Q50 Data on AVITI

Characterizing and addressing error modes to improve sequencing accuracy

Semyon Kruglyak[1], Andrew Altomare[1], Mark Ambroso[1], Vivian Dien[1], Bryan Lajoie[1], Kelly N. Wiseman[1], Shawn Levy[1], and Matthew Kellinger[1]

[1]Element Biosciences, San Diego, CA

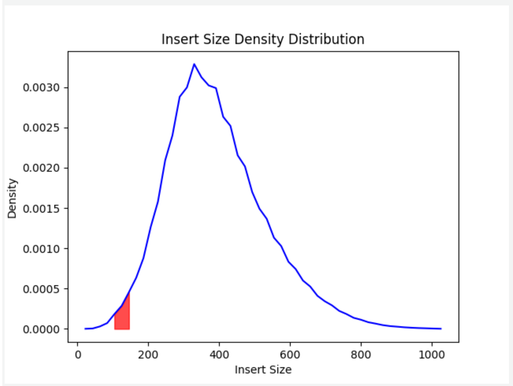
Element Biosciences

Background

The accuracy of a sequencing platform has traditionally been measured by the %Q30, or percentage of data exceeding a basecall accuracy of 99.9%. Improvements to accuracy beyond Q30 may be beneficial for certain applications such as the identification of low frequency alleles or the improvement of reference genomes. Here we demonstrate how we achieved over 70% Q50 (99.999% accuracy) data on the AVITI™ sequencer. This level of accuracy required us to not only improve sequencing quality but also to mitigate library preparation errors and analysis artifacts.

Methods

Avidity base chemistry (ABC) separates the stepping along the DNA template strand from the resolving base calls via Avidite binding. Independent optimization of these processes resulted in improved accuracy. The optimizations included changes to reagent concentrations, reaction contact times, and fit-for-purpose enzymes. One capability enabled by the ABC is dark cycling. Dark cycling omits the Avidite binding and imaging steps, enabling us to skip any portion of a DNA fragment. One of the many applications of dark cycling is described in the results section.

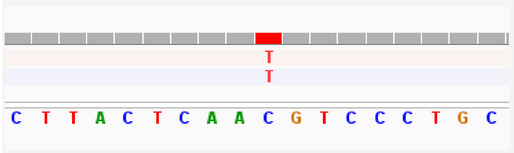


We leveraged the E. coli model system and focused on short insert fragments to stratify errors. The model system offers the following benefits for error characterization:

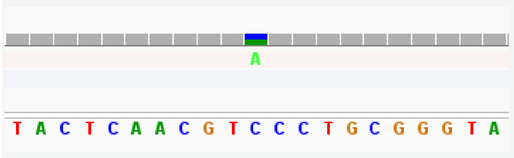
- Haploid genome → no heterozygous variants
- Strain-specific reference genome → few reference errors
- Relatively few repeat regions → simplified alignment

In left tail of the insert length distribution (inserts < 150 bp) in a sequencing library, R1 and R2 overlap completely, enabling error stratification, leveraging the fact that the same molecule is sequenced twice.

We focus errors assigned to high-quality scores because (1) most of the data is concentrated in high-quality bins, and (2) understanding the source of these errors is the key to attaining Q50.



R1 and R2 both make high-quality calls (Q>29) that agree with each other but disagree with the reference. This is likely a library preparation errors, e.g. caused by deamination damage.



R1 and R2 both make a high-quality call (Q>39) but R1 disagrees with the reference. This is almost certainly a sequencing error in R1.

Analysis artifacts can appear as high-quality errors. We remove analysis artifacts from the E. coli sequencing data as follows:

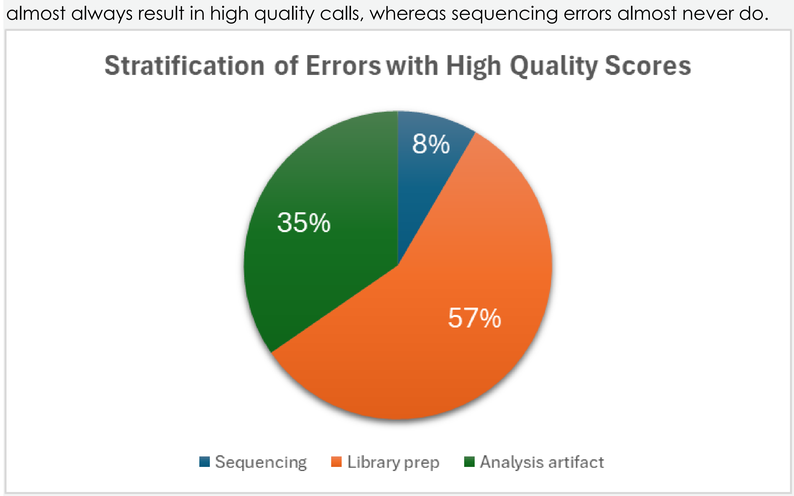
- Identify difficult-to-align regions using GenMap with 30 kmer size. Discard reads that overlap such regions (2.5% of the E. coli reference). The display on the left show multiple high-quality mismatches in low MAPQ regions

GenMap: ultra-fast computation of genome mappability
Christopher Pockrandt, Mai Alzamel, Costas S Iliopoulos, Knut Reinert
Bioinformatics, Volume 36, Issue 12, June 2020, Pages 3687–3692, https://doi.org/10.1093/bioinformatics/btaa222

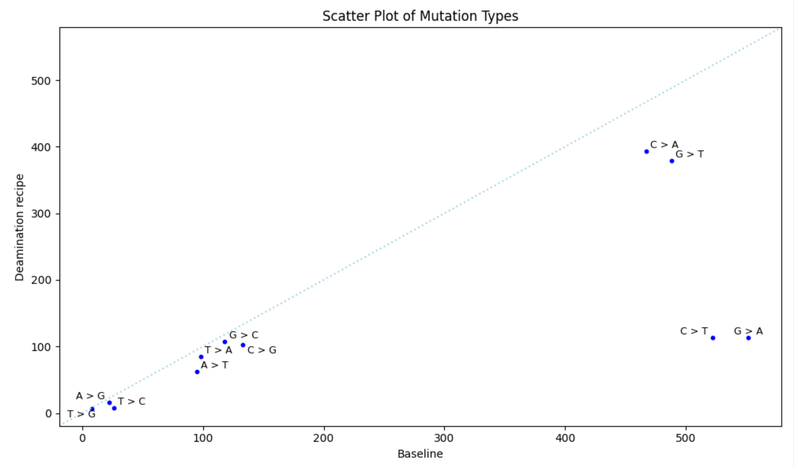
- Identify positions in the genome where high-quality calls across sequencing runs and technologies disagree with the published reference. Discard reads overlapping such sites (10 additional sites). The display on the left shows an example of one such site. The apparent high-quality errors are the result of the mismatch between the published reference and the input DNA.

Results

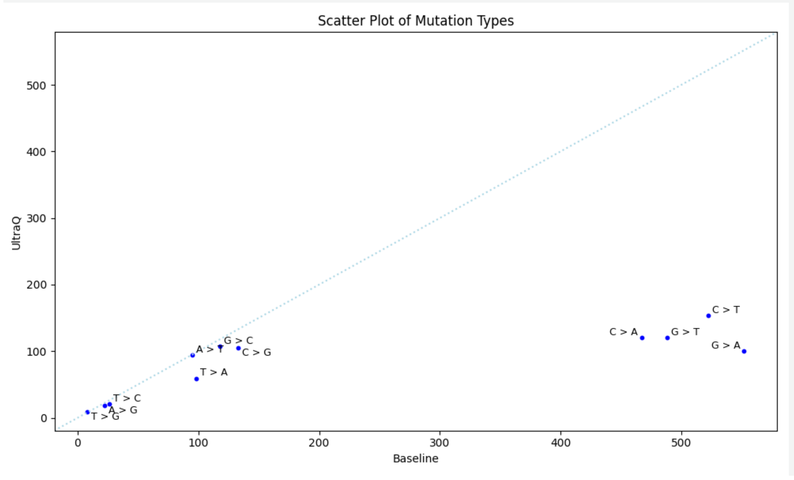
Below is a pie chart of errors (prior to the reference modifications). Sequencing errors make up a small fraction of the total, and library prep errors are the biggest contributors. This is only true when conditioning on high Q scores. Library prep errors almost always result in high quality calls, whereas sequencing errors almost never do.



Given the prevalence of library preparation errors among the high-quality errors, we made several changes to mitigate this error class. First, a well known cause of library preparation errors is deamination, where a C nucleotide becomes a U and is sequenced as T, resulting in a C>T mismatch (or G to A depending on the presented orientation). Two modifications were made to address deamination: (1) recipe changes such as the use of NaOH denaturation rather than high temperature and (2) the addition of the USER enzyme (NEB) to cut any fragments with deaminated bases. The figure below shows the impact. The run with the modifications is on the Y-axis.



The modifications led to a large drop in C > T (G > A) errors as desired. However, the relative instance of G > T (C > A) errors remained. We determined that these errors primarily occurred in the early cycles of R2. We hypothesize that they are the result of the end repair step that follows fragmentation. We therefore used dark cycling through the first 15 bases of R2, followed by 150 regular cycles. The next figure shows the results.

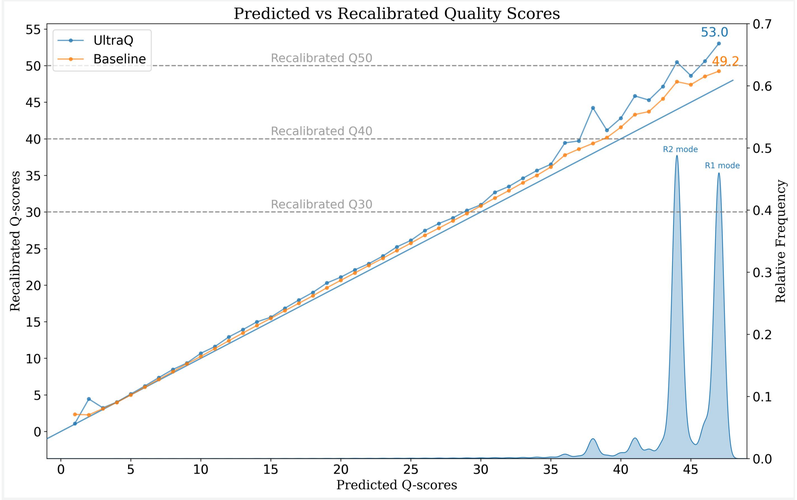


The run labeled UltraQ includes the changes aimed at minimizing reads with deamination damage. It also includes the dark cycling of the beginning of R2. The G > T (C > A) errors are now significantly reduced and in line with other error types.

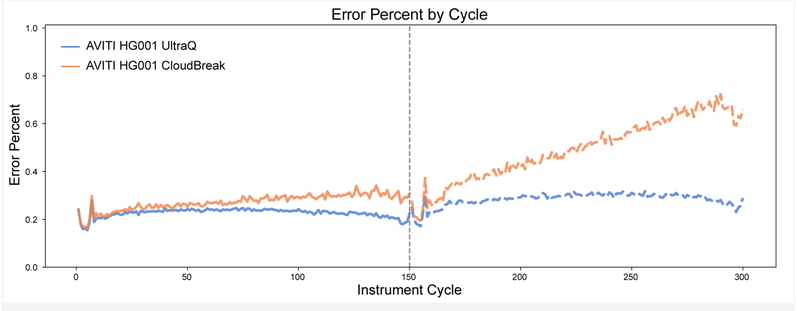
The table below compares key metrics between our current chemistry and the Ultra Q modifications. A greater fraction of the data is assigned to high quality scores. The average mismatch rate is driven by low-quality sequencing errors.†

Sequencing metrics	Current chemistry	UltraQ chemistry
Per cycle phasing	0.15%	0.08%
Per cycle prephasing	0.04%	0.01%
Recalibrated %Q30	96.7%	98.5%
Recalibrated %Q40	88.2%	96.3%
Recalibrated %Q50	0%	78.1%
Average mismatch rate†	0.128%	0.056%

The QQ plot shows recalibrated quality scores and demonstrates a high fraction of Q50 data. The histogram below the QQ plot shows the distribution of scores for the entire data set.



When applied to human genome sequencing, the UltraQ recipe shows significantly reduced mismatch rate relative to our current chemistry. The mismatch rate is driven by low quality sequencing errors and includes human variation as well as any analysis artifacts.



Discussion

In this study, we demonstrated a high percentage of Q50 data through optimizations to ABC sequencing. Notably, addressing library preparation errors and analysis artifacts was critical to reaching Q50. The study has important limitations: it focuses on model organisms and short inserts. However, we expect the high-level conclusions and improvements to apply to human sequencing with standard insert lengths. The QQ plot that shows recalibrated quality scores above Q50 was derived from the entire fragment length distribution. Also, significant improvements in average mismatch rate are observed in human. In a follow up study, we plan to perform recalibration based on human genome sequencing based on the haploid CHM13 cell line, where the T2T reference and the lack of variants are expected to significantly reduce analysis artifacts.

Visit **elementbiosciences.com** for more information.