



Selective multiplexed enrichment for the detection and quantitation of low-fraction DNA variants via low-depth sequencing

Ping Song¹, Sherry X. Chen¹, Yan Helen Yan^{1,2}, Alessandro Pinto³, Lauren Y. Cheng¹, Peng Dai¹, Abhijit A. Patel⁴ and David Yu Zhang^{1,2} ✉

DNA sequence variants with allele fractions below 1% are difficult to detect and quantify by sequencing owing to intrinsic errors in sequencing-by-synