Published online 26 July 2008

Nucleic Acids Research, 2008, Vol. 36, No. 16 ***e105***
doi:10.1093/nar/gkn425

Substantial biases in ultra-short read data sets from high-throughput DNA sequencing

Juliane C. Dohm[1], Claudio Lottaz[2], Tatiana Borodina[1] and Heinz Himmelbauer[1,*]

[1]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin and [2]Institute for Functional Genomics, Computational Diagnostics, University of Regensburg, Josef-Engert-Str. 9, 93053 Regensburg, Germany

Received December 21, 2007; Revised June 16, 2008; Accepted June 19, 2008

ABSTRACT

Novel sequencing technologies permit the rapid production of large sequence data sets. These technologies are likely to revolutionize genetics and biomedical research, but a thorough characterization of the ultra-short read output is necessary. We generated and analyzed two Illumina 1G ultra-short read data sets, i.e. 2.8 million 27mer reads from a *Beta vulgaris* genomic clone and 12.3 million 36mers from the *Helicobacter acinonychis* genome. We found that error rates range from 0.3% at the beginning of reads to 3.8% at the end of reads. Wrong base calls are frequently preceded by base G. Base substitution error frequencies vary by 10- to 11-fold, with A > C transversion being among the most frequent and C > G transversions among the least frequent substitution errors. Insertions and deletions of single bases occur at very low rates. When simulating re-sequencing we found a 20-fold sequencing coverage to be sufficient to compensate errors by correct reads. The read coverage of the sequenced regions is biased; the highest read density was found in intervals with elevated GC content. High Solexa quality scores are over-optimistic and low scores underestimate the data quality. Our results show different types of biases and ways to detect them. Such biases have implications on the use and interpretation of Solexa data, for *de novo* sequencing, re-sequencing, the identification of single nucleotide polymorphisms and DNA methylation sites, as well as for transcriptome analysis.

INTRODUCTION

The DNA sequencing field has experienced a major boost with the emergence of novel sequencing technologies. Several systems are currently on the market, including Illumina's Solexa instrument, the Applied Biosystems' Sequencing by Oligonucleotide Ligation and Detection (SOLiD) technology, and the GS FLX instruments from Roche/454 Life Sciences. The Polony cyclic sequencing by synthesis technology is to be launched (1).

These technologies allow sequence determination much quicker and cheaper than the dideoxy chain terminator method presented by Sanger in 1977 (2). The main difference between Sanger sequencing output and the output of the new technologies is an increased read number, associated with a decrease in the length of individual reads.

To achieve high throughput, the new approaches apply different strategies. 454 Life Sciences has adapted pyrosequencing to a microbead format to sequence 400 000 DNA fragments simultaneously, resulting in a per-run dataset of 100 Mbp with reads averaging 250 bp. SOLiD sequencing also uses templates immobilized onto microbeads. Here, the sequence of the template DNA is decoded by ligation assays involving oligonucleotides labeled with different fluorophores. The SOLiD read length is currently 25–35 bases, and 2–3 Gbp of data can be collected during an 8-day run. Solexa sequencing is based on amplifying single molecules attached to the surface of a flow cell to generate clusters of identical molecules, followed by sequencing using fluorophore-labeled reversible chain terminators. Solexa sequencing proceeds a base at a time and read length depends on the number of sequencing cycles. Current Illumina sequencing instrumentation achieves read lengths of 36 bases. The Solexa flow cell is composed of eight separately loadable lanes. Since each lane has a capacity of about 5 million reads, > 40 million reads can be generated in a run of 3 days, equivalent to > 1.3 Gbp.

The adoption of high-throughput sequencing will revolutionize molecular biology research, similar to the invention of the polymerase chain reaction (PCR) twenty years ago (3). 454 pyrosequencing short (~100 bp) reads generated on Roche GS20 instruments (now replaced by GS FLX) were successfully used for the de novo sequencing of small genomes and BACs as well as for transcript discovery and characterization (4–9). De novo genomic sequencing succeeded even when ultra-short (27–36 bp) reads generated by Solexa sequencing were employed for

*To whom correspondence should be addressed. Tel: +49 30 8413 1354; Fax: +49 30 8413 1380; Email: himmelbauer@molgen.mpg.de

a small genome (10). For the human genome, ultra-short reads were applied in studies on chromatin analysis (11,12).

