White Paper

DNA Error Quantification

Introduction

Elegen's virtual cloning process provides a new class of DNA product, delivering long, high-purity DNA with market leading turnaround time. We accomplish low error rates in long DNA production by using innovative technology and processes that circumvent the bottlenecks in traditional DNA production. To establish confidence and assurance of DNA quality, we have developed a method to quantify error rates based on unique molecular identifiers (UMIs)[1,2,] and we describe the method and our observations below.

Sequencing stands out among DNA error-detection technologies because it not only counts the errors, but also provides the location and identity of the error. However, the sensitivity to errors in sequencing is limited by the rate of errors introduced by sequencing. An error in a molecule must be at higher level than sequencing noise or it cannot be attributed confidently to the input DNA. The error rate of DNA synthesis is reported to range from 1:1000 bp for standard commercial phosphoramidite oligosynthesis products to 1:10^{10} bp for clonal DNA[3–5]. Current sequencing technologies have error rates of 1:10—1:1000 bp, all above the expected error rate of DNA products, making it impossible to measure error rates in DNA products directly with sequencing[6–10].

To quantify error rates with sequencing, we use UMIs to simultaneously capture the error rate in Elegen's long DNA product and elevate it above sequencing noise (Fig. 1). We label each molecule in a population using a highly diverse set of sequence tags, collect a sample of uniquely tagged molecules, and amplify the sample. The process creates millions of copies of each original molecule, and the UMI records the identity of the parent molecule in the original sample for every molecule in the amplification product. By sequencing the labeled, amplified sample, each read can be attributed to a parent molecule in the original sample by the UMI. If the parent molecule had a synthesis error, the reads attributed to that molecule by sequence tags will have the error at a higher level than the sequencing error rate, and the DNA error can be identified with confidence.

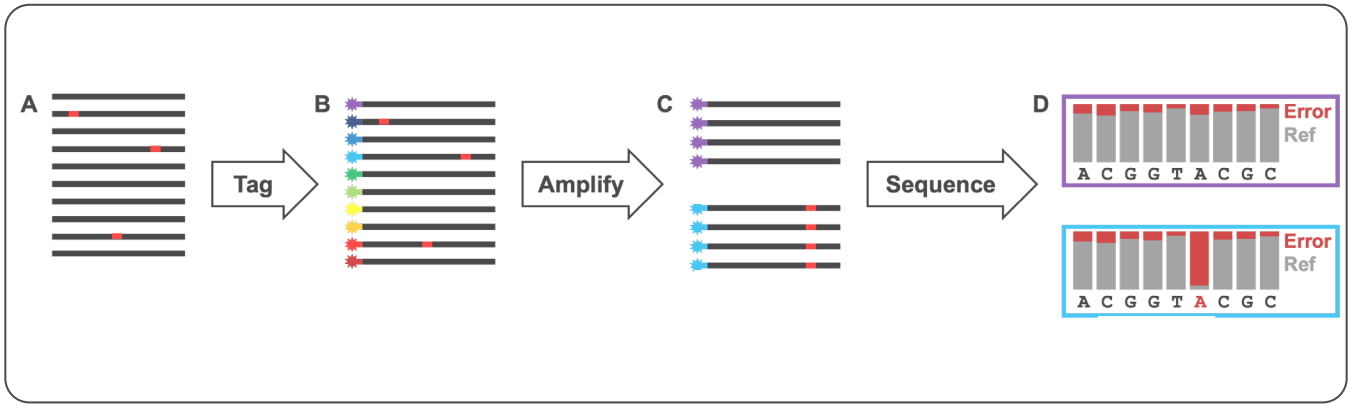

Figure 1. Elegen's method to quantify low abundance errors with sequencing. A) A population of DNA has low abundance errors. B) The molecules are labeled with UMIs. C) A tagged sample is amplified. D) Sequencing and analysis of tagged molecules reveal low abundance errors at levels above the sequencing error rate.

In this method, error-detection sensitivity is set by the sample size of tagged molecules rather than the sequencing error rate. As more molecules are sampled, the sensitivity and precision of the measurement increases (see Fig. 2, SI Methods).
Selecting ~100 tagged molecules that are typical gene length (1-5 Kbp) gives sensitivity to error rates of 1:10^4—1:10^5 bp. Increasing the sample size to 1000 molecules raises the sensitivity to 1:10^5—1:10^6 bp for the same gene lengths (Fig. 2A).

The sample size of tagged molecules also determines the precision of the error rate measurement, defined as the width of the 95% confidence interval. For example, a population error rate of 1:10^4 bp has a 95% confidence interval of an order of magnitude at sample sizes less than 100 (Fig 2B, SI Methods). The interval decreases to less than 2-fold with a sample size of 1000 molecules.

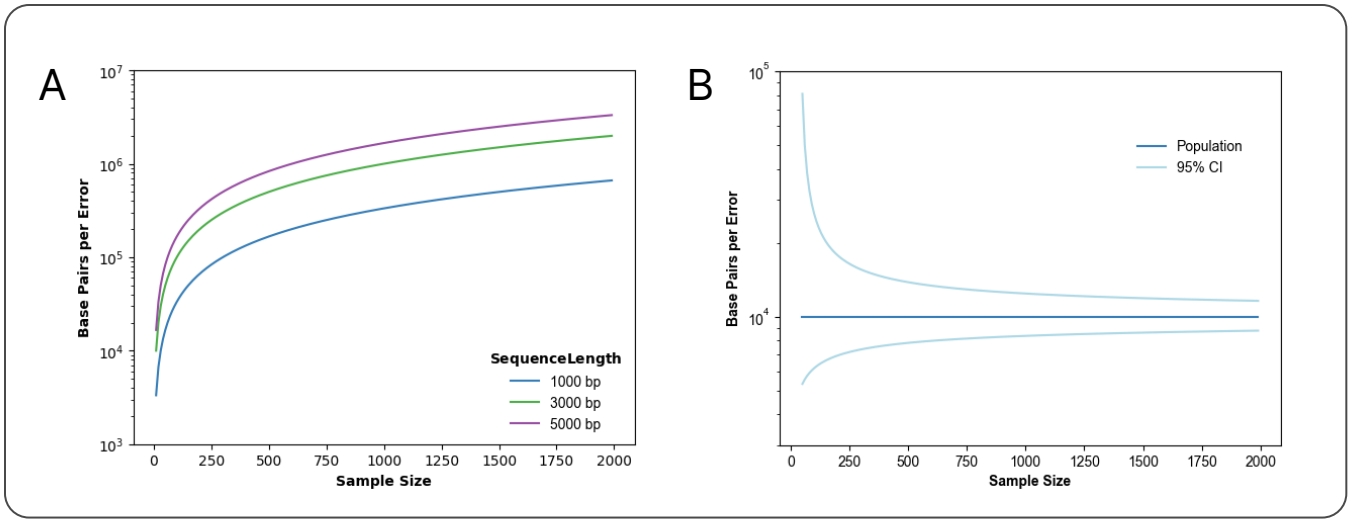

Figure 2. Error rate measurement dependence on the sample size of UMI-tagged molecules. A) The sensitivity of error rate measurement is a function of sample size and sequence length. B) The precision of error rate measurements, expressed as the width of the 95% CI, as a function of sample size.

Using this method, we have quantified the error rate for a large sample of virtually cloned DNA sequences synthesized for customers.

Methods

Unique Molecular Identifier (UMI) design. The UMI has the sequence YNNNRYNNNRYNNNRYNNNRYNNNR. The sequence was designed to perform well with Oxford Nanopore (ONT) sequencing technology and for a large range of sample sizes. The UMI characteristics that facilitate reading by ONT are the sequence pattern and length. The pattern of repeating units of YNNR prevents UMIs from having homopolymer sequences that are poorly read by ONT[8,10,11]. The length and sequence diversity of the UMI increase the confidence in aligning reads to the UMI. As an illustration, the UMI sequences in a sample of 1000 sequences have a mean edit distance of 13.4, which means that on average 13 or more bases need to be changed for one UMI to be equivalent to any other UMI sequence in the sample (SI Fig. 1). No UMIs have an edit distance of 3 or lower when compared to all other UMIs, and an edit distance of 3 is the high end of the expected errors for ONT sequencing. The high sequence diversity in the large UMI pool minimizes ambiguity in assigning reads to a UMI.

The design parameters to improve sequencing reliability also ensures that for the sample sizes we used, each UMI is assigned to a single sample molecule. The large number of mixed bases consists of 10^{12} different UMI sequences. The diversity of sequences means that for samples with as many as 10^5 molecules, the probability p is less than 10^{-5} for an UMI to be shared among more than one molecule (SI Methods). The low probability of shared UMIs across the large range of sample sizes allows us to use a large range of sample sizes and supports the fundamental assumption in this method of one molecule per UMI.

UMI-library preparation. An aliquot of each virtually cloned DNA sequence was mixed with a primer containing the UMI, polymerase, and PCR master mix. The reaction mixture was denatured, annealed, and the primer extended in a single cycle.

Following clean-up, the tagged molecules were diluted and a sample of molecules was collected for use in the following steps. The median sample size was 387 molecules (Q1 = 349, Q3 = 454). The sample was amplified with primers that bind outside of the product and UMI sequence, creating many copies of each UMI-tagged molecule. The sample amplification product was cleaned up and then prepared for sequencing following the sequencing vendor protocol.

Sequencing. The UMI libraries were prepared for sequencing using materials and methods provided by ONT. Sequencing adapters were added to the libraries and the sequencing solution was prepared using the version 10 chemistry Adapter Ligation Kit (LSK-110). The sequencing solutions were applied to Minion or Flongle flow cells with the version 9.4 pore. The signal-level fast5 files were base called with guppy using the high-accuracy models for the version 9.4 pore.

Sequencing analysis. Base-called sequencing data in fastq format was input to the analysis pipeline. The pipeline was executed using a combination of custom functions, python modules, and open-source sequencing tools. The open-source tools used in the pipeline are indicated in parentheses. First, UMI sequences within the sample were determined by compiling the candidate UMI sequences and filtering based on several criteria. The candidate UMI sequence pool is created by trimming the UMI region from all reads (cutadapt)[12], filtering out potential UMI sequences that did not have the expected length and YNNNR base pattern, clustering the remaining potential UMI sequences (usearch)[13], and removing potential UMI sequences that were within a minimum edit distance of each other. The final UMI sequences were then used to bin the full sequencing reads (bbmap)[14], where each bin contained all the reads with the same UMI, and thus represent a unique, single parent molecule from the original sample. The bins were then filtered to those with at least 25 reads and at least 25% of reads in each direction.
For the passing bins, the reads were used to create a consensus sequence (minimap2, racon, medaka)[15–17], and the consensus was compared to the expected reference sequence. The differences between the consensus and expected sequence were counted and recorded. Across all bins, the number and type of errors were summarized to give the error rate for the synthesized DNA, as well as information about the type of errors observed.

Pipeline validation with simulated data. We used simulated data to test the performance of the analysis pipeline. In silico, we added substitutions, deletions, and insertions in known locations to a reference sequence, followed by addition of UMIs and other accessory sequence introduced during library preparation to create a virtual sample of DNA molecules. The collection of known sequences was then used to simulate reads from ONT using the open-source software NanoSim[18]. This software trains a sequencing error model from real data, and then applies that model to a list of sequences to generate fastq files with insertions, deletions, substitutions, and other sequencing type errors at rates observed in real ONT data. For the simulation, each read name stored the identity of the original molecule and allowed tracking how the pipeline performed from beginning to end.

The pipeline was tested on a simulated data set with 5000 UMIs and ~ 50 reads per UMI. The pipeline went through all the steps described above, and the outcomes were compared to the known inputs. Over 99% (4962) of the expected 5000 UMIs were found, and less than 3% of reads were misassigned to UMI bins. The output error rate was within 0.85% of the input error rate. The test with simulated data demonstrated that the analysis pipeline handles data as expected, and error introduced by the analysis pipeline is negligible compared to sampling uncertainty.

Pipeline validation with reference DNA material. The analysis method was applied to a reference DNA sample to test the ability to quantify a known error rate. Reference genomic DNA was acquired (Horizon Discovery, Tru-Q 7, 1.3% Tier Reference Standard). A G to A substitution and a 15-base deletion in the EGFR gene (GRCh38 coordinates 7:55174014) had vendor quantified allelic frequencies of 16.7% and 1%, respectively. The alleles were amplified within a 4 kb amplicon, which was then treated with the UMI tagging and library preparation described above. The data were analyzed with the described method, and the expected variant was at a rate of 18 ± 6% (26 of 144 UMIs) and 0 + 3% (0 of 144 UMIs), consistent with the vendor quantified values. The agreement between our method and the benchmarked methods used by the vendor[19] supports that our method accurately returns the population error rate for DNA products.

Sequences to be characterized. 45 sequences produced by our commercial protocol for Elegen's Early Access Program were selected for error rate measurement. We chose sequences spanning characteristics including length, overall GC content, local GC content, and repeats. For each of these characteristics, we selected sequences that are in typical ranges as well as several that approach the boundaries of our sequence acceptance criteria, including local GC% and repeat length (see Fig. 3).

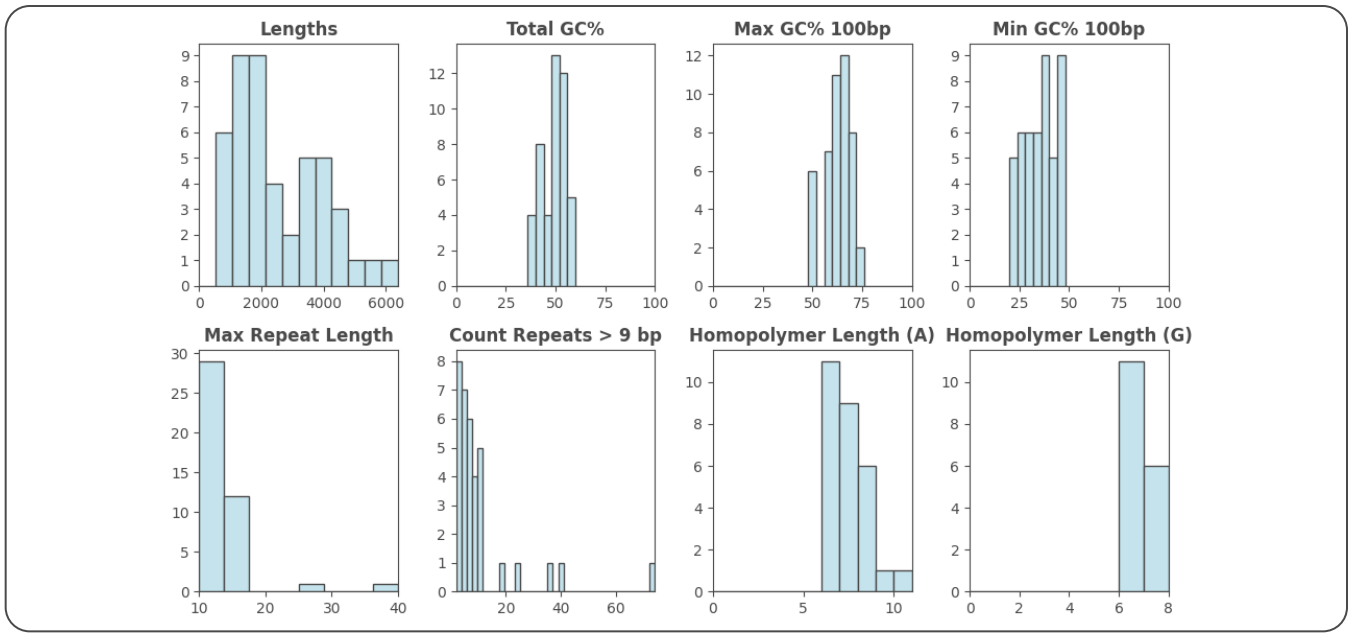

Figure 3. Characteristics of the sequences for which the per base error rate was quantified.

Results

Across the diversity of customer sequences that we synthesized and measured the error rate, the median error rate per product was 1:77 Kbp (see Table 1). Elegen products are divided into two sequence complexity classes, Elegen Standard Complexity (previously known as Tier 1) and Elegen High Complexity (previously known as Tier 2), and the complexity class correlates with the error rate. The median error rates were 1:79 Kbp for Standard Complexity sequences and 1:63 Kbp for High Complexity sequences. The median percentage of error-free molecules in a synthesized product is 97.5%, with Standard and High Complexities having similar median values. The distribution of observed error rates is shown in Figure 4 A&C. The individual values and the corresponding 95% confidence intervals are shown in Figure 4 B&D.

	Kbp per Error			Error Free Molecules		
Complexity	Q1	Median	Q3	Q1	Median	Q3
Standard	44.2	79.4	194.2	96.7	97.6	98.4
High	52.9	62.6	245.8	92.7	97.5	97.7
All	44.5	76.5	206.7	96.6	97.5	98.4

Table 1. Observed 1st quartile, median, and 3rd quartile error rates, and percent error-free molecules for 45 customer sequences.

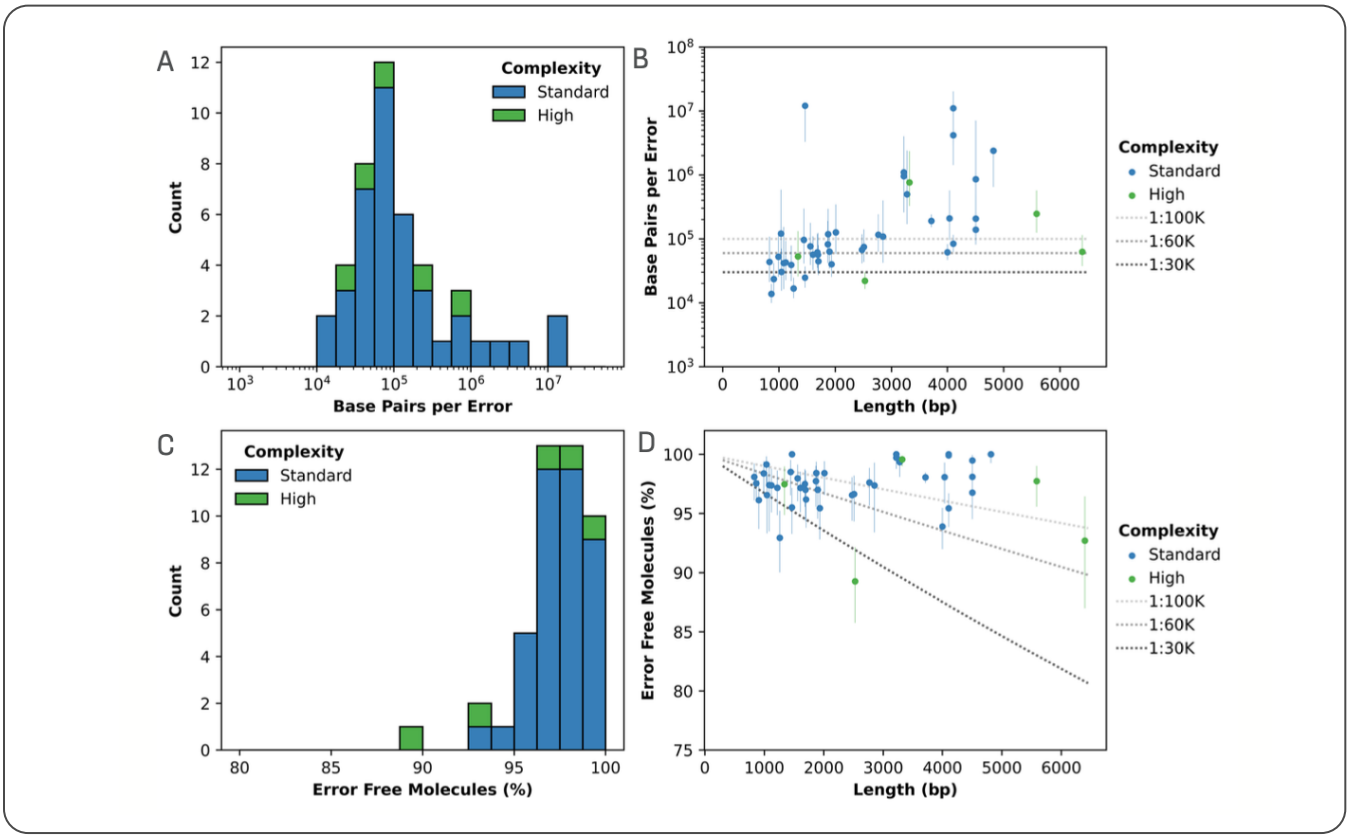

Figure 4. Observed error rates in customer products that were part of Elegen's Early Access Program. A) The distribution of error rates shown as base pairs per error. B) Individual product error rates shown as a function of sequence length. The error bars represent the 95% confidence interval assuming a binomial distribution. C) The observed distribution of percent error-free molecules. D) The observed error-free percentage as a function of sequence length with error bars marking the 95% confidence interval. The dotted lines show how a given error rate corresponds to the percentage of error free molecules.

Across all the synthesized sequences, 605 errors were observed in 4×10^4 UMIs and 1×10^8 bases. The global error rate across this set of molecules is 1:159 Kbp (± 13 Kbp), and the overall percent error-free molecules is 98.4% (± 0.1%).

Discussions and Conclusions

The described error quantification method provides deep insight into the quality of DNA synthesized by the Elegen virtual cloning process. The sequencing-based method provides the type, position, and count of all error types including sense, missense, and frameshift mutants, structural variation, chimeras, and large deletions or insertions.

We here demonstrate a median error-rate of 1:79 Kbp (1.3×10^{-5} per base) for Standard Complexity sequences. This is an order of magnitude lower than the error-rates often quoted for synthetic products (e.g., 1:3.5 Kbp or 1:5 Kbp). We also show the distribution of observed error rates, since a single number does not well represent the quality of a DNA product across sequence variation and batches. Some of the products have error rates that approach typical quoted clonal error rates. The highest error rates are still several fold better than other currently available non-clonal synthetic products, and the median error rate is more than 20x better.

While error rate is commonly used to quantify or communicate DNA quality, the fraction of molecules that are error-free is another measure of DNA quality that directly relates to successful cloning. A clone is generally the product of a single DNA molecule entering a cell, and the number of clones needed to find the correct sequence is proportional to the fraction of sequence perfect molecules

Our approach of measuring DNA errors quantifies the number of error free molecules in a sample because linkages between errors is retained across the length of the molecule. Our products have a median error-free percentage of 97.5%, and 95% of all products have an error-free percentage of 92% or higher (Fig. 4D). This means that across the entire set of sequences, a single clone would have the correct sequence in >95% of trials and two clones would give the correct sequence in >99.5% of all cases.

Others have shown that the ONT sequencing error is systematically elevated for homopolymers relative to other sequences motifs[8]. Any error-detection method utilizing ONT sequencing is lower confidence for errors in homopolymers; however, for this study, the homopolymers are within the length that ONT sequencing can call with confidence (see Fig. 3).

By using this method, we have demonstrated that the error rates of the DNA from Elegen's virtual cloning process are orders of magnitude lower than is expected for non-clonal synthetic DNA, with significant improvement on the turnaround time of clonal DNA. The observed error rates are low enough to allow direct utilization in many applications, while applications which require clonality will benefit from the faster turnaround time and significantly reduced number of clones required to find a perfect sequence.

References

1. Sloan, D. B., Broz, A. K., Sharbrough, J. & Wu, Z. Detecting rare mutations and DNA damage with sequencing-based methods. Trends Biotechnol. 36, 729–740 (2018).
2. Salk, J. J., Schmitt, M. W. & Loeb, L. A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat. Rev. Genet. 19, 269–285 (2018).
3. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. Genetics 148, 1667–1686 (1998).
4. Xiong, A.-S. et al. A simple, rapid, high-fidelity and cost-effective PCR-based two-step DNA synthesis method for long gene sequences. Nucleic Acids Res. 32, e98 (2004).
5. Sequeira, A. F., Brás, J. L. A., Guerreiro, C. I. P. D., Vincentelli, R. & Fontes, C. M. G. A. Development of a gene synthesis platform for the efficient large scale production of small genes encoding animal toxins. BMC Biotechnol. 16, 86 (2016).
6. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinforma. 3, lqab019 (2021).
7. Davis, E. M. et al. SequencErr: measuring and suppressing sequencer errors in next-generation sequencing data. Genome Biol. 22, 37 (2021).
8. Delahaye, C. & Nicolas, J. Sequencing DNA with nanopores: Troubles and biases. PLOS ONE 16, e0257521 (2021).
9. Lang, D. et al. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. GigaScience 9, giaa123 (2020).
10. Weirather, J. L. et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research 6, 100 (2017).
11. Karst, S. M. et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat. Methods 18, 165–169 (2021).
12. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12 (2011).
13. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461 (2010).
14. BBMap.SourceForgehttps://sourceforge.net/projects/bbmap/.
15. Li, H. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37, 4572–4574 (2021).
16. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746 (2017).
17. Medaka. https://github.com/nanoporetech/medaka (2022).
18. Yang, C., Chu, J., Warren, R. L. & Birol, I. NanoSim: nanopore sequence read simulator based on statistical characterization. GigaScience 6, gix010 (2017).
19. Tru-Q 7 (1.3% Tier) Reference Standard. https://horizondiscovery.com/en/reference-standards/products/truq-7-13-tier-reference-standard#description.

Supporting Information

Figures

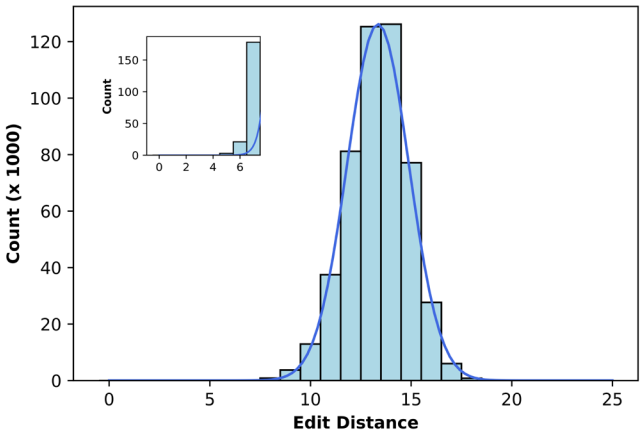

Figure 1. UMI sequence characteristics. The edit distance distribution for a sample of 1000 UMI sequences with all sequences compared to all other sequences (4.99 × 105 unique comparisons). The edit distance distribution is approximately normal (blue line). Edit distance is a measure of the minimum number of changes to a sequence is required for one sequence to become another, allowing inserts and deletions.

Methods

Sensitivity calculation. The error rate sensitivity is calculated by modeling the per base error rate as a binomial distribution of all bases in which an error is or is not observed. The lower limit of the observed error rate is 0 of n observed bases. The 95% CI, where 1- α = 0.95, X = 0 observed errors, and p represents the per base error rate, is expressed:

$$Pr(X=0)=(1-p)^n \leq 0.05 \quad (1)$$

Using Taylor's formula to approximate (ln(1-p)=-p) for p close to 0 gives:

$$n(-p) \leq \ln(0.05) \quad (2)$$

With ln(0.05)=-2.996, the upper limit of the confidence interval is approximately equal to:

$$p=3/n \quad (3)$$

The number of observed bases is the product of the sequence length (l) and the sample size of UMI-tagged molecules (N). The error rate detection limit (E), represented as base pairs per error, is then given by:

$$E= Nl/3 \quad (4)$$

For errors well within the sensitivity limit, the sample size correlates with the precision of the measurement. This can be shown using the binomial distribution confidence interval, where p is the per base error rate, n is the number of observed bases, and z is the critical z-score:

$$p \pm z\sqrt{((p(1-p))/n)} \quad (5)$$

Shared UMI calculation. The probability of shared UMIs is calculated with the following equation, where is p is the probability, and Pr(X = x) represents the probability mass function (pmf) for each UMI being represented x times:

$$p=(1-(Pr(X=1)+Pr(X=0)))/(1-Pr(X=0)) \times 100 \quad (6)$$