Title:

Platform-independent method for detecting errors in metagenomic sequencing data: DRISEE.

Kevin P. Keegan, William L. Trimble, Jared Wilkening, Andreas Wilke, Travis Harrison, Mark D'Souza, Folker Meyer

Abstract:

Conventional methods to quantify sequencing error have intrinsic features that limit their applicability to shotgun metagenomic data. The increasing diversity and proliferation of high-throughput technologies have made the need for universal error-estimation methods particularly acute. We introduce DRISEE, a platform-independent method to assess sequencing error in shotgun metagenomic data, and utilize it to discover previously uncharacterized error in *de novo* sequence data from two widely used technologies.

ANL/MCS-P1924-0811

Main Text:

Accurate quantification of sequencing error is essential to sequence-dependent investigations, making it possible to distinguish observations of genuine interest from background noise. Sequence-based experimental inferences, particularly those related to the identification and characterization of features (protein or 16s rRNA coding regions, regulatory elements, etc.) are greatly affected by the presence of sequencing errors[1]. Errors in metagenomic amplicon-based sequencing have led to grossly inflated estimates of taxonomic diversity[2-4]. While methods such as denoising[3] have been developed to address the issue of noise/error in *amplicon-based* metagenomic sequencing[5,6], no analogous techniques have been reported to account for noise/error in the context of *shotgun-based* metagenomic sequencing. Limitations inherent to methods used to assess *de novo* sequencing error are largely to blame. At present, two methods are commonly used: (1) Reference-genome-based methods utilize previously completed genomes as a standard to which *de novo* data can be compared, but such methods are applicable only to genomic data. (2) Score-based methods rely on sophisticated platform-dependent (frequently proprietary) error models to produce base calls, each with an accompanying fidelity estimate or score, but offer no information regarding error type (i.e. substitution, insertion, deletion), and cannot account for errors introduced by procedures that are common to nearly every experimental sequencing protocol (e.g. the use of different DNA/RNA extraction procedures to process a group of samples). Each of these methods requires extensive reference data (in the form of reference genomes and/or error models), and neither is wholly suited to platform-independent analysis of error in shotgun-based

metagenomic data (see Supplemental Text for more detailed descriptions of reference-genome and score-based methods).

The limitations of reference-genome and score-based methods inspired the creation of Duplicate Read Inferred Sequencing Error Estimation (DRISEE). DRISEE exploits *artifactually duplicated reads* (nearly identical reads that share a completely identical prefix, present with abundances that greatly exceed chance expectations, even when a modest level of biological duplication is taken into account)[7,8] to explore sequencing error (see Methods). Sequence variation within a group of artifactually duplicated reads is more likely to be the product of technical artifact(s) (i.e. sample processing and/or sequencer errors) than a reflection of genuine diversity in the originally sampled population or culture. Based on this premise, DRISEE utilizes prefix identical clusters of artifactually duplicated reads to create internal standards (consensus sequences) to which each individual duplicate read is compared. Sequencing error is determined as a function of the variation that exists within clusters of artifactually duplicated reads. This strategy is platform-independent and can be used to quantify error in metagenomic or genomic samples with respect to error frequency and type. DRISEE identifies reads using stringent requirements for prefix length and abundance that are extremely unlikely to occur unless the sequences have been artifactually duplicated. For the work presented here, a prefix length of 50 bases and a minimum abundance of 20 reads was used; chance occurrence was ≈ 4E-32 (see Methods). The Methods present a detailed, workflow-based description of DRISEE analysis.

Initial validations of DRISEE with simulated data showed nearly perfect correlations between known and DRISEE-based error estimates (Supplementary Fig. 1a,

R^2 = 0.99). Additional validations with real genomic sequencing data exhibit good correlation with error estimates produced with conventional reference-based analyses[8] of the same samples (Supplementary Fig. 1b, R^2 =0.89, excluding outliers). In further trials, DRISEE was applied to genomic and metagenomic shotgun data produced by two widely utilized sequencing technologies, 454 and Illumina (n=242 genomic 454, n=65 metagenomic 454, n=10 genomic Illumina, and n=159 metagenomic Illumina samples), 476 samples in all. Less than half of the individual samples (*n*=169) exhibit DRISEE-based errors consistent with the reported range of second-generation sequencing errors (0.25–4%)[2,8-12]. The majority of samples (*n*=307) exhibit DRISEE-based errors that fall outside the range of reported sequencing errors (error < 0.25%, *n*=73; error > 4%, *n*=234; avg ± stdev = 12.63 ± 15.12) (Supplementary Fig. 2). To explore this result, we obtained FASTQ data via SRA (http://www.ncbi.nlm.nih.gov/sra) for subsets of DRISEE-analyzed samples: 20 of the 65 metagenomic 454 samples and 12 of the 159 Illumina metagenomic samples. Per base DRISEE and Phred[13]-based errors for these samples were calculated and compared (see Methods). Whereas Phred values exhibit nearly indistinguishable trends between the 454 and Illumina data, DRISEE error profiles differ for each technology (Figs. 1a and 1b).

DRISEE produces error estimates that are routinely higher than those derived from score-based (Figures 1a & 1b) or reference-genome-based (Supplementary Fig. 1b) methods. We attribute this observation to fundamental differences among the methods: Phred scores estimate the fidelity with which a given technology has made a base call. This measure is an exclusive product of error introduced by the actual act of sequencing. No information regarding error introduced by procedures that precede sequencing is

provided. As an example, amplification is commonly used to produce sufficient quantities of material for sequencing from an initial sample. Even with high fidelity enzymes (e.g. Taq or Φ29 DNA polymerase) each amplification product may contain errors, deviations from the original biological template. Successive amplification(s) propagate previous errors and introduce new ones, leading to an increasingly divergent population of reads. Deviations from the original biological template constitute a type of error that is undetectable with score-based methods (see Supplementary Fig. 3). Reference-genome-based methods can detect such errors but cannot be applied to metagenomic data, and they suffer from additional artifacts introduced by biased selection procedures that can deflate error estimates. In particular, conventional reference-genome-based methods frequently discard duplicate reads and reads that fail to meet arbitrary standards with respect to identity/alignment with the relevant reference genome. This strategy can constrain calculated errors to artificially low values. The most error prone reads, those that do not align well with the reference genome and would contribute significantly to observed errors, are not considered.

We also used DRISEE to provide data regarding error type. Figure 2 presents all error types together (total error) as well as a breakdown of each error type (A,T,C, and G substitutions and insertion/deletion errors) observed across metagenomic 454 (65 samples) and Illumina (159 samples) data. The results are consistent with previous observations in genomic shotgun sequencing: Illumina data are dominated by substitution-based errors[14], whereas 454 data exhibit a majority of insertion/deletion errors[11] (Figs. 2a and 2b). No other method provides estimates with respect to error type in metagenomic shotgun data.

DRISEE provides a more complete estimate of sequencing error than is possible with score-based methods, one that accounts for error introduced at any procedural step in a sequencing protocol, from collection of a biological sample to extraction of DNA/RNA, intermediary processing of the extracted material and, finally, sequencing itself (see Supplementary Fig. 3). Distinct profiles were observed when samples produced with the same technology were grouped by experiment, suggesting the presence of technological or lab-dependent errors (Fig. 1c). Even finer distinctions are observable among the error profiles for single samples taken from the same experiment (Fig. 1d). Based on its unbiased treatment of duplicate reads—reads are not screened against an external standard—DRISEE estimates of error are much less constrained by identity based filters than conventional reference-genome-based methods and are therefore capable of measuring a much broader range of errors.

Arguably, DRISEE has some limitations. At present, it is not applicable to eukaryotic data, sequences with low complexity, and/or known sequences that may exhibit an unusually high level of biological repetition, particularly amplicon ribosomal RNA data. These types of data are likely to meet DRISEE requirements for prefix length and abundance, but represent real biological variation that could be misinterpreted by DRISEE as sequencing error. Moreover, DRISEE operates on artifactually duplicated reads—an approach that works well with current platforms such as 454 and Illumina but may require procedural modifications (such as the intentional inclusion of highly abundant sequence standards) if future developments eliminate artifactually duplicated reads.

In summary, DRISEE provides accurate assessments of sequencing error of metagenomic and genomic data, accounting for error type as well as frequency. DRISEE error profiles can be used to explore correlations between sequencing error and metadata, allowing investigators to differentiate experimentally meaningful trends from artifacts introduced by previously uncharacterized sequencing error. Traditional score-based and reference-genome-based methods do not allow for such observations with respect to shotgun metagenomic data. DRISEE also offers the advantage that it requires no data other than an input FASTA file. Moreover, DRISEE considers error independent of sequencing platform, without prior knowledge. These characteristics make DRISEE a promising method—particularly with respect to the enormous quantities of shotgun-based metagenomic data that are anticipated in the near future.

DRISEE is available via MG-RAST. We also provide code that allows users to perform DRISEE analyses independent of MG-RAST (ftp://ftp.metagenomics.anl.gov/DRISEE/).

Methods:

Overview:

Duplicate Read Inferred Sequencing Error Estimation (DRISEE) is a method that can be applied to sequence data produced from any sequencing technology. It provides an error profile that can be used to explore the sequencing error, as well as biases in error, that are present in a single sequencing run or any group of sequencing runs. The latter capability enables the user to produce error profiles specific to a particular sequencing technology, sample preparation procedure, or sequencing facility—in short, to any quantified variable (i.e., metadata) related to the sequencing of a given set of samples.

DRISEE exhibits several desirable characteristics that are not found in the most widely utilized methods to quantify sequencing error: *reference-genome-based methods* that rely on comparison to standard sequences (generally a published sequenced genome): and quality *score-based* methods that rely on sophisticated, platform-dependent models of error to derive base calls with affiliated confidence estimates (Q or Phred scores) for each sequenced base. DRISEE can be applied to metagenomic or genomic data produced with any sequencing technology and requires no prior knowledge (i.e., reference genomes or platform–dependent error models).

DRISEE relies on the occurrence of artifactually duplicated reads—nearly identical sequences that exhibit abundances that greatly exceed expectations of chance, even when a modest amount of possible biological duplication is taken into account. Illumina and 454 platforms exhibit a well documented[7,8], but poorly understood, propensity to produce large numbers of duplicate reads. DRISEE utilizes this artifact as a means to create internal sequence standards that can be used to assess error within a single sample, or across multiple samples. We identify duplicate reads as those that

exhibit an identical prefix (prefix = the first l bases of a read) at some threshold abundance (n) that exceeds chance expectations, even those that take biological duplication into account. The precise values of l (prefix length) and n (prefix abundance) can be varied to accommodate the scale of any sequencing technology. In the work presented here, bins (groups) of duplicate reads were used to calculate error values if they exhibited and identical prefix length (l) of 50 bases with an abundance (n) of 20 or more reads. This requirement is stringent enough to justify assumptions of biological and statistical uniqueness; indeed, such an occurrence is extremely unlikely:

$$=1 \quad *4 \quad = 120*450 = 4E\text{-}32$$

where p is the probability that a prefix of length l (50bp) will be observed n (20) times; 4 represents the number of possible bases (A, T, C, and G). Even in data that are Illumina scale (on the order of 1 million reads per run), a chance observation of 20 reads that exhibit the same 50bp prefix is so unlikely as to be safely deemed improbable.

DRISEE exhibits a universality that other methods lack, but only if the data under consideration meet a few criteria: (1) Data must be in FASTA format. (2) There must be enough duplicate reads to safely infer that they are the product of artifact and not of real biological variation. (3) Input sequence data should not be culled, trimmed, or modified in any way by sequencer processing software. (4) Data under consideration should be the product of random (i.e. shotgun) sequencing. (5) Amplicon data—specifically, directed sequencing of amplicon ribosomal RNA data, are not suitable for DRISEE analysis; reads start with highly conserved regions (primer target sites) followed by regions that exhibit a large degree of real biological variation (the hypervariable regions) that would be misinterpreted as error by DRISEE.

Typical DRISEE workflow (see Supplementary Figure 4)

DRISEE was designed primarily as a means to assess metagenomic sequencing error in samples submitted to MG-RAST. DRISEE is available via MG-RAST-based analyses of shotgun metagenomic data. We also provide MG-RAST independent code that will allow users to perform DRISEE analyses on their own. Below we outline the steps in a typical DRISEE workflow. All analyses presented in the accompanying manuscript utilized the default values and arguments mentioned below.

1. *Sequence data in FASTA format*

In an effort to keep DRISEE as platform-independent as possible, it considers data in nearly ubiquitous FASTA format. Tools are described elsewhere for conversion to FASTA (http://maq.sourceforge.net/fq_all2std.pl). DRISEE specifically avoids use of Phred scores or any other platform-dependent error metric as a means to sort, cull, trim, or assess reads. Typical input includes all sequenced reads produced in a single sequencing run or sample.

2. *Check for random sequencing*

DRISEE was designed to consider sequencing data generated by random (shotgun) procedures. While it can be used to explore variation in amplicon-based data, such variation cannot be equated with sequencing error. Amplicon-based data are identified through either available metadata or the application of a tool that identifies amplicon and/or randomly sequenced samples based on an analysis of their prefix entropy (i.e. the Shannon[15] entropy of the distributions of prefixes of successive lengths). Samples that exhibit a nonrandom sequence patterns in their prefix, consistent with expectations of amplicon samples, are excluded from consideration by DRISEE.

3. *Screen reads for length and the presence of ambiguous base calls*

Sequence data undergo a mild two-step filtering procedure[2,9] to remove two types of reads:

A. Reads that exhibit uncharacteristic lengths (the default setting is to remove those that exhibit lengths farther than two standard deviations from the mean).

B. Reads with ambiguous base calls (the default setting is to remove reads that contain any ambiguous base calls).

4. *Bin reads based on the presence of identical prefixes*

A list of unique prefixes in the screened FASTA file is generated. All remaining reads are then grouped (or binned) according to their prefix. Any prefix length can be used; the current default is 50 bases.

5. *Screen bins by abundance*

Groups (or bins) of prefix identical reads are sorted with respect to read abundance. Bins with a minimum number of reads are processed further. Any minimum can be used; the current default is 20 reads. DRISEE can consider all such bins in a sample, or can consider any arbitrary number of such bins. DRISEE analyses presented in the accompanying manuscript considered no more than 1,000 randomly selected bins for each sample. DRISEE can use all reads in a given bin, or any arbitrary number of reads that is equal to or larger than the required minimum number of reads. DRISEE analyses presented in the accompanying manuscript considered all reads for bins that contained between 20 and 1,000 reads. A random selection of 1,000 reads was considered for any bin that had more than 1,000 reads. Results derived from complete DRISEE analyses (consideration of all reads in all bins that met DRISEE requirements) and analyses that

used a maximum bin limit of 1,000 with a read limit of 1,000 exhibited little appreciable difference (data not shown). A variety of alternative selection and/or bootstrap procedures are possible with DRISEE analyses.—these are not discussed here..

6. *Screen for eukaryotic or other suspect content*

Bins can now be screened for eukaryotic content, sequences with low complexity, and/or known sequences that may exhibit an unusually high level of biological repetition (16s rRNA-based, sequences with low complexity, eukaryotic sequences etc.). Bins that contain such sequences should be excluded from further consideration.

7. *End-trim reads*

Reads within a bin can be trimmed to a uniform length; the default operation is to trim reads to the length of the shortest read in their respective bin. Trimming can be performed by using other criteria or can be avoided altogether.

8. *Construct consensus sequences via multiple alignment*

All reads for a given bin undergo multiple alignment (UCLUST[16] – with additional custom scripts). The multiple alignments are used to generate a consensus (i.e. Steiner) sequence. In the first iteration, a seed sequence is randomly chosen from all reads in the bin. In the second and all subsequent iterations, the consensus sequence generated from the previous iteration is used as the seed. This process is iterated until it achieves convergence (defined as three consecutive iterations with no change in cluster identity) or once a pre-determined iteration limit has been achieved (by default 10; in practice, bins rarely require more than the first two to three iterations to achieve convergence).

9. *Compare bin reads with the bin consensus sequence*

Each individual read in a bin is compared, base-by-base, with the final consensus sequence for that bin. Only the non-prefix portion of the reads is considered (bases 51 and onwards in the work presented in the accompanying manuscript). At each position (with respect to the consensus sequence) a read is scored as one of six matches (A, T, C, G, insertion, or deletion) or one of six mismatches (A substitution, T substitution, C substitution, G substitution, insertion, or deletion). Insertions and deletions that span more than one base are scored as multiple mismatches (e.g. a deletion of three bases is scored as three mismatches).

10. *Construct bin-level DRISEE profiles*

Deviations and matches for all reads in a bin are tallied with respect to position in the consensus sequence. The consensus position indexed table of matches and mismatches for the bin represent its DRISEE profile.

11. *Construct run- or sample-level DRISEE profiles*

DRISEE profiles for all considered bins in a given run can be combined (summed) to produce a DRISEE error profile for the entire run or sample (see Supplementary Fig. 1 for an example).

12. *Sample Group level DRISEE profile construction*

DRISEE profiles for all bins in a group of runs can be combined (summed) to produce a DRISEE error profile for any group of runs/samples.

DRISEE profiles and the data they contain can be visualized directly, or processed further to generate detailed analyses of the information they contain.

Construction and Analysis of Simulated Data

Datasets with known rates of error were generated with Metasim[17]. Each artificial data set contained 100,000 sequences (average length of 252bp) randomly selected from the first 2,500 bases of the *E. coli* K-12 genome (Refseq accession number NC_000913) and randomly corrupted with errors of known type and frequency. Data sets represented a variety of substitution (0 to 2.3%) and insertion and deletion (0 to1.5% each) rates. Data sets contained substitution only, insertion and deletion only, or a combination of error types. Simulated data sets were processed with DRISEE as if they were real data sets. DRISEE detected errors were compared with known simulated rates, with respect to frequency and type (see Supplementary Fig. 1a).

Comparative Snalyses of Real Genomic Sequencing Data

Real genomic sequencing data were acquired from a previously reported study that used reference-genome-based methods to explore sequencing error in several single genome sequencing samples[8]. DRISEE was applied to data sets only if they met the criteria used with all other data considered in the accompanying manuscript: one or more bin(s) of reads that exhibit 20 or more reads with an identical 50-base prefix. A total of 12 samples obtained from SRA (http://www.ncbi.nlm.nih.gov/sra) met these criteria: SRR013470, SRR001574, SRR007446, SRR013433, SRR013137, SRR000266, SRR006411, SRR018125, SRR017783, SRR013477, SRR013431, SRR013382. These samples were analyzed with the reference-genome-based method described in Niu *et al.*[8] and with DRISEE. Errors derived from the originally reported reference-genome-based method and DRISEE were compared (see Supplementary Fig. 1b).

Data Sources

Unless otherwise indicated, data sets examined in this study were obtained via SRA or MG-RAST. Supplementary Table 2 contains a complete list of sequence data used in the accompanying manuscript. Datasets are referenced by their SRA (http://www.ncbi.nlm.nih.gov/sra), MG-RAST (http://metagenomics.anl.gov/), or both identifiers/accession numbers.

Sofware availability

An MG-RAST independent version of DRISEE code can be downloaded via ftp (ftp://ftp.metagenomics.anl.gov/DRISEE/) (see Supplementary Table 2).

SUPPLEMENTARY TEXT

A Word on Reference-Genome and Score-Based Error Methods

Reference-genome-based methods compare sequenced reads to preexisting standards (published genomes). Samples are typically cultured from a clonal isolate for which a reliable reference is available. Where the best available reference is a related strain or species, real biological variation can be mistaken for sequencing error[8,9,11,18]. Shotgun metagenomic data are excluded from consideration by such methods because of their diversity (samples contain a broad spectrum of species) and the absence of adequate reference data (many species have no appropriate reference genome(s), and reference *metagenomes* do not exist).

Score-based methods use an alternative approach. Sequencer signals are compared with sophisticated, frequently proprietary, probabilistic models that attempt to account for platform-dependent artifacts, generating base calls, each with an affiliated quality (Phred or Q) score providing an estimate of error frequency, but no information about error type. Although these methods can be applied to metagenomic data, they present a challenge to investigations that require knowledge regarding error type. For example, similarity-based gene annotation is extremely sensitive to frame-shifting insertion/deletion errors but only moderately affected by substitutions[1]; in this instance the ratio of insertion and/or deletions to substitutions provides critical information, unattainable with conventional Q scores. The absence of information regarding error type is an even greater concern in light of documented platform-dependent biases in sequencing error: Illumina-based sequencing exhibits high substitution rates[14], whereas 454 technologies exhibit a preponderance of insertion/deletion errors[11]; identical Q scores

from these two technologies are likely to represent different types of error, rendering nominally comparable scores incomparable[10,11,19-23].

The limitations of reference-genome and score-based methods are exacerbated by the increasing democratization (http://www.technologyreview.com/biomedicine/26850/) of high-throughput sequencing technologies and the rapid proliferation of projects[24-27] (www.1000genomes.org, www.commonfund.nih.gov/hmp, www.earthmicrobiome.org) that employ them. This includes an increasing trend toward meta-analyses (studies that consider data from multiple sources) to examine collections of samples that can exhibit a diverse technical provenance[7,28-30]. Meaningful comparisons of technically diverse sequence data require accurate and platform-independent measures of sequencing error, such that *bona fide* observations can be differentiated from background noise.

