Sage Products Powering New Methods, 2018-19

sage science

SageELF

Title

Highly-accurate long-read sequencing improves variant detection and assembly of a human genome

Authors

Aaron M. Wenger[1]†, Paul Peluso[1]†, William J. Rowell[1], Pi-Chuan Chang[2], Richard J. Hall[1], Gregory T. Concepcion[1], Jana Ebler[3,4,5], Arkarachai Fungtammasan[6], Alexey Kolesnikov[2], Nathan D. Olson[7], Armin Töpfer[1], Michael Alonge[8], Medhat Mahmoud[9], Yufeng Qian[1], Chen-Shan Chin[6], Adam M. Phillippy[10], Michael C. Schatz[8], Gene Myers[11], Mark A. DePristo[2], Jue Ruan[12], Tobias Marschall[3,4], Fritz J. Sedlazeck[9], Justin M. Zook[7], Heng Li[13], Sergey Koren[10], Andrew Carroll[2], David R. Rank[1]*, Michael W. Hunkapiller[1]*

Abstract

The major DNA sequencing technologies in use today produce either highly-accurate short reads or noisy long reads. We developed a protocol based on single-molecule, circular consensus sequencing (CCS) to generate highly-accurate (99.8%) long reads averaging 13.5 kb and applied it to sequence the well-characterized human HG002/NA24385. We optimized existing tools to comprehensively detect variants, achieving precision and recall above 99.91% for SNVs, 95.98% for indels, and 95.99% for structural variants. We estimate that 2,434 discordances are correctable mistakes in the high-quality Genome in a Bottle benchmark. Nearly all (99.64%) variants are phased into haplotypes, which further improves variant detection. *De novo* assembly produces a highly contiguous and accurate genome with contig N50 above 15 Mb and concordance of 99.998%. CCS reads match short reads for small variant detection, while enabling structural variant detection and *de novo* assembly at similar contiguity and markedly higher concordance than noisy long reads.

https://www.biorxiv.org/content/10.1101/519025v2 --posted January 23, 2019

Figure 1

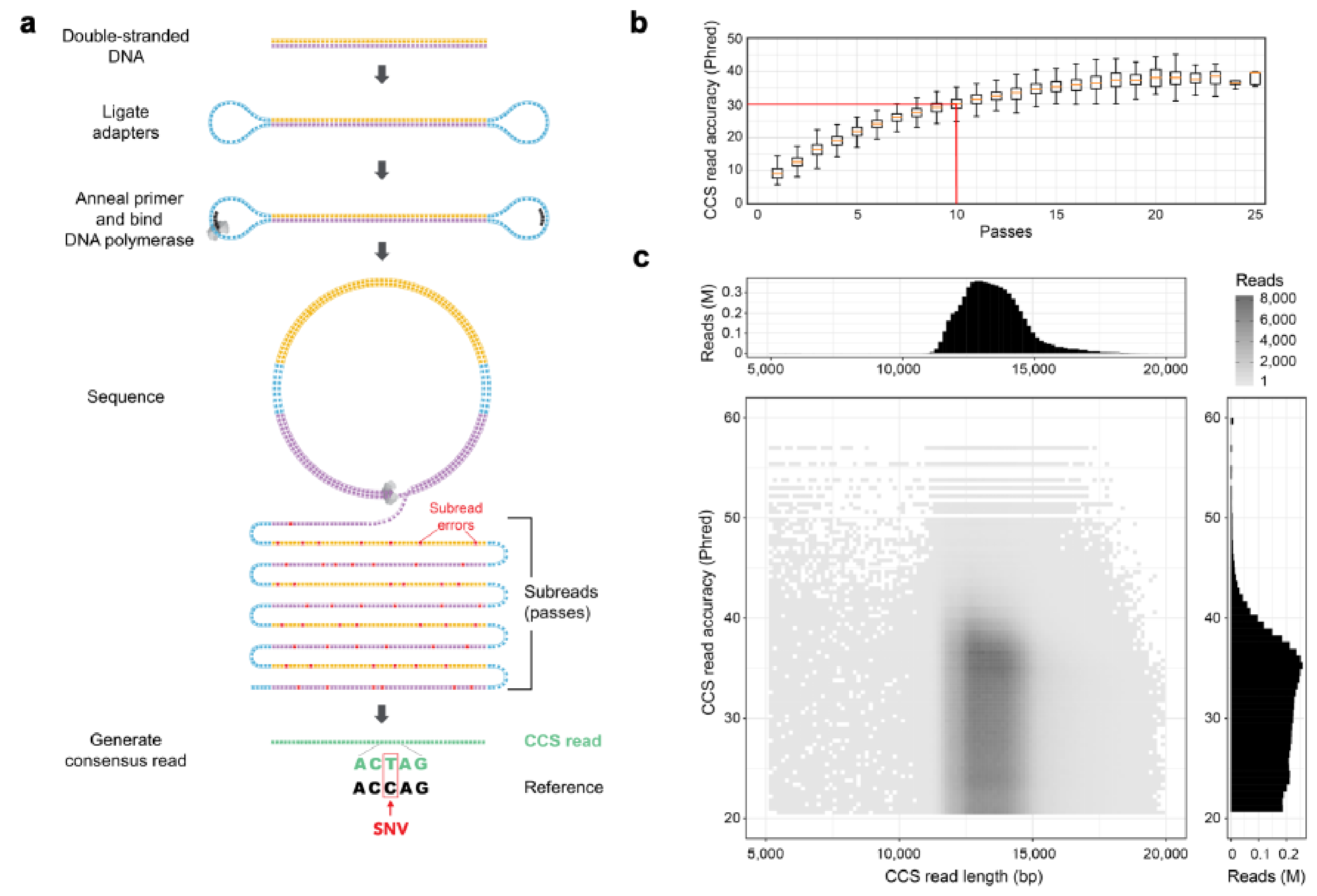
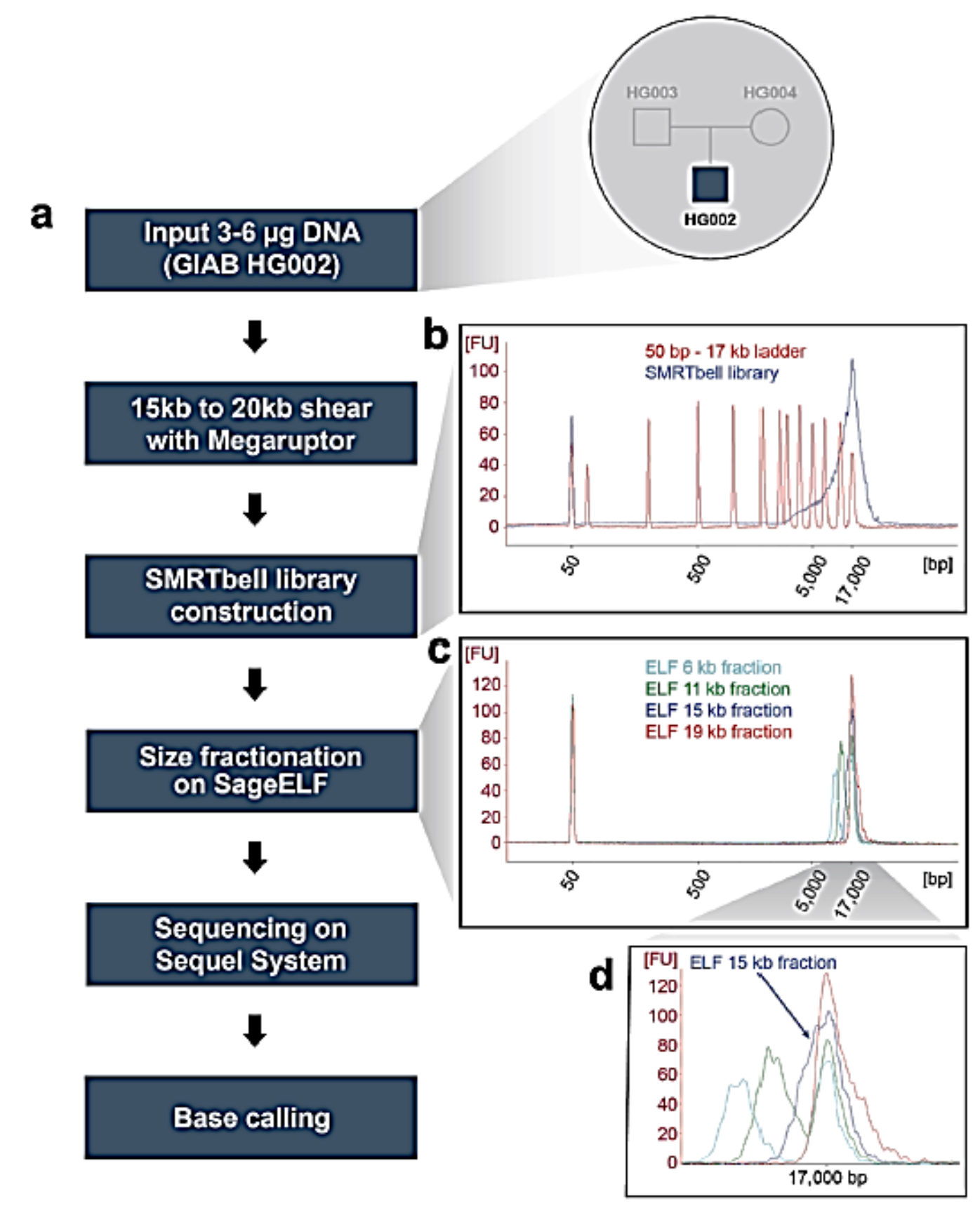


Figure 1. Sequencing HG002 with highly-accurate, long reads. (a) Circular consensus sequencing (CCS) derives a consensus read from multiple passes of a single template molecule, producing accurate reads from noisy individual subreads (passes). (b) Predicted accuracy of CCS reads with different numbers of passes, for sequencing of the human male HG002. At 10 passes, the median read achieves Q30 predicted accuracy. (c) Length and predicted accuracy of CCS reads.



SageELF

Supplementary Figure 1. CCS protocol. (a) Sample preparation and sequencing workflow. (b) BioAnalyzer trace for the SMRTbell library, sheared to target 15-20 kb fragments. "FU" is fluorescence units. (c) BioAnalyzer trace for ELF fractions of the SMRTbell library. (d) The fraction centered around 15 kb was used for sequencing.

Pippin HT

SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE

CANCER

Enhanced detection of circulating tumor DNA by fragment size analysis

Florent Mouliere[1,2]*†, Dineika Chandrananda[1,2]*, Anna M. Piskorz[1,2]*, Elizabeth K. Moore[1,2,3]*, James Morris[1,2], Lise Barlebo Ahlborn[4,5], Richard Mair[1,2,6], Teodora Goranova[1,2], Francesco Marass[1,2,7,8], Katrin Heider[1,2], Jonathan C. M. Wan[1,2], Anna Supernat[1,2,9], Irena Hudecova[1,2], Ioannis Gounaris[1,2,3], Susana Ros[1,2], Mercedes Jimenez-Linan[2,3], Javier Garcia-Corbacho[10], Keval Patel[1,2], Olga Østrup[5], Suzanne Murphy[1,2], Matthew D. Eldridge[1,2], Davina Gale[1,2], Grant D. Stewart[2,3,11], Johanna Burge[2,11], Wendy N. Cooper[1,2], Michiel S. van der Heijden[12,13], Charles E. Massie[1,2,14], Colin Watts[15], Pippa Corrie[3], Simon Pacey[3,14], Kevin M. Brindle[1,2,16], Richard D. Baird[17], Morten Mau-Sørensen[4], Christine A. Parkinson[1,2,3,18,19], Christopher G. Smith[1,2], James D. Brenton[1,2,3,18,19]‡§, Nitzan Rosenfeld[1,2]‡§

Existing methods to improve detection of circulating tumor DNA (ctDNA) have focused on genomic alterations but have rarely considered the biological properties of plasma cell-free DNA (cfDNA). We hypothesized that differences in fragment lengths of circulating DNA could be exploited to enhance sensitivity for detecting the presence of ctDNA and for noninvasive genomic analysis of cancer. We surveyed ctDNA fragment sizes in 344 plasma samples from 200 patients with cancer using low-pass whole-genome sequencing (0.4×). To establish the size distribution of mutant ctDNA, tumor-guided personalized deep sequencing was performed in 19 patients. We detected enrichment of ctDNA in fragment sizes between 90 and 150 bp and developed methods for in vitro and in silico size selection of these fragments. Selecting fragments between 90 and 150 bp improved detection of tumor DNA, with more than twofold median enrichment in >95% of cases and more than fourfold enrichment in >10% of cases. Analysis of size-selected cfDNA identified clinically actionable mutations and copy number alterations that were otherwise not detected. Identification of plasma samples from patients with advanced cancer was improved by predictive models integrating fragment length and copy number analysis of cfDNA, with area under the curve (AUC) >0.99 compared to AUC <0.80 without fragmentation features. Increased identification of cfDNA from patients with glioma, renal, and pancreatic cancer was achieved with AUC > 0.91 compared to AUC < 0.5 without fragmentation features. Fragment size analysis and selective sequencing of specific fragment sizes can boost ctDNA detection and could complement or provide an alternative to deeper sequencing of cfDNA.

Mouliere *et al., Sci. Transl. Med.* **10**, eaat4921 (2018) 7 November 2018

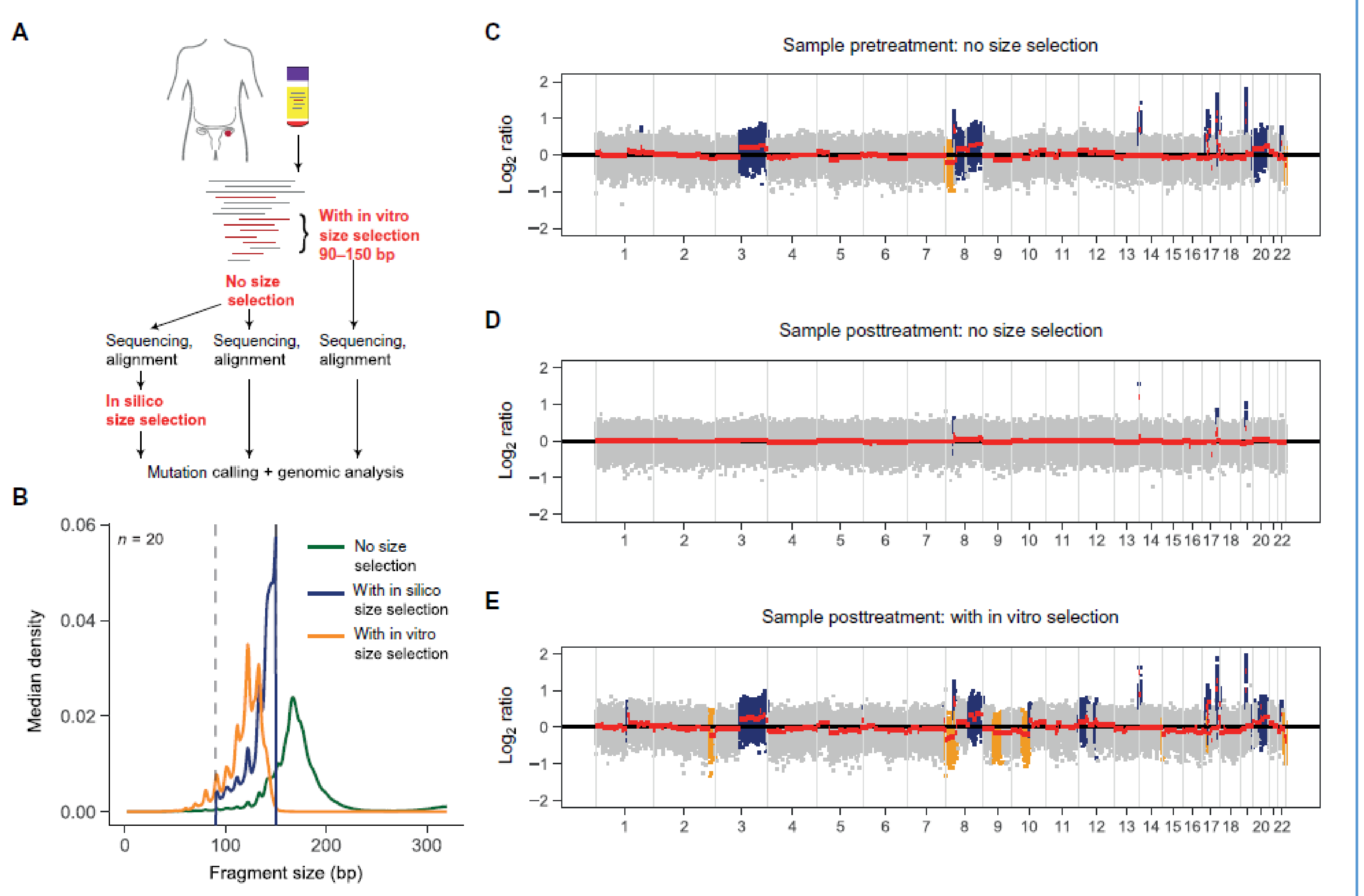


Fig. 3. Enhancing the tumor fraction from plasma sequencing with size selection. (**A**) Plasma samples collected from patients with ovarian cancer were analyzed in parallel without size selection or using either in silico or in vitro size selection. (**B**) Accuracy of the in vitro and in silico size selection determined on a cohort of 20 healthy controls. The size distribution before size selection is shown in green, after in silico size selection (with sharp cutoff at 90 and 150 bp) in blue and after in vitro size selection in orange. Vertical lines indicate 90 and 150 bp. (**C**) SCNA analysis with sWGS from plasma DNA of a patient with ovarian cancer collected before initiation of treatment, when ctDNA MAF was 0.271 for a *TP53* mutation as determined by tagged-amplicon deep sequencing (TAm-Seq). Inferred amplifications are shown in blue and deletions in orange. Copy number neutral regions are shown in gray. (**D**) SCNA analysis of a plasma sample from the same patient as in (C), collected 3 weeks after treatment start. The MAF for the *TP53* mutation at this time point was 0.068, and sWGS revealed only limited evidence of copy number alterations (before size selection). (**E**) Analysis of the same plasma sample as in (D) after in vitro size selection of fragments between 90 and 150 bp in length. The MAF for the *TP53* mutation increased to 0.402 after in vitro size selection, and SCNAs were apparent by sWGS. More SCNAs were detected in comparison to (C) and (D) (for example, in chr2, chr9, and chr10). SCNAs were also detected in this sample after in silico size selection (fig. S7).

Materials and Methods

In vitro size selection. Between 8-20 ng of DNA were loaded into a 3% agarose cassette (HTC3010, Sage Bioscience), and size selection was performed on a PippinHT (Sage Bioscience) according to the manufacturer's protocol.

PippinHT