nature biotechnology

Article

https://doi.org/10.1038/s41587-023-01750-7

Sequencing by avidity enables high accuracy with low reagent consumption

Received: 15 August 2022

Accepted: 15 March 2023

Published online: 25 May 2023

Check for updates

A list of authors and their affiliations appears at the end of the paper

We present avidity sequencing, a sequencing chemistry that separately optimizes the processes of stepping along a DNA template and that of identifying each nucleotide within the template. Nucleotide identification uses multivalent nucleotide ligands on dye-labeled cores to form polymerase–polymer–nucleotide complexes bound to clonal copies of DNA targets. These polymer–nucleotide substrates, termed avidites, decrease the required concentration of reporting nucleotides from micromolar to nanomolar and yield negligible dissociation rates. Avidity sequencing achieves high accuracy, with 96.2% and 85.4% of base calls having an average of one error per 1,000 and 10,000 base pairs, respectively. We show that the average error rate of avidity sequencing remained stable following a long homopolymer.

Avidity sequencing chemistry enables a diversity of applications that include single-cell RNA sequencing (RNA-seq) and whole-human-genome sequencing. For the human sample HG002, avidity sequencing reached a single-nucleotide polymorphism (SNP) F1 score of 0.9958 and small-indel F1 score of 0.9954.

Over the past 15 years, highly parallel sequencing methods have enabled a broad set of applications[1–8]. Multiple technologies have been introduced during this time, each having various strengths and limitations[9]. The technologies vary by accuracy, read length, run time and cost. The most widely used method uses highly parallel and accurate short-read sequencing, described in ref. 10 and termed sequencing by synthesis (SBS).

The SBS methodology sequences DNA by controlled (that is, one at a time) incorporation of modified nucleotides[11]. The modifications consist of a 3′ blocking group and a dye label[12,13]. The blocking group ensures that only a single nucleotide is incorporated, and the dye label enables identification of each nucleotide following an imaging step. The blocking group and label are subsequently removed, completing the sequencing cycle. The cycle is repeated with the incorporation of the next blocked and labeled nucleotide. Incorporation of the modified nucleotide meets two objectives: to advance the polymerase along the DNA template and to differentially label the incorporated nucleotide for base identification. Although combination of the two processes is efficient, it prevents independent optimization of the processes. High-yielding and rapid incorporation requires micromolar concentrations of nucleotides to drive the polymerizing reaction[14–18]. The alternative, of allowing longer incorporation times, results in longer cycle times that have an additive effect over 300 cycles of stepwise sequencing.

We present a different sequencing chemistry, termed avidity sequencing, that separates and independently optimizes the controlled incorporation and nucleotide identification steps to achieve increased base-calling accuracy relative to SBS while reducing the concentration of key reagents to nanomolar scale. To advance this approach, we first had to overcome the technical challenge of signal persistence. For example, a potential strategy for separation of the steps described above could be to first incorporate a 3′ blocked but unlabeled nucleotide and then to bind a complementary labeled nucleotide to the subsequent base in the template for base identification. This approach is problematic because the dissociation rate for single nucleotides from a polymerase–template complex is large, and the polymerase–nucleotide complex does not remain stable throughout imaging unless prohibitively high concentrations of nucleotides are present in the bulk solution. To overcome this challenge, we used avidity.

Avidity refers to the accumulated strength of multiple affinities of individual noncovalent binding interactions, which can be achieved when multivalent ligands tethered in close proximity simultaneously bind to their targets[19]. Coincident binding increases ligand affinity and residence time[20]. As an example of the potential impact of avidity on both affinity and decreased dissociation rate, Zhang et al.[21] demonstrated that, by changing a monomeric to a pentameric nanobody, it is possible to decrease dissociation rates by three to four orders of magnitude. Our approach was to use avidity for nucleotide detection within the sequencing chemistry (Fig. 1). We demonstrate here that avidity sequencing achieves accuracy, surpassing an average of one error per 10,000 base pairs (bp) (*Q*40), and enables a diversity of applications that include single-cell RNA-seq and whole-human-genome

e-mail: mprevite@elembio.com

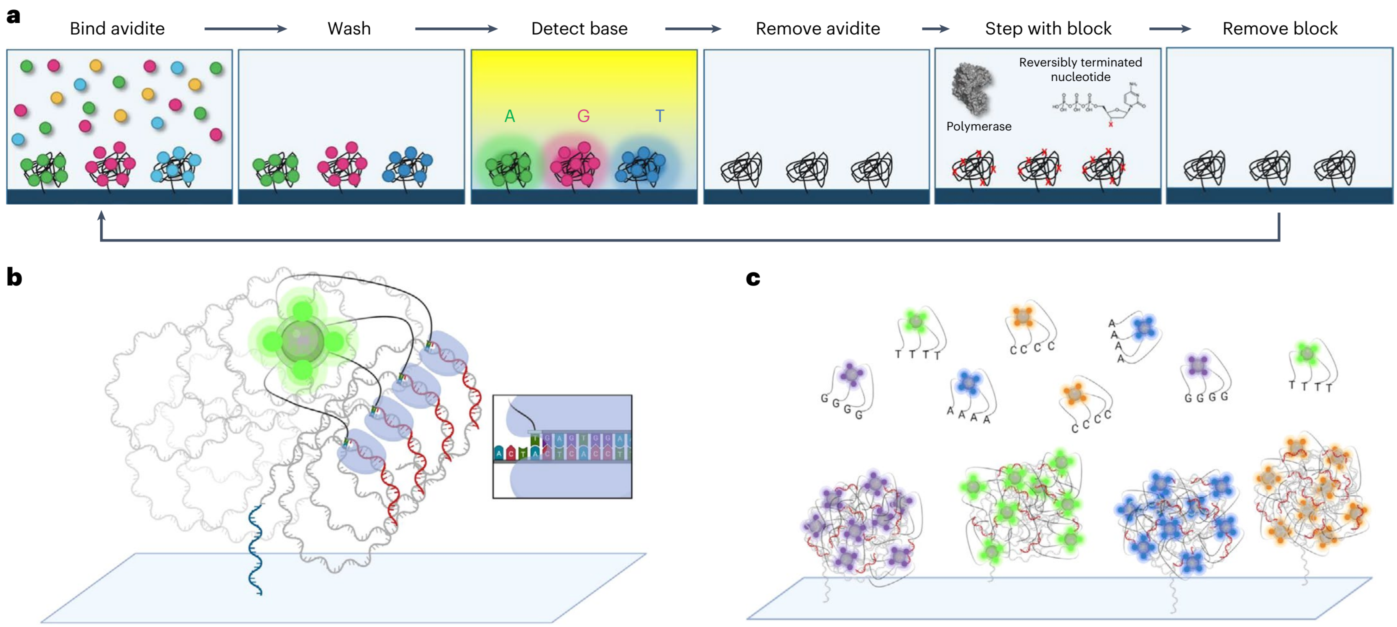


Fig. 1 | Avidity sequencing workflow and scheme. a, Sequencing by avidity. A reagent containing multivalent avidite substrates and an engineered polymerase are combined with DNA polonies inside a flowcell. The engineered polymerase binds to the free 3′ ends of the primer-template of a polony and selects the correct cognate avidite via base-pairing discrimination. The multivalent avidite interacts with multiple polymerases on one polony to create avidity binding that reduces the effective K_d of the avidite substrates 100-fold compared with a monovalent dye-labeled nucleotide, allowing productive binding of nanomolar concentrations. Multiple polymerase-mediated binding events per avidite ensure a long signal persistence time. Imaging of fluorescent, bound avidites enables base classification. Following detection, avidites are removed from the polonies. Extension by one base using an engineered polymerase incorporates an unlabeled, blocked nucleotide. A terminal 3′ hydroxyl is regenerated on the DNA strand, allowing repetition of the cycle. **b**, Rendering of a single avidite bound to a DNA polony via polymerase-mediated selection. The initial surface primer used for library hybridization and extension during polony formation is shown in blue. Sequencing primers (red) are shown annealed to the single-strand DNA polony (gray). Each arm of the avidite (black) connects the avidite core containing multiple fluorophores (green) to a nucleotide substrate. The polymerase bound to the sequencing primer selects the correct nucleotide to base pair with the templating base (inset). The result is multiple base-mediated anchor points noncovalently attaching the avidite to the DNA polony. **c**, Rendering of multiple DNA polonies with template-specific avidites bound during the binding step of the cycle (polymerase not shown for simplicity). Many avidites bind to each DNA polony generating a fluorescent signal during detection. Multiple long, flexible polymer linkers connect the core to the nucleotide substrates.

sequencing. We also demonstrate an improved ability of this chemistry to sequence through homopolymer sequences.

Results

Before sequencing, DNA fragments of interest were circularized and captured on the surface of a flowcell. Clonal copies of DNA fragments were then created through rolling circle amplification, generating approximately 1 billion concatemers on the flowcell surface[22–25]. The resulting concatemers, referred to as polonies using the original term coined by Church and collaborators[26], were used as the DNA substrate for sequencing. In contrast to the DNA nanoballs developed by Complete Genomics, polonies are amplified on-instrument following library hybridization to the flowcell[27]. This approach simplifies user workflow and eliminates the possibility that DNA fragments may interact in solution during the amplification process. We then constructed the avidite: a dye-labeled polymer with multiple, identical nucleotides attached. In the presence of a polymerase, the avidite was able to bind multiple complementary nucleotides specifically in concatemer copies of a DNA fragment within a polony. A polymerase and a mixture of four avidites, each corresponding to a particular label and nucleotide, were applied to the flowcell and used for base discrimination. The avidite was not incorporated, but provided a stable complex while enabling removal under specifically formulated wash conditions. Removal of the avidite left no modifications in the synthesized strand. The avidites decreased the required concentration of reporting nucleotides by 100-fold relative to single-nucleotide binding, yielded negligible dissociation rates and obviated the need to have nucleotides present in the bulk solution. A low avidite concentration leads to reduced use of fluorophores relative to the strategy of using high-concentrations of dye-labeled nucleotides. The advent of the avidite enabled us to separate the process of stepping along the DNA template from the process of identifying each nucleotide, and to optimize each for quality and reagent consumption. Figure 1a shows a complete cycle of avidity sequencing, Fig. 1b depicts a single avidite interacting with multiple DNA copies within a polony and Fig. 1c shows many avidites specifically bound to several polonies on the surface. Additional detail on the structure of one version of an avidite is provided in Extended Data Fig. 1.