References

1. K. J. Hoff, *BMC Genomics* 10, 520 (2009).
2. V. Kunin, A. Engelbrektson, H. Ochman et al., *Environ Microbiol* 12 (1), 118 (2010).
3. C. Quince, T. P. Curtis, and W. T. Sloan, *The ISME journal* 2 (10), 997 (2008).
4. C. Quince, A. Lanzen, T. P. Curtis et al., *Nat Methods* 6 (9), 639 (2009).
5. J. G. Caporaso, J. Kuczynski, J. Stombaugh et al., *Nature methods* 7 (5), 335 (2010).
6. S. M. Huse, D. M. Welch, H. G. Morrison et al., *Environ Microbiol* 12 (7), 1889 (2010).
7. V. Gomez-Alvarez, T. K. Teal, and T. M. Schmidt, *ISME J* 3 (11), 1314 (2009).
8. B. Niu, L. Fu, S. Sun et al., *BMC Bioinformatics* 11, 187 (2010).
9. S. M. Huse, J. A. Huber, H. G. Morrison et al., *Genome Biol* 8 (7), R143 (2007).
10. M. Margulies, M. Egholm, W. E. Altman et al., *Nature* 437 (7057), 376 (2005).
11. A. R. Quinlan, D. A. Stewart, M. P. Stromberg et al., *Nat Methods* 5 (2), 179 (2008).
12. Y. Sun, Y. Cai, L. Liu et al., *Nucleic acids research* 37 (10), e76 (2009).
13. B. Ewing and P. Green, *Genome research* 8 (3), 186 (1998).
14. J. C. Dohm, C. Lottaz, T. Borodina et al., *Nucleic Acids Res* 36 (16), e105 (2008).
15. C. E. Shannon, *MD Comput* 14 (4), 306 (1997).
16. Robert C. Edgar, UCLUST v3.0 (2011).
17. D. C. Richter, F. Ott, A. F. Auch et al., *PLoS One* 3 (10), e3373 (2008).
18. M. P. Cox, D. A. Peterson, and P. J. Biggs, *BMC Bioinformatics* 11, 485 (2010).
19. H. C. Bravo and R. A. Irizarry, *Biometrics* 66 (3), 665 (2010).
20. P. J. Cock, C. J. Fields, N. Goto et al., *Nucleic Acids Res* 38 (6), 1767 (2010).
21. T. D. Harris, P. R. Buzby, H. Babcock et al., *Science* 320 (5872), 106 (2008).
22. W. C. Kao, K. Stevens, and Y. S. Song, *Genome Res* 19 (10), 1884 (2009).
23. K. J. McKernan, H. E. Peckham, G. L. Costa et al., *Genome Res* 19 (9), 1527 (2009).
24. D. H. Huson, A. F. Auch, J. Qi et al., *Genome Res* 17 (3), 377 (2007).
25. V. M. Markowitz, N. N. Ivanova, E. Szeto et al., *Nucleic Acids Res* 36 (Database issue), D534 (2008).
26. M. J. Pallen, N. J. Loman, and C. W. Penn, *Curr Opin Microbiol* 13 (5), 625 (2010).

27. R. Seshadri, S. A. Kravitz, L. Smarr et al., *PLoS Biol* 5 (3), e75 (2007).
28. E. A. Dinsdale, R. A. Edwards, D. Hall et al., *Nature* 452 (7187), 629 (2008).
29. S. G. Tringe, C. von Mering, A. Kobayashi et al., *Science* 308 (5721), 554 (2005).
30. C. von Mering, P. Hugenholtz, J. Raes et al., *Science* 315 (5815), 1126 (2007).

Acknowledgements
This work was supported in part by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357. The authors would like to thank Dionysios A. Antonopoulos, Gail W. Pieper, and Jennifer F. Salazar for comments on the manuscript, as well as Elizabeth M. Glass and the MG-RAST development team for technical support.

Argonne License to be removed at time of publication
The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

Affiliations
Argonne National Laboratory, Argonne IL, United States.
Kevin P. Keegan, William L. Trimble, Jared Wilkening, Andreas Wilke, Travis Harrison, Mark D'Souza, Folker Meyer

University of Chicago, Chicago IL, United States.
Kevin P. Keegan, Jared Wilkening, Andreas Wilke, Travis Harrison, Mark D'Souza, Folker Meyer

Institute for Genomics and Systems Biology, Chicago IL, United States.
Kevin P. Keegan, Jared Wilkening, Andreas Wilke, Travis Harrison, Mark D'Souza, Folker Meyer

Contributions
K.P.K., W.L.T., J.W., A.W., T.H., M.D. wrote several of the individual software components of DRISEE and performed code optimization. K.P.K., W.L.T., and F.M. designed research and helped to author this manuscript. J.W., A.W. offered comments on this manuscript. K.P.K. and W.L.T. performed data analyses. K.P.K. was the principal developer of the DRISEE method and all related software and was the principal author of this manuscript.

Competing financial interests
The authors have no competing financial interests.

Corresponding author
Correspondence to: Folker Meyer (folker@anl.gov)

Supplementary Table 1: Excerpt from a whole sample/run DRISEE error profile table

ID 4462612.3

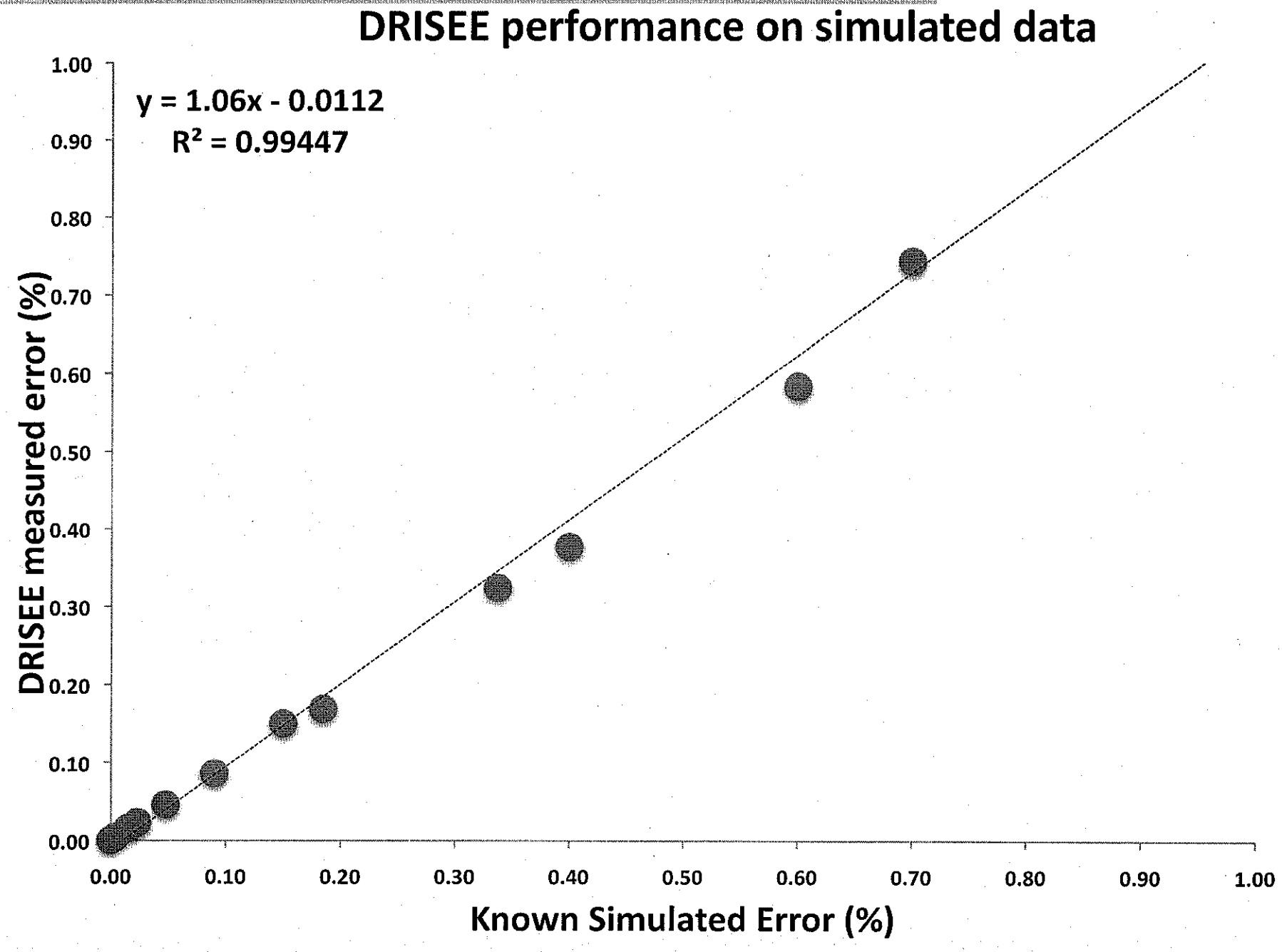
Summary:

A_subst	T_subst	C_subst	G_subst	InDel	Total_err	prefix_length = 50
0.1436	0.0961	0.1441	0.0907	0.0053	0.4798	

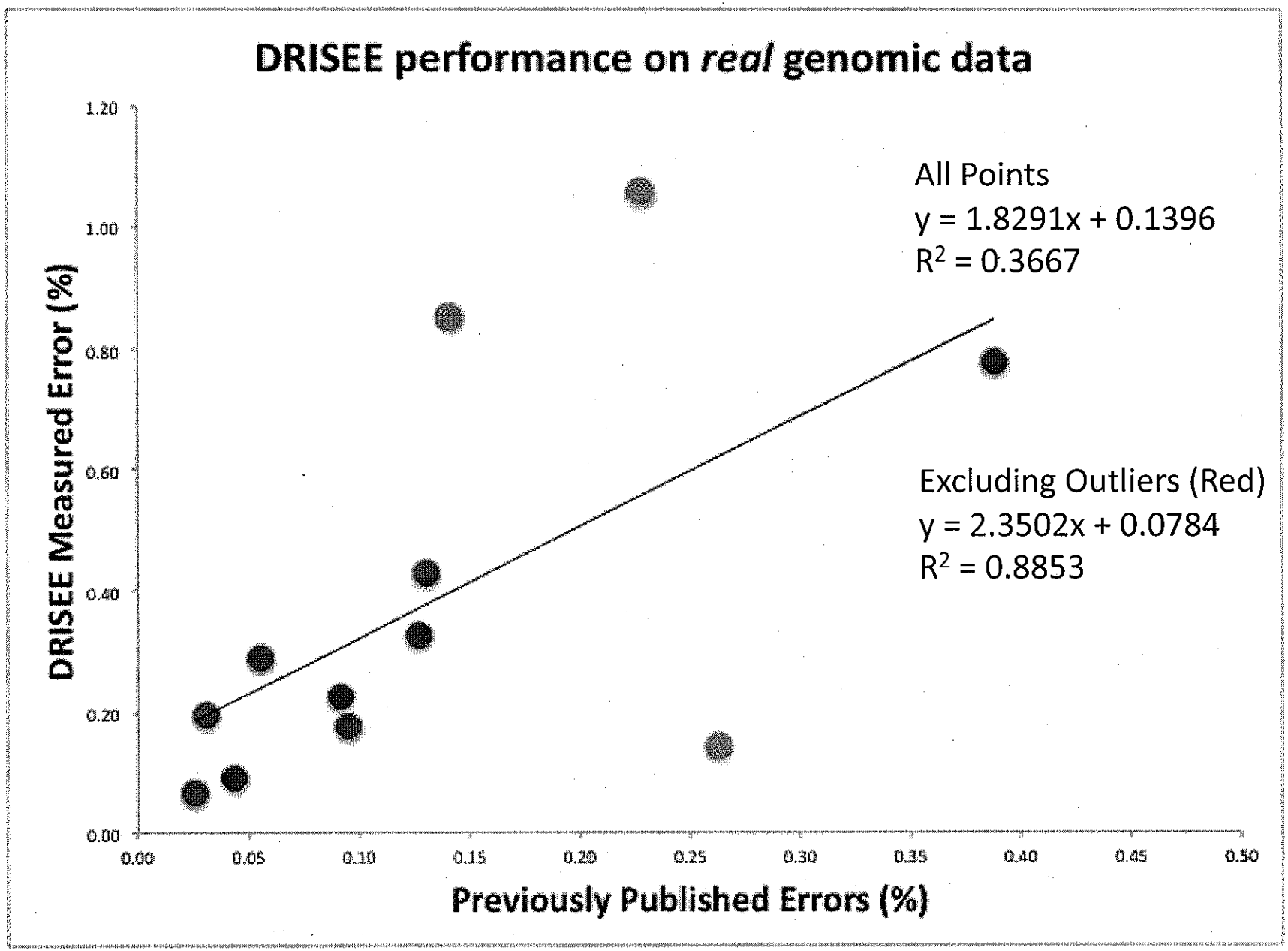
bp counts

Consensus Position	Match Consensus					Do Not Match Consensus				
	A	T	C	G	InDel	A	T	C	G	InDel
1	8646	5849	6119	11508	0	0	0	0	0	0
2	15800	6418	2782	7122	0	0	0	0	0	0
3	6897	15562	2685	6978	0	0	0	0	0	0
4	8820	7475	10374	5453	0	0	0	0	0	0
5	10454	7590	2645	11433	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
75	6320	5899	2543	2546	0	5	17	26	1	0
76	5760	5997	2548	2476	0	75	64	67	88	0
77	5912	5884	2431	2579	0	9	3	37	19	0
78	5228	6069	2414	2650	0	38	28	33	53	1
79	5511	5402	2550	2664	0	43	15	27	0	0
...	...	...	...	...	...	...	...	...	...	...

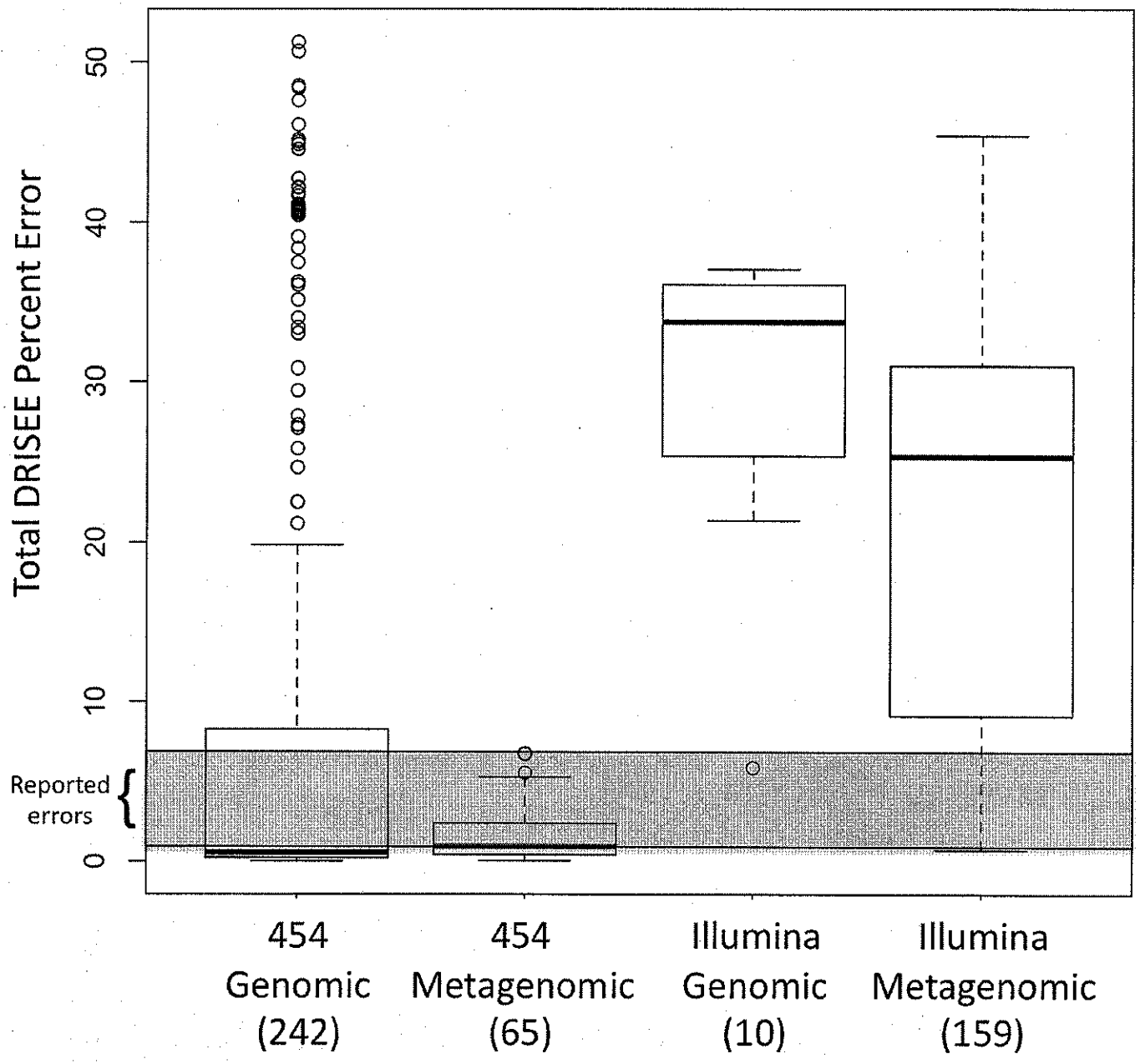
Supplementary Figure 1a: DRISEE performance on simulated data



Supplementary Figure 1b: DRISEE performance on real genomic data



Supplementary Figure 2: Total DRISEE errors of genomic and metagenomic data produced by 454 and Illumina technologies



Supplementary Figure 3:
Ability of different methods to detect error introduced by procedural steps in a typical sequencing protocol

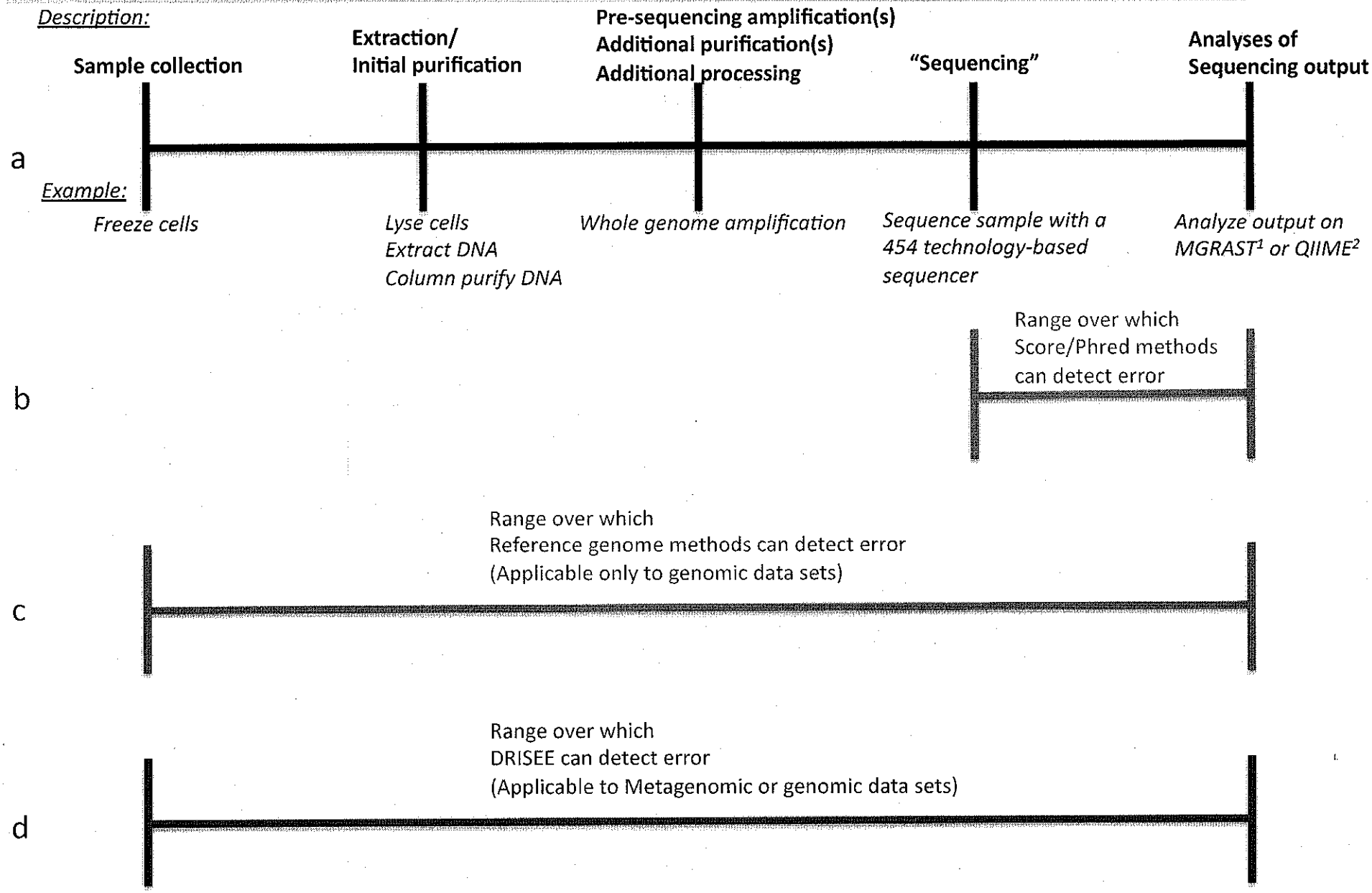


Figure 1: DRISEE error profiles for metagenomic sequencing data sets

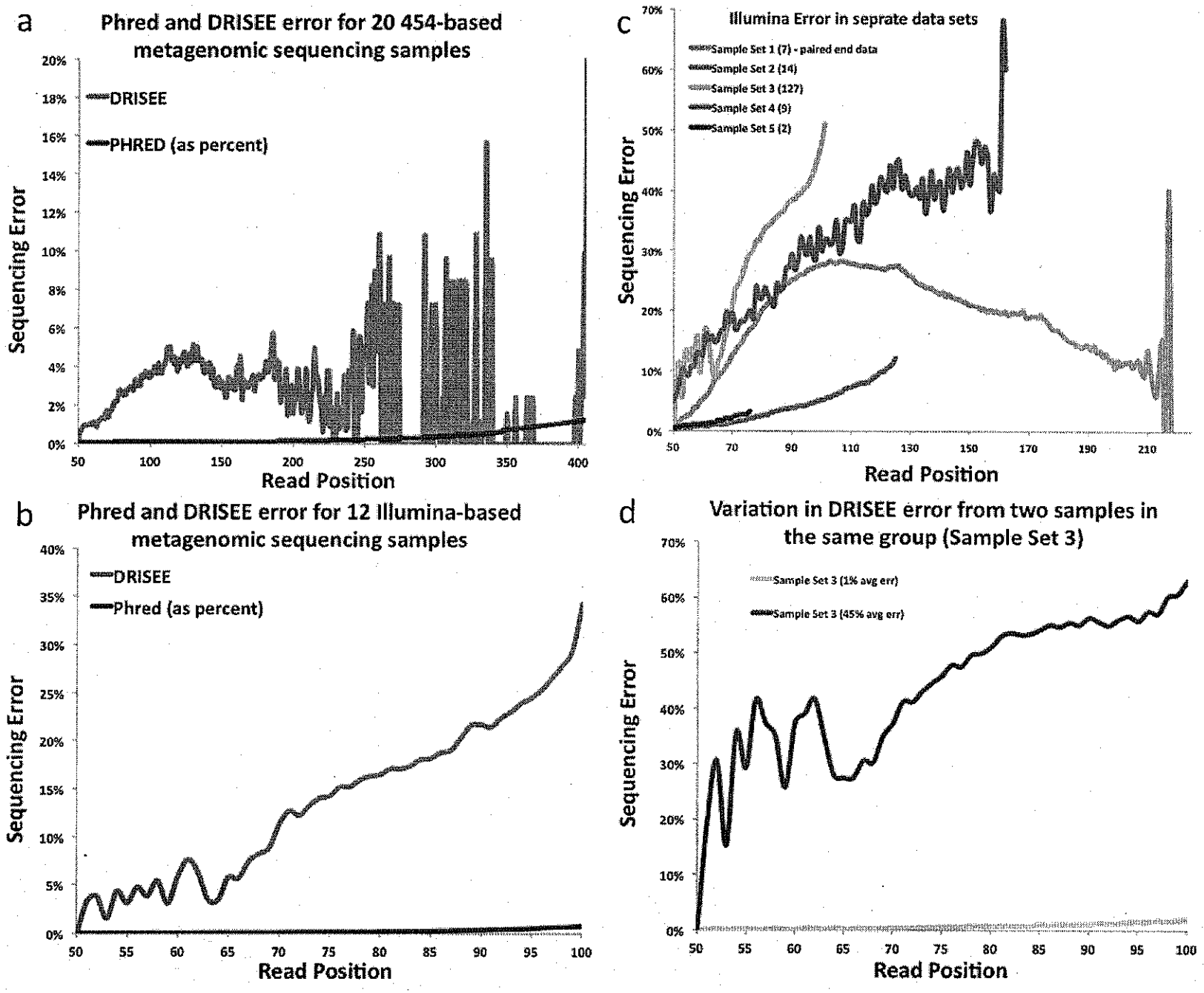
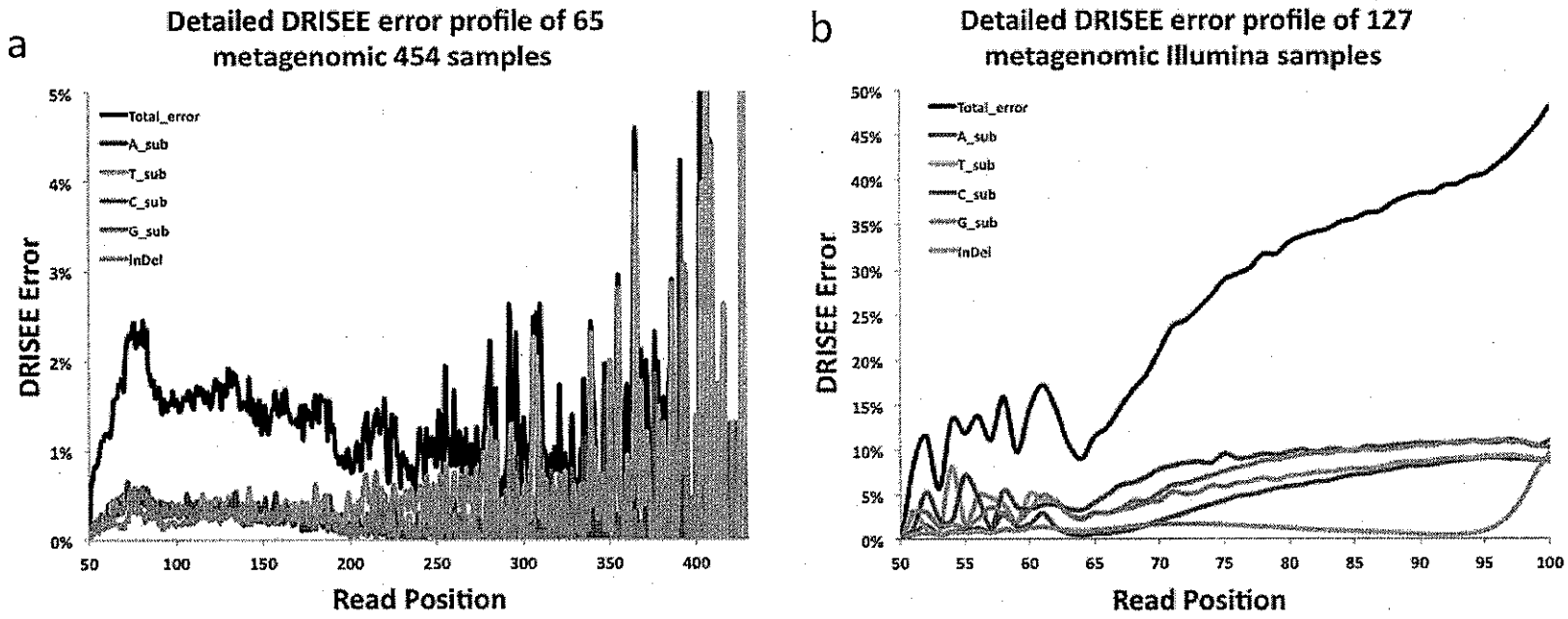
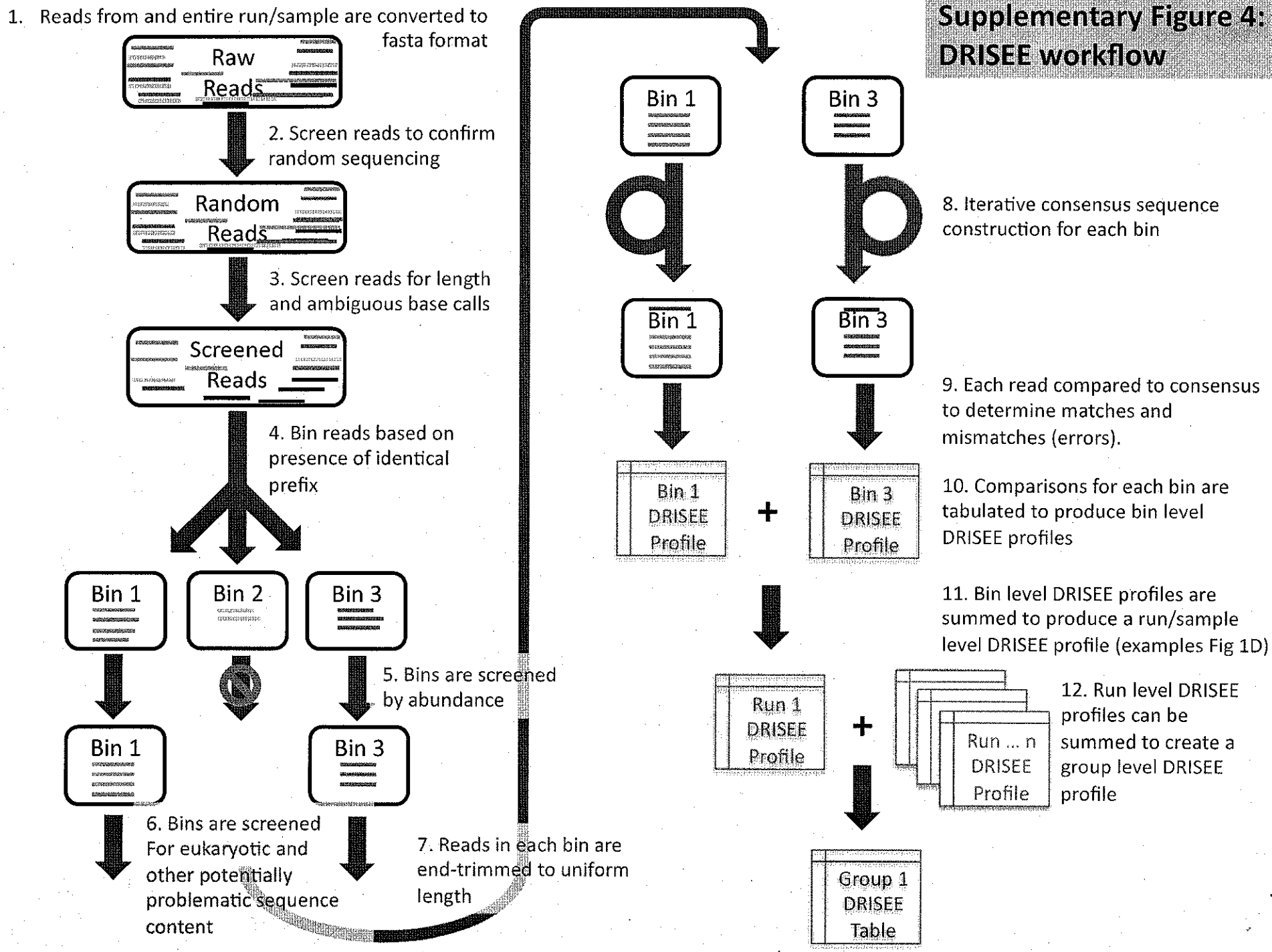


Figure 2: DRISEE calculated Errors, separated by error type, for 454 and Illumina metagenomic samples



Supplementary Figure 4: DRISEE workflow



(a)

MG-RAST ID	SRA
4443703.3	SRR000673
4443706.3	SRR000684
4443707.3	SRR000678
4443708.3	SRR000690
4443709.3	SRR000696
4443711.3	SRR001663
4443718.3	SRR000287
4443719.3	SRR000288
4443721.3	SRR000282
4443723.3	SRR000283
4443765.3	SRR001663
4444077.3	ERR010489
4444083.3	ERR010496
4445065.3	ERR010500
4445066.3	ERR010494
4445067.3	ERR010492
4445068.3	ERR010498
4445069.3	ERR010483
4445070.3	ERR010486
4445081.3	ERR010482

(b)

MG-RAST ID	SRA
4462766.3	SRR061437
4462771.3	SRR061442
4462772.3	SRR061443
4462773.3	SRR061444
4462775.3	SRR061446
4462777.3	SRR061448
4462780.3	SRR061451
4462782.3	SRR061453
4462783.3	SRR061454
4462784.3	SRR061455
4462786.3	SRR061457
4462788.3	SRR061459

(c)

MG-RAST ID	SRA
4460638.3*	
4460639.3*	
4460640.3*	
4460641.3*	
4460642.3*	
4460643.3*	
4460644.3*	
4460638.3*	
4460638.3*	
4460642.3*	
4460642.3*	
4460643.3*	
4460643.3*	
4460644.3*	
4460644.3*	
4460639.3*	
4460639.3*	
4460640.3*	
4460640.3*	
4460641.3*	
4460641.3*	
	SRR061914
	SRR061471
	SRR061478
	SRR061534
	SRR061577
	SRR061591
	SRR061535
	SRR061576
	SRR062031
	SRR061476
	SRR061477
	SRR061488
	SRR062010
	SRR062064
	SRR061485
	SRR061468
	SRR062030
	SRR061466
	SRR061469
	SRR061479

SRR062007
SRR062069
SRR061538
SRR061935
SRR061943
SRR061938
SRR061539
SRR061976
SRR062068
SRR061461
SRR061464
SRR061467
SRR061462
SRR061941
SRR062027
SRR061922
SRR061489
SRR061518
SRR061490
SRR062006
SRR061956
SRR062053
SRR061519
SRR061492
SRR061590
SRR061560
SRR062026
SRR061491
SRR061514
SRR061500
SRR061444
SRR061486
SRR061437
SRR061484
SRR061465
SRR061498
SRR062065
SRR061940
SRR061923
SRR061454
SRR061502
SRR061470
SRR062044

SRR061522
SRR061939
SRR061501
SRR061960
SRR061480
SRR061562
SRR061455
SRR061460
SRR061447
SRR061474
SRR061445
SRR061902
SRR061440
SRR061515
SRR062048
SRR061453
SRR062045
SRR061587
SRR062049
SRR061523
SRR061457
SRR061482
SRR061499
SRR061481
SRR061541
SRR061586
SRR061463
SRR061475
SRR061918
SRR062011
SRR061957
SRR061561
SRR061915
SRR061977
SRR061483
SRR061472
SRR061458
SRR061446
SRR062013
SRR061919
SRR061695
SRR061694
SRR061449

SRR061934
SRR061494
SRR061497
SRR061441
SRR061493
SRR061487
SRR061495
SRR061451
SRR061503
SRR061496
SRR061448
SRR062052
SRR061903
SRR061961
SRR061556
SRR061540
SRR061557
SRR061456
SRR061459
SRR061443
SRR061442
4460167.3
4460175.3
4460260.3
4460261.3
4460262.3
4460265.3
4460266.3
4460267.3
4460268.3
4465820.3
4465823.3

* Data sets were used as me

Sample Set	Notes
Sample Set 1	Merged
Sample Set 1	Merged
Sample Set 1	Merged
Sample Set 1	Merged
Sample Set 1	Merged
Sample Set 1	Merged
Sample Set 1	Merged
Sample Set 2	unmerged 1 of 2
Sample Set 2	unmerged 2 of 2
Sample Set 2	unmerged 1 of 2
Sample Set 2	unmerged 2 of 2
Sample Set 2	unmerged 1 of 2
Sample Set 2	unmerged 2 of 2
Sample Set 2	unmerged 1 of 2
Sample Set 2	unmerged 2 of 2
Sample Set 2	unmerged 1 of 2
Sample Set 2	unmerged 2 of 2
Sample Set 2	unmerged 1 of 2
Sample Set 2	unmerged 2 of 2
Sample Set 2	unmerged 1 of 2
Sample Set 2	unmerged 2 of 2
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	
Sample Set 3	

(d)

MG-RAST ID	SRA	Notes
	SRR061442	avg err = 1%
	SRR061914	avg err = 45%

Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3

Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3

Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 3
Sample Set 4
Sample Set 4
Sample Set 4
Sample Set 4
Sample Set 4
Sample Set 4
Sample Set 4
Sample Set 4
Sample Set 5
Sample Set 5

erged (Sample 1) and umerged (Sample 2) paired end