However, working with large data sets of short reads involves difficulties, especially due to wrong base calls. To exploit the full prospects of the novel technologies there is the need to know as much as possible about biases in the output data sets, especially with respect to errors. Previous studies focused on the 454 technology (13) or dealt with the prospects of short read sequencing as such (14). Here, we characterize two Solexa read data sets: 12.3 million 36mer reads (trimmed to 32 bases) from the *Helicobacter acinonychis* genome and 2.8 million 27mer reads from a *Beta vulgaris* bacterial artificial chromosome (BAC) clone. We analyze these reads and detect biases with respect to error positions, error rates, erroneous base calls and their neighboring bases and single base insertions or deletions. We determine the compensation of erroneous base calls by correct base calls depending on the sequencing coverage. We analyze read start positions, the read coverage along the target sequence, and dependencies of read coverage and local sequence characteristics. Finally, we assess the reliability of quality values for wrong and correct base calls.

METHODS

Solexa sequencing

Helicobacter acinonychis. DNA was fragmented by nebulization as described in the Solexa protocol (www.illumina.com). *Beta vulgaris* DNA was sheared for 1 h with a UTR200 sonication device (Hielscher Ultrasonics GmbH) at 100% amplitude and 0.5 cycle mode. Fragmented DNA was further processed as described previously (10). Sequencing was carried out by running 27 or 36 cycles, respectively, on the Illumina 1G sequencing instrument. The Goat module (Firecrest v.1.8.28 and Bustard v.1.8.28 programs) of the Solexa pipeline v.0.2.2.3 (for *Helicobacter* data set) and v.0.2.2.5 (for *Beta* data set) were used for image deconvolution and quality value calculation. Parameterization was auto-generated by the pipeline (see Supplementary Data for intensity plots and run parameters, i.e. frequency cross-talk matrix, offsets, phasing). Set up configuration was used as installed by Illumina's technical staff. The *Helicobacter* data set was collected from three lanes of two flow cells. The *Beta* data set was generated in a single lane from a further flow cell.

Data analysis

We developed various Perl scripts to extract and process information from ELAND output files (Gerald module v.1.27 of the Solexa pipeline) and to find positions of reads that can be aligned more than once to the reference sequence without mismatches (the positions of those reads are not reported by ELAND). We wrote Perl scripts for the detection of deletions and insertions of single nucleotides in otherwise error-free reads and for the analysis of quality values per base call. Plots were generated with the statistical computing environment R (www.R-project.org) or OpenOffice Calc (www.openoffice.org). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org) or OpenOffice Calc (www.openoffice.org).

Data availability

Solexa read data are available from the SHARCGS project website at http://sharcgs.molgen.mpg.de.

RESULTS

We previously generated 12 288 791 36mer reads from *Helicobacter acinonychis* on an Illumina 1G sequencing device (10). The *Helicobacter* genome is 1.55 Mbp in size and has a GC content of 38%. A high-quality reference sequence for *Helicobacter* is available (GenBank NC_008229) (15). We ran the ELAND software on the read data set (trimmed by the last four bases, because ELAND processes the first 32 bases only) and selected the 8 389 548 32mer reads that ELAND reported to be uniquely matched against the *Helicobacter* reference sequence with zero, one or two mismatches (labeled U0, U1 or U2, respectively, see Figure 1b). Additionally, we generated a 27mer read data set for the sugar beet (*Beta vulgaris*) bacterial artificial chromosome (BAC) clone ZR-47B15. The data set consists of 2 788 286 reads, 2 156 266 of which were labeled U0, U1 or U2 in the ELAND output (Figure 1a). The Sanger reference sequence in finished quality of this BAC insert consists of 109 563 bases with 34.85% GC (Dohm *et al.*, manuscript submitted for publication). For all uniquely matched reads, ELAND reports the match position in the reference sequence as well as the error position(s) in the read.

Start positions of reads and read distribution on the target sequence

The preparation of Solexa sequencing libraries involves the fragmentation of the DNA, followed by the adaptor ligation, pre-amplification for material enrichment and amplification within the flow cell prior to sequencing. In order to detect whether the steps preceding sequencing show biases, we analyzed the first bases of a read and the bases that flank the read start position on either side. Of all possible 27mer tuples (*Beta*) and 32mer tuples (*Helicobacter*), 99.8 and 98.8% are unique, respectively. We therefore assume that potential biases are representative for the data set.

We calculated the frequency of 2- to 10-base tuples enclosing the starting point for 8 389 548 uniquely matched *Helicobacter* reads and for 2 156 266 uniquely matched *Beta* reads relative to the frequency of these tuples in the reference sequences. Since the bases in the reads are subject to errors, we used for both sides the bases of the corresponding region in the reference sequence.

A general sequence bias for the immediate vicinity of the read start position could not be deduced from the two data sets. The results for the *Beta* data set did not suggest any tendencies (Supplementary Figure 1a). The results for the reads from *Helicobacter* showed a weak tendency towards T being the most frequent base call to the left and

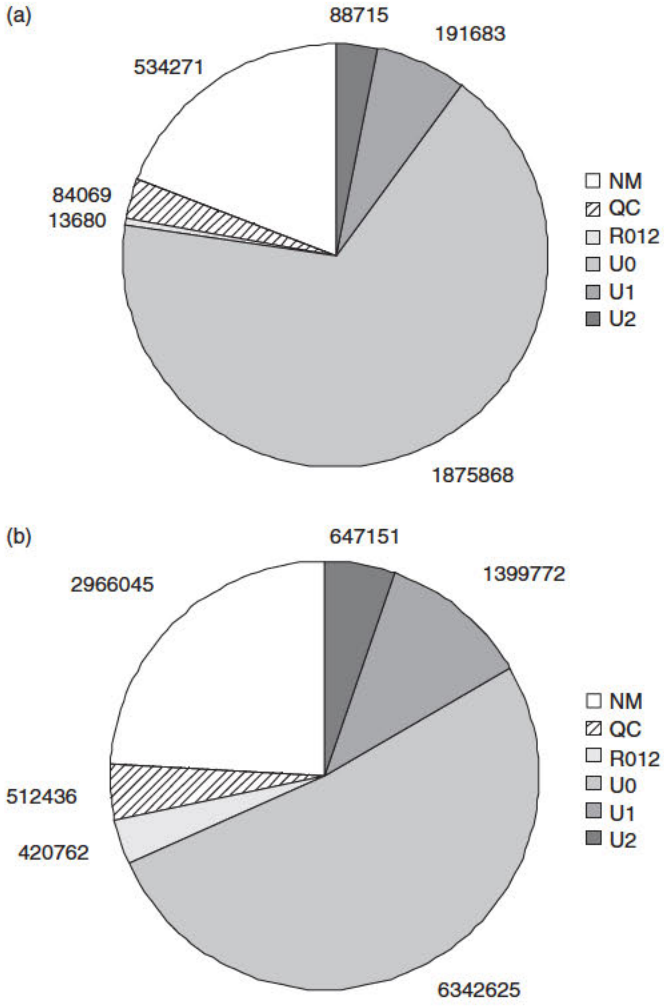

Figure 1. Pie charts of the read analysis with ELAND. The ELAND categories are: QC: no matching done because of low quality of the read (more than two positions with quality score 5), NM, no match found; U0, unique exact match found; U1, unique match with one error; U2, unique match with two errors; R0, multiple exact matches found; R1, multiple matches with one error; R2, multiple matches with two errors. The categories R0, R1, R2 are shown as a single entity. (**a**) ELAND categorizations for 27mer reads from *Beta vulgaris* clone ZR 47B15 (2 788 286 in total). (**b**) ELAND categorizations for 32mer reads from *Helicobacter acinonychis* (12 288 791 in total, trimmed by the last four base calls of the original 36mer data).

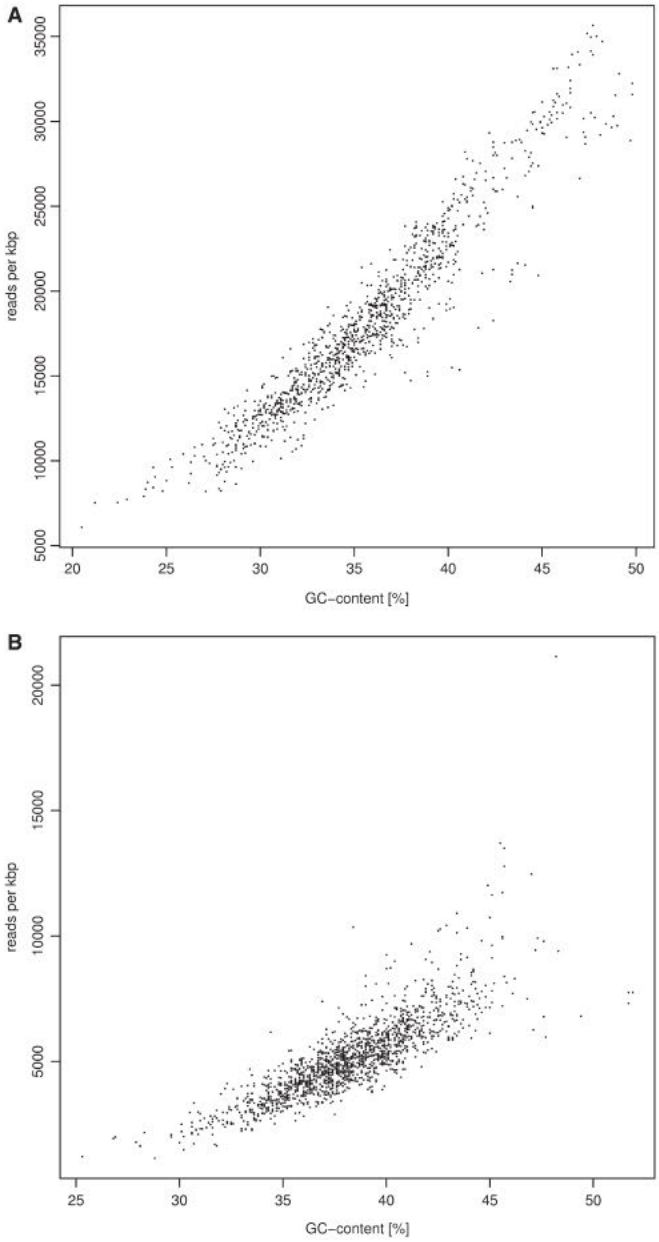

Figure 2. Correlation of the Solexa read coverage and GC content. (**a**) 27mer reads generated from *Beta vulgaris* BAC ZR 47B15 (**b**) 32mer data set from the *Helicobacter acinonychis* genome. Each data point corresponds to the number of reads recorded for a 1 kbp window (shift of 100 bp in *Beta* and 1 kbp in *Helicobacter*).

to the right of the read start position (Supplementary Figure 1b). Since two different fragmentation methods were used, sonication for *Beta* and nebulization for *Helicobacter*, the results may indicate method-inherent properties.

However, by analysing sequence characteristics and number of reads starting in a sliding window of 1 kbp in width, we found a correlation of read coverage and GC content in both data sets (Figure 2). In regions of elevated GC content the number of reads was increased. For instance, windows with a GC content of 40% contain almost twice as many reads as windows with 30% GC in the *Beta* data set. Thus, while the vicinity of 10 bp was not

points, there is a strong preference towards GC-rich regions in 1 kbp sliding windows. Since both templates show the correlation of read coverage and GC content, the shift to GC rich regions seems to be a general feature of the current pre-sequencing procedure. A similar finding was reported by Hillier *et al.* (16).

The overall coverage considering matching reads only is 165-fold in the *Helicobacter* data set (185-fold for 36mer reads) and 465-fold in the *Beta* data set. The distribution of matching reads along the reference sequences is shown

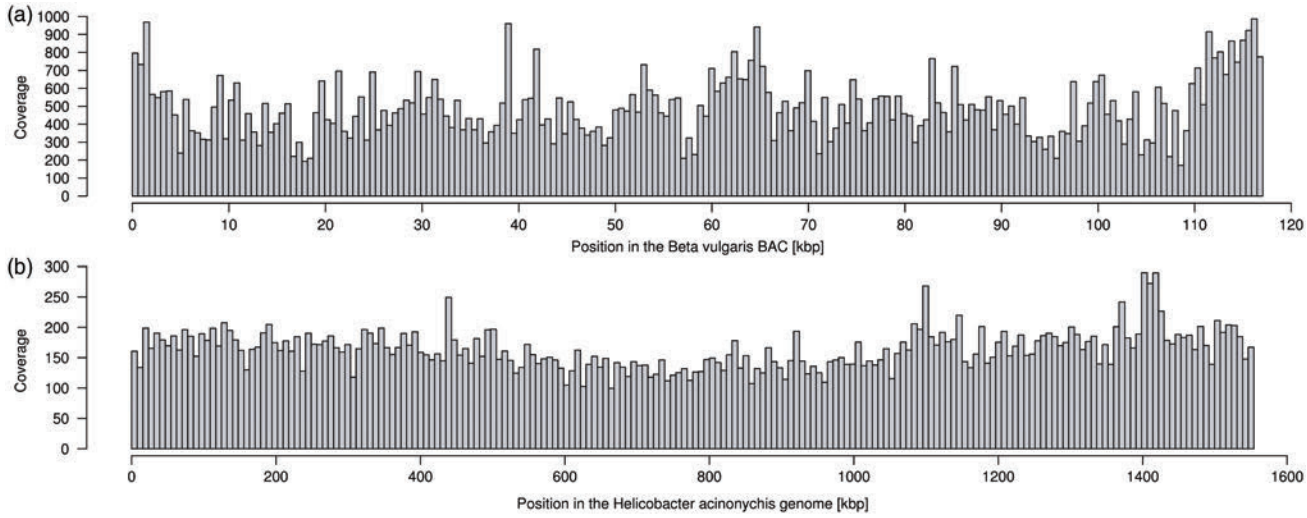

Figure 3. Distribution of Solexa reads along the reference sequences considering unique match positions reported by ELAND (zero, one or two mismatch bases) and reads with more than one match position (no mismatch bases) detected with a Perl script. (**a**) Read distribution along the *Beta vulgaris* BAC sequence (with cloning vector pBeloBACII). 2 166 892 27mer reads were matched against the finished sequence (enclosed by the cloning vector,~117 kbp in total). The read coverage was calculated in 200 consecutive 0.58 kbp windows. (**b**) Read distribution along the 1.55 Mbp *Helicobacter* genome, based on 8 700 113 32mer reads. The local coverage is shown in 200 consecutive windows of 7.77 kbp.

size 7.77 kbp for *Helicobacter* (Figure 3a) and of size 0.58 kbp for *Beta* (Figure 3b). The coverage varied by a factor of 13 and 3.8, respectively, ranging from 49- to 652-fold for *Helicobacter* and from 238- to 897-fold for *Beta* (Table 1). We tested whether the distributions shown in Figure 3 are compatible with a uniform distribution of reads across the target sequences. We have applied a χ^2-test (goodness of fit) to reject the hypothesis that reads have the same probability to fall into equally sized regions of the target sequence ($P < 1e^{-10}$ even when dividing target sequences in only five regions).

There is a number of 'gap' positions in the target sequences where no read starts from. However, since there are no gaps larger than read length all positions of the target sequence are covered (Supplementary Table 1).

Table 1. Proportion of reference sequence and coverage ranges (based on ELAND U0, U1, U2, R0 matched reads and reads with single indels)

Beta		Helicobacter	
Coverage	BAC (%)	Coverage	Genome (%)
200 300	4.27	<100	3.53
300 400	23.93	100 150	26.06
400 500	25.64	150 200	42.28
500 600	23.93	200 250	21.49
600 700	12.82	250 300	4.44
700 800	4.27	300 350	1.29
800 900	5.13	>350	0.90

Distribution of error positions along reads

We selected all ELAND U1 and U2 reads, i.e. 28 0173 *Beta* reads and 2 046 923 *Helicobacter* reads (cf. Figure 1), to analyze the occurrence of errors per position. We performed two types of calculations. Firstly, we calculated the fraction of wrong base calls at each read position considering wrong base calls only. Secondly, we calculated per-base error rates, i.e. the fraction of wrong base calls per position considering all base calls. The result is shown in Figure 4. The number of occurrences of wrong bases is increased at the first position. Rising from the lowest error rate at the second position, the highest error rate is observed at the last positions of the read [similar observation reported in (16)]: 2.5 and 2.9% of the errors in the data sets of *Beta* and *Helicobacter*, respectively, were found at read position 1, and 11.8% of errors were recorded at the last read position (position 27 in the *Beta* data set and position 32 in the *Helicobacter* data set, Figure 4a). The per-base error rates range from 0.3% to 3.8% (Figure 4b) resulting in an average error rate of 0.6% for the *Beta* data set and 1.0% for the *Helicobacter* data set. Note that only uniquely matched reads with less than three substitution errors are considered.

In re-sequencing projects, sequencing errors can be compensated by high-coverage sequencing. In a re-sequencing project, the reads are aligned against a reference sequence. Wherever a mismatch between sequencing data and the reference is observed, a polymorphism is postulated. In order to avoid spurious detection of polymorphisms due to sequencing errors, a consensus between several reads at each position of the reference is common practice. Here, we simulate re-sequencing at different depth by randomly choosing the appropriate number of reads from our two data sets and counting wrong and correct base calls [five (*Helicobacter*) or ten (*Beta*) simulations per data point]. An error was considered as compensated when at least one correct base call for the same position existed. A correct base call and the reference sequence hold the majority over one wrong base call, i.e. x wrong base calls at the same position can be compensated by x correct base calls (plus reference sequence)

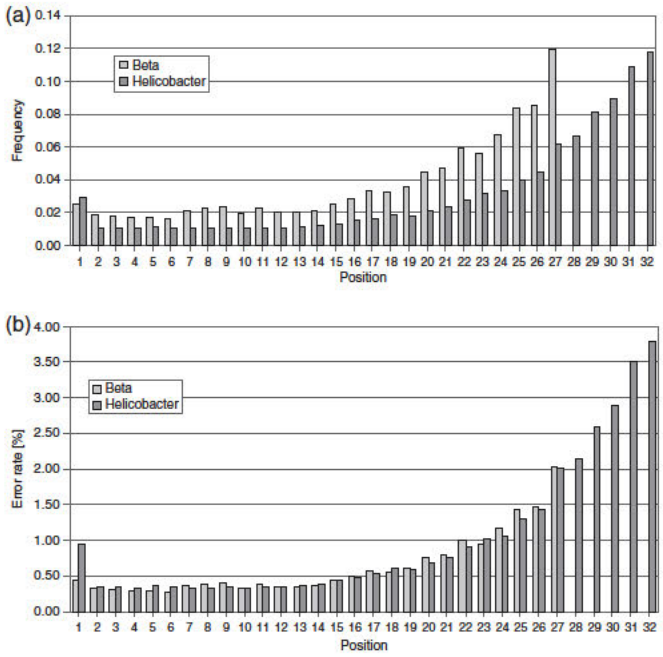

Figure 4. Frequency of wrong base calls in Solexa reads depending on the position along the read (27mer reads from *Beta vulgaris* and 32mer reads from *Helicobacter*). (**a**) Error frequency per position calculated from considering wrong base calls only. The highest error frequency is observed at the read 3′ end. (**b**) Per base error rates (overall error frequency per position considering all base calls).

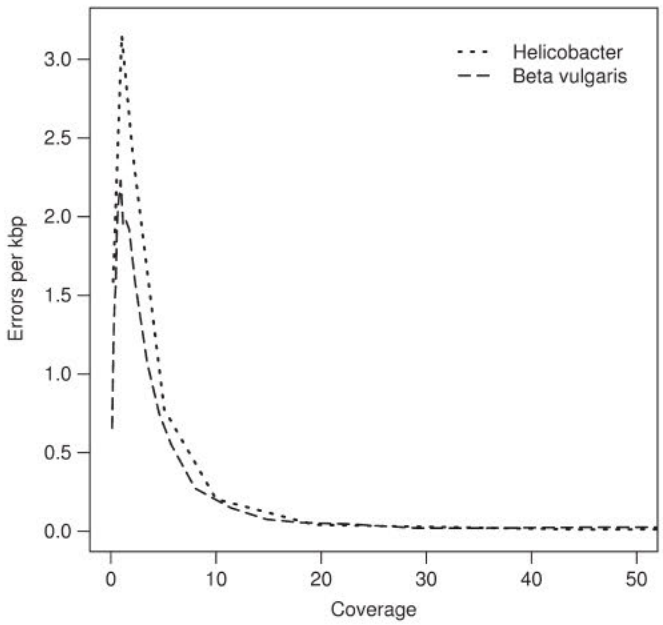

Figure 5. Compensation of sequencing errors by deep sequencing in re sequencing projects. The average number of errors per kbp is shown for different levels of coverage. For coverages below 2, reads are unlikely to overlap and compensation of sequencing errors is rare (thus, sequencing errors accumulate when the coverage is increased). For coverages above 3 fold the number of uncompensated errors drops rapidly with the increase of coverage.

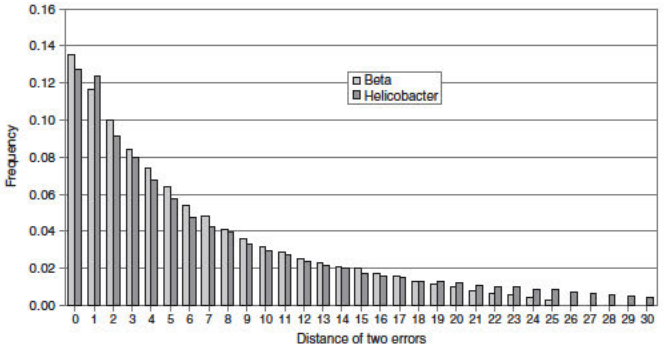

Figure 6. Distance between two errors on a read in the *Helicobacter* and *Beta vulgaris* data sets. '0' indicates that the erroneous base calls are next to each other.

We plotted the dependency of sequencing coverage and error compensation in Figure 5 (range of simulation results: see Supplementary Figure 2). Increasing the sequencing coverage results in a rapid decrease of uncompensated errors. At a coverage of 20-fold the average number of errors per kilo base pair is close to zero and does not decrease any further. However, such estimates are likely to change with improvements of the sequencing technology, as less coverage will be sufficient for reduced error rates.

Analysis of reads containing two errors

ELAND reported 88 753 reads containing two errors in the *Beta* data set, corresponding to 4.1% of all uniquely matched reads. In *Helicobacter*, 647 151 reads contained two errors (7.7% of all uniquely matched reads). We analyzed the distance between erroneous bases and found a preference for small distances between errors (Figure 6). In 25% of reads that contained two errors the erroneous bases were either at adjacent positions or separated by one base. This observation does not contradict the assumption that errors occur independently according to their position-specific probability. The heatmap in Supplementary Figure 3 illustrates the occurrence of two errors relative to the positions in the read. As expected from the per-base error rates, two-error occurrences are concentrated at the 3′ end of reads and are therefore close together. In addition, error pairs also occur with increased frequency at read positions 1 and 2. We provide even stronger evidence for the independence of error positions in two-error reads in Supplementary Figure 4.

Although error positions seem to be independent in reads with two errors, there is evidence that errors accumulate in reads more easily than expected. We deduce this from the ratios of the observed and expected number of reads containing one and two errors respectively: Given the determined error rates per position (for the *Helicobacter* data set) we expect 3.5 times more correct reads (U0) than reads with one error (U1), but we observe 4.5 times more U0 than U1; we expect 19.8 times more correct reads than reads with two errors (U2), but we

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.