Avidity sequencing overcomes the kinetic challenges of generating a signal by incorporation of a dye-labeled monovalent nucleotide. In bulk solution, incorporation of a dye-labeled nucleotide is limited by a specificity constant (k_{cat}/K_m) that governs the observed rate of productive nucleotide binding and incorporation[28]. A specificity constant of $0.54 \pm 0.22\ \mu M^{-1} s^{-1}$ for monovalent dye-labeled nucleotides using an engineered polymerase was observed resulting from a maximum rate of incorporation (k_{pol}) of $0.86 \pm 0.14\ s^{-1}$ and an apparent dissociation constant K_d ($K_{d,app}$) of $1.6 \pm 0.6\ \mu M$ (Fig. 2a). This apparent K_d reflects the K_m of a kinetic system not in equilibrium rather than the true K_d of the nucleotide substrate[29]. To achieve complete product turnover, this high apparent K_d can be overcome either by using increased concentrations of fluorescent nucleotide substrate or allowing longer incorporation time for completion of the reaction. Both paths used to overcome this substrate limitation have the undesirable consequence of either high cost or long cycle time. Together, the use of avidity substrates and DNA polonies containing many copies of substrate DNA in close proximity overcomes the limitations of incorporating a monovalent dye-labeled nucleotide.

Using binding of the four labeled avidites for base identification established a binding equilibrium that reached saturation based on

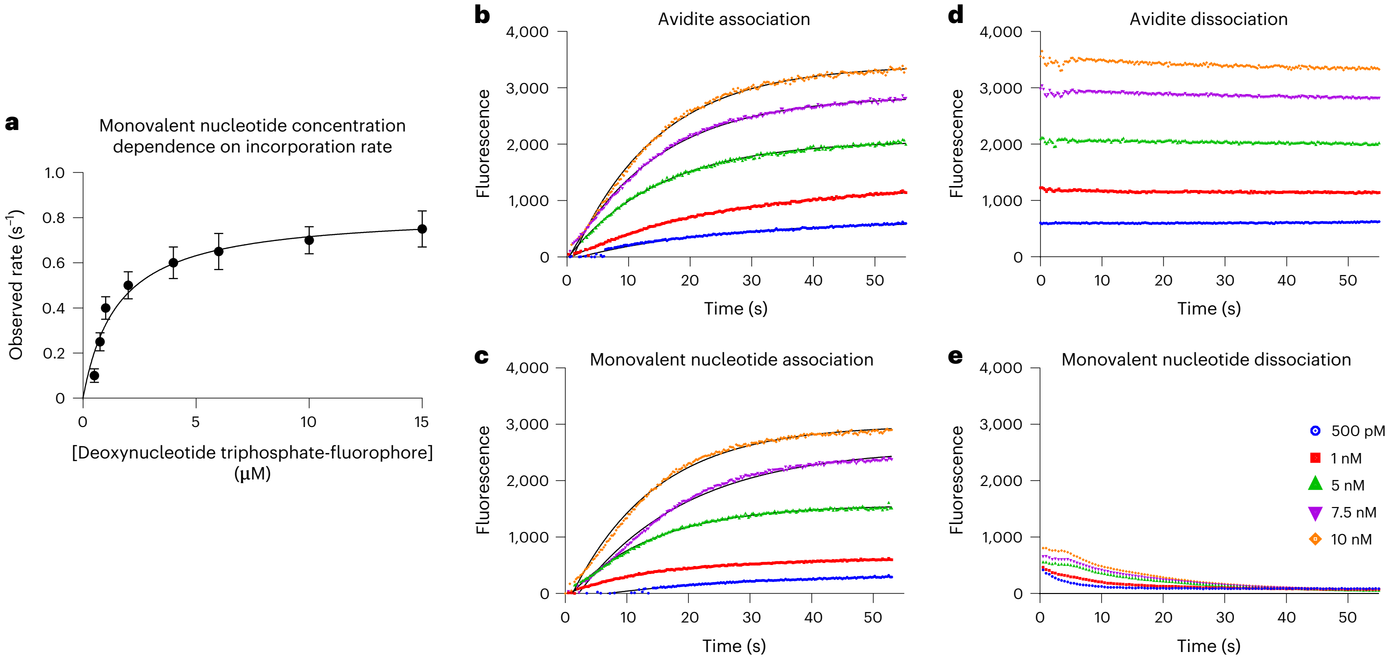


Fig. 2 | Nucleotide and avidite binding kinetics. a, Monovalent fluorophore-labeled nucleotide concentration dependence of the observed rate of incorporation. Time series were performed at each concentration and fit to a single exponential equation to derive a rate. Observed rates were plotted as a function of concentration and fit to a hyperbolic equation, deriving a value of $k_{pol} = 0.86 \pm 0.14\ s^{-1}$ and $K_{d,app} = 1.6 \pm 0.6$ μM. **b**,**c**, Real-time association kinetics of signal generation resulting from reacting multivalent avidite substrates (**b**) and monovalent nucleotides (**c**) with DNA polonies. **d**,**e**, Real-time measurement of signal decay following flow cell washing for imaging of multivalent avidite substrates (**d**) and monovalent nucleotides (**e**).

substrate concentration within 30 s to generate signal, rather than relying on catalysis. The binding kinetics of this interaction were monitored using real-time data collection to observe avidites binding to polonies with an association rate ($k_{on,avidite}$) of 271 ± 82 $nM^{-1} s^{-1}$ (Fig. 2b). This observed association occurred within the limit of error of a single fluorescently labeled monovalent nucleotide (Fig. 2c). Major differences were observed in the dissociation kinetics of avidite substrates versus monovalent nucleotides. Avidite substrates bound to the DNA polonies tightly with no measurable dissociation over the >1-min timescale needed for imaging and base calling (Fig. 2d). This is in sharp contrast to fluorescently labeled monovalent nucleotides, which dissociated rapidly during the wash step following binding and then continued to dissociate during imaging (Fig. 2e). The negligible dissociation rate resulted in decreased K_d of more than two orders of magnitude for avidites compared with monovalent nucleotides. With near-zero avidite dissociation rates, a persistent signal was achieved without the presence of free avidites in bulk solution, eliminating background. Without avidity, dissociation kinetics with monovalent nucleotides showed a fourfold signal decrease at the beginning of imaging due to rapid dissociation, as a result of disruption of the binding equilibrium during reagent exchange (Fig. 2e).

Sequencing instrumentation

Avidity sequencing was performed on the AVITI commercial sequencing system. Briefly, the instrument is a four-color optical system with two excitation lines of approximately 532 and 635 nm. The four-color system is created using an objective lens, multiple tube lenses and multiple cameras for simultaneous imaging of four spectrally separated colors. The detection channels for emission are centered at approximately 553, 596, 668 and 716 nm, respectively. Reagents are delivered using a selector valve and syringe pump to perform reagent cycling. The instrument contains two fluidics modules and a shared imaging module, enabling parallel utilization of two flowcells. Subsequent to image collection, data were streamed through an onboard processing unit that performs image registration, intensity extraction and correction, base calling and quality score assignment (Methods).

Accuracy of avidity sequencing

To evaluate the accuracy of avidity sequencing, 20 sequencing runs were performed using a well-characterized human genome. Sequencing data were used to train quality tables according to the methods of Ewing et al.[30], but with modified predictors. Quality tables were then applied to independent sequencing runs. Figure 3 shows the data quality obtained in a representative run not used for training. Quality scores were well calibrated across the entire range, meaning that predicted quality matched observed quality as determined by alignment to a known reference. Combined over reads 1 and 2, 96.2% of base calls were >$Q30$ (an average of one error per 1,000 bp) and 85.4% >$Q40$, with a maximum of $Q44$, or approximately one error in 25,000 bases. For comparison, a publicly available PCR-free NextSeq 2000 dataset was downloaded from the Illumina public demo set repository (https://basespace.illumina.com/datacentral) and a publicly available NovaSeq 600 dataset (https://console.cloud.google.com/storage/browser/brain-genomics-public/research/sequencing/fastq). The NextSeq 2000 and NovaSeq 6000 datasets had 90.1% and 92.7% of data >$Q30$, respectively, and none of the base calls exceeded $Q40$.

To obtain an additional measure of accuracy, we used the same datasets to compute the percentage of k-mers ($k = 1, 2, 3$) containing at least one mismatch after alignment to a well-characterized reference. Known SNP sites were masked before the comparison. When compared with NextSeq 2000 and NovaSeq 6000, we found that AVITI had the highest accuracy across four out of four 1-mers, 16 out of 16 2-mers and 58 out of 64 3-mers (Extended Data Fig. 2).

Homopolymer sequencing

Sequencing through long homopolymers has posed challenges for multiple sequencing technologies[31,32]. Although SBS improves homopolymer sequencing relative to flow-based technologies,

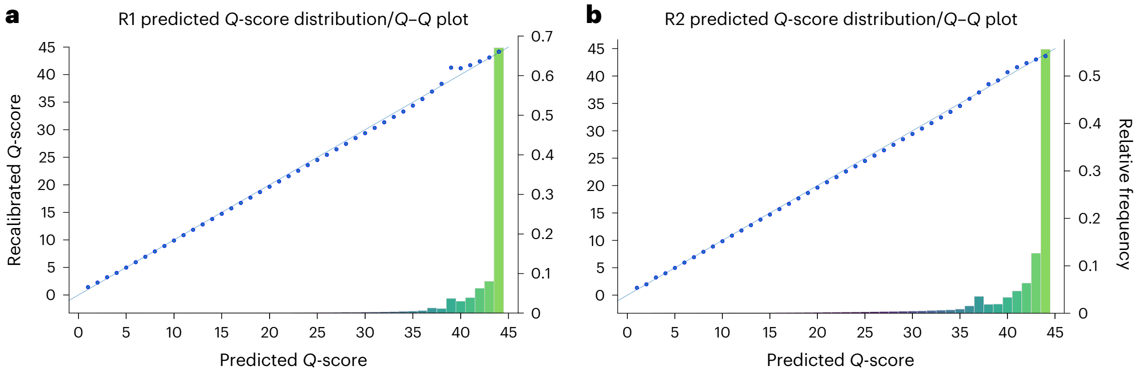


Fig. 3 | Predicted and observed quality scores for a 2 × 150-bp sequencing run of human genome HG002. a, Read 1 (R1). **b**, Read 2 (R2). Points on the diagonal indicate that predicted scores match observed scores. The histograms show that the majority of the data points are >*Q*40.

the error rates of reads that pass through long homopolymer regions increase substantially[33]. Correction algorithms have been proposed to circumvent the inherent challenges with base-calling post-homopolymer repeats[34], but the exact cause has not been fully established in the literature. In contrast to SBS, avidity sequencing leverages rolling circle amplification, polymerases evolved to accommodate the avidite complex formation and a separate polymerase evolved for efficient incorporation of unlabeled and 3′ blocked nucleotides. We evaluated the impact of these differences on sequencing through long homopolymers. Specifically, homopolymers of length 12 or more nucleotides were used to assess the accuracy of reads before and after homopolymer regions. Figure 4 shows the results comparing avidity sequencing with SBS, averaged across the ~700,000 homopolymer loci of length 12 or more. Average error rate of avidity sequencing remained stable following a long homopolymer (controlling for the fact that post-homopolymer stretch occurs in later cycles of a read). By contrast, the error rate of SBS reads increased by more than a factor of five following homopolymer stretches. Extended Data Fig. 3 shows the histogram of pairwise error rate differences between avidity sequencing and SBS for all long homopolymer loci. The avidity sequencing error rate outperformed SBS in >97% of cases and the magnitude of difference is correlated with homopolymer length (Fig. 5). Extended Data Fig. 4 shows representative loci from the 95th, 50th and fifth percentiles of the histogram.

Single-cell RNA-seq

To demonstrate sequencing performance across common applications, single-cell RNA expression libraries were prepared and sequenced. Two libraries from a reference standard consisting of human peripheral blood mononuclear cells were generated using the 10X Chromium instrument. The two libraries contain RNA from roughly 10,000 and 1,000 cells, respectively. Following circularization, the libraries were sequenced to generate paired-end reads with read lengths of 28 and 90 for reads 1 and 2, respectively, as recommended by the vendor. The analysis was done using CellRanger (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation). Because this reference standard is used by 10X Genomics to evaluate sequencing performance, a set of metrics and guidelines to assess sequencing results is provided along with the biological material. Extended Data Table 1 shows each metric, the guideline values from 10X Genomics and the performance of each sequenced library. All metrics were within the guide ranges, and metrics pertaining to sequencing quality exceeded the thresholds provided.

Whole-human-genome sequencing

Another common application is human-whole-genome sequencing. This application challenges sequencer accuracy to a greater extent than measurement of gene expression because the latter requires only accurate alignment while the former depends on nucleotide accuracy to resolve variant calls. To demonstrate performance for this application, the well-characterized human sample HG002 was prepared for sequencing using a Covaris shearing and PCR-free library preparation method and sequenced with 2 × 150-bp reads. The run generated 1.02 billion passing filter paired-end reads with a duplicate rate of 0.58% (0.11% classified as optical duplicates by Picard (https://broadinstitute.github.io/picard/)). To underscore the impact of low duplicates, we compared the number of input reads with genomic coverage (Extended Data Fig. 5).

A FASTQ file with the base calls and quality scores was downsampled to 35-fold coverage and used as an input into the DNAScope analysis pipeline from Sentieon. SNP and indel calls achieved F1 scores of 0.995 and 0.996, respectively. Extended Data Table 2 shows variant-calling performance for SNPs and small indels on the GIAB-HC regions. Sensitivity, precision and F1 scores are shown. The performance on SNPs and indels is comparable. Extended Data Fig. 6 shows the F1 score for SNPs and indels across all GiaB stratifications with at least 100 variants in the truth set.

Extensibility of avidity sequencing

To assess the extensibility of avidity chemistry we continued a sequencing run beyond 150 bp to generate a 1 × 300 dataset from an *Escherichia coli* library. To achieve this we used both an optimized polymerase and an optimized reagent formulation. Figure 6a shows quality scores as a function of sequencing cycle. Because quality scores were not trained to these lengths, the scores are approximate. Figure 6b shows the *E. coli* error rate as a function of cycle number based on alignment to the known reference strain. The error rate of the final cycle was 1.9% and that at cycle 150 was 0.1%. Error calculations were based on the vast majority of the data with a pass filter rate for the run of >99.6% and Burrows–Wheeler aligner (BWA) settings aimed at strongly discouraging soft clipping (no cycles with soft clipping >0.04%). The enzymes and formulations developed for this run will be leveraged as we continue to identify extensions and improvements.

Discussion

We present a sequencing chemistry that achieves improved quality and lower reagent consumption by independent optimization of nucleotide incorporation and signal generation. Although other chemistries have proposed the separation of incorporation and signal generation[35], the avidite concept benefits from the fact that multiple nucleotides on the avidite bind multiple copies of the DNA template within a polony, which decreases dissociation rate constant and the labeled reagent concentration requirement for base classification. Furthermore, the avidite construct is modular. The core can be swapped for a different substrate. Both number and type of dye molecules are configurable, and many

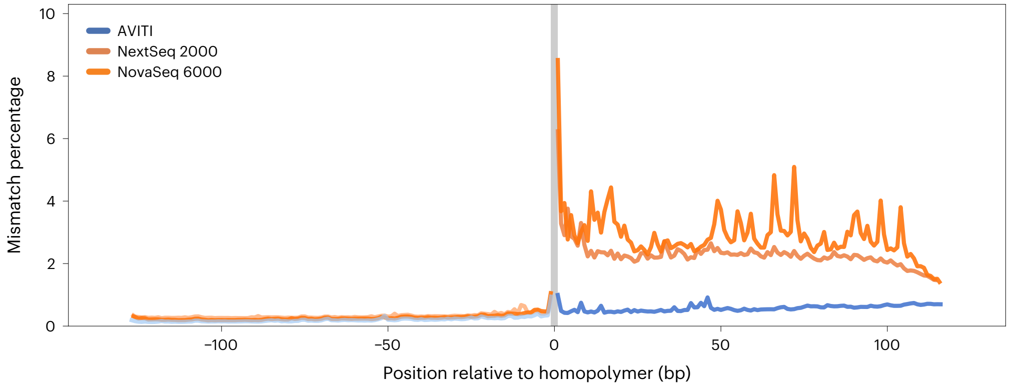


Fig. 4 | Post-homopolymer performance across platforms. Mismatch percentages of AVITI, NovaSeq 6000 and NextSeq 2000 reads before and after homopolymers of length 12 or greater.

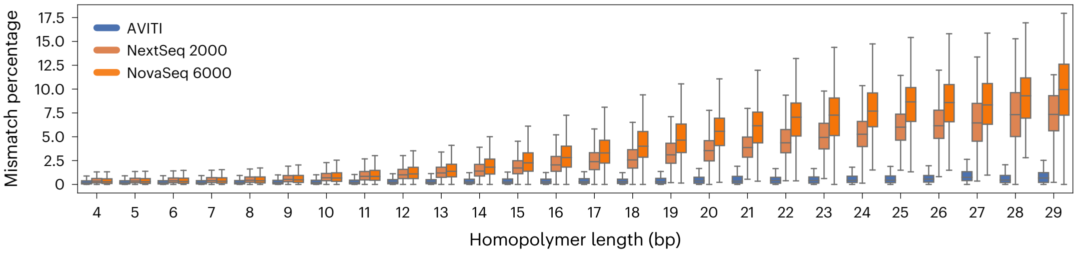


Fig. 5 | Comparison of mismatch rate following homopolymers of length between four and 29. Mismatch percentage difference between avidity sequencing and SBS increases with homopolymer length. The box plot shows median, quartiles and whiskers, which are 1.5× interquartile range.

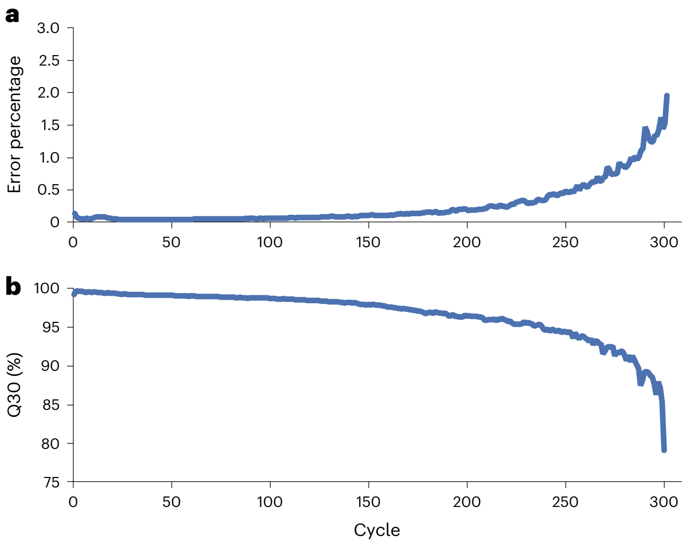


Fig. 6 | Performance of a 300-cycle *E. coli* sequencing run. a, Percentage $Q30$ by cycle. Overall $Q30$ percentage exceeds 96% and end of read has 85% $Q30$. **b**, *E. coli* error rate as a function of cycle. Alignment settings strongly discourage soft clipping, and >99% of reads pass filter. Final cycle error rate was 0.019.

types of linkers can be used. The changes are straightforward to implement and do not require modification of the polymerase responsible for binding the nucleotides attached to the linkers. The modular design speeds technology improvement because each component can be optimized in parallel for increased signal, decreased cycle time, lower reagent concentration or any other potential axis of improvement.

The avidity chemistry described above has been implemented as part of a benchtop sequencing solution. The accuracy of the sequencer was demonstrated by training a quality model on human sequencing data, which shows that in the majority of bases in an independent human-whole-genome sequencing run is >$Q40$. The high level of accuracy probably results from (1) the use of an engineered high-fidelity polymerase, (2) synergistic binding of multiple nucleotides on a single avidite to ensure only the correct cognate avidite binds to the polony and (3) a binding disadvantage for out-of-phase DNA copies within a polony that lack other out-of-phase neighbors to serve as avidity substrates. Future work will be required to investigate the relative contribution of each mechanism proposed above. In addition to overall accuracy improvements, the chemistry retains good performance in reads containing long homopolymers. The sequencer can be used in a wide range of applications, as exemplified by results for single-cell RNA-seq and for whole-human-genome sequencing. In both cases, reference standards were sequenced so that the quality of result could be assessed. The single-cell data exceeded the quality metric guidelines provided by 10X Genomics (https://www.10xgenomics.com/compatible-products?query=&page=1). The human genome variant-calling results showed high sensitivity and precision for both SNPs and small indels[36]. The two benchmarking studies were selected due to the availability of well-characterized samples and because they represent very different use cases. However, these are only examples and other applications have been demonstrated, including whole-genome sequencing for rare disease[37], low-pass sequencing with imputation[38] and single-cell sequencing of DNA and RNA[39]. Although the current implementation of avidity-based sequencing already achieves high accuracy and broad applicability, there are many improvement directions being explored. In addition to the initial demonstration of longer reads shown here, further quality improvements, shorter cycle times and higher densities are under development.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-01750-7.

References

1. Levy, S. E. & Myers, R. M. Advancements in next-generation sequencing. *Annu. Rev. Genomics Hum. Genet.* **17**, 95–115 (2016).
2. van Dijk, E. L. et al. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418–426 (2014).
3. Yohe, S. & Thyagarajan, B. Review of clinical next-generation sequencing. *Arch. Pathol. Lab. Med.* **141**, 1544–1557 (2017).
4. Zhang, Y. et al. Single-cell RNA sequencing in cancer research. *J. Exp. Clin. Cancer Res.* **40**, 81 (2021).
5. Ekblom, R. & Galindo, J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1–15 (2011).
6. Morozova, O. & Marra, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255–264 (2008).
7. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5**, 16–18 (2008).
8. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
9. Hu, T. et al. Next-generation sequencing technologies: an overview. *Hum. Immunol.* **82**, 801–811 (2021).
10. Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
11. Chen, F. et al. The history and advances of reversible terminators used in new generations of sequencing technology. *Genomics Proteomics Bioinformatics* **11**, 34–40 (2013).
12. Tsien, R. P., Fahnestock, M. & Johnston, A. J. DNA sequencing. International patent WO1991006678A1 (1990).
13. Zavgorodny, S. et al. 1-Alkylthioalkylation of nucleoside hydroxyl functions and its synthetic applications: a new versatile method in nucleoside chemistry. *Tetrahedron Lett.* **32**, 7593–7596 (1991).
14. Joyce, C. M. et al. Fingers-closing and other rapid conformational changes in DNA polymerase I (Klenow fragment) and their role in nucleotide selectivity. *Biochemistry* **47**, 6103–6116 (2008).
15. Kati, W. M. et al. Mechanism and fidelity of HIV reverse transcriptase. *J. Biol. Chem.* **267**, 25988–25997 (1992).
16. Kuchta, R. D. et al. Kinetic mechanism of DNA polymerase I (Klenow). *Biochemistry* **26**, 8410–8417 (1987).
17. Xia, S. & Konigsberg, W. H. RB69 DNA polymerase structure, kinetics, and fidelity. *Biochemistry* **53**, 2752–2767 (2014).
18. Yang, G. et al. Steady-state kinetic characterization of RB69 DNA polymerase mutants that affect dNTP incorporation. *Biochemistry* **38**, 8094–8101 (1999).
19. Rudnick, S. I. & Adams, G. P. Affinity and avidity in antibody-based tumor targeting. *Cancer Biother. Radiopharm.* **24**, 155–161 (2009).
20. Vauquelin, G. & Charlton, S. J. Exploring avidity: understanding the potential gains in functional affinity and target residence time of bivalent and heterobivalent ligands. *Br. J. Pharmacol.* **168**, 1771–1785 (2013).
21. Zhang, J. et al. Pentamerization of single-domain antibodies from phage libraries: a novel strategy for the rapid generation of high-avidity antibody reagents. *J. Mol. Biol.* **335**, 49–56 (2004).
22. Fire, A. & Xu, S. Q. Rolling replication of short DNA circles. *Proc. Natl Acad. Sci. USA* **92**, 4641–4645 (1995).
23. Liu, D. et al. Rolling circle DNA synthesis: small circular oligonucleotides as efficient templates for DNA polymerases. *J. Am. Chem. Soc.* **118**, 1587–1594 (1996).
24. Rubin, E. et al. Convergent DNA synthesis: a non-enzymatic dimerization approach to circular oligodeoxynucleotides. *Nucleic Acids Res.* **23**, 3547–3553 (1995).
25. Sabanayagam, S. T., Masasi, J., Hatch, A. & Cantor, C. Nucleic acid assays and methods of synthesis. US patent US20020076716A1 (1999).
26. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
27. Drmanac, R. et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
28. Michaelis, L. et al. The original Michaelis constant: translation of the 1913 Michaelis–Menten paper. *Biochemistry* **50**, 8264–8269 (2011).
29. Tsai, Y. C. & Johnson, K. A. A new paradigm for DNA polymerase specificity. *Biochemistry* **45**, 9675–9687 (2006).
30. Ewing, B. et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
31. Loman, N. J. et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).
32. Foox, J. et al. Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. *Nat. Biotechnol.* **39**, 1129–1140 (2021).
33. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.* **3**, lqab019 (2021).
34. Heydari, M. et al. Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. *BMC Bioinformatics* **20**, 298 (2019).
35. Drmanac, S. et al. CoolMPS™: advanced massively parallel sequencing using antibodies specific to each natural nucleobase. Preprint at *bioRxiv* https://doi.org/10.1101/2020.02.19.953307 (2020).
36. Olson, N. D. et al. PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom.* **2**, 100129 (2022).
37. Biswas, P. et al. Avidity sequencing of whole genomes from retinal degeneration pedigrees identifies causal variants. Preprint at *medRxiv* https://doi.org/10.1101/2022.12.27.22283803 (2022).
38. Li, J. H. et al. Low-pass sequencing plus imputation using avidity sequencing displays comparable imputation accuracy to sequencing by synthesis while reducing duplicates. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.07.519512 (2022).
39. Olsen, T. R. et al. Scalable co-sequencing of RNA and DNA from individual nuclei. Preprint at *bioRxiv* https://doi.org/10.1101/2023.02.09.527940 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sinan Arslan[1,2], Francisco J. Garcia[1,2], Minghao Guo[1,2], Matthew W. Kellinger[1,2], Semyon Kruglyak[1,2], Jake A. LeVieux[1,2], Adeline H. Mah[1,2], Haosen Wang[1,2], Junhua Zhao[1,2], Chunhong Zhou[1,2], Andrew Altomare[1], John Bailey[1], Matthew B. Byrne[1], Chiting Chang[1], Steve X. Chen[1], Byungrae Cho[1], Claudia N. Dennler[1], Vivian T. Dien[1], Derek Fuller[1], Ryan Kelley[1], Omid Khandan[1], Michael G. Klein[1], Michael Kim[1], Bryan R. Lajoie[1], Bill Lin[1], Yu Liu[1], Tyler Lopez[1], Peter T. Mains[1], Andrew D. Price[1], Samantha R. Robertson[1], Hermes Taylor-Weiner[1], Ramreddy Tippana[1], Austin B. Tomaney[1], Su Zhang[1], Minna Abtahi[1], Mark R. Ambroso[1], Rosita Bajari[1], Ava M. Bellizzi[1], Chris B. Benitez[1], Daniel R. Berard[1], Lorenzo Berti[1], Kelly N. Blease[1], Angela P. Blum[1], Andrew M. Boddicker[1], Leo Bondar[1], Chris Brown[1], Chris A. Bui[1], Juan Calleja-Aguirre[1], Kevin Cappa[1], Joshua Chan[1], Victor W. Chang[1], Katherine Charov[1], Xiyi Chen[1], Rodger M. Constandse[1], Weston Damron[1], Mariam Dawood[1], Nicole DeBuono[1], John D. Dimalanta[1], Laure Edoli[1], Keerthana Elango[1], Nikka Faustino[1], Chao Feng[1], Matthew Ferrari[1], Keith Frankie[1], Adam Fries[1], Anne Galloway[1], Vlad Gavrila[1], Gregory J. Gemmen[1], James Ghadiali[1], Arash Ghorbani[1], Logan A. Goddard[1], Adriana Roginski Guetter[1], Garren L. Hendricks[1], Jendrik Hentschel[1], Daniel J. Honigfort[1], Yun-Ting Hsieh[1], Yu-Hsien Hwang Fu[1], Scott K. Im[1], Chaoyi Jin[1], Shradha Kabu[1], Daniel E. Kincade[1], Shawn Levy[1], Yu Li[1], Vincent K. Liang[1], William H. Light[1], Jonathan B. Lipsher[1], Tsung-li Liu[1], Grace Long[1], Rui Ma[1], John M. Mailloux[1], Kyle A. Mandla[1], Anyssa R. Martinez[1], Max Mass[1], Daniel T. McKean[1], Michael Meron[1], Edmund A. Miller[1], Celyne S. Moh[1], Rachel K. Moore[1], Juan Moreno[1], Jordan M. Neysmith[1], Cassandra S. Niman[1], Jesus M. Nunez[1], Micah T. Ojeda[1], Sara Espinosa Ortiz[1], Jenna Owens[1], Geoffrey Piland[1], Daniel J. Proctor[1], Josua B. Purba[1], Michael Ray[1], Daisong Rong[1], Virginia M. Saade[1], Sanchari Saha[1], Gustav Santo Tomas[1], Nicholas Scheidler[1], Luqmanal H. Sirajudeen[1], Samantha Snow[1], Gudrun Stengel[1], Ryan Stinson[1], Michael J. Stone[1], Keoni J. Sundseth[1], Eileen Thai[1], Connor J. Thompson[1], Marco Tjioe[1], Christy L. Trejo[1], Greg Trieger[1], Diane Ni Truong[1], Ben Tse[1], Benjamin Voiles[1], Henry Vuong[1], Jennifer C. Wong[1], Chiung-Ting Wu[1], Hua Yu[1], Yingxian Yu[1], Ming Yu[1], Xi Zhang[1], Da Zhao[1], Genhua Zheng[1], Molly He[1] & Michael Previte[1] ✉

[1]Element Biosciences, San Diego, CA, USA. [2]These authors contributed equally: Sinan Arslan, Francisco J. Garcia, Minghao Guo, Matthew W. Kellinger, Semyon Kruglyak, Jake A. LeVieux, Adeline H. Mah, Haosen Wang, Junhua Zhao, Chunhong Zhou. ✉e-mail: mprevite@elembio.com

Methods

Solution measurements of nucleotide incorporation

Solution measurements of nucleotide kinetics were performed using commercially available dATP-Cy5 (Jena Bioscience, catalog no. NU-1611-CY5-S). DNA substrates for solution kinetic assays were prepared by annealing a 5′FAM-labeled primer oligo (purchased from IDT) and high-performance liquid chromatography-purified (5′-CGAGCCGTCCAACCTACTCA-3′) with a template oligo (5′-ACGACCATGTTGAGTAGGTTGGACGGCTCG-3′). Annealing was performed with 10% excess template oligo in the annealing buffer using a PCR machine to heat oligos to 95 °C, followed by slow cooling to room temperature over 60 min. Solution kinetics were performed by mixing a preformed enzyme–DNA complex with fluorescent nucleotide and $MgSO_4$ using a RQF3 Rapid Quench Flow (KinTek Corp.). The enzyme used was an engineered variant of *Candidatus altiarchaeales* archaeon. The final reaction was conducted in 25 mM Tris pH 8.5, 40 mM NaCl and 10 mM ammonium chloride at 37 °C. Extension products were separated from unextended primer oligos by capillary electrophoresis using a 3500 Series Genetic Analyzer (ThermoFisher) to achieve single-base resolution. Products were quantified and fit to a single exponential equation. The observed rates as a function of nucleotide concentration were then fit to a hyperbolic equation to derive apparent K_d ($K_{d,app}$) and rate of polymerization (k_{pol}).

Avidite synthesis and construction

Initial research scale avidites were constructed by dissolving 5 mg of 10 kD 4-arm-PEG-SG (Laysan Bio, catalog no. 4arm-PEG-SG-10K-5g) in 100 µl of 95% organic solvent (for example, ethanol) and 5 mM MOPS pH 8.0 to make a 50 mg ml^{-1} solution (5 mM), 19 µl of which was combined with 1.5 µl of 10 mM dATP-NH_2 (7-deaza-7-propargylamin′-2′-deoxyadenosin′-5′-triphosphate; Trilink, catalog no. N-2068) and 8.0 µl of 3.75 mM 2 kD Biotin-PEG-NH_2 (Laysan Bio, catalog no. Biotin-PEG-NH2-2K-1g) in 95% organic solvent (for example, ethanol) and 5 mM MOPS pH 8.0. After mixing, 5 mM 10 kD 4-arm-PEG-SG was added. The final composition was 0.50 mM dA-NH_2, 1.0 mM biotin-PEG-NH2 (2 kD), 0.25 mM 4-arm-PEG-NHS, 85.5% organic solvent (for example, ethanol) and 4.5 mM MOPS pH 8.0. Following 1,000-rpm incubation at 25 °C for 90 min, the reaction volume was adjusted to 100 µl by the addition of MOPS pH 8.0. Purification was performed using a Biorad Biospin P6 column pre-equilibrated in 10 mM MOPS pH 8.0. The purified dATP-PEG–biotin complex was mixed with Zymax Cy5 Streptavidin (Fisher Scientific, catalog no. 438316) in a 2.5:1 volumetric ratio and allowed to equilibrate for 30 min at room temperature.

Real-time measurement of avidite association and dissociation

Real-time measurement of avidite binding kinetics was performed using an Olympus IX83 microscope at 545 and 635 nm excitation (Lumencor Light Engine) set to an approximate power density of about 1 W cm^{-2}, with an Olympus objective (catalog no. UCPLFLN20XPH) and a Semrock BrightLine multiband laser filter set (catalog no. LF405/488/532/635) containing a matching quad band exciter, emitter and dichroic. Flow rates of 60 µl s^{-1} were used for reagent exchanges. Circular PhiX libraries were introduced to AVITI flow cells, hybridized in 3× SSC buffer for 5 min at 50 °C and cooled to room temperature. Amplification reagents were introduced into the flow cell to perform rolling circle amplification and amplify genomic DNA. The instrument was paused following polony generation and priming and the flowcell moved to the microscope. Custom control software was written to control all peripheral hardware and synchronize data collection with flow of materials into the sample. Data collection (4 fps) was triggered by flow of the avidity mix and collected for 55 s. Polonies in the field were localized by a spot-finding algorithm, and background-corrected intensities were extracted versus time. Experiments were performed at 0.5 pM, 1 nM, 7.5 nM and 10 nM avidite or monovalent dye-labeled nucleotide concentrations. Substrates at the respective concentrations were combined with 100 nM engineered enzyme variant of *C. altiarchaeales* archaeon in the avidity on rate assay buffer formulation (25 mM HEPES pH 8.8, 25 mM NaCl, 0.5 mM EDTA, 5 mM strontium acetate, 25 mM ascorbic acid and 0.2% Tween-20). Avidites and nucleotides were labeled with Alexa Fluor 647. Higher-concentration data collection was limited by the ability to detect polony intensity from free avidite intensity at elevated concentrations. Off-rate measurements were performed by binding avidites to flowcell polonies, followed by washing with avidity on rate assay buffer and triggering of data collection.

Genomic DNA and next-generation sequencing library preparation

Human DNA from cell line sample HG002 was obtained from the Coriell Institute. Linear next-generation sequencing library construction was performed using a KAPA HyperPrep library kit (Roche, catalog no. 07962363001) according to published protocols. Finished linear libraries were circularized using the Element Adept Compatibility kit (catalog no. 830-00003). Final circular libraries were quantified by quantitative PCR with the standard and primer set provided in the kit. Circular library DNA was denatured using sodium hydroxide and neutralized with excess Tris pH 7.0 before dilution. Denatured libraries were diluted to 8 pM in hybridization buffer before loading onto the sequencing cartridge.

Single-cell 3′ gene expression library circularization

Single-cell RNA-seq libraries were prepared from two lots of peripheral blood mononuclear cell suspension (10,000 and 1,000 cells) using the Chromium Next GEM Single Cell 3′ Kit v.3.1 (catalog no. 1000268). Each library was quantified and individually processed for sequencing using the Adept Library Compatibility Kit (catalog no. 830-00003). Processed libraries were pooled and sequenced with 28 cycles for read 1, 90 for read 2 and index reads.

Sequencing instrument and workflow

Sequencing results were obtained with commercialized formulations of avidites, enzymes and buffers. Element Bioscience's AVITI commercial system (catalog no. 88-00001) was used for all sequencing data. AVITI 2 × 150 kits were loaded on the instrument (catalog no. 86-00001). Primary analysis was performed onboard the AVITI sequencing instrument, and FASTQ files were subsequently analyzed using a secondary analysis pipeline from Sentieon.

Sequencing primary analysis

Four images were generated per field of view during each sequencing cycle, corresponding to the dyes used to label each avidite. An analysis pipeline was developed that uses the images as input to identify the polonies present on the flowcell and to assign to each polony a base call and quality score for each cycle, representing the accuracy of the underlying call. The analysis approach has steps similar to those described in ref. 25. Briefly, intensity is extracted for each polony in each color channel; intensities are then corrected for color cross-talk and phasing and normalized to make cross-channel comparisons. The highest normalized intensity value for each polony in each cycle determines the base call. In addition to assigning a base call, a quality score corresponding to call confidences is also assigned. The standard Q-score definition is utilized where the Q-value is defined as $Q = -10 \times \log\_10p$, where p is the probability that the base call is an error. Q-score generation follows the approach of Ewing et al., with modified predictors[21], and is encoded using the phred+33 ASCII scheme. The predictors used for quality score training are (1) maximum intensity per polony across color channels; (2) clarity of each polony (defined as $(A+1)/(B+1)$, where A is the highest intensity across color channels and B is the second highest); (3) the sum of phasing and prephasing estimates; and (4) the median clarity value taken across the 10% of the lowest-intensity polonies. The sequence of base call assignments and quality scores across the cycles constitutes the output of the run. These data are represented in standard FASTQ format for compatibility with downstream tools.

Quality score assessment

To assess the accuracy of quality scores (Fig. 3), the FASTQ files were aligned with BWA to generate BAM files. GATK BaseRecalibrartor was then applied to the BAM, specifying files of publicly available known sites to exclude human variant positions.

K-mer error analysis

The same run used to generate recalibrated quality scores was analyzed via custom script for all *k*-mers of size 1, 2 and 3. The computation is based on 1% of a 35X genome to ensure adequate sampling of each *k*-mer. For example, each 3-mer is sampled at least 850,000 times (average 6.7 million). This figure is based on a publicly available run from each platform. For the instances of each *k*-mer, the percentage mismatching a variant-masked reference was computed. The same script was applied to a publicly available NovaSeq dataset for HG002 and a publicly available NextSeq 2000 dataset for HG001 (Demo Data for HG002 were not available). We tabulated the number of *k*-mers in which the percentage incorrect was lowest for AVITI among the three platforms compared.

Homopolymer analysis

A BED file provided by National Institute of Standards and Technology (NIST) genome-stratifications v.3.0, containing 673,650 homopolymers of length >11, was used to define regions of interest for homopolymer analysis (GRCh38_SimpleRepeat_homopolymer_gt11_slop5). Reads overlapping these BED intervals (using samtools view -L and adjusting for slop5) were selected for accuracy analysis. Reads with any of the following flags set were discarded: secondary, supplementary, unmapped or reads with mapping quality of 0. Reads were oriented in the 5′→3′ direction and split into three segments: preceding the homopolymer, overlapping it and following it. The mismatch rate for each read segment was computed, excluding N-calls, softclipped bases and indels. For example, if a 150-bp read (aligned on the forward strand) contained a homopolymer in positions 100–120, the first 99 cycles were used to compute the error rate before the homopolymer and the last 30 to compute error rate following the homopolymer. Reads were discarded if the sequence either preceding or following the homopolymer was <5 bp in length. All reads were then stacked into a matrix according to their positional offset relative to the homopolymer, and error rate per post-offset was computed.

Average error rate was computed for avidity sequencing runs and for publicly available data from multiple SBS instruments, for comparison. Differences oin mismatch percentage, across all BED intervals, between AVITI and NovaSeq were plotted in a histogram and examples showing various percentiles within the distribution were chosen for display via Integrative Genomics Viewer.

Publicly available datasets for NovaSeq were obtained from the Google Brain Public Data repository on Google Cloud (https://console.cloud.google.com/storage/browser/brain-genomics-public/research/sequencing/fastq). Publicly available NextSeq 2000 data were obtained from Illumina Demo Data on BaseSpace (https://basespace.illumina.com/datacentral).

Single-cell gene expression data analysis

Following sequencing, Bases2Fastq software was used to generate FASTQ files for compatible upload into 10X Cloud and subsequent analysis with the 10X Genomics Cell Ranger analysis package. Data visualization of single-cell gene expression profiling was generated using 10X Genomics Loupe Browser.

Whole-genome sequencing analysis

A FASTQ file with base calls and quality scores was downsampled to 35× raw coverage (360,320,126 input reads) and used as an input into Sentieon BWA followed by Sentieon DNAscope[40]. Following alignment and variant calling, variant calls were compared with the NIST genome in Bottle Truth Set v.4.2.1 via the hap.py comparison framework to derive total error counts and F1 scores[41]. The results are computed based on the 3,848,590 SNV and 982,234 indel passing variant calls made by DNAScope.

1 × 300 Data generation

An *E. coli* library was prepared using enzymatic shearing and PCR amplification. The library was then sequenced for 300 cycles using new enzymes for stepping along the DNA template and for avidite binding. The reagent formulation with increased enzyme and nucleotide concentrations during the stepping process was used to improve stepping performance. The contact times for avidite binding and exposure were both reduced without performance losses, to decrease cycle time over the 600 cycles of sequencing. The displays show only 299 cycles of data, because cycle 300 was used only for prephasing correction. To minimize soft clipping during alignment the following inputs were used in the call to BWA–MEM: -E 6,6 -L 1000000 -S.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The avidity sequencing datasets described in the paper are available for download via the AWS CLI in the public bucket s3://avidity-manuscript-data/, pending upload to the sequence read archive under BioProject PRJNA869673. Publicly available datasets for NovaSeq were obtained from the Google Brain Public Data repository on Google Cloud (https://console.cloud.google.com/storage/browser/brain-genomics-public/research/sequencing/fastq). Publicly available NextSeq 2000 data were obtained from Illumina Demo Data on BaseSpace (https://basespace.illumina.com/datacentral).

Code availability

Scripts used for analysis are available via GitHub (https://github.com/Elembio/AvidityManuscript2023).

References

40. Freed, D. et al. The Sentieon Genomics Tools—a fast and accurate solution to variant calling from next-generation sequence data. Preprint at *bioRxiv* https://doi.org/10.1101/115717 (2017).
41. Krusche, P. et al. Author correction: Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 567 (2019).

Acknowledgements

We thank J. Puglisi and T. Ben-Yehezkel for valuable comments and discussion during the writing of the paper.

Author contributions

The author list is divided into three sections, each in alphabetical order. Authors in the first section made equal contributions to the critical elements of the technology and paper development. Authors in the second category made specific technology contributions described within the paper. Authors in the third group helped to develop some aspects of the underlying technology that culminated in the final product. M.H. and M.P. shared in the intellectual supervision of the work.

Competing interests

All authors are current or former employees of Element Biosciences. All authors may hold stock options in the company.

Additional information

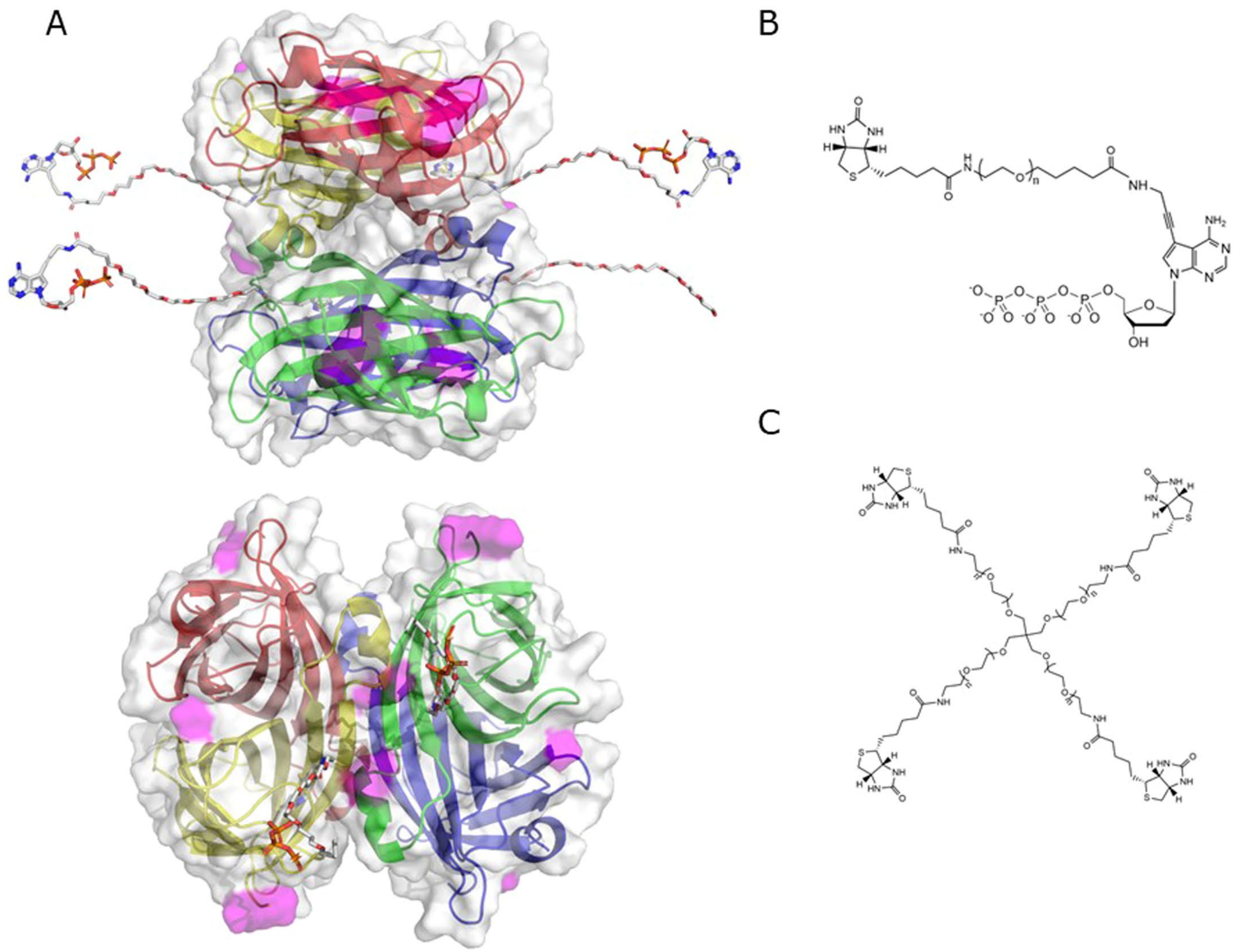
Extended data is available for this paper at https://doi.org/10.1038/s41587-023-01750-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01750-7.

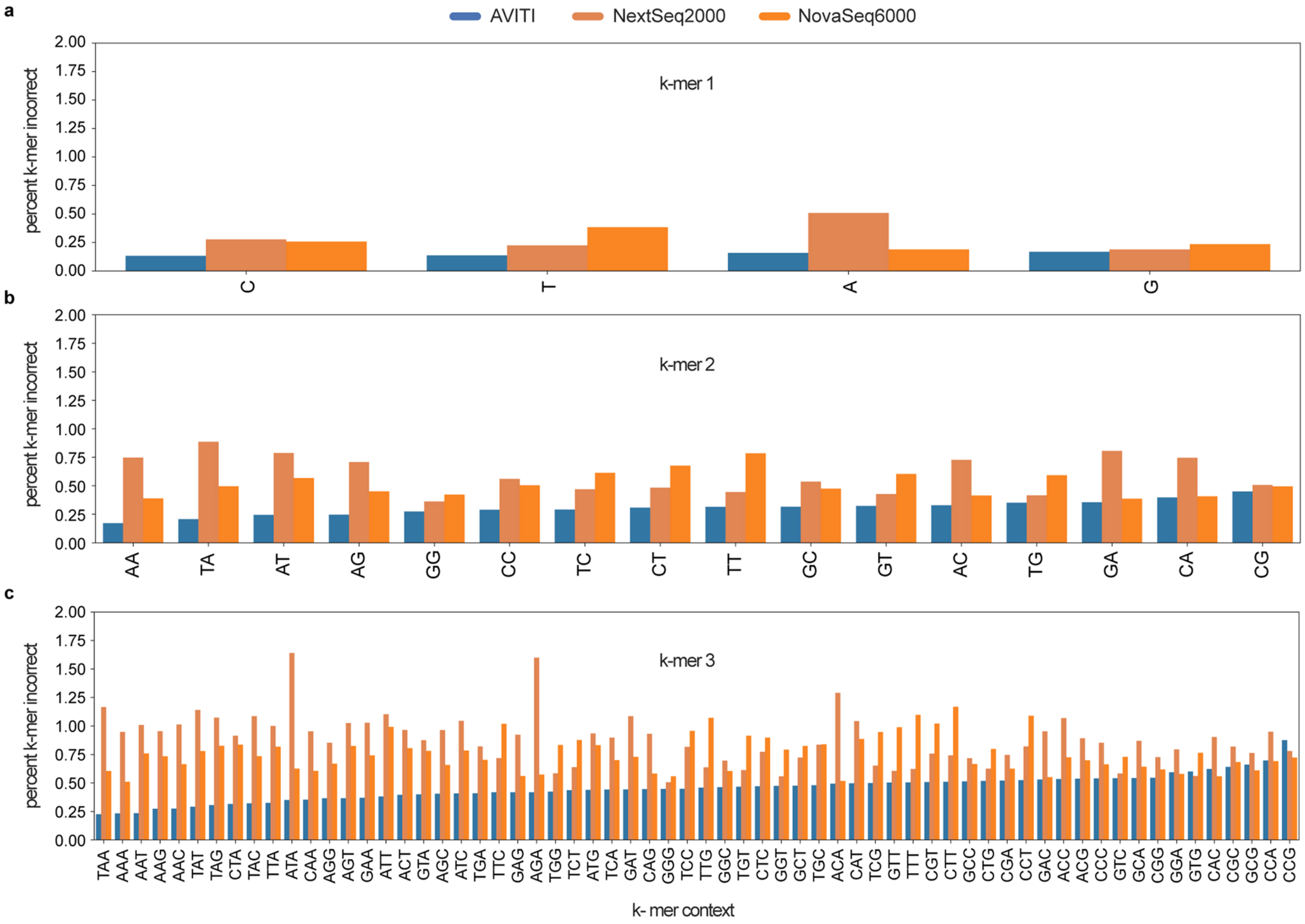
Correspondence and requests for materials should be addressed to Michael Previte.

Peer review information *Nature Biotechnology* thanks Michael Quail, Kenneth Beckman, Nathanael Olson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

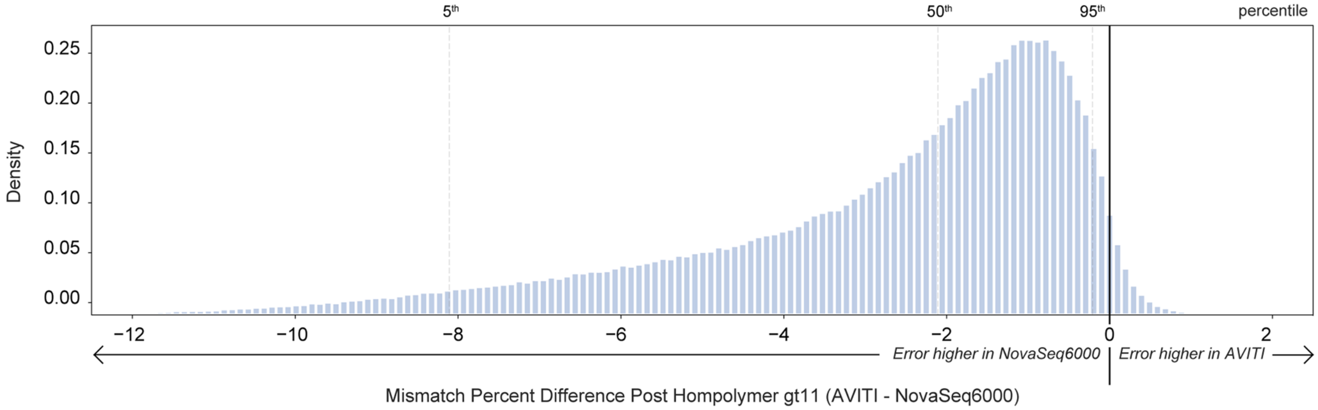
Reprints and permissions information is available at www.nature.com/reprints.



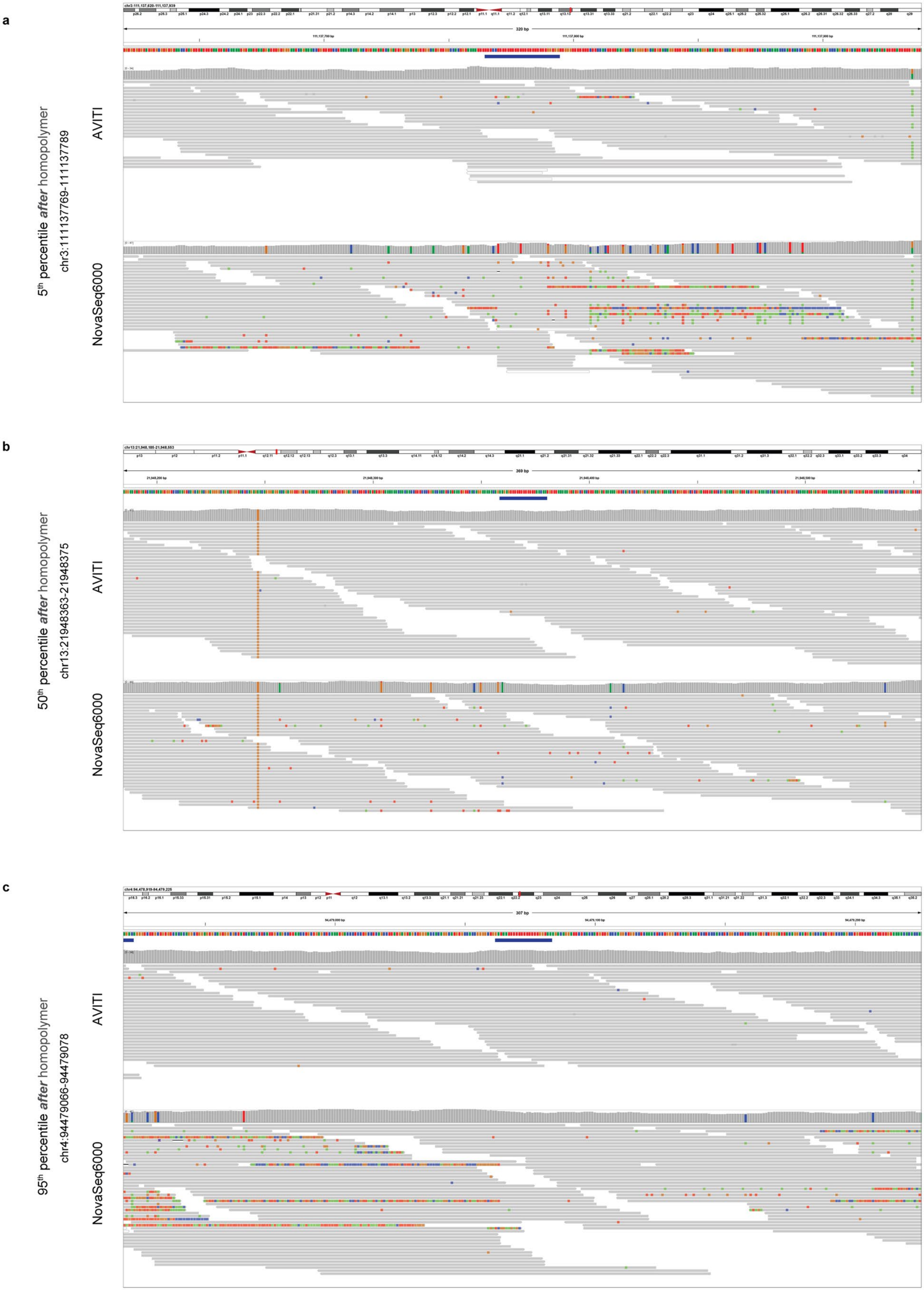
Extended Data Fig. 1 | Model of an avidite. (**a**) side and top views of a modeled avidite. The protein core consists of fluorophore labeled streptavidin. The monomers of tetrameric streptavidin are colored red, blue, green, and yellow. Dye conjugation sites through lysine-NHS chemistry are denoted in the surface rendering as magenta. Fluorophores are not pictured. Avidite arms are associated via a biotin interaction with the core streptavidin protein. Arms are mixed stoichiometrically to achieve averages of three nucleotide containing arms and one linker to additional cores. Molecules conjugated to have been shortened in this representation. (**b**) Structure of an avidite arm. (**c**) Structure of the 4-arm linker connecting avidite cores.



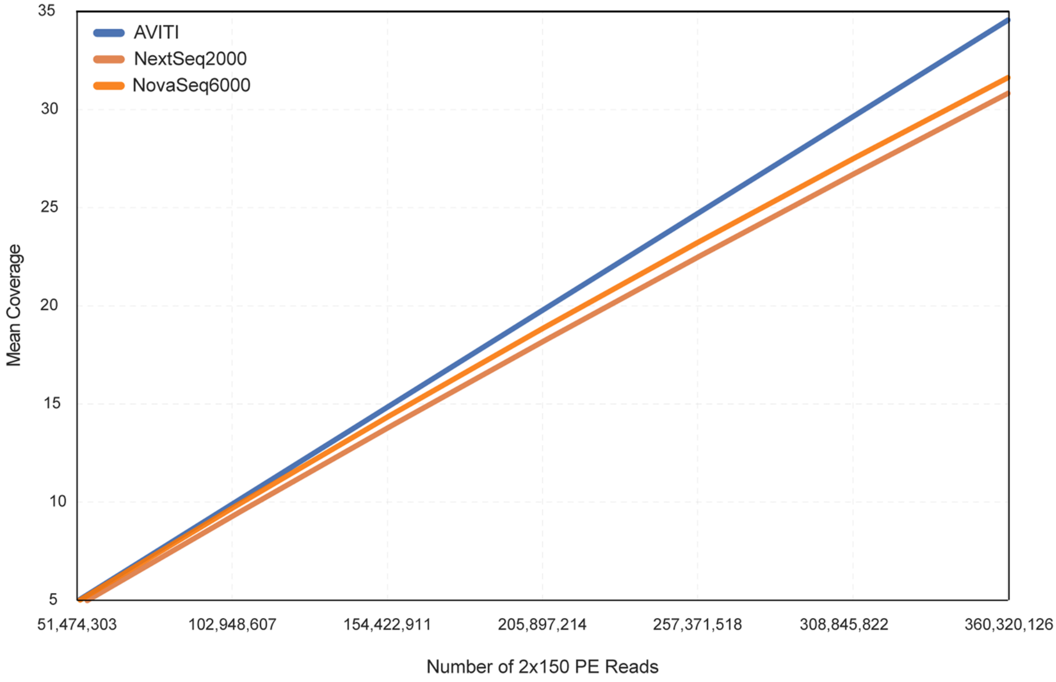
Extended Data Fig. 2 | Percentage of instances that a k-mer contained at least one mismatch compared across 3 instruments. Panels **a**, **b**, and **c** display 1-mers, 2-mers, and 3-mers, respectively. The bars are sorted by AVITI contexts from most to least accurate.



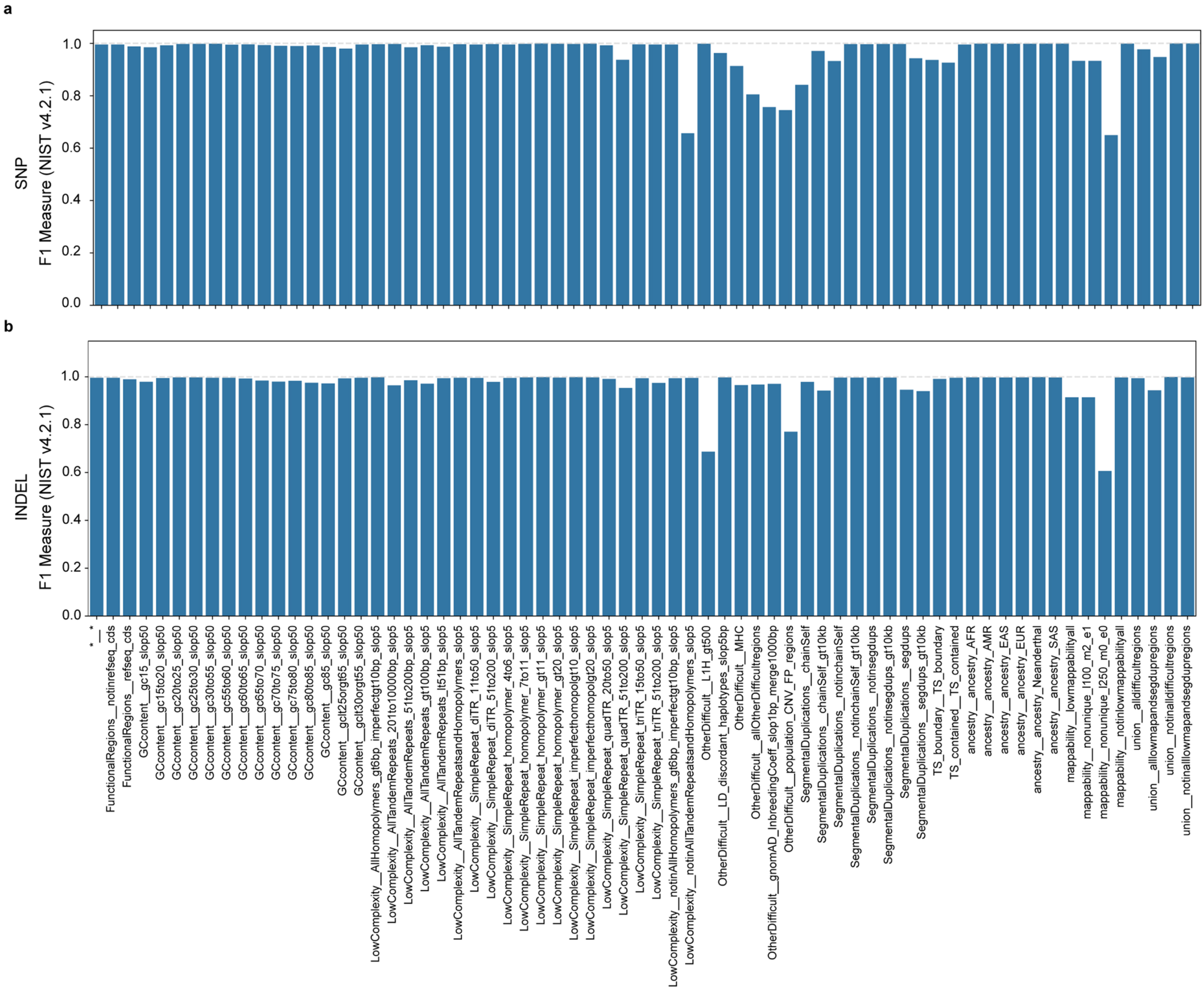
Extended Data Fig. 3 | Histogram of pairwise error differences. Difference was selected as the metric to cancel the effects of human variants from the mismatch percent.



Extended Data Fig. 4 | IGV display of homopolymer loci at the 5th, 50th, and 95th percentile of AVITI minus NovaSeq mismatch percent (corresponding to the dashed lines of Extended Data Fig. 3). The red bar at the top indicates the homopolymer. Colors within the IGV read stack correspond to mismatches and softclipping. Only mismatches contribute to the error rate calculation and softclipped bases are ignored.



Extended Data Fig. 5 | Comparison of read number vs genomic coverage computed via Picard for PCR-free whole genome data. AVITI most closely matches the 45-degree line due to the low duplicate rate.

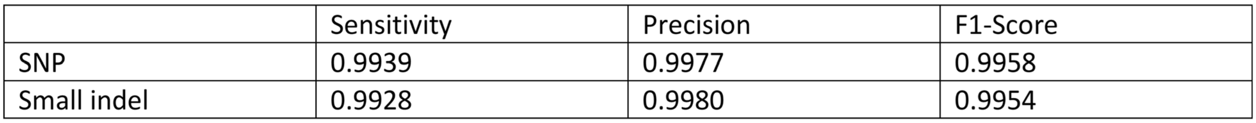


Extended Data Fig. 6 | F1 Score of SNPs and indels across GiaB stratifications. F1 score for SNPs and indels stratified by all GiaB regions with at least 100 variants in the 4.2.1 truth set of sample HG002.

Extended Data Table 1 | Single cell expression: CellRanger metric values for 10 K cell and 1 K cell libraries from the PBMC reference

CellRanger v7.0 Metric	Performance expectation	AVITI 10K cells	AVITI 1K cells
Valid barcodes	>90%	97.5%	97.5%
Reads mapped confidently to exonic regions	>50%	53.0%	53.8%
Read mapped confidently to transcriptome	>40%	74.7%	77.8%
Fraction reads in cells	>80%	95.5%	92.6%
Q30 bases in barcode	>85%	99.5%	99.5%
Q30 bases in RNA read	>75%	98.6%	98.8%
Mean reads per cell	>50,000	61,326	68,766
Median genes per cell	>1700	2,910	2,951
Total genes detected	N/A	23,863	29,679
Estimated number of cells	+/-20%	8,513	922

Extended Data Table 2 | Variant calling performance for HG002 on GIAB-HC regions



	Sensitivity	Precision	F1-Score
SNP	0.9939	0.9977	0.9958
Small indel	0.9928	0.9980	0.9954

nature portfolio

Corresponding author(s): Michael Previte

Last updated by author(s): 3/7/2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Kinetic data for Figure 2A was collected using RQF3 Rapid Quench flow (Kintek corporation). Real-time measurements for Figure 2B-E were collected on an Olympus IX83 microscope equipped with 545 and 637 lines (Lumencor), Semrock brightline multiband laser filter set (LF405/488/532/635) containing matching quad band exciter, emitter and dichroic. Flow was induced by a syringe pump pulling reagents across an AVITI flow cell at a rate of 60 ul/s. Prior to injection of reagents, real-time data was collected on an Andor sCMOS camera at 4 frames/s. All sequencing data was collected on the AVITI commercial instrument.
Data analysis	Kinetic data was analyzed and fit using conventional non-linear regression. All error bounds were propagated in the analysis and are reflected in figure 2 panel A. Reported kcat and Kd,app were obtained by fitting to a hyperbolic equation using no constrains other than the error reported for each point. Primary analysis of the collected data was performed on the AVITI instrument according to similar steps described on Whiteford et al. (25) FASTQ were generated using the bases2fastq software toolkit (version 1.1.1). Tools and scripts supporting bioinformatic analysis of this manuscript can be found at the following repo located on github - https://github.com/Elembio/AvidityManuscript2023. Single cell RNA was performed using CellRanger (version 7.0.1). Whole genome sequencing analysis was performed by first down-sampling the input FASTQ to 35X raw coverage (360,320,126, 2x150 input reads), and then aligning, de-duplicating and sorting using sentieon bwa (version 202112.02). The BAM was then used as input to Sentieon DNAscope (version 202112.02) in addition to a element specific ML model (SentieonDNAscopeModelElementBio0.3.model) to produce a VCF. Following alignment and variant calling, the variant calls were benchmarked using hap.py (version hap.py-0.3.14) to the NIST genome in a bottle truth set v4.2.1 across all regions to derive total error counts and F1 scores. To assess the accuracy of quality scores shown in Fig. 3, the aligned BAMS were processed using GATK BaseRecalibrartor (version gatk4:4.2.0.0—0), and specifying publicly available known sites files to exclude human variant positions (HG002 NIST v4.2.1 bed/vcf,

1000G_phase1.snps.high_confidence.hg38, dbsnp_144.hg38). The resulting predicted and recalibrated q-scores were plotted.
To compute the mismatch percentage of AVITI, NovaSeq 6000, and NextSeq 2000 reads before and after homopolymers of length 12 or greater, a BED file provided by NIST genome-stratifications v3.0, containing 673,650 homopolymers of length greater than 11 was used to define the regions of interest for the homopolymer analysis (GRCh38_SimpleRepeat_homopolymer_gt11_slop5). Reads that overlapped these BED intervals (using samtools version 1.1.1) were selected for accuracy analysis. Reads with any of the following flags set were discarded (secondary, supplementary, unmapped or reads with mapping quality of 0). Reads were oriented in the 5' -> 3' direction, and split into 3 segments, preceding the homopolymer, overlapping the homopolymer, and following the homopolymer. The mismatch rate for each read-segment was computed, excluding N-calls, softclipped bases and indels. For example, if a 150 bp read (aligned on the forward strand) contains a homopolymer in positions 100-120, then the first 99 cycles were used to compute the error rate prior to the homopolymer, and the last 30 cycles were used to compute the error rate following the homopolymer. Reads were discarded if either the sequence preceding or following the homopolymer was less than 5bp in length (accounting for the GIAB slop used). All reads were then stacked into a matrix, according to their positional offset relative to the homopolymer, and error rate per pos-offset was computed.
The average error rate was computed for avidity sequencing runs and for publicly available data from multiple SBS instruments, for comparison. The differences of mismatch percentages, across all BED intervals, between AVITI™ and NovaSeq were plotted in a histogram and examples showing various percentiles within the distribution were chosen for display via IGV.

The interval-error.tsv and offset-error.tsv files can be found in the following directory: https://github.com/Elembio/AvidityManuscript2023/tree/main/data/homopolymer-error/GRCh38_SimpleRepeat_homopolymer_gt11_slop5
To compute the mismatch percent difference between avidity sequencing and SBS across homopolymer lengths, the four GIAB supplied homopolymer bed files were combined, and duplicates were removed (4to6, 7to11, gt11, gt20), producing a new bed file representing all homopolymer of size 4 to inf. The box plot shows median, quartiles, and the whiskers are 1.5*IQR.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The avidity sequencing data sets described in the manuscript are available for download via the AWS CLI using the following command:
aws s3 ls --no-sign-request s3://avidity-manuscript-data/

Samples and FASTQ have been accessioned in SRA under BioProject PRJNA869673.

Bioinformatic tools and scripts can be found on the following github repo: https://github.com/Elembio/AvidityManuscript2023

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	There were no human research participants in this study.
Population characteristics	There were no human research participants in this study.
Recruitment	There were no human research participants in this study.
Ethics oversight	There were no human research participants in this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sequencing calibration studies were performed on 20 samples. Single cell studies were performed on multiple samples that generated consistent results, but a single example was used for this particular study. To determine k-mer errors, a million k-mers of each length were used to determine percent mismatch. For the homopolymer analysis, ~700,000 loci were used. For GiaB stratifications, we selected context classes with at least 100 variants.
Data exclusions	There was no data excluded (Filtered data is excluded from the sequencing runs).
Replication	We checked that all presented runs are representative by looking at no fewer than 20 sequencing runs. For analyses such as homopolymer and k-mer accuracy, sample size calculations are based on the number of relevant loci within a run. There were no failures to replicate.
Randomization	The study performed was validating first principles studies such as enzyme kinetics to validate the hypotheses of avidity chemistry, thus sample randomization would not be necessary. Sequencing data was performed on known samples and comparative metrics to known reference samples also obviates the need for randomization of the studies as the known reference samples are a widely known control.
Blinding	The study performed was validating first principles studies such as enzyme kinetics to validate the hypotheses of avidity chemistry, thus blind would not be necessary. Sequencing data was performed on known samples and comparative metrics to known reference samples also obviates the need for blind studies as the known reference samples are a widely known control.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging