (or non paired-end) dataset.

As a first test of our classification algorithm we ran 100 iterations in which we generated training and test sets by randomly dividing the “SNPs” and “Systematic errors” sets into halves (from each the “SNPs” halves 169 instances were randomly selected in order to have the same number of systematic errors and SNPs in the training and test sets). In each iteration we generated a feature matrix for the training and test sets, learned the coefficients of the logistic regression classifier from the training set, and classified the instances of the test set, recording the percentage of instances that were classified correctly (as either systematic errors or heterozygous sites). The distribution of the percentage of instances classified correctly from the 100 iterations had a mean of 99.0% and a standard deviation of 0.005.

A strong characteristic of systematic errors is that the differences from the reference have a strong bias to occur on either the forward or reverse direction. We tested the ability to classify locations using the same logistic regression classifier but using only the directionality bias feature: $u_l = (q_{l1} - q_{l2})$. When running 100 iterations of training and testing as before using this classifier, the distribution of the percentage of instances classified correctly had a mean of 72.1% and a standard deviation of 0.021. Therefore, a significant amount of precision is gained when making use of all six features in the classification process. This is mostly due to an increase in the recall rate of the classifier, where SNPs that are annotated as systematic errors when using only the directionality bias criterion are recognized as SNPs when making use of all features.

A main purpose when designing SysCall was to be able to distinguish systematic errors from heterozygous sites in datasets of lower coverage than that available to us (35.4x). To evaluate

SysCall's performance on different coverage depths, we simulated experiments of lower coverage by randomly sampling a given percentage from the initial set of reads. For each of 20%, 40%, 60% and 80% (resulting in coverage of 7x, 14x, 21x, and 28x respectively), we ran 100 iterations where in each iteration we randomly chose the given percentage from our reads, refined our set of locations to those with at least one base-call differing from the reference and proceed as in the previous test: divide the locations into a training and test set (the number of instances in each being half of the smaller sized set), compute features, train, classify, and record the percentage of instances classified correctly. The results for these tests, together with the results for the same tests when using only the directionality bias feature for classification are shown in Figure 7. In this iterative approach the classifier was trained on each sequencing depth separately because at different depths the different features may be indicative to different extents. For example, at high sequencing depths the paired t-test statistic and the frequency of error on each direction may have a more significant affect than at lower sequencing depths, where the sequence motif is more informative. Given a new dataset, SysCall computes the mean coverage for the given dataset and uses the model parameters learned from the training set with coverage closest to that observed (out of five training sets with coverages as specified in Figure 7).

SysCall's running time for classification is instantaneous, and the running time for feature assembly depends on the number of sequenced reads in the experiment, currently taking 10 seconds per 100,000 reads, and is trivially parallelizable. SysCall is available at `http://www.cs.berkeley.edu/~meromit/SysCall`.

Presence of Systematic Errors in other datasets

In order to verify that systematic errors are not specific for the methyl-Seq procedure we looked for evidence of systematic errors in other high throughput datasets. We believe systematic error will be extremely important to correct for in RNA-Seq experiments, in which one attempts to annotate both heterozygous sites and expression levels to derive allele specific expression estimates. We therefore looked for systematic errors in the RNA-Seq dataset from Ambion human Brain Reference by Illumina (accession SRA012427), on chromosome 1. Since this dataset did not contain overlapping paired reads we could not annotate *error-pairs*. Instead, we used directionality bias of the base-calls different from the reference to annotate systematic error. We could do so because the coverage in this dataset is high (at transcripts that are highly expressed). For each of the 857,570 locations covered by at least 10 forward and 10 reverse reads we conducted a chi-square test, testing for association between occurrence of mismatches and directionality of sequencing. Under a Bonferroni correction for a 0.05 significance level, we found 991 systematic errors. Thus we have approximately 1 in 1000 sites that are shown to be systematic errors. The method used here, using directionality bias, is statistically weaker than the method with which we identified systematic errors from the methyl-Seq experiment, where we used overlapping mate-pairs to identify base-call errors. The fact that the frequency of identified systematic errors in the RNA-Seq dataset is as high as in the methyl-Seq dataset implies that there are more systematic errors present in the RNA-Seq data than in the methyl-Seq data; this could be due to this dataset being produced by an older version of Illumina's GA.

We also looked at new Illumina data generated by the HiSeq 2000 machines. We analyzed new data from [3] (accession ERX012204) from the exome of chromosome 1. We aligned reads to the reference genome with Bowtie and refined our analysis to the 848,742 sites that were covered by at least 10 reads in each direction. When conducting the same statistical test as for the RNA-Seq data, only 2 sites were determined as statistically significant with respect to the differences from the reference being present on one of the sequencing directions. However, testing for directionality bias of mismatches in this way has little power, and many strong systematic errors are missed by this method (Figure 8). This results in many locations that are not detected by this method as systematic errors but would be wrongly annotated as heterozygous sites due to their characteristics. We therefore annotated a set of candidate heterozygous sites as those locations with at least 10% of the base-calls being different from the reference sequence and with at least 5 differences from the reference, resulting in a set of 1,712 locations. Running SysCall on this set, 316 locations were classified as systematic errors.

We then looked at an Illumina high throughput dataset generated by the 1000 genomes project [3]. Specifically, we analyzed the high-coverage data from sample NA19240 at the first 2 megabases from chromosome 20. Excluding from our analysis partially mapped reads, we were left with a coverage of 23.8x. Out of the 935,490 locations covered by at least 10 forward and 10 reverse reads, 1,812 locations had more than 10 percent of base-calls different from the reference with at least 5 differences total. Of these sites SysCall classified 154 as systematic errors. When annotating SNPs in the 1000 genomes project a filtering step was applied, detailed in sections 5.1.1 and 5.2.1 of the supplementary information of [3], designed specifically to filter out locations in which the base-calls different from the reference are not evenly distributed between the forward oriented and reverse oriented reads. The filtering step applied in [3] to avoid calling systematic errors as SNPs can decrease the number of false-positive SNP calls, but relies on having a sufficient number of reads from each strand and makes use only of the strand-specific characteristic of systematic errors. As we have shown, distinguishing between systematic errors and heterozygous sites can be greatly improved by taking additional evidence into account.

Conclusions

We have identified systematic error in Illumina sequence that is prevalent in different types of datasets, and that does not appear to be easily correctible during base-calling. This systematic error has significant implications for SNP calling, especially at low coverage [9]. Moreover, while increasing the extent of coverage enables the detection of rare variants in population studies and low expression rates in transcriptome studies, it also reveals locations of weaker systematic errors (locations at which there is a small accumulation of base-call errors). Thus, the problem of distinguishing systematic error from true heterozygous sites persists regardless of the extent of coverage. We detected this type of error, and could thoroughly characterize it, thanks to a dataset with overlapping paired-end reads and with very high coverage. Making use of our characterization we have designed and implemented a classifier to correct for systematic errors at much lower coverage depths. We have shown that by using the different characteristics in the prediction process we gain significant increase in performance over using directionality bias alone.

Although we have provided a preliminary characterization of systematic error, with further work and additional data it may be possible to better identify sequences associated with error. In particular, it should be possible to identify and characterize systematic error resulting from other sequencing technologies. Although such a comprehensive assessment is beyond the scope of this study, we have looked at RNA-Seq SOLiD data from [12] and have identified statistically significant systematic error. At the same time, we believe that as sequencing technology improves systematic errors should decrease, and we have observed this to be the case based on the Illumina samples we have investigated. Sequence from two years ago shows higher systematic error rates than recently sequenced data. Nevertheless, we believe that systematic error is a continuing characteristic of Illumina sequence.

Methods

methyl-Seq Experiments

The human sample was collected with IRB approval. Immediately after phlebotomy, leukocytes were isolated by Ficoll centrifugation. B cells were isolated from the leukocyte fraction with an indirect magnetic labeling system for the isolation of untouched B cells which yields highly pure B cell preparations (Miltenyi). DNA was extracted by standard methods, and digested overnight with HpaII (NEB). HpaII cuts the sequence CCGG; methylation of the central cytosine on one or both strands protects the sequence from digestion with HpaII [4]. HpaII fragments 50-300 bp in length were isolated on an agarose gel. A paired-end sequencing library was constructed with the standard Illumina kit, and sequenced on an Illumina GAIIX to collect reads of 76 bases, resulting in 15,598,990 read pairs. Read pairs that did not terminate at CCGG restriction sites were removed, leaving 14,205,350 read pairs. The reads were mapped to the human reference genome (hg18) using Bowtie [6] as single end reads allowing 3 mismatches and requiring that the alignments be unique. Those that did not align were removed and the remaining reads were mapped again, this time as paired end reads with a mismatch limit of 2. The higher mismatch limit of 3 was used in the initial alignment step to avoid having reads with more base-call errors preferentially pass the uniqueness requirement. This produced 6,939,310 aligned read pairs mapped to 313,789 distinct locations.

The same procedure was followed for the second methyl-Seq experiment from monocyte DNA. The experiment generated 14,432,723 read pairs, of which 7,265,035 were ultimately mapped to 274,230 distinct locations.

Systematic error probability upper bound

To calculate the upper bound on the systematic error probabilities given the quality scores, we used the following computation. Let n be the coverage depth, and $W = \{p_1, p_2, \ldots, p_n\}$ s.t. $p_i \geq p_{i+1}$ be the set of base-call error probabilities for those base-calls given their quality scores, and X be a random variable denoting the number of base-call errors. Then

$$\begin{aligned}
&P(X = k|W, n) \\
&= \sum_{\{p_{i_1},\ldots,p_{i_k}\}\subseteq W} p_{i_1}p_{i_2}\cdots p_{i_k}(1-p_{i_{k+1}})(1-p_{i_{k+2}})\cdots(1-p_{i_n}) \\
&\leq \binom{n}{k} p_1p_2\cdots p_k(1-p_{k+1})(1-p_{k+2})\cdots(1-p_n).
\end{aligned}$$

We then found the upper bound for $P(X \geq x|W)$ by summing over all k equal to or greater than x, the number of observed base-call errors.

Authors contributions

FM, MS and LP designed the research, FM and MS conducted the research, DB, JD and DM contributed the datasets analyzed and FM, MS and LP wrote the manuscript.

Acknowledgements

We thank Professor Yun Song and Dr. Wei-Chun Kao from UC Berkeley for the phiX174 dataset and the associated NaiveBayesCall output.

References

[1] Gavin E. Crooks, Gary Hon, John-Marc M. Chandonia, and Steven E. Brenner. WebLogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, June 2004.

[2] Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105, September 2008.

[3] RM Durbin et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

[4] Richard M Harland. Inheritance of DNA methylation in microinjected eggs of Xenopus laevis. . *Proc Natl Acad Sci USA*, 79(7):2323–2327, 1982.

[5] Wei-Chun Kao and Yun Song. naiveBayesCall: An Efficient Model-Based Base-Calling Algorithm for High-Throughput Sequencing. In Bonnie Berger, editor, *Research in Computational Molecular Biology*, volume 6044 of *Lecture Notes in Computer Science*, chapter 15, pages 233–247. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2010.

[6] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25+, 2009.

[7] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[8] Mingyao Li, Isabel X. Wang, Yun Li, Alan Bruzel, Allison L. Richards, Jonathan M. Toung, and Vivian G. Cheung. Widespread rna and dna sequence differences in the human transcriptome. *Science*, 2011.

[9] N Malhis and SJ Jones. High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics*, 26:1029–1035, 2010.

[10] K Namamura, T Oshima, T Morimoto, S Ikeda, H Yoshikawa, et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, pages 1–13, 2011.

[11] Rasmus Nielsen. Genomics: In search of rare human variants. *Nature*, 467(7319):1050–1051, October 2010.

[12] A Roberts, C Trapnell, J Donaghey, JL Rinn, and L Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12:R22, 2011.

[13] James T. Robinson, Helga Thorvaldsdottir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nat Biotech*, 29(1):24–26, Jan 2011.

[14] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.

[15] M Taub, HC Bravo, and RA Irizarry. Overcoming bias and systematic errors in next generation sequencing data. *Genome Medicine*, 2:87, 2010.

[16] C Trapnell, BA Williams, G Pertea, A Mortazavi, GK, MJ van Baren, SL Salzberg, BJ Wold, and L Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511–515, May 2010.

[17] K Zhang et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nature Methods*, pages 613–618, 2009.

Figures

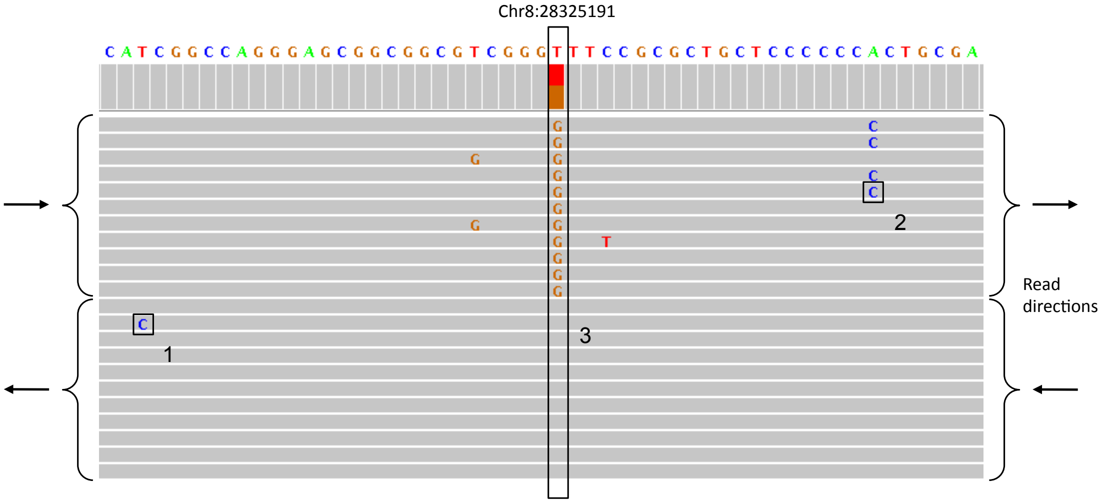
Chr8:28325191
C A T C G G C C A G G G A G C G G C G G C G T C G G G T T T C C G C G C T G C T C C C C C C A C T G C G A
2
3
1
Read directions

Figure 1: **Types of errors**. A screenshot from the IGV browser [13] showing three types of error in reads from an Illumina sequencing experiment: (1) A random error likely due to the fact that the *position* is close to the end of the read. (2) Random error likely due to *sequence specific* error-in this case a sequence of Cs are probably inducing errors at the end of the low complexity repeat. (3) *Systematic error*: although it is likely that the GGT sequence motif and the GGC motifs before it created phasing problems leading to the errors, the extent of error cannot be explained by a random error model. In this case, all the base calls in one direction are wrong as revealed by the 11 overlapping mate-pairs. In particular, all differences from the reference genome are base-call errors, verified by the mate-pair reads, which do not differ from the reference. Given the background error rate, the probability of observing 11 *error-pairs* at a single location, given that 11 mate-pair reads overlap the location, is 1.5×10^{-26}. Moreover, given the presence of such errors at a single location, the probability that all of the errors occur on the same strand (i.e., on the forward mate pair) is $\frac{1}{1024} = 0.00098$. Note that the IGV browser made an incorrect SNP call at the systematic error site (colored bar in top panel).

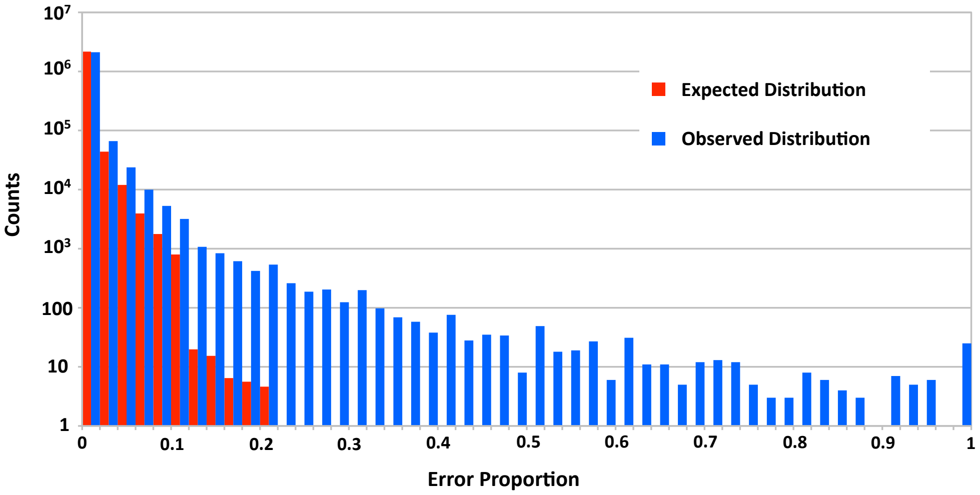

Figure 2: **Proportion of base call errors at systematic error sites**. The observed (blue) number of locations with high base-call error frequencies significantly exceeds the expected amount (red).

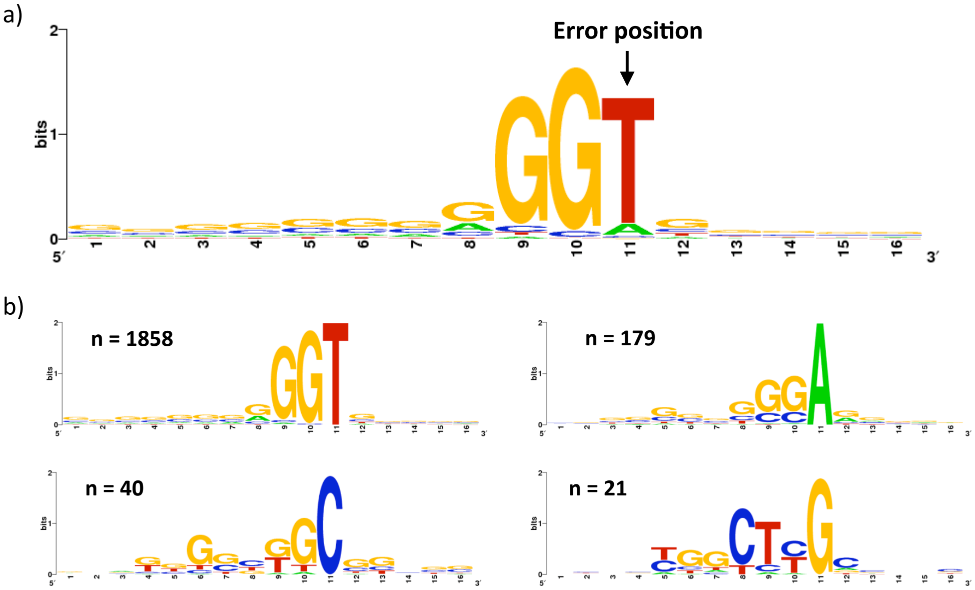

Figure 3: **Sequence motifs at systematic error sites**. (a) The motifs around systematic errors reveal a strong enrichment for instances preceded by an occurrence of GG and for the error to occur at locations where the reference genome is T. (b) categorized by the nucleotide at the error location. The number of systematic errors in each subset is denoted by n.

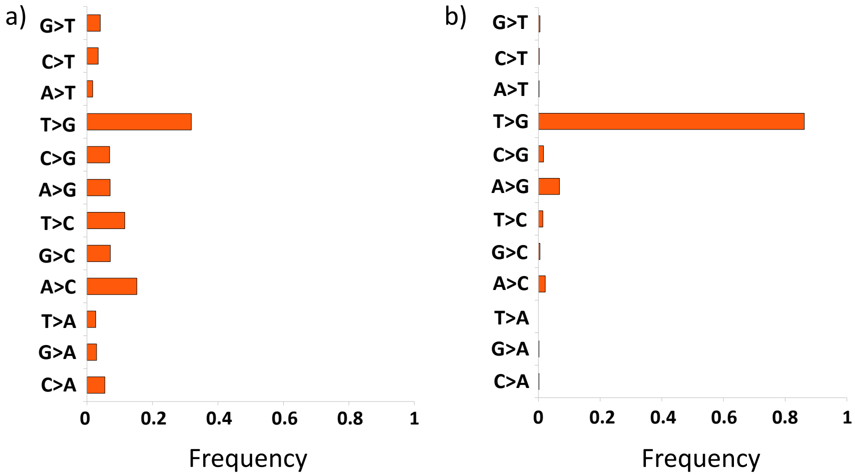

Figure 4: **Base substitutions of systematic errors**. Frequency of different base substitutions in (a) all errors (b) systematic errors.

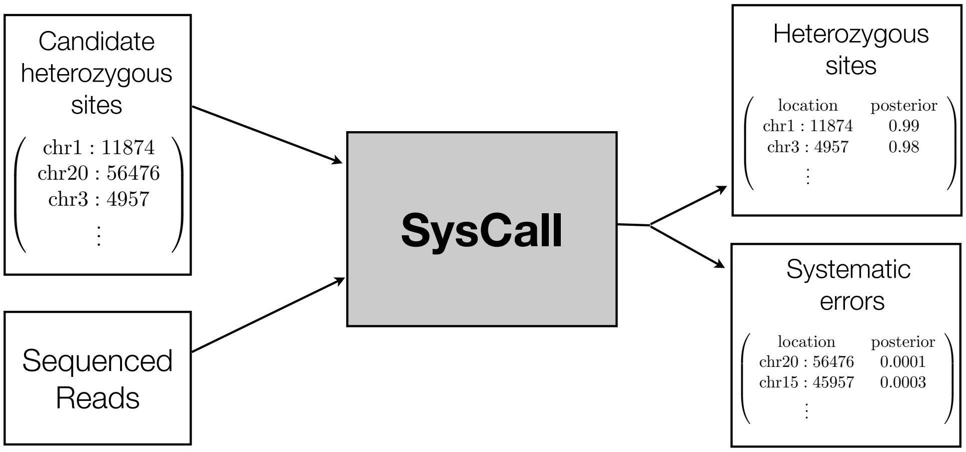

Figure 5: **Using SysCall to distinguish heterozygous sites from systematic errors**. SysCall takes as input a list of genomic locations indicating candidate heterozygous sites and the reads sequenced sequenced from the experiment (in sam format) and divides the initial candidate list into lists of annotated heterozygous sites and systematic errors, along with the annotated posterior probability of every location being a true heterozygous site.

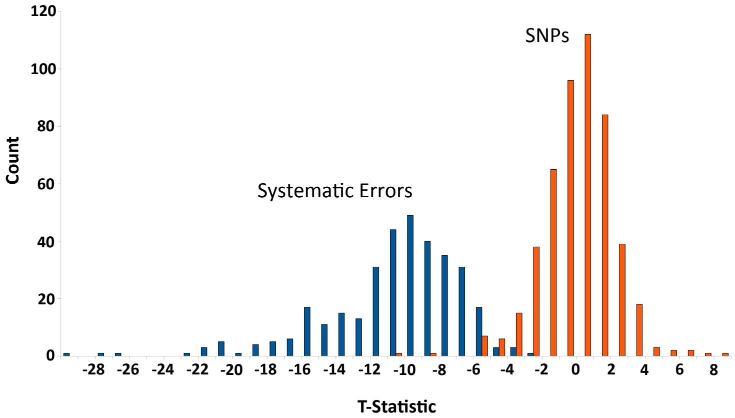

Figure 6: **The paired t-test statistic helps distinguish true SNPs from systematic errors**. The paired t-test ($PT(w_0, w_1)$) was computed for the "SNPs" and "Systematic errors" sets used for training SysCall. The histogram of paired t-test for the "SNPs" set (red) is centered around 0 (mean: 0.0024, std: 2.035), indicating that the quality scores at those locations were similar to their neighboring quality scores. The histogram of the "Systematic errors" set (blue) formed an almost disjoint distribution (mean: -10.505, std: 3.919).

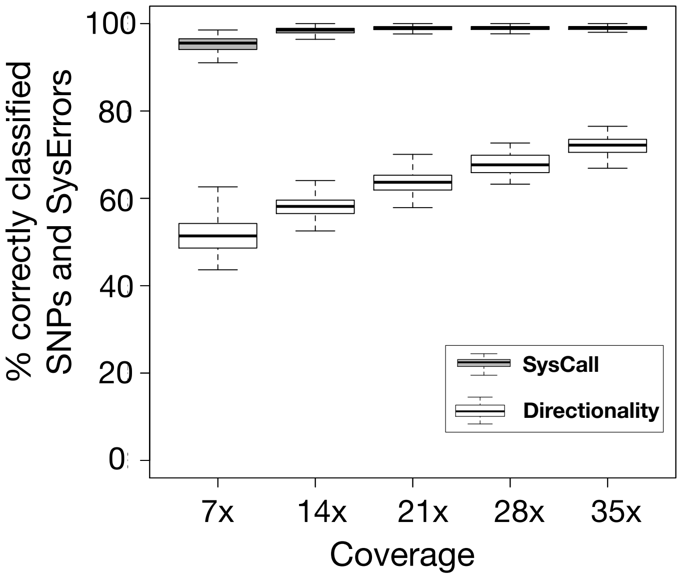

Figure 7: **Performance of SysCall**. Proportion of correctly classified instances for our classifier (grey) and for a classifier considering only the feature of difference in error frequency (white) for different sequencing coverage.

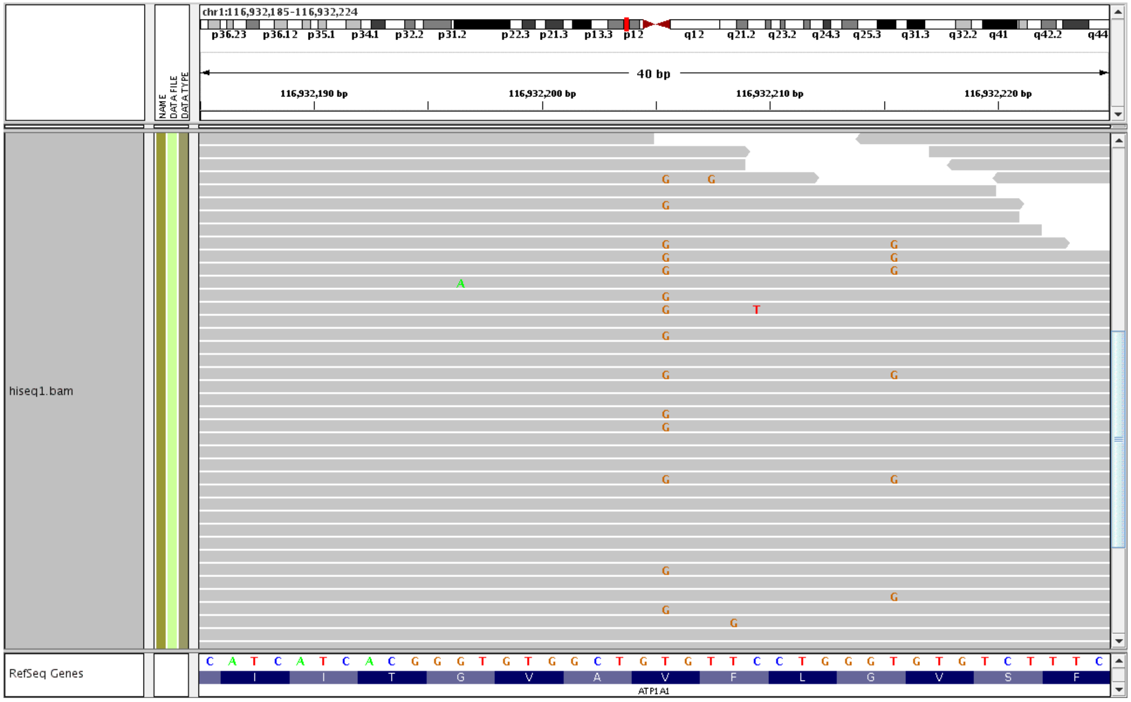
chr1:116,932,185-116,932,224
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q12 q21.2 q23.2 q24.3 q25.3 q31.3 q32.2 q41 q42.2 q44
NAME
DATA FILE
DATA TYPE
40 bp
116,932,190 bp
116,932,200 bp
116,932,210 bp
116,932,220 bp
hiseq1.bam
RefSeq Genes
C A T C A T C A C G G G T G T G G C T G T G T T C C T G G G T G T G T C T T T C
I I T G V A V F L G V S F
ATP1A1

Figure 8: **Example Systematic Error in HiSeq Data**. These systematic errors are not statistically significant under the test for directionality bias (after correcting for multiple hypotheses), which demonstrates the weakness of this test.