

Guardant - EXHIBIT 2023
Tempus AI, Inc. v. Guardant Health, Inc.
IPR2025-01434

CODEC enables ‘single duplex’ sequencing

Jin H. Bae^{†,1}, Ruolin Liu^{†,1}, Erica Nguyen,¹ Justin Rhoades,¹ Timothy Blewett,¹ Kan Xiong,¹ Douglas Shea,¹ Gregory Gydush,¹ Shervin Tabrizi,^{1,2,3} Zhenyi An,¹ Sahil Patel,¹ G. Mike Makrigiorgos,⁴ Todd R. Golub,¹ and Viktor A. Adalsteinsson^{*1}

¹*Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA*

²*Koch Institute for Integrative Cancer Research at MIT, Cambridge, Massachusetts, USA*

³*Harvard Radiation Oncology Program, Boston, Massachusetts, USA*

⁴*Dana-Farber Cancer Institute, Boston, Massachusetts, USA*

[†]*Equal contribution*

**Correspondence should be addressed to V.A.A. (email: viktor@broadinstitute.org)*

(Dated: June 11, 2021)

Abstract

Detecting mutations as rare as a single molecule is crucial in many fields such as cancer diagnostics and aging research but remains challenging. Third generation sequencers can read a double-stranded DNA molecule (a ‘single duplex’) in whole to identify true mutations on both strands apart from false mutations on either strand but with limited accuracy and throughput. Although next generation sequencing (NGS) can track dissociated strands with Duplex Sequencing, the need to sequence each strand independently severely diminishes its throughput. Here, we developed a hybrid method called Concatenating Original Duplex for Error Correction (CODEC) that combines the massively parallel nature of NGS with the single-molecule capability of third generation sequencing. CODEC physically links both strands to enable NGS to sequence a single duplex with a single read pair. By comparing CODEC and Duplex Sequencing, we showed that CODEC achieved a similar error rate (10^{-6}) with 100 times fewer reads and conferred ‘single duplex’ resolution to most major NGS workflows.

Introduction

Discovering extremely low-level mutations as rare as within a single double-stranded DNA molecule (a ‘single duplex’) is crucial to finding diagnostic[1, 2], predictive[3, 4], and prognostic[5, 6] biomarkers, understanding cancer evolution[7, 8] and somatic mosaicism[9, 10], and studying infectious diseases[11, 12] and aging[13, 14]. Third generation sequencing technologies (e.g., PacBio, Oxford Nanopore Technologies) in principle make it possible to sequence each single DNA duplex in whole to resolve true mutations on both strands apart from false mutations on either strand, but, in practice, lack the required accuracy and throughput[15, 16]. Next generation sequencing (NGS), on the other hand, continues to offer superior read accuracy and throughput[17], but is not configured to sequence single duplexes—at least not without severely compromising its throughput or utility.

NGS affords high throughput by reading short, clonally amplified DNA fragments in massively parallel fluorescence analysis. Yet, its accuracy is limited by the need to dissociate Watson and Crick strands of each DNA duplex. Without a complementary strand for comparison, errors introduced on either strand due to base damage[18], PCR[19], and sequencing[20] can be disguised as real mutations (Fig. 1a). While it is possible to use unique molecular identifiers (UMIs) to separately track both strands of each DNA molecule and compare their sequences to detect true mutations on both strands of each duplex[21, 22], it does not solve the underlying limitation of NGS: duplex dissociation. For example, Duplex Sequencing[23] tags double-stranded UMIs on each original duplex to trace them back after PCR and NGS. By forming a duplex consensus between reads assigned to the Watson and Crick strands of each original duplex, Duplex Sequencing achieves 1,000-fold or higher accuracy (error rate below 10^{-6}) and can thus resolve true mutations within single DNA duplexes. However, recovering both strands among up to 10 billion other strands on an NGS flow cell (e.g., Illumina

NovaSeq) requires 100-fold excess reads[24], which invariably diminishes the throughput of NGS and severely limits its applicability.

To date, a few methods have sought to overcome the high inefficiency of Duplex Sequencing. Duplex Proximity Sequencing (Pro-Seq)[25] uses a polyethylene glycol linker to link 5′-ends of an original Watson strand and a copied Crick strand of a duplex to avoid hairpin formation for whole-genome sequencing (WGS). However, concatenating two strands with the opposite directions blocks DNA amplification which is necessary for most applications. CypherSeq[26] generates a circularized duplex followed by rolling circle amplification, but the lack of asymmetry between the two strands obscures whether both strands were actually sequenced. Some technologies such as o2n-seq[27] and Circle Sequencing[28] are compatible with PCR but only link a single strand of each duplex and thus, lack the ability to create a duplex consensus. BotSeqS[29, 30] uses dilution instead of linking to increase the chance of recovering both strands, but by doing so it only sequences 0.001% of the input DNA. Despite the need for sequencing single duplexes with high accuracy and throughput, there has been no such method with universal applicability. We thus reasoned that linking the information of both strands before dissociation could make NGS capable of reading single DNA duplexes with high accuracy and throughput.

Here, we developed a method that combines the massively parallel nature of NGS and the single-molecule capability of third generation sequencing to sequence both strands of each DNA duplex with single read pairs. In this hybrid approach called Concatenating Original Duplex for Error Correction (CODEC), each molecule becomes self-sufficient for forming a duplex consensus via NGS (Fig. 1a). By using the opposite strand as a template for extension instead of directly linking them, CODEC physically concatenates the sequence information of Watson and Crick strands into a single strand without forming a strong hairpin structure (Fig. 1b). Any differences between concatenated sequences

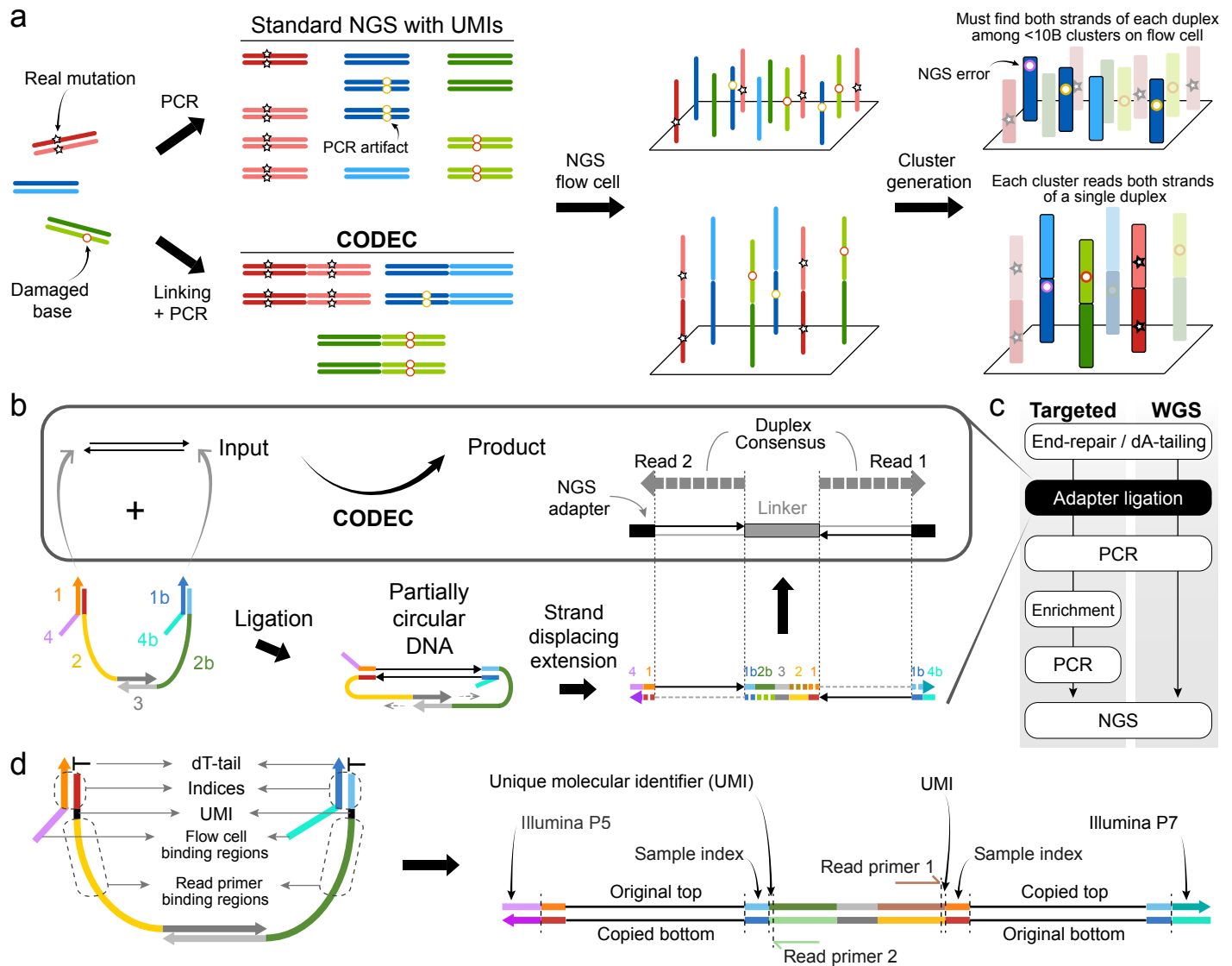


FIG. 1. Overview of Concatenating Original Duplex for Error Correction (CODEC). (a) Standard NGS workflows involve dissociation of DNA duplex, which loses the intrinsic property of DNA that encodes genetic information twice. Both strands of a duplex can be tracked through unique molecular identifiers (UMIs) to identify false mutations caused by base damage, PCR, and NGS errors, but finding them among <10 billion other strands costs throughput, highlighted by blue clusters. CODEC physically links each duplex before dissociation, ensuring each library molecule retains information of both strands. (b) CODEC links the sequence information of an original duplex into a single strand. As a result, each pair of NGS reads becomes self-sufficient for forming a duplex consensus (box). It utilizes the adapter complex instead of a duplex adapter for ligation, followed by strand displacing extension. (c) CODEC modifies the ligation step of ligation-based NGS workflows. (d) CODEC adapter complex is prepackaged with all of the components needed for Illumina NGS. Unlike standard NGS libraries, CODEC reads outward to sequence a UMI, an index, and an insert together. No indexed primers are required as indices and flow cell binding regions (P5 and P7) are added by the ligation.

95 would indicate either non-canonical base pairing created by
 96 nucleobase damage or an alteration confined to one strand
 97 of the original DNA duplex, or an error introduced during
 98 PCR amplification or sequencing. We tested CODEC with
 99 different sample types and NGS workflows, and confirmed
 100 that it suppressed both single nucleotide variants (SNV) and
 101 indel errors as accurately as Duplex Sequencing but with 100-
 102 fold fewer reads, thereby conferring ‘single duplex’ resolution
 103 to NGS.

104 Results

105 **CODEC adapter complex and workflow.** The CODEC
 106 structure can be built by a streamlined workflow using a
 107 commercial ligation-based NGS preparation kit and CODEC
 108 adapter complex. First, a typical duplex adapter was replaced
 109 with the adapter complex consisting of four oligonucleotides,

110 containing all elements required for NGS. We rationally de-
 111 signed double-stranded segments of the adapter to hold the
 112 whole complex based on DNA hybridization thermodynam-
 113 ics (**Supplementary Figure S1a**) and introduced single-
 114 stranded segments to mitigate bending stiffness of rigid double
 115 helix (**Supplementary Figure S1b**). After adapter ligation
 116 closes both ends of an input molecule, strand displacing exten-
 117 sion initiates at remaining 3’-ends to elongate each strand by
 118 using the opposite strand as a template. The resulting struc-
 119 ture is two original strands concatenated with the CODEC
 120 linker in the middle and NGS adapters on both sides. The
 121 molecular process depicted in **Fig. 1b** is integrated into the
 122 adapter ligation step of commercial NGS library construction
 123 kits (**Fig. 1c**).

124 To fully utilize the concatenated holder, we also relo-

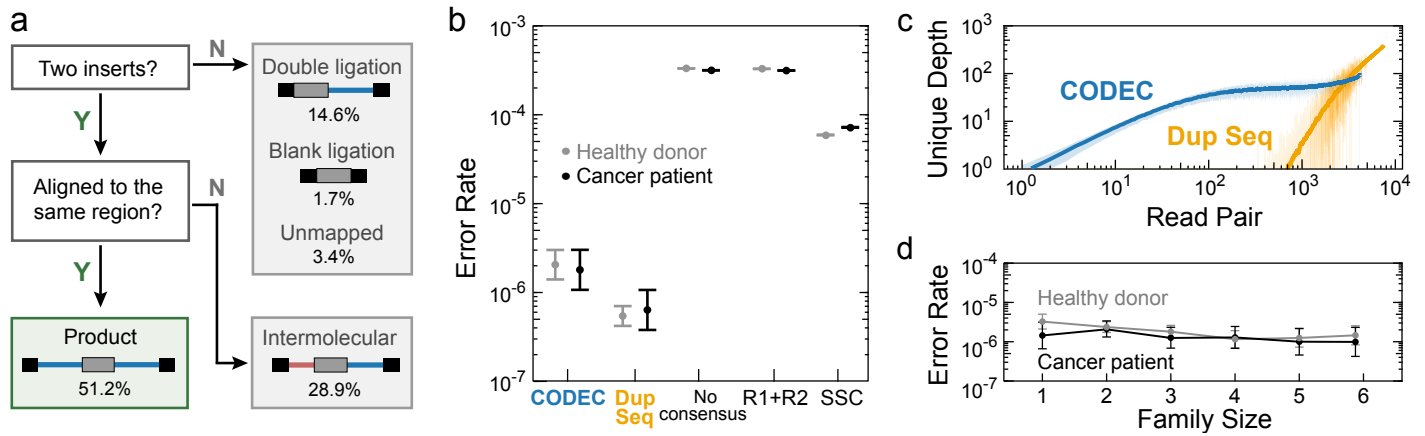


FIG. 2. Proof-of-concept. (a) Ratios of the correct CODEC product and byproducts which have been named after how they were likely created. (b) Error rates of CODEC, Duplex Sequencing, and other consensus methods including typical paired-end read (R1+R2) and single strand consensus (SSC). Target enrichment with a pan-cancer gene panel was performed on cell-free DNA (cfDNA) of two individuals. Error bars indicate 95% binomial confidence intervals. (c) Recovery of unique original duplexes per captured region in healthy donor cfDNA against the amount of sequencing. Solid lines show moving averages and shades indicate standard deviations. (d) CODEC error rates at each family size, which is the number of raw reads with the same UMI and start-stop positions.

125 cated the NGS library components (Fig. 1d). In contrast to
 126 the conventional Illumina structure with the NGS read primer
 127 binding sites on the outer side, we moved the binding sites
 128 to the CODEC linker in the middle and sequenced outward
 129 to prevent reading molecules without the linker (Supple-
 130 mentary Figure S1c). Having the binding sites at conven-
 131 tional locations had resulted in poor Quality Scores, which
 132 we attributed to template hopping in cluster amplification
 133 (Supplementary Figure S2a), whereas moving the bind-
 134 ing sites to the linker overcame this issue (Supplementary
 135 Figure S2b). Sample indices, which are typically located
 136 outer to the read primer binding sites and read separately
 137 from the inserts, were moved right next to the inserts. By
 138 adding the indices during adapter ligation and reading them
 139 with the inserts in a single step, CODEC suppressed index
 140 hopping even better than the gold standard of using unique
 141 dual indices [31, 32] (0.056% vs. 0.16%). We designed sets
 142 of 4 sample indices that collectively have all four bases at
 143 every position to ensure high base diversity for proper cluster
 144 identification, phasing correction, and chastity filtration
 145 (Supplementary Figure S3). Because indexed primers
 146 were no longer needed, we were able to include Illumina P5
 147 and P7 segments in the adapter complex and use them as
 148 universal primer binding regions.

149 **Proof-of-concept.** We first confirmed that the CODEC
 150 workflow could create the intended NGS library structure
 151 by converting fragmented human genomic DNA (gDNA)
 152 from peripheral blood mononuclear cells into a CODEC-
 153 NGS library and sequencing it. Due to the novel structure of
 154 CODEC reads, we created a user-friendly analysis pipeline
 155 called CODECsuite to process the data (see Methods). We
 156 found that more than half of the reads showed the correct
 157 structure (Fig. 2a). Meanwhile, the major byproducts ap-
 158 peared to have been created when an input duplex was either
 159 ligated to two different adapter complexes (“double ligation”)
 160 or no adapter complex (“blank ligation”), or when strand
 161 displacing extension occurred between two ligated products
 162 (“intermolecular”) (Supplementary Figure S4). Yet, al-
 163 most 90% of byproducts still retained information on one
 164 side of a duplex just like standard NGS, suggesting that the

165 byproducts may still yield useful data.

166 We next explored whether the fragments with the correct
 167 CODEC structure could provide comparable error rates to
 168 Duplex Sequencing using significantly fewer reads. To assess
 169 this, we performed a head-to-head comparison. Because
 170 Duplex Sequencing requires high sequencing depth per locus,
 171 we ran target enrichment with a pan-cancer panel on NGS
 172 libraries prepared with each method, built from 20 ng cell-free
 173 DNA (cfDNA) from a cancer patient and a healthy donor.
 174 We found that the mean CODEC error rate of two individuals
 175 (1.9×10^{-6}) was similar to that of Duplex Sequencing ($5.9 \times$
 176 10^{-7}) (Fig. 2b) with no statistically significant difference in
 177 sequence contexts of errors except for C:G>T:A in a healthy
 178 donor (Supplementary Figure S5a), which we believe
 179 could be resolved using an improved end-repair method [30, 33]
 180 (Supplementary Figure S5b). Additionally, when error
 181 rates were plotted as a function of distance from either end
 182 of a fragment, we saw elevated error rates from CODEC and
 183 Duplex Sequencing data toward the fragment ends of duplex
 184 consensus, consistent with prior reports of error propagation
 185 in end-repair [30, 33] (Supplementary Figure S6). This
 186 observation reassures that reading a single CODEC fragment
 187 is equivalent to reading two Duplex Sequencing fragments
 188 from each strand and affirms the need to trim 12 base pairs
 189 (bp) from both ends of each original DNA duplex in silico [24].

190 To further confirm that the error suppression potential of
 191 CODEC is uniquely enabled by reading both strands of the
 192 original DNA duplex together, as opposed to simply forming
 193 a consensus of forward and reverse reads, we then compared
 194 error rates of three additional methods from the same NGS
 195 data: no consensus, paired-end reads consensus (R1+R2, col-
 196 lapses read 1 and read 2), and single strand consensus (SSC,
 197 collapses reads from the same original strand). Interestingly,
 198 the error rate gap between the no consensus and R1+R2
 199 was negligible (Fig. 2b), suggesting that many errors are
 200 physically present in NGS library molecules, and could have
 201 been introduced during library amplification, or when each
 202 library molecule undergoes bridge amplification for cluster
 203 generation (Fig. 1a). Although SSC was more accurate than
 204 R1+R2 and the no consensus reads, without a consensus of

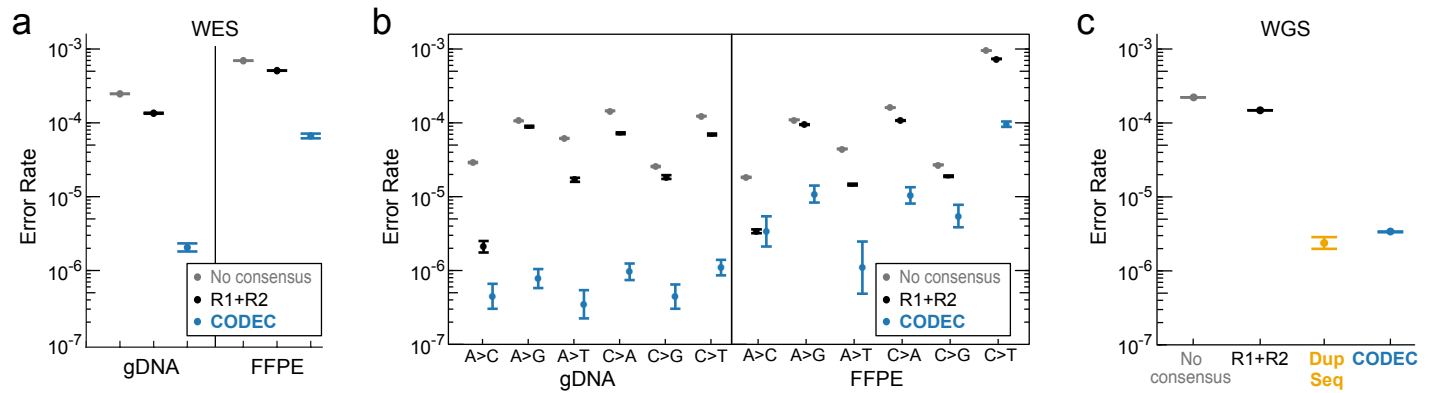


FIG. 3. Error rates of whole-exome sequencing (WES) and whole-genome sequencing (WGS). (a) Error rates of CODEC on formalin-fixed paraffin-embedded (FFPE) and matching normal samples of a cancer patient. (b) Errors in (a) broken down by sequence context. (c) Error rates of WGS with Duplex Sequencing and CODEC performed side by side.

205 Watson and Crick strands, its error rate was 23-fold higher
 206 than that of CODEC. The fact that reading the same strand
 207 multiple times does not contribute as much as duplex cons-
 208 sensus implies the intrinsic limitation of other sequenc-
 209 ing technologies[27, 28].

210 We next explored the number of reads required to uncover
 211 the same number of unique DNA duplexes. When we used
 212 UMIs as well as start and stop mapping positions of each
 213 molecule to collapse all reads to unique original duplexes, we
 214 found that Duplex Sequencing could not start reassembling
 215 duplexes until receiving 700 reads (Fig. 2c). In contrast,
 216 CODEC started to reassemble 350-fold earlier. The gap
 217 between required reads was maximized when recovering a
 218 smaller number of duplexes, suggesting that CODEC could
 219 be uniquely capable of sequencing broad genomic regions
 220 with shallow depth. Notably, even a single paired-end read
 221 of CODEC was highly accurate (Fig. 2d), as each CODEC
 222 read is self-sufficient to form a duplex consensus. Our results
 223 suggest that CODEC confers the accuracy of duplex sequenc-
 224 ing from single paired-end reads and thus sequences more
 225 DNA duplexes using substantially fewer reads.

226 **CODEC confers the accuracy of duplex sequencing**
 227 **to WGS and WES.** We next sought to determine whether
 228 CODEC could enable human whole-exome and whole-genome
 229 ‘duplex’ sequencing, which would otherwise be impractical due
 230 to high cost. To assess this, we applied CODEC whole-exome
 231 sequencing (WES) to gDNA and formalin-fixed paraffin-
 232 embedded (FFPE) samples from a cancer patient, whose
 233 samples had been tested in our prior publication[24]. We
 234 found that CODEC reduced the sequencing error rates of
 235 both samples, with 100-fold improvement for gDNA(Fig.
 236 3a). Analyzing the sequence context of the errors revealed
 237 that CODEC improved accuracy across all types of SNV
 238 (Fig. 3b), suggesting that the capability of CODEC to
 239 suppress errors is not limited to specific contexts. Of note,
 240 there were more C>T errors in FFPE samples due to deami-
 241 nation artifacts[34], which we believe could be resolved with
 242 improved end-repair methods[30, 33].

243 Next, we applied CODEC and Duplex Sequencing to WGS
 244 of the pilot genome NA12878 of the Genome in a Bottle
 245 Consortium (GIAB)[35]. For a fair comparison, we assigned
 246 the same amount of sequencing to each method although
 247 Duplex Sequencing could not recover many unique duplexes.
 248 The error rates of both Duplex Sequencing (2.38×10^{-6}) and

249 CODEC (3.37×10^{-6}) were much lower than that of the no
 250 consensus reads (2.2×10^{-4}) or R1+R2 (1.48×10^{-4}) (Fig.
 251 3c). This result confirms that CODEC is as accurate as
 252 Duplex Sequencing under the same conditions. The error
 253 rates of each sequence context showed that CODEC has a
 254 similar error profile to Duplex Sequencing (Supplementary
 255 Figure S7).

256 Depth of coverage analysis for WGS further demonstrated
 257 that CODEC achieved 160-fold greater unique duplex depth
 258 than Duplex Sequencing. On the GIAB v3.3.2 hg19 high
 259 confidence genomic region (2.6B bases), CODEC had a mean
 260 unique duplex depth of 3.96 with 320M raw reads, whereas
 261 Duplex Sequencing had only 0.025 mean depth even with
 262 35% more raw read output (431M reads), because most reads
 263 did not find their matching strand of the original duplex
 264 (Fig. 4a). Thus, we concluded that Duplex Sequencing is
 265 not appropriate for WGS and treated Duplex Sequencing
 266 WGS data as standard WGS data without generating duplex
 267 consensus after this point. In contrast, CODEC covered each
 268 base with four unique duplexes on average, confirming the
 269 strength of resolving single duplexes.

270 **CODEC pushes the frontiers in secondary analysis**
 271 **applications.** Achieving the error rate of Duplex Sequencing
 272 in WGS/WES gives CODEC the ability to push the limits of
 273 many secondary analysis applications. One such application
 274 is benchmarking the whole genome small germline variant
 275 calling (SNV + indel). To test the potential of CODEC at
 276 low coverage as implied in Figure 2c, we compared CODEC
 277 data of the aforementioned NA12878 sample against standard
 278 NGS (R1+R2) at coverages ranging from 1x to 10x, while
 279 acknowledging that state-of-the-art germline calling usually
 280 requires 30x depth. GATK4 was used for variant calling and
 281 followed by the GIAB best practice for benchmarking small
 282 germline variants[35]. CODEC showed 90% fewer false posi-
 283 tives (FP) than standard WGS with R1+R2 at a cost of 5%
 284 higher false negatives (FN) across all downsampled depths
 285 (Fig. 4b, Supplementary Table S1). By downsampling
 286 NGS data, we also observed how FP and FN are affected
 287 by the depth. The lower level of FP in CODEC was the
 288 expected result, considering its lower error rate. Its FN levels
 289 were slightly higher than that of standard WGS, probably
 290 because the lower library conversion efficiency resulted in
 291 higher duplication rate, but the difference between FN rates
 292 of CODEC and standard WGS became smaller as the cov-

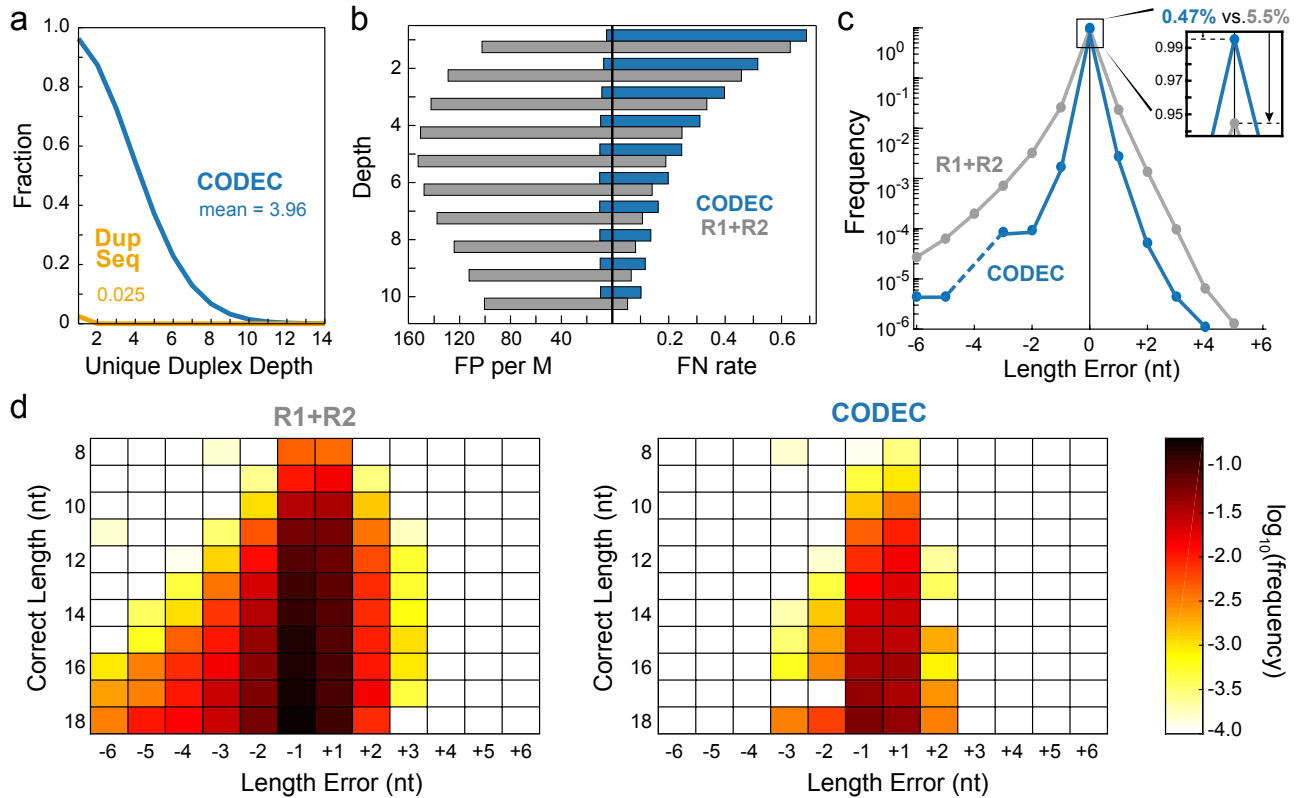


FIG. 4. **In-depth comparison of WGS results.** (a) Fractions of each unique duplex depth of CODEC and Duplex Sequencing. (b) False positives and false negatives of CODEC and R1+R2 when downsampled to lower depths. (c) Summarized indel error frequency at mononucleotide microsatellites. (d) Indel error frequency at mononucleotide microsatellites with different lengths from 8 to 18 nucleotides.

293 erage decreased. Meanwhile, the advantage of having low
 294 FP became more significant at the lower coverage, implying
 295 that applications with shallow depth could benefit more from
 296 using CODEC.

297 Considering CODEC's performance for indel detection
 298 at low coverage, we thought that CODEC could improve
 299 the sequencing accuracy of microsatellites (MS), which are
 300 well-known mutation hot spots. Indeed, when the reference
 301 sequences of the MS in NA12878 were compared between
 302 CODEC and standard NGS results, CODEC showed lower
 303 frequencies of both insertion and deletion errors than stan-
 304 dard WGS at mononucleotide MS from 8 to 18 nucleotides
 305 (Fig. 4c). The ratio of CODEC reads with incorrect MS
 306 lengths was 0.47%, which was 12 times lower than that of
 307 standard WGS. Such lower frequencies were consistently ob-
 308 served across mononucleotide MS of varied lengths (Fig. 4d).
 309 These findings imply that CODEC could be used to read
 310 the repeat numbers/copy numbers of MS sites for detecting
 311 microsatellite instability (MSI). MSI has been shown to be
 312 a predictive marker of response to cancer immunotherapy
 313 but remains challenging to detect at low frequency such as
 314 from liquid biopsy samples[36]. Tracing mutations in MS is
 315 also useful for tracing cell lineages and evolution[37]. The
 316 improvements in the secondary applications we have shown
 317 highlight what CODEC could enable by sequencing a single
 318 duplex within each NGS cluster.

Discussion

319
 320 By physically linking both strands of each DNA duplex,
 321 CODEC enables each NGS cluster to have single duplex
 322 resolution like third generation sequencers. Unlike Duplex

323 Sequencing which requires dissociating duplexes and recov-
 324 ering them back to form a duplex consensus, CODEC dis-
 325 tinguishes real mutations from errors with similarly high
 326 accuracy but with 100-fold fewer reads. We first showed the
 327 proof-of-concept of our approach using cDNA enriched by a
 328 pan-cancer panel, followed by testing its consistency across
 329 other major NGS workflows (e.g., WES and WGS) and sam-
 330 ple types (e.g., FFPE and germline DNA). To present more
 331 uses of CODEC, we also showed that it suppressed FP espe-
 332 cially at shallow sequencing depth and reduced indel errors
 333 at MS sites.

334 In a head-to-head comparison, we showed that CODEC
 335 is as accurate as Duplex Sequencing but with a much lower
 336 sequencing requirement, which has been a major limitation
 337 of Duplex Sequencing. Because an error rate is affected by
 338 multiple factors other than a sequencing technology itself,
 339 any direct comparison requires everything else to be the
 340 same. We used the same experimental and computational
 341 protocols whenever applicable, including input samples and
 342 mass, reagents, target regions, definition of an error, and
 343 analysis pipelines for precise comparison.

344 Because CODEC redefines standard NGS with a novel
 345 molecular structure, there may still be room for improvement
 346 in its use with target selection protocols including hybrid
 347 capture, multiplexed amplicon, and mutation enrichment
 348 sequencing[38]. We are also working to improve CODEC's
 349 conversion efficiency. The CODEC adapter complex is at-
 350 tached through two consecutive ligations: a bimolecular li-
 351 gation followed by a unimolecular ligation. Unlike typical
 352 bimolecular adapter ligation where increasing adapter con-
 353 centration also increases conversion efficiency, unimolecular

ligation could be less favorable when the adapter concentration is too high. Consequently, the current version of CODEC adapter complex needs balancing between two ligations. We are currently developing another version of CODEC that circumvents two consecutive ligations.

Although conventional end-repair/dA-tailing of a commercial kit was used throughout this work, the accuracy can be further improved if a new end-repair method is adopted before CODEC. Recent studies[30, 33] have reported that base damage on overhangs and single-stranded breaks of original DNA duplexes can lead errors on one strand to be copied to both strands. It was also indirectly observed in this work that error rates were generally higher toward the ends of DNA fragments (Supplementary Figure S6). While such errors appear on duplex consensus and result in false mutations, new end-repair methods prevent the error propagation, and we believe that even higher accuracy will be attainable when CODEC is combined with new end-repair methods[30, 33].

Reading a single CODEC fragment is equivalent to reading both strands of an original duplex, which eliminates the need to read the same locus multiple times. The low error rate of CODEC at 1x read depth opens possibilities for various applications across fields from diagnostics to bioinformatics. One example is discovering rare somatic mutations with a limited number of reads, which has a higher chance of finding a true mutation when the error rate gets lower[39]. Another example is shotgun metagenomic sequencing for microbiome analysis, where suppressing false SNVs with CODEC would prevent incorrect taxonomic classifications and inaccurate evaluation of microbial diversity[40]. In de novo assembly, lower error rates contribute to more contiguous assembly in de Bruijn graph paradigm and faster process in overlap-layout-consensus paradigm[41].

In summary, CODEC transforms standard NGS instruments into massively parallel ‘single duplex’ sequencers by concatenating both strands of each original DNA duplex. This strategy enables SNV and indel detection as accurate as Duplex Sequencing, even in cases where Duplex Sequencing is not possible due to low throughput. We thus believe that CODEC could be broadly enabling for many important biomedical applications such as detecting early-stage cancer or minimal residual disease from liquid biopsies, clinically actionable mutations from liquid or tumor biopsies, clonal hematopoiesis of indeterminate potential (CHIP) from blood samples, somatic mosaicism in normal tissue samples, and beyond.

Methods

DNA samples and oligonucleotides. Cell-free DNA of patient 315 from cohort 05-246 and both FFPE and gDNA of patient 95 from cohort 05-055 were from another study[24]. NA12878 was purchased from Coriell. All samples were stored in low TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8) and were fragmented by Covaris ultrasonicator to have a mean size of 150 bp except cfDNA. All oligonucleotides for CODEC were synthesized by Integrated DNA Technologies (IDT) and went through PAGE purification (See **Supplementary Table S2** for their sequences). The adapter for Duplex Sequencing was custom-ordered for the Broad Institute by IDT.

CODEC. The CODEC adapter complex was prepared by diluting four 100 μ M oligonucleotides to 5 μ M with low TE buffer and 100 mM NaCl, followed by heating at 85 °C for 3 minutes, cooling with -1 °C/min to 20 °C, and incubating at room temperature for 12 hours. Mastercycler X50 (Eppendorf) and MAXYMum Recovery PCR tubes (Axygen) were

used for the annealing. The annealed adapter complex was kept at -20 °C for future use. We used NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, NEB) and followed the manufacturer’s manual with several exceptions:

1. ligation time was increased to 1 hour, 5 μ M adapter complex was diluted with adapter dilution buffer (10 mM Tris-HCl, 1 mM EDTA, 10 mM NaCl, pH 8) to 500 nM before use and replaced NEB adapter,
2. 3 μ L of 5’-deadenylase (NEB) were added to ligation reaction,
3. strand displacing extension (sample 40 μ L, 10x buffer 10 μ L, 0.2 mM dNTP, polymerase 1 μ L, H₂O up to 100 μ L) was performed with phi29 DNA polymerase (New England Biolabs) at 30 °C for 20 minutes, followed by standard AMPure XP (Beckman Coulter) clean up with 0.75x volume ratio,
4. KAPA HiFi HotStart ReadyMix and xGen Library Amplification Primer Mix (IDT) were used for PCR by following the manufacturer’s manuals with 2 minutes of extension,
5. and AMPure XP clean up with 0.75x volume ratio was performed twice after the PCR.

Libraries for standard NGS and Duplex Sequencing were prepared as described elsewhere[24]. All Library preparations were performed on twin.tec PCR Plates LoBind 250 μ L (Eppendorf). Library quantitation was performed with Qubit dsDNA HS kit (Invitrogen) paired with Bioanalyzer DNA High Sensitivity chips (Agilent).

Enrichment. Both pan-cancer and WES enrichment was performed with xGen Hybridization and Wash kits and xGen Blocking Oligos (IDT), following the manufacturer’s manual. For capture probes, xGen Pan-cancer Panel (IDT, 800 kb) and custom WES panel for the Broad Institute by Twist Bioscience were used.

Sequencing. Standard NGS and Duplex Sequencing were performed with Illumina HiSeq 2500 Rapid Run (300 cycles) for a pan-cancer panel and WGS. CODEC was performed with Illumina HiSeq 2500 Rapid Run (500 cycles) for a pan-cancer panel and WGS, and NovaSeq SP (500 cycles) for WGS and WES. The extra cycles were used to confirm the CODEC structure.

CODEC data processing. Due to the unique CODEC read structure, we developed CODECsuite (available at <https://github.com/broadinstitute/CODECsuite>) to process CODEC data (**Supplementary Note**). CODECsuite is written in C++14 and python3.7 and we use snakemake6.0.3[42] as the workflow management system. CODECsuite consists of 4 major steps: demultiplexing, adapter trimming, consensus calling and computing accuracy. The first 3 steps are specific to CODEC data. The workflow also involves other standard tools such as BWA[43], Fgbio and GATK[44]. Illumina bcl2fastq was used to generate fastq files (with -R -o, no -sample-sheet because CODECsuite will demultiplex), but is not included in the suite. To speed up the data processing, we recommend splitting the fastq files in batches and processing them in parallel. In this study, using 40 batches, the preprocessing (demultiplexing and adapter trimming) of 800M NovaSeq reads took just a few hours in a HPC environment where each batch was executed using a single CPU and 8G RAM. After demultiplexing and adapter removal, we mapped the raw reads using BWA(0.7.17-r1188) against human reference hg19. Fgbio (<https://github.com/fulcrumgenomics/fgbio>) was then used to collapse the PCR duplicates and to form essentially single-strand consensus (SSC) reads. These SSC reads were then mapped to the reference genome using BWA again. Next, the duplex consensus reads between R1 and R2 were generated from the SSC alignments. We filtered a consensus base if any of the bases from R1 or R2 has base quality less than 30. The duplex consensus reads were aligned to the reference genome using BWA and the subsequent alignments were indel realigned using GATK3 (<https://hub.docker.com/r/broadinstitute/gatk3>).

Duplex Sequencing data processing. Duplex Sequencing data processing used in this study has been described elsewhere[24, 38]. Briefly, Fgbio was used to generate duplex consensus and to filter the consensus reads. The entire workflow and more details are available at the CODECsuite github. Read families with at least 2 copies of each strand were required for generating duplex consensus except for Duplex Sequencing WGS, which relaxed the requirement to 1 copy of each strand to get the best possible duplex recovery.

Duplex recovery and downsample to certain family sizes. Two custom python scripts were used to generate Figure 2c and 2d, respectively. For duplex recovery, we subsampled the pre-consensus family-

489 assigned reads (after Fgbio GroupReadsByUmi) per target at log spaced
490 fractions starting from 10^{-4} (np.logspace(-4, 0, 30)) and calculated the
491 number of duplex formed at each downsample fraction. In this study,
492 this allowed us to understand situations when only limited sequencing
493 was given (e.g., < 100 read pairs). To understand the impact of family
494 size on error rate, we wrote another python script for downsampling. In
495 our sample, the number of duplex consensus having the exact family
496 sizes (number of pre-collapsed raw reads) were limited and thus gave
497 less confident results. Thus, we took advantage of families with strictly
498 larger family sizes and downsample them to the target family size. We
499 also sought to maintain an equal or close ratio between the number of
500 reads from each strand.

501 **Error rates in capture sequencing.** Throughout the article, we
502 defined the error rate as substitution error rate at the base level after
503 mapping to the reference genome (hg19). We used the substitution error
504 rate for calculating the general error rates because Illumina sequencers
505 usually generate 100-fold less indel errors[45] and this definition is com-
506 pliant with what other studies have reported[30]. For panel sequencing
507 with match normal, we used Miredas to calculate the error rate in
508 concordance with our previous work[24]. The duplex BAMs from both
509 cfDNA and matched normal samples were generated in the same way
510 and were applied to the same set of filters: 1. no secondary and supple-
511 mentary alignments; 2. Mapq ≥ 60 ; 3. Levenshtein distance (L-distance)
512 between the reads excluding soft clipping and reference genome ≤ 5 and
513 number of non N-base L-distance ≤ 2 ; 4. Excluding bases within 12 bp
514 distance from both fragment ends. In order not to confuse errors with
515 real mutations, we pre-computed the germline SNVs and using GATK4
516 (HaplotypeCaller[46]) from the Duplex Sequencing normal samples as
517 they have higher on-target ratio and hence coverage (89% vs 40% of
518 CODEC). For the patient sample, we found three somatic SNVs (median
519 VAF=0.26, range 0.24 - 0.28) in the captured regions (**Supplementary**
520 **Table S3**) using MuTect[39]. Those somatic mutations (patient sample
521 only) and germline mutations were masked when calculating the error
522 rates. The error rates were only reported for cfDNA samples and the
523 match normal were used for filtering possible germline (failed to call
524 or did not pass quality filter by HaplotypeCaller) and CHIP. Thereby
525 we also masked any SNV positions where there were at least 1 duplex
526 read support in match normal samples as CHIP can occur at very low
527 mutation frequency. Finally, the specificity checks[24] were performed
528 on cfDNA samples to remove substitutions that may rise from alignment
529 errors.

530 **Error rate in whole genome sequencing.** The WGS error rate was
531 computed similarly to capture data, except for a few differences. 1, We
532 used 'codec accuracy', a C++ program, as a replacement for Miredas
533 due to its speed improvement. 2, We used v3.3.2 GIAB NA12878 high
534 confidence VCF and BED[35] file as germline masks and evaluation
535 regions. 3, there was no match normal. 4, we forwent specificity checks
536 as it is also very slow for large genomes.

537 **Germline SNV and small indel calling in downsampled WGS.**
538 We merged the HiSeq 2500 Rapid Run and NovaSeq SP CODEC data
539 to evaluate germline variant calling. The merged CODEC and standard
540 WGS NA12878 samples were downsampled to 1 to 10x (step size 1x)
541 median coverage in the high confidence regions using GATK Downsam-
542 pleSam. Next, we ran GATK4.1.4.1 best practices pipeline via Cromwell
543 and Terra workflow (available at [web resources](#)) and computed on the
544 Google Cloud Platform. We used RTG [vcfeval](#) to calculate False Posi-
545 tives (FP) and False Negatives (FN) for SNVs and indels (< 50 bp)
546 without penalizing genotyping error (if heterozygous variants are called
547 as homozygous and vice versa) using v3.3.2 high confidence VCF and
548 BED file as input. We then calculated FP per million bases by normal-
549 izing against the high confidence region size and FN ratio by dividing
550 FN by the total number of true variants.

551 **Microsatellite instability detection.** The full-coverage CODEC
552 consensus BAM and full-coverage standard NGS R1R2 consensus BAM
553 on NA12878 were compared against each other to demonstrate CODEC
554 ability to correct PCR stutter errors and thus to reduce background
555 noise for MSI detection. MSIsensor-pro[47] was used to scan the hg19
556 for homopolymers of size 8 - 18 nt. Since MSIsensor-pro does not have
557 mapping quality or secondary alignments filters, we pre-filtered the
558 BAM using SAMtools[48] by requiring mapq ≥ 60 and no secondary or
559 supplementary alignments. And then it was used again to count the
560 number of reads that support different lengths of homopolymer at those
561 pre-selected sites. We removed any homopolymer sites that overlap or

562 are in close proximity (+/-5 bp) with any germline variants. After that,
563 the reference lengths of the homopolymer sites were considered as true
564 lengths. And observed length distributions from reads were compared
565 against truth. The results were generated from chromosome 1 only.

566 **Code availability.** CODECsuite and examples and tutorials including
567 how to regenerate the figures in the manuscript are available at
568 the github site <https://github.com/broadinstitute/CODECsuite>.
569 The end-to-end workflow is available at
570 <https://github.com/broadinstitute/CODECsuite/tree/master/snakefile>.

571 **Data availability.** CODEC data and Duplex Sequencing data will be
572 available on dbGAP.
573

References

- 574 [1] Lennon, A. M. et al. Feasibility of blood testing combined with PET-
575 CT to screen for cancer and guide intervention. *Science* 369, eabb9601
576 (2020).
- 577 [2] Deveson, I. W. et al. Evaluating the analytical validity of circulating
578 tumor DNA sequencing assays for precision oncology. *Nat. Biotechnol.*
579 (2021) doi:10.1038/s41587-021-00857-z.
- 580 [3] Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in
581 cancer. *Nature* 575, 299–309 (2019).
- 582 [4] Beaubier, N. et al. Integrated genomic profiling expands clinical options
583 for patients with cancer. *Nat. Biotechnol.* 37, 1351–1360 (2019).
- 584 [5] Griffith, O. L. et al. The prognostic effects of somatic mutations in
585 ER-positive breast cancer. *Nat. Commun.* 9, 3476 (2018).
- 586 [6] Jamal-Hanjani, M. et al. Tracking the Evolution of Non-Small-Cell
587 Lung Cancer. *N. Engl. J. Med.* 376, 2109–2121 (2017).
- 588 [7] Gerlinger, M. et al. Intratumor Heterogeneity and Branched Evolution
589 Revealed by Multiregion Sequencing. *N. Engl. J. Med.* 366, 883–892
590 (2012).
- 591 [8] Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature*
592 578, 122–128 (2020).
- 593 [9] D’Gama, A. M. & Walsh, C. A. Somatic mosaicism and neurodevelop-
594 mental disease. *Nature Neuroscience* 21, 1504–1514 (2018).
- 595 [10] Serra, E. G. et al. Somatic mosaicism and common genetic variation
596 contribute to the risk of very-early-onset inflammatory bowel disease.
597 *Nat. Commun.* 11, 995 (2020).
- 598 [11] Blauwkamp, T. A. et al. Analytical and clinical validation of a microbial
599 cell-free DNA sequencing test for infectious disease. *Nat. Microbiol.* 4,
600 663–674 (2019).
- 601 [12] Ménard, D. et al. A Worldwide Map of Plasmodium falciparum K13-
602 Propeller Polymorphisms. *N. Engl. J. Med.* 374, 2453–2464 (2016).
- 603 [13] Brazhnik, K. et al. Single-cell analysis reveals different age-related
604 somatic mutation profiles between stem and differentiated cells in
605 human liver. *Sci. Adv.* 6, (2020).
- 606 [14] Bick, A. G. et al. Inherited causes of clonal haematopoiesis in 97,691
607 whole genomes. *Nature* 586, 763–768 (2020).
- 608 [15] Wenger, A. M. et al. Accurate circular consensus long-read sequencing
609 improves variant detection and assembly of a human genome. *Nat.*
610 *Biotechnol.* 37, 1155–1162 (2019).
- 611 [16] Karst, S. M. et al. High-accuracy long-read amplicon sequences using
612 unique molecular identifiers with Nanopore or PacBio sequencing. *Nat.*
613 *Methods* 18, 165–169 (2021).
- 614 [17] Shendure, J. et al. DNA sequencing at 40: past, present and future.
615 *Nature* 550, 345–353 (2017).
- 616 [18] Arbeithuber, B., Makova, K. D. & Tiemann-Boege, I. Artfactual muta-
617 tions resulting from DNA lesions limit detection levels in ultrasensitive
618 sequencing applications. *DNA Res.* 23, 547–559 (2016).
- 619 [19] Potapov, V. & Ong, J. L. Examining Sources of Error in PCR by
620 Single-Molecule Sequencing. *PLoS One* 12, 1–19 (2017).
- 621 [20] Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age:
622 ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*
623 17, 333–351 (2016).
- 624 [21] Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B.
625 Detection and quantification of rare mutations with massively parallel
626 sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 108, 9530–9535 (2011).
- 627 [22] Kivioja, T. et al. Counting absolute numbers of molecules using unique
628 molecular identifiers. *Nat. Methods* 9, 72–74 (2012).
- 629 [23] Schmitt, M. W. et al. Detection of ultra-rare mutations by next-
630 generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 109, 14508–14513
631 (2012).
- 632 [24] Parsons, H. A. et al. Sensitive Detection of Minimal Residual Disease
633 in Patients Treated for Early-Stage Breast Cancer. *Clin. cancer Res.*
634 26, 2556–2564 (2020).
- 635 [25] Pel, J. et al. Duplex Proximity Sequencing (Pro-Seq): A method to
636 improve DNA sequencing accuracy without the cost of molecular bar-
637 coding redundancy. *PLoS One* 13, 1–19 (2018).
- 638 [26] Gregory, M. T. et al. Targeted single molecule mutation detection with
639 massively parallel sequencing. *Nucleic Acids Res.* 44, e22 (2016).
- 640 [27] Wang, K. et al. Ultrasensitive and high-efficiency screen of de novo
641 low-frequency mutations by o2n-seq. *Nat. Commun.* 8, 15335 (2017).
- 642 [28] Lou, D. I. et al. High-Throughput DNA sequencing errors are reduced

- 643 by orders of magnitude using Circle Sequencing. *Proc. Natl. Acad. Sci.* 682
644 U. S. A. 110, 19872–19877 (2013). 683
- 645 [29] Hoang, M. L. et al. Genome-wide quantification of rare somatic muta- 684
646 tions in normal human tissues using massively parallel sequencing. 685
647 *Proc. Natl. Acad. Sci. U. S. A.* 113, 9846–9851 (2016). 686
- 648 [30] Abascal, F. et al. Somatic mutation landscapes at single-molecule 687
649 resolution. *Nature* 593, 405–410 (2021). 688
- 650 [31] Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes in- 689
651 accuracies in multiplex sequencing on the Illumina platform. *Nucleic* 690
652 *Acids Res.* 40, e3–e3 (2012). 691
- 653 [32] Costello, M. et al. Characterization and remediation of sample index 692
654 swaps by non-redundant dual indexing on massively parallel sequencing 693
655 platforms. *BMC Genomics* 19, 332 (2018). 694
- 656 [33] Xiong, K. et al. Duplex-Repair enables highly accurate sequencing, 695
657 despite DNA damage. *bioRxiv* (2021) doi:10.1101/2021.05.21.445162. 696
- 658 [34] Kim, S. et al. Deamination Effects in Formalin-Fixed, Paraffin- 697
659 Embedded Tissue Samples in the Era of Precision Medicine. *J. Mol.*
660 *Diagnostics* 19, 137–146 (2017).
- 661 [35] Zook, J. M. et al. An open resource for accurately benchmarking small 698
662 variant and reference calls. *Nat. Biotechnol.* 37, 561–566 (2019). 699
- 663 [36] Yu, F. et al. NGS-based identification and tracing of microsatellite 700
664 instability from minute amounts DNA using inter-Alu-PCR. *Nucleic* 701
665 *Acids Res.* 49, e24–e24 (2021).
- 666 [37] Woodworth, M. B., Girsakis, K. M. & Walsh, C. A. Building a lineage 702
667 from single cells: Genetic techniques for cell lineage tracking. *Nat. Rev.* 703
668 *Genet.* 18, 230–244 (2017).
- 669 [38] Gydush, G. et al. MAESTRO affords ‘breadth and depth’ for mutation 704
670 testing. *bioRxiv* (2021) doi:10.1101/2021.01.22.427323. 705
- 671 [39] Cibulskis, K. et al. Sensitive detection of somatic point mutations in 706
672 impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219 707
673 (2013).
- 674 [40] May, A., Abeln, S., Crielgaard, W., Heringa, J. & Brandt, B. W. Un- 708
675 raveling the outcome of 16S rDNA-based taxonomy analysis through 709
676 mock data and simulations. *Bioinformatics* 30, 1530–1538 (2014).
- 677 [41] Limasset, A., Flot, J. F. & Peterlongo, P. Toward perfect reads: Self- 710
678 correction of short reads via mapping on de Bruijn graphs. *Bioinform-*
679 *atics* 36, 1374–1381 (2020).
- 680 [42] Mölder, F. et al. Sustainable data analysis with Snakemake. 711
681 *F1000Research* 10, 33 (2021).
- 682 [43] Li, H. & Durbin, R. Fast and accurate long-read alignment with 712
683 Burrows-Wheeler transform. *Bioinformatics* 26, 589–595 (2010).
- 684 [44] DePristo, M. A. et al. A framework for variation discovery and geno- 713
685 typing using next-generation DNA sequencing data. *Nat. Genet.* 43, 714
686 491–498 (2011).
- 687 [45] Schirmer, M. et al. Insight into biases and sequencing errors for amplicon 715
688 sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 43, 716
689 e37 (2015).
- 690 [46] DePristo, M. A. et al. A framework for variation discovery and geno- 717
691 typing using next-generation DNA sequencing data. *Nat. Genet.* 43, 718
692 491–498 (2011).
- 693 [47] Jia, P. et al. MSIsensor-pro: Fast, Accurate, and Matched-normal- 719
694 sample-free Detection of Microsatellite Instability. *Genomics. Pro-*
695 *teomics Bioinformatics* 18, 65–71 (2020).
- 696 [48] Li, H. et al. The Sequence Alignment/Map format and SAMtools. 720
697 *Bioinformatics* 25, 2078–2079 (2009).

Acknowledgements

The authors acknowledge the Gerstner Family Foundation for its generous support. This study was also supported in part by SPARC award from the Broad Institute.

Author contributions

Conception and design: V.A.A., J.H.B.
Method development: V.A.A., J.H.B., R.L.
Data acquisition: J.H.B., E.N.
Data analysis: R.L.
Data interpretation: V.A.A., Z.A., J.H.B., T.B., G.G., R.L., E.N., S.P., J.R. D.S., S.T., K.X.
Writing and review of the manuscript: V.A.A., J.H.B., T.R.G., R.L., G.M.M.

Competing interests

The authors have filed a patent application on this method. V.A.A. is a member of the scientific advisory boards of AGCT GmbH and Bertis Inc. T.R.G. has advisor roles at Foundation Medicine, GlaxoSmithKline, and Sherlock Biosciences.