MICROBIAL GENOMICS

RESEARCH ARTICLE

Quail *et al.*, *Microbial Genomics* 2024;10:001228

DOI 10.1099/mgen.0.001228

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Identifying the best PCR enzyme for library amplification in NGS

Michael A. Quail[1,*], Craig Corton[1], James Uphill[1], Jacqueline Keane[2] and Yong Gu[1]

Abstract

Background. PCR amplification is a necessary step in many next-generation sequencing (NGS) library preparation methods [1, 2]. Whilst many PCR enzymes are developed to amplify single targets efficiently, accurately and with specificity, few are developed to meet the challenges imposed by NGS PCR, namely unbiased amplification of a wide range of different sizes and GC content. As a result PCR amplification during NGS library prep often results in bias toward GC neutral and smaller fragments. As NGS has matured, optimized NGS library prep kits and polymerase formulations have emerged and in this study we have tested a wide selection of available enzymes for both short-read Illumina library preparation and long fragment amplification ahead of long-read sequencing.

We tested over 20 different hi-fidelity PCR enzymes/NGS amplification mixes on a range of Illumina library templates of varying GC content and composition, and find that both yield and genome coverage uniformity characteristics of the commercially available enzymes varied dramatically. Three enzymes Quantabio RepliQa Hifi Toughmix, Watchmaker Library Amplification Hot Start Master Mix (2X) 'Equinox' and Takara Ex Premier were found to give a consistent performance, over all genomes, that mirrored closely that observed for PCR-free datasets. We also test a range of enzymes for long-read sequencing by amplifying size fractionated *S. cerevisiae* DNA of average size 21.6 and 13.4 kb, respectively.

The enzymes of choice for short-read (Illumina) library fragment amplification are Quantabio RepliQa Hifi Toughmix, Watchmaker Library Amplification Hot Start Master Mix (2X) 'Equinox' and Takara Ex Premier, with RepliQa also being the best performing enzyme from the enzymes tested for long fragment amplification prior to long-read sequencing.

INTRODUCTION

In barely over a decade next-generation sequencing (NGS) has transformed the biological sciences and is now used for a diverse range of applications on a diverse range of species and sample types. From the early days of NGS, the concept of bias was recognized. No existing sequencing technology can 'read' a genome from start to end so DNA/RNA are fragmented to a size that can be read by the sequencing platform, a library of those fragments is prepared and original sample sequence is reconstituted following sequencing of the resulting mixture of library fragments. Ideally all of the constituent parts of the genome would be sequenced with equal representation, though in practice this is never the case. During the library prep and sequencing process there are several points that can introduce a bias into the representation, though perhaps the single most-bias introducing step is PCR. Sequencing libraries are comprised of a mixture of fragments of different size, GC and repeat content that together represent the original sample. Smaller, more GC neutral, fragments that do not contain any secondary structure will amplify more efficiently than larger, high GC, high AT fragments and those that are capable of forming secondary structures. Over multiple cycles any bias will be amplified so the number of PCR cycles should be kept to a minimum, indeed for the most even coverage sequencing dataset PCR-free libraries are recommended [3]. Whilst advantageous, PCR-free methods are not always practicable due to the high DNA mass input requirements, necessitating the use of PCR amplification during library prep.

Received 19 December 2023; Accepted 26 March 2024; Published 05 April 2024

Author affiliations: [1]Wellcome Sanger Institute, Hinxton, Cambs., CB10 1SA, UK; [2]Department of Medicine, University of Cambridge, Cambridge, Cambs., CB2 1TN, UK.

***Correspondence:** Michael A. Quail, mq1@sanger.ac.uk

Keywords: amplification; barcode; illumina; indexing; next-generation sequencing; PCR; polymerase.

Abbreviations: bp, base pairs; LCI, low coverage index; NGS, next-generation sequencing; PCR, polymerase chain reaction; Tm, melting temperature.

All datasets have been deposited in the ENA read archive. Human genome datasets under accession number ERP141249 (https://www.ebi.ac.uk/ena/browser/view/PRJEB56321), and microbial genomes under accession number ERP141224 (https://www.ebi.ac.uk/ena/browser/view/PRJEB56300).

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.Fourteen supplementary figures and three supplementary tables are available with the online version of this article.

001228 © 2024 The Authors

This is an open-access article distributed under the terms of the Creative Commons Attribution License.

Impact Statement

PCR amplification is a central step in many NGS protocols. This can introduce extreme bias in the data and result in overrepresentation of some sequences and underrepresentation or even loss of other sequences. Our original publication a decade ago, identified Kapa HiFi as the enzyme that gave the least bias and that gave the most even representation. This has been widely adopted. Here we identify three enzymes that significantly outperform Kapa HiFi, which should enable researchers across the globe produce better NGS data.

Since its invention in 1986 PCR has proven to be a convenient and effective way to selectively amplify genomic loci of interest and is the most common approach used for targeted selection ahead of both short- and long-read sequencing. Whilst the maximal fragment size sequenceable in a contiguous manner using short-read sequencing is around 600 bp, Pacific Bioscience and Oxford Nanopore Technologies sequencing platforms can generate reads in excess of 100 kb. This capability has led to long-read sequencing approaches being preferred for de novo sequencing. However, the long-read technologies require significant amounts of input DNA meaning amplification during long-read library preparation may be necessary. Here we compare enzymes for their ability to amplify long DNA fragments and compare yield and sequencing performance.

In 2011 we published a study that identified Kapa HiFi as the best enzyme for Illumina library amplification steps as it gave the most even coverage across a set of microbial genomes of diverse GC content [4]. Whilst most enzymes gave relatively even coverage over the more GC neutral genomes used (*Salmonella Pullorum* and *Staphylococcus aureus*) large differences in coverage representation were observed with libraries generated using different enzymes for the GC-rich genome of *Bordetella pertussis* and the AT-rich genome of *Plasmodium falciparum*.

Ten years on, we performed a similar study using many currently available and newly developed high-fidelity polymerase mixes/NGS amplification formulations. There are a great many commercially available PCR enzymes. Enzymes were chosen for this evaluation through discussion with individual enzyme manufacturers and suppliers to identify their enzymes that they recommended for NGS, these were mostly hotstart formulations of type II enzymes that have been developed to amplify complex templates and introduce fewer errors than standard Taq polymerases.

DATA SUMMARY

The following reference genome sequences were used for analysis

Human NA12878. GenBank: GCF_000001405.26

Bordetella pertussis Tohama I, GenBank: ASM400897v1

Escherichia coli MG1655. GenBank: U00096.3

Clostridioides difficile 630. NCBI Reference Sequence: NC_009089.1

Plasmodium falciparum 3D7. Genbank: GCA_000002765.3

Saccharomyces cerevisiae S288C. ATCC 204508

METHODS

Genomic DNA

Human NA12878 CEPH/UTAH PEDIGREE 1463 DNA was purchased from Coriell Camden, NJ.

Bordetella pertussis Tohama I (ATCC-BAA-589D-5), *Escherichia coli* MG1655 (ATCC-700926D-5), *Clostridioides difficile* 630(ATCC-BAA-1382DQ), *Plasmodium falciparum 3D7* (ATCC-PRA-405D) genomic DNA were obtained from ATCC via LGC standards, Teddington, UK.

Saccharomyces cerevisiae S288C genomic DNA (69240–3) was purchased from Merck, Gillingham, UK.

Illumina library construction

DNA (0.5 µg in 100 µl of 10 mM Tris-HCl, pH 8.5) was sheared in an AFA microtube using a Covaris S2 device (Covaris), with the following settings: for 200 bp fragments (duty cycle 20, intensity 5, 200 cycles/burst, 90 s) or for 500 bp fragments (duty cycle 20, intensity 5, 200 cycles/burst, 30 s).

Sheared DNA was purified by binding to an equal volume of AMPure XP beads (A63881, Beckman Coulter, Brea, CA) and eluted in 50 µl of 10 mM Tris-HCl, pH 8.5. End-repair, A-tailing and short Truseq adapter ligation* were performed using NEBNext UltraII (E7645L, New England Biolabs, Ipswich, MA). After ligation, excess adapters and adapter dimers were removed by Ampure XP clean-up with a 0.9:1 ratio of beads to sample, eluting in 50 µl of 10 mM Tris-HCl, pH 8.5.

Yield of adapter ligated fragments was ascertained by fluorimetric quantification using qubit high-sensitivity DNA reagents (Q32851, ThermoFisher, Waltham, MA).

These adapter ligated fragments were used as pre-PCR template for each of the PCR conditions tested. To avoid variation within starting material these pre-PCR templates were stored in aliquots and for each experimental phase, wherein comparisons were made, a single common pre-PCR template was used for all reactions.

*Oligo sequences are supplied in Table S1, available in the online version of this article.

PCR-free libraries were prepared using the same approach starting with 500 ng of genomic DNA input and ligating 'IDT for Illumina' full-length unique dual indexed adapters.

PCR

Each PCR was performed using 1 ng of template (Illumina truseq adapter ligated sheared DNA), 0.1 nmol each of unique dual indexed i7 and i5 barcoding Illumina PCR primers (Table S1) and 25 µl of 2×enzyme premix.

Unless indicated, all PCR reactions were performed for 14 cycles on an MJ Tetrad 4 thermocycler that had been validated for accuracy using a Driftcon thermocycler calibration instrument (Cyclertest, Landgraaf, the Netherlands).

Cycling conditions used for each enzyme are detailed in Table S2, with annealing at 60°C for 15 s, and with denaturation and extension parameters as recommended by each manufacturer.

PCR reaction products were cleaned and size selected using a 0.7:1 ratio of AMPure XP SPRI beads (Beckman Coulter) to sample, according to the manufacturers' protocol with elution in 30 µl of EB buffer.

Illumina sequencing

Prior to sequencing libraries were quantified by real-time PCR, using the SYBR Fast Illumina Library Quantification Kit (Kapa Biosystems cat. no. KK4834).

Libraries were pooled in an equimolar fashion whilst correcting for genome size to facilitate equal sequence coverage from each.

Samples were sequenced on an Illumina Novaseq 6000 instrument with 150 paired end read length and v1.5 chemistry.

Long fragment amplification and pacific biosciences sequencing

Two aliquots of *S. cerevisiae* DNA S288C were sheared using a Megaruptor 3 (Diagenode, NJ); aliquot 1:5 µl diluted to 150 µl volume with EB buffer was sheared at speed 30 and then 31 to tighten peak, aliquot 2:10 µg diluted to 310 µl volume with EB buffer was sheared at speed 31. Both aliquots of sheared DNA were purified and concentrated using a 1:1 ratio of PacBio AMPure SPRI beads with elution in 30 µl EB. The amount of sheared DNA was measured by fluorimetry using qubit high-sensitivity DNA quantification reagents. Aliquot 1 had 3.1 µg, and was size selected on Sage Sciences ELF instrument aiming for a maximum fragment size capture around 20 kb. Aliquot 2 had 4.8 µg, and 2.4 µg was size selected on Sage Sciences Blue pippin instrument with selection set to range 10–50 kb. Size selected fractions were purified using a 1:1 ratio of PacBio AMPure SPRI beads with elution in 50 µl EB. Prior to amplification Illumina adapters were added to the ends of each fragment to enable primer annealing. End-repair, A-tailing and short Truseq adapter ligation were performed using NEBNext UltraII as above. After ligation, excess adapters and adapter dimers were removed by Ampure XP clean-up with a 0.9 : 1 ratio of beads to sample, eluting in 50 µl of 10 mM Tris-HCl, pH 8.5. Yield of adapter ligated fragments was ascertained by fluorimetric quantification using qubit high-sensitivity DNA reagents; ELF fractionated DNA was 10.2 ng μl^{-1}, Blue pippin fractionated DNA was 3.04 ng μl^{-1}. Each was diluted to 1 ng μl^{-1} with EB and 1 µl of these dilutions used as template for long range PCR. Size fractionation of DNA is recommended prior to long-read sequencing as smaller fragments sequence with higher efficiency, due to smaller size reduced yield and ultimately shorter sequence reads that yield more fragmented assemblies. Here we have fractionated using both Sage Science ELF and Blue Pippin Instruments (https://sagescience.com/) as both were available to us and we wished to assess the performance of each. Each of these platforms allows preparative electrophoresis of DNA eluting user size ranges of fragments upto 40 kb. The Blue pippin allows collection of a user defined size range of DNA fragments on upto four DNA samples per run. The ELF instrument size fractionates a single sample into 12 size fractions, upto 40 kb.

After PCR amplified products were concentrated with a 1:1 ratio of PacBio AMPure SPRI beads with elution in 7 µl EB, 1 µl of which was used for QC and 5 µl for PacBio barcoded overhang adapter amplicon library prep with a different barcoded adapter being used for each successful PCR reaction [5].

PacBio libraries were quantified using qubit high-sensitivity DNA reagents and where possible ligated samples were equimolar pooled, weak samples were pooled in their entirety. Pooled barcoded amplified fragments were sequenced using a Pacbio Sequel IIe instrument using binding kit v2.2 and sequencing chemistry 2.0, library was loaded at 80 pM.

Data processing and analysis

After sequencing, reads were mapped to each genome reference sequence using Minimap2 [6]. SAMtools [7] was then used to generate pileup and coverage information from the mapping output.

The quality of the sequence data was assessed using FastQC v0.11.9.

For human genome sequence data fastq files were automatically aligned to reference GRCh38 and were mapped using bwa version 0.7.17-r1188 and the command bwa mem -t 12 p -Y -K 100000000<reference.fa> <read1.fastq.gz> <read2.fastq.gz> and duplicates were marked using biobambam2 version 2.0.79.

The resulting CRAM files were converted to BAM, sorted and indexed with samtools v.1.15.1. The stats and graphs were produced using samtools v.1.15.1.

All bams were subsampled to approx. 33X using samtools v 1.15.1 and seed 42.

Variants were called using GATK HaplotypeCaller v3.5 with special options '-stand_call_conf 2 -stand_emit_conf 2 A BaseQualityRankSumTest -A ClippingRankSumTest -A Coverage -A FisherStrand -A LowMQ -A RMSMappingQuality -A ReadPosRankSumTest -A StrandOddsRatio -A HomopolymerRun -A TandemRepeatAnnotator' [8].

For each dataset the overlap of variants on chromosomes 1–22 with the GIAB (Genome in a Bottle) v4.2.1 benchmark dataset was calculated using bcftools isec version 1.15.1. The true positive (TP) value was calculated as the number of variant sites that were identified in both the sample and and the GIAB benchmark dataset. The false negative (FN) was calculated as the number of variant sites identified in the GIAB dataset but not identified in the sample. The sensitivity was calculated as TP/(TP +FN) or the number variant sites found in the sample as a percentage of all the variant sites in the giab benchmark dataset [9].

PacBio HiFi read assembly was performed on 30×downsized coverage using IPA [10] assembler within SMRTlink v10.2.

RESULTS

Initial evaluation of enzymes for Illumina library amplification

Each enzyme was assessed for its ability to amplify genomic Illumina adapter ligated library fragments of an expected average insert size of approximately 500 bp, from a set of four microbial genomes with differing GC content:- *Bordetella pertussis*, 67.7% GC; *Escherichia coli,* 50.8% GC; *Clostridioides difficile*, 29.1% GC; and *Plasmodium falciparum,* 19.3% GC. For each enzyme tested, 1 ng of pre-PCR library fragments from each genome was amplified using manufacturers recommended denaturation and extension times, with annealing at 60°C for 15 s and 14 cycles. Unique dual indexed P7 and P5 amplification primers were used to avoid index hopping [11, 12].

For UDI oligonucleotides used, see Table S1.

Details of enzymes used and cycling conditions are listed in Table S2.

After 0.7×Ampure bead cleanup the yield of each library was assessed using fluorimetric measurement (Fig. S1). There were surprising differences in the yields obtained from the enzymes tested with some giving relatively little library. These were repeated with fresh enzyme on a different PCR block, always with the same outcome. Quantabio RepliQa and sparQ, Kapa HiFi, Invitrogen Platinum Superfi II, Thermo Collibri and Phusion U multiplex PCR mastermix, Tools Ultra, Biotool Univerase and Agilent Herculase gave good yields.

Barcoded libraries were pooled in a pseudoequimolar manner according to genome size and run on an Illumina Novaseq 6000 SP or S4 flowcell lane to give >30×coverage of each genome. To fairly compare results, datasets were randomly trimmed to contain reads representing 30×coverage. We tabulated the depth of coverage seen at each position of the genome and calculated the fraction of each genome (referred to as low coverage index) that was covered to a depth of less than 15×, i.e. half the mean coverage. Most datasets had <5% low coverage with the GC neutral *E. coli* genome but higher degrees of low coverage were observed for less base-balanced (either AT or GC rich) genomes with a lot of enzymes tested, with the extremely AT-rich genome of *P. falciparum* posing the biggest challenge (Fig. 1). Quantabio RepliQa, Kapa HiFi, and Collibri had <5% low coverage index with all four genomes.

Further evaluation of enzymes for Illumina library amplification

To test reproducibility further replicate libraries were made from the better performing enzymes, with the addition of Watchmaker Genomics Equinox library amplification mastermix, and Takara Ex Premier (these new formulations had been previously unavailable

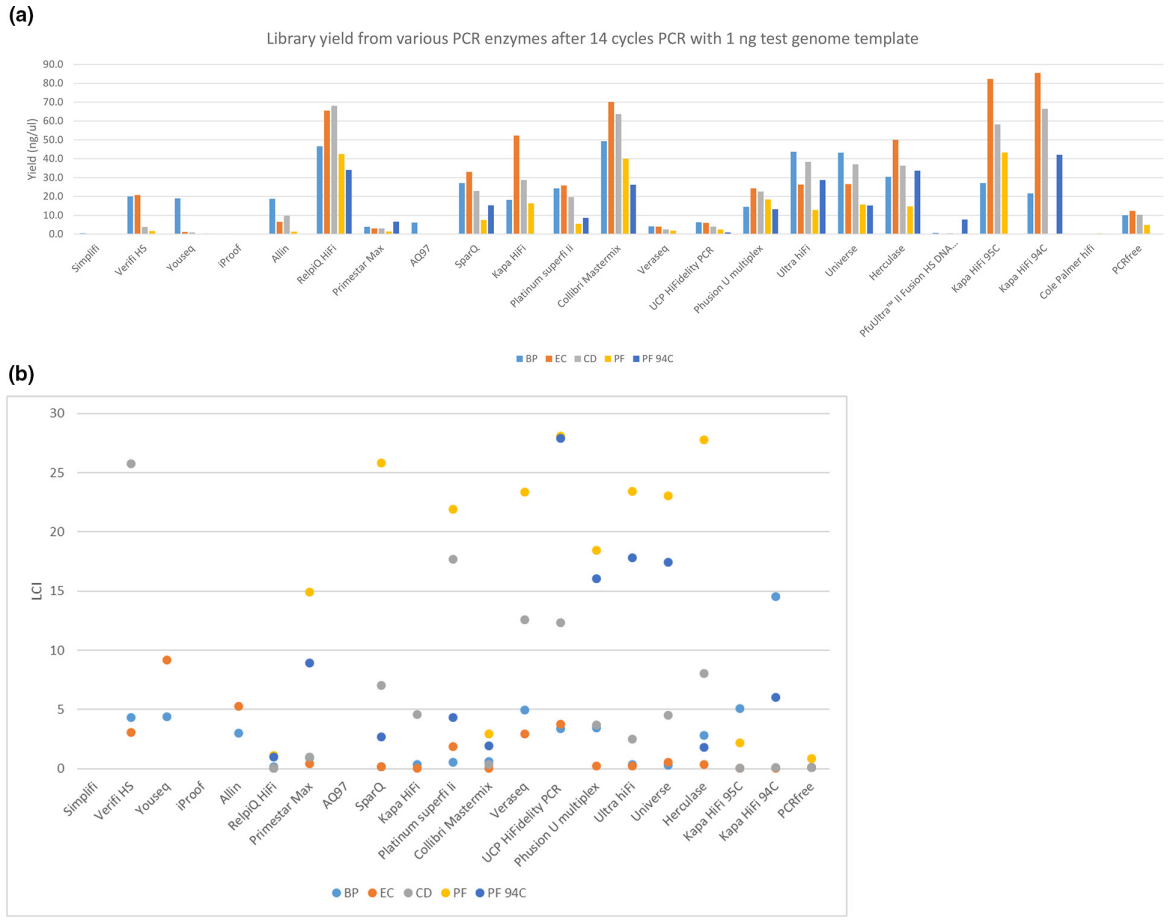


Fig. 1. Low coverage index (fraction of genome covered at <50% mean coverage) obtained after 14 cycles of amplification with 1 ng of each test microbial genome; BP: *Bordetella pertussis*, EC: *Escherichia coli*, CD: *Clostridioides difficile*, and PF: *Plasmodium falciparum*. For each PF was also amplified using a 94°C denaturation temperature, 'PF 94C'.

for testing), under a variety of different cycling conditions; Table S2. With this selected group, yields were quite high with all templates (Figs S2 and S3).

Again the low coverage index was calculated for each dataset and enzymes/conditions ranked from low to high LCI for each genome (Fig. 2).

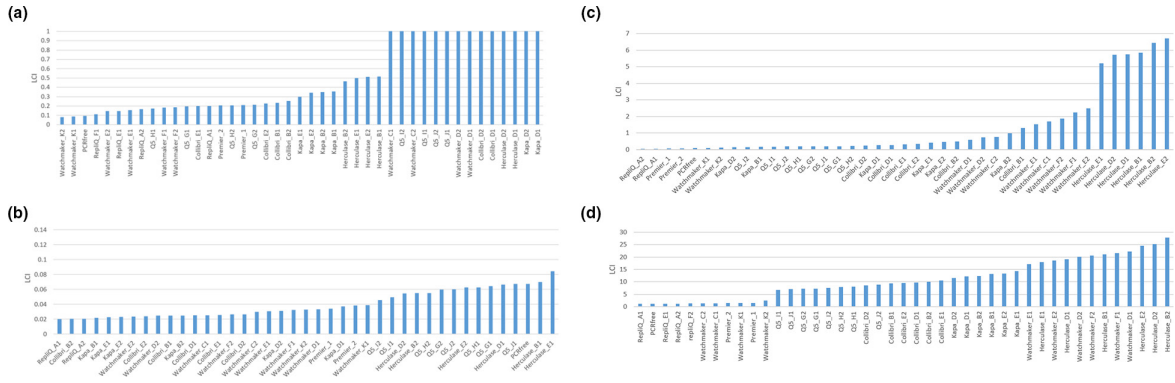


Fig. 2. Ranked low coverage index (fraction of genome covered at <50% mean coverage) values obtained after 14 cycles of amplification with 1 ng of each test microbial genome; (a) *Bordetella pertussis*, (b) *Escherichia coli*, (c) *Clostridioides difficile*, and (d) *Plasmodium falciparum*.

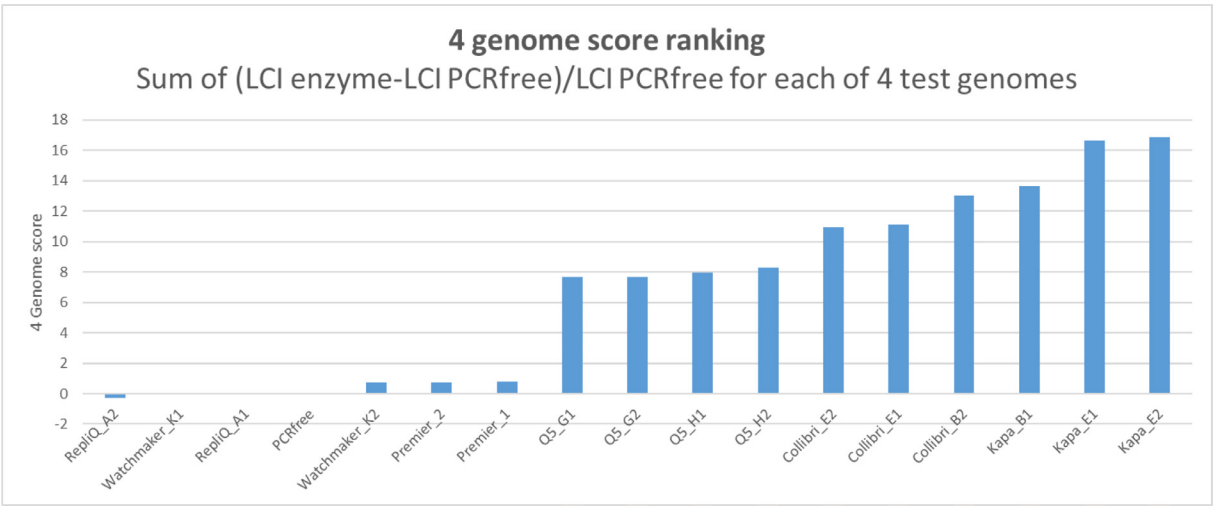


Fig. 3. Multi-genome average low coverage index (fraction of genome covered at <50% mean coverage) ranking taking the sum of the LCI values for all four genomes compared to that obtained from PCR free data.

Whilst some enzymes perform better in certain genomic contexts RepliQa, Watchmaker Equinox, and Takara Ex Premier, give good coverage uniformity with all genomes. To assess the average low coverage index across all genomes we calculated the sum of the low coverage values compared to coverage from PCR-free libraries (Fig. 3) illustrating that these three enzymes have minimal bias each giving coverage uniformity similar to that seen with PCR-free libraries.

The end result of the more even coverage obtained with RepliQa and Equinox relative to other enzymes could be clearly seen in the more challenging GC- or AT-rich regions where RepliQa, Equinox and PCR-free had good coverage in GC-rich (locally 100% GC) regions of *B. pertussis* and also in AT-rich (locally <4% GC) regions of *P. falciparum* (Fig. 4).

Sequencing data was also obtained for human genome template amplification. Again this showed RepliQa, Watchmaker Equinox and Takara Ex Premier to have the most even genome coverage reflected in the lowest LCI values (Fig. S4).

It has been observed that some enzymes are inhibited with magnetic beads that are commonly used in NGS workflows, e.g. SPRI magnetic bead cleanup and size selection, and streptavidin bead capture of biotin labelled DNA fragments. With SPRI cleanup

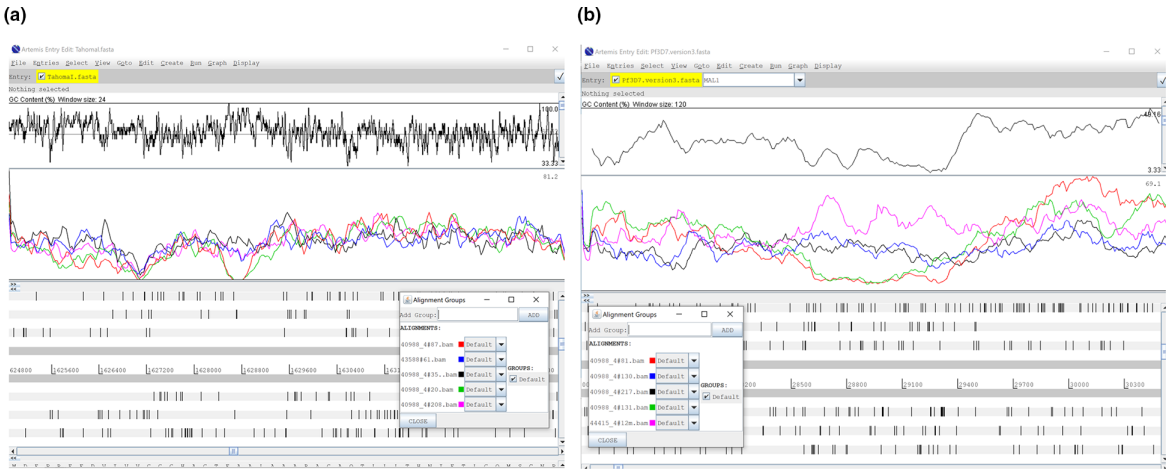


Fig. 4. Screenshot from Artemis genome browser [7] over regions of (a) *B. pertussis* and (b) *P. falciparum* genomes. In each the top panel is a GC content plot with the maxima and minima GC content of each region displayed at the far right. The bottom panel is an open read frame plot and the middle panel plots read coverage over each base. The inset panel on each denotes the enzyme used for each coloured line. In (a) RepliQa (pink), Watchmaker (blue) and PCR-free data (black) coverage is unaffected whereas with Kapa HiFi (red) and Collibri (green) coverage drops to near zero in the GC-rich region in the region in the centre of the plot. Likewise in (b) RepliQa (blue), Watchmaker (black) and PCR-free data (pink) coverage is unaffected whereas with Kapa HiFi (red) and Collibri (green) coverage drops to near zero in the AT-rich region in the region in the centre of the plot.

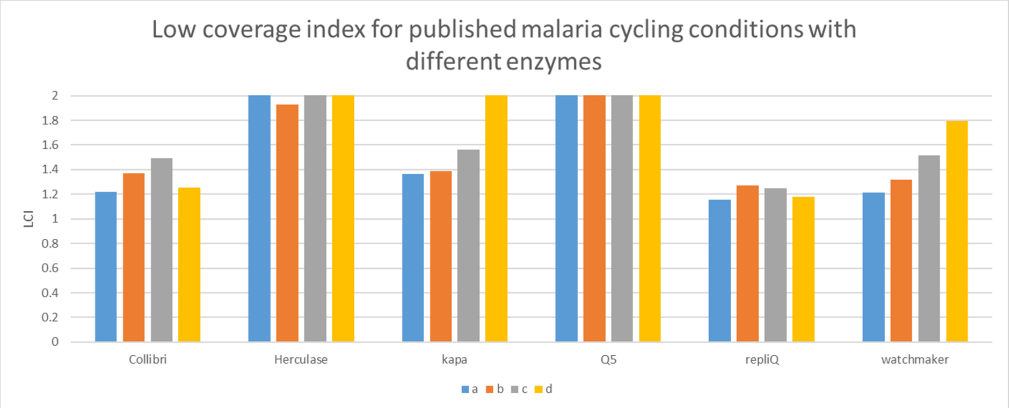
Table 1. Indel true positive, false negative and sensitivity values for NA12878 libraries prepared with either approx. 200 or 500 bp inserts with different PCR enzymes

Sample ID	No. SNPs (TP)	No. SNPs (FN)	SNP sensitivity (%)	No. indels (TP)	No. indels (FN)	Indel sensitivity (%)
repliQ_200 bp	3264914	93588	97.21	468524	67965	87.33
kapa HiFi_200 bp	3271956	86546	97.42	457018	79471	83.86
kapa HiFi_500 bp	3322273	36229	98.92	470902	65587	85.19
herculase_500 bp	3293130	65372	98.05	464692	71797	86.62
repliQ_500 bp	3333633	24869	99.26	502494	33995	93.66
Watchmaker_500 bp	3333145	25357	99.25	490866	45623	91.50
Premier_500 bp	3330695	27807	99.17	503501	32988	93.85

carryover of beads after elution is commonplace and some protocols employ 'with bead' approaches where increased yield is obtained when the beads are not removed after the final elution step [13]. Streptavidin conjugated magnetic beads are also used in NGS protocols for selection of biotin labelled library fragments, most commonly in hybrid capture target enrichment procedures [14], and due to the extremely strong affinity of streptavidin for biotin such methods require PCR amplification of bead bound library fragments. To test if amplification by the enzymes used in the second phase of this study are inhibited by such beads amplification was carried out without beads, with a volume of washed beads in water equivalent to equal sample volume or with the template bound to 50 µl of washed streptavidin beads (Dynabeads MyOne Streptavidin T1, Thermo, cat no. 65602). All of these enzymes tested were found to be unaffected by the presence of Ampure or streptavidin magnetic beads, apart from Q5 (Fig. S5).

Accuracy and utility of the human genome sequences obtained with each enzyme were assessed by comparing each dataset with the NA12878 reference genome and variant list (see Methods). Both numbers of indels and SNPs detected were slightly greater with 500 bp mean inserts compared to 200 bp. The three enzymes with the highest sensitivity for SNP and indel detection in the microbial reference genomes were QuantaBio RepliQa, Watchmaker Equinox and Takara Ex Premier (Table 1). These enzymes were found to replicate more SNPs and indels with greater reproducibility, compared to Kapa HiFi at rates that are comparable to those seen in the Precision FDA Truth challenge when using 50×PCR-free datasets [15].

There may be times in a high-throughput lab or within a clinical sequencing lab when fast turnaround is required when rapid PCR may be desired. Quantabio promote RepliQa for its short extension times. When we tested the enzymes in phase 2 with increasing



Condition	Publication	Initial extension	Denaturation	Anneal	Extend	Polish
a	Lopez-Barragan	94°C - 2min	94°C - 45 sec	60°C - 30 sec	60°C - 2min	60°C - 3min
b	Quail	98°C - 2min	98°C - 20 sec	60°C - 30 sec	60°C - 2min	60°C - 3min
c	Oyola	94°C - 2min	94°C - 15 sec	60°C - 30 sec	65°C - 2min	65°C - 5min
d	Aird	98°C - 3min	98°C - 80 sec	60°C - 30 sec	60°C - 2min	60°C - 10min

Fig. 5. Low coverage index (fraction of genome covered at <50% mean coverage) for amplification of 1 ng P falciparum Illumina library template with a range of enzymes using PCR conditions described by Lopez-Barragan [14] (blue), Quail (orange) [4], Oyola [12] (grey) and Aird [15] (yellow).

Table 2. Assembly statistics from PacBio HiFI 30×genome coverage from the various PCR enzyme datasets

Enzyme	Template	PCR cycles	Reads	Yield	Fold coverage	Sample name	Polished contigs	Max. contig length	Mean contig length	N50 contig length	Sum of contig lengths	No. of circular contigs
repliQa	ELF	15	95403	874550481	72.88	repliQAx15	20	1529563	609561	834867	12191225	2
repliQa	Bluepippin	15	67874	851073366	70.92	repliQaBx15	22	1092081	551212	746483	12126668	2
repliQa	Bluepippin	12	39211	508550579	42.38	repliQaBx12	22	1533220	553116	784885	12168568	2
Terra	Bluepippin	15	69375	811670287	67.64	terraBx15	23	1097683	528785	785063	12162064	3
Terra	ELF	15	132515	851869210	70.99	terraAx15	33	1092332	368810	509984	12170747	2
Q5	Bluepippin	15	76603	523297797	43.61	Q5B×15	35	880290	345193	474198	12081766	2
LongAmp	Bluepippin	15	69634	626495361	52.21	longampBx15	39	934826	310307	499795	12102002	1
Universe	Bluepippin	15	51436	470686450	39.22	universeBx15	104	604982	111981	161100	11646056	2
Watchmaker Equinox	Bluepippin	15	132438	1,120,679,413	93.39	WmBx15	123	386404	94855	134608	11667272	2
Universe	ELF	15	119121	599671617	49.97	universeAx15	147	421584	80022	116594	11763238	1
Watchmaker Equinox	ELF	15	89058	419979074	35.00	WmAx15	189	312222	62704	94017	11851164	3
SuperFi II	Bluepippin	15	85578	903200782	75.27	superfiBx15	196	292580	56206	80346	11016479	2
Promega Go Taq Long	Bluepippin	12	52834	591119208	49.26	promegaBx12	207	262329	51862	72104	10735587	4
SuperFi II	ELF	15	96073	536211718	44.68	superfiAx15	217	262033	51414	76546	11157036	2

extension times (5, 15, 30 or 60 s) it was observed that Collibri, Q5, RepliQa and Watchmaker Equinox enzymes gave near maximal yield after just 5 s of extension whereas Herculase and Kapa HiFi yields increased with extension time (Fig. S6).

The genome of the malaria parasite *Plasmodium falciparum* has an extremely low GC content of 19.3% [16] and has been shown to be one of the most challenging genomes to amplify and sequence [3, 4, 17]. There have been several papers published for this genome detailing methods to minimize the biases introduced by PCR and sequencing including PCR-free library approaches [3, 18] and optimized PCR protocols [4, 11, 19, 20]. In this study when using these approaches we find that the fraction of the genome at less than 50% of mean coverage could be decreased even further (Fig. 5) though the most successful reduction was achieved by using a different approach for different enzymes. The lowest LCI was achieved using RepliQa with denaturation at 94C and extension at 60°C as described by Lopez-Barragan, though near similar LCI values were also obtained under these conditions using Collibri and Watchmaker Equinox enzymes. Following on from this we tested RepliQa under a range of denaturation and extension temperature combinations and found that these conditions could not be improved upon (results not shown).

Long-range amplification for long-read sequencing

Long-range PCR is a common approach for generation of material for long-read sequencing. Many users have found this to be even more challenging with low yield and a bias towards smaller fragments during amplification. To test the suitability of PCR enzymes for this application we prepared size fractionated adapter ligated yeast genome fragments adding Illumina adapters to enable amplification using the same primers as used in the rest of this study.

Sheared *S. cerevisiae* DNA was size fractionated using Sage Sciences ELF or Bluepippin instruments yielding modal fragment sizes of 21.6 and 13.3 kb, respectively (Fig. S7). After adapter ligation 1 ng of each of these were used as a template for long range PCR with a range of enzymes using manufacturers recommended cycling conditions (Table S2).

Initially, 12 cycles of PCR was used, but with most enzymes that generated little or no product (data not shown) so PCR was repeated for 15 cycles after which time amplicons of the expected size were observed with most enzymes (Fig. S8), though yields varied widely (Table S3). The long-range PCR products were then prepared for Pacific Biosciences HiFi sequencing using manufacturers' recommended amplicon library prep protocol and barcoded adapters. Sequencing yields and coverage obtained are summarized in Fig. S9 and Table 2. Due to extremely low yields after PCR, products from some enzymes gave insufficient yield to obtain significant coverage.

For those amplification product libraries that gave >30× genome coverage low coverage index was calculated. The lowest LCI (indicating more even genome coverage was obtained with RepliQa followed by Terra polymerase (Fig. S10).

Long-range PCR can often preferentially amplify smaller templates such that after multiple cycles the amplification reaction can be dominated by such shorter amplicons. Bluepippin size selected templates amplified by RepliQa and terra polymerase gave the longest average subread lengths (library insert size) of approximately 12 kb (Figs 11 and 12). With the larger 21 kb ELF fractionated template the majority of reads were obtained from shorter amplification products. RepliQa gave the largest fraction of 20 kb subreads (Fig. S13).

By comparing the PacBio HiFi data with the sequence of the *S. cerevisiae* S288C genome reference the error profile of the library generated after amplification with each enzyme could be determined. Terra, LongAmp and Promega Go Long are Taq based polymerase formulations and as a result were observed to give higher rates of particularly mismatch errors compared to the other enzymes that possess proofreading activity. NEB Q5 gave the lowest error rates (Fig. S14).

As might be expected those enzymes that gave the most even genome coverage also gave the best assembly statistics when 30×normalized coverage reads were assembled in the SMRTlink portal (Table 2). Here RepliQa followed by Terra polymerase gave the most contiguous assemblies. The *S. cerevisiae* genome is known to have 16 chromosomes [21] and additional circular chromosomal elements have been reported [22], therefore assemblies from material amplified using these enzymes has given near complete contiguity with the sum of contig lengths matching that expected for the yeast genome and with ELF fractionated fragments amplified for 15 cycles with RepliQa assembling into just 20 contigs.

DISCUSSION

It has been estimated that when sequencing the human genome 30× average genome coverage is required to give local base coverage at >15× so that both homozygous and heterozygous variants can be accurately detected [23, 24]. Indeed for the UK Biobank sequencing project 95% coverage at >15× was a key sequence dataset QC metric [25].

Here we compare sequence datasets from different PCR enzymes based on low coverage index, the percentage of the genome that is covered to <15× when using 30×datasets.

We have used 14 cycles PCR with 1 ng of template in order to exacerbate any biases so to be able to differentiate between the enzymes used. Even with 14 cycles and a standardized input there was a broad range of yields. With those enzymes that gave

a higher yield, fewer PCR cycles could have been used and indeed here we could have overamplified, possibly introducing bias. However, we obtained similar low coverage index values after 10 cycles PCR for these enzymes (results not shown).

The results presented demonstrate that there are distinct differences between PCR enzymes in the yield and evenness of amplification of fragments prior to sequencing. This further confirms that PCR can be a source of bias in genomic data and illustrates that the user should consider which enzyme is used for these applications, particularly for GC biassed templates where coverage bias is more pronounced. Some enzymes work well, giving even coverage and low bias, in some situations, but few are capable of unbiased amplification of both GC- and AT-rich templates. Here, we have demonstrated that RepliQa, Watchmaker Equinox and Takara Ex Premier can amplify extremes of GC content with coverage bias similar to PCR-free data and better uniformity than Kapa HiFi that we previously reported to give the best performance [4]. Surprisingly there was considerable variation in yield between enzymes, again illustrating that the user should carefully choose the enzyme used as utilizing a low efficiency enzyme may give low yield, requiring the user to employ more cycles of PCR compared to other enzymes and accentuate the bias even more.

The performance of RepliQa, Equinox and also Takara Ex Premier reported here is impressive, with coverage uniformity after 14 cycles of PCR using 1 ng template close to that obtained from PCRfree libraries made with 500 ng input, even over wide extremes of GC content.

Here the genome of the malaria parasite has been a particular challenge. It is extremely AT rich with long stretches close to 100% AT [16]. Plasmodium causes a severe disease burden especially in sub-Saharan Africa, in the 2021 World Malaria report there were 241 million malaria cases and 627 000 malaria deaths worldwide in 2020. As a result its genome is a major focus for researchers who often face challenges with human host DNA contamination and as those infected are mostly young children from which only low volume blood samples can be taken, only low amounts of DNA are often available. Using previously published modifications to cycling conditions we report extremely even genome coverage from just 1 ng template using RepliQa, Watchmaker Equinox and Takara Ex Premier.

Sequencing of human genomes, particularly for variant discovery for academic and clinical outcomes, is a major application of NGS. The human genome is much larger and more complex than the microbial genomes used above and so the lead enzymes from this study were also tested for their ability to amplify limiting amounts of human genome template. DNA from the genome in a bottle standard NA12878 was used for this as it is probably the most sequenced human genome and has a good reference and validated list of variants [26]. Here the RepliQa, Watchmaker Equinox and Takara Ex Premier all performed remarkably well giving even coverage (low LCI) as well as high degrees of SNP and Indel precision all of which are better than the performance seen with Kapa HiFi. There are slight differences in performance with RepliQa and Watchmaker having greater SNP precision (Table 1) and RepliQa and Takara Ex Premier slightly better performance on Indels, meaning that from data generated here RepliQa would be the enzyme of choice as it gives superior performance both for Indels and SNPs.

In this study we have shown these enzymes to be capable of amplification of the complex genomes and genomes with extreme GC composition, to give uniformity of coverage similar to that obtained in PCR free library datasets. As a result we would expect these enzymes to perform well on other genomes irrespective of complexity and GC content, especially when using pure genomic DNA as used here. It should be noted that RepliQa is also reported to be tolerant to many known PCR inhibitors that can be present in crude extracts and has been shown to efficiently amplify targets from soil [27] and insects [28]. Given that these enzymes give efficient amplification of both high and low GC content loci, though not tested here, it is expected that they would be particularly suited for amplification of metagenomes and microbial communities containing many different species at different percentages and with widely varying GC content genomes.

PCR amplification is also used ahead of long-read sequencing either as a means of targeted sequencing or for low input template preparation. Standard ONT and PacBio library prep methods are PCR-free and require over 1 µg input DNA though both sequencing companies have protocols for low input library prep that involves amplification. We have compared a variety of polymerases for their ability to amplify 1 ng of adapter ligated long template DNA and found the majority of enzymes to be quite inefficient for amplification of yeast genome fragments of 13.3 and 21 kb and highlight a small number of enzymes that are capable of such reactions. The enzymes which performed best resulting in near complete genome coverage were RepliQa and Terra polymerase, though being a derivative of Taq Terra had higher rates of substitution error. Both these enzymes gave superior performance to Takara Primestar GXL reported by [29] to be the best enzyme for long template amplification [29], indicating that better enzymes are continuing to be developed.

Having better performing PCR enzymes will enhance the performance of NGS approaches, particularly for long-read sequencing and yield the most contiguous assemblies and most complete analysis of genomic variants. At the time of this report RepliQa appears to give the best outcome for a variety of genomes and applications.

Funding information
This work was supported by the Wellcome Trust [grant number 098051].

Acknowledgements
The authors thank the Wellcome Trust Sanger Institute core sequencing and informatics teams. Steven Leonard and IDT for design of barcode sequences.

Author contributions
M.Q. performed the experiments and performed primary data analysis. M.Q. designed the experiments. C.C. and J.U. performed the PacBio library prep and sequencing work. M.Q. wrote the manuscript. Y.G.U. and J.K. carried out bioinformatics analysis.

Conflicts of interest
MQ is a member of the New England Biolabs panel of Key Opinion Leaders. The other authors have no competing interests.

Ethical statement
Not applicable. The only human DNA used in this study was from a commercial source.

References

1. Bronner IF, Quail MA. Best practices for illumina library preparation. *Curr Protoc Hum Genet* 2019;102:e86.
2. Bronner IF, Quail MA, Turner DJ, Swerdlow H. Improved protocols for illumina sequencing. *Curr Protoc Hum Genet* 2014;80:11–42.
3. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, *et al*. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* 2009;6:291–295.
4. Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, *et al*. Optimal enzymes for amplifying sequencing libraries. *Nat Methods* 2011;9:10–11.
5. Biosciences P. https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-SMRTbell-Libraries-using-PacBio-Barcoded-Overhang-Adapters-for-Multiplexing-Amplicons.pdf
6. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.
7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al*. Genome project data processing S: the sequence alignment/map format and Samtools. *Bioinformatics* 2009;25:2078–2079.
8. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, *et al*. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* 2019;37:561–566.
9. GIAB. https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv4.2.1/GRCh38
10. DunnC, Sovic I. IPA Hifi genome assembler. *GitHub* 2022.
11. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, *et al*. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* 2018;19:332.
12. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, *et al*. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *Mol Biol*2017.
13. Fisher S, Barry A, Abreu J, Minie B, Nolan J, *et al*. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* 2011;12:R1.
14. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, *et al*. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 2009;27:182–189.
15. Challenge PFT. https://precision.fda.gov/challenges/truth/results
16. Gardner MJ, Hall N, Fung E, White O, Berriman M, *et al*. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 2002;419:498–511.
17. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, *et al*. Optimizing illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics* 2012;13:1.
18. Chappell L, Ross P, Orchard L, Russell TJ, Otto TD, *et al*. Refining the transcriptome of the human malaria parasite *Plasmodium falciparum* using amplification-free RNA-seq. *BMC Genomics* 2020;21:395.
19. López-Barragán MJ, Quiñones M, Cui K, Lemieux J, Zhao K, *et al*. Effect of PCR extension temperature on high-throughput sequencing. *Mol Biochem Parasitol* 2011;176:64–67.
20. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, *et al*. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 2011;12:R18.
21. Fisk DG, Ball CA, Dolinski K, Engel SR, Hong EL, *et al*. *Saccharomyces cerevisiae* S288C genome annotation: a working hypothesis. *Yeast* 2006;23:857–865.
22. Møller HD, Parsons L, Jørgensen TS, Botstein D, Regenberg B. Extrachromosomal circular DNA is common in yeast. *Proc Natl Acad Sci U S A* 2015;112:E3114–E322.
23. Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics* 2013;14:195.
24. Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* 2014;15:247.
25. Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, *et al*. The sequences of 150,119 genomes in the UK Biobank. *Nature* 2022;607:732–740.
26. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, *et al*. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 2014;32:246–251.
27. Santiago JM, Fox J-P, Guzmán SM, Rossi L. Effect of fabric mulch ground covers on lemon trees rhizosphere microbiome in florida flatwood soils. *Front Soil Sci* 2023;3:3.
28. Stein F, Wagner S, Bräsicke N, Gailing O, Moura CCM, *et al*. A non-destructive high-speed procedure to obtain DNA barcodes from soft-bodied insect samples with a focus on the dipteran section of schizophora. *Insects* 2022;13:679.
29. Jia H, Guo Y, Zhao W, Wang K. Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. *Sci Rep* 2014;4:5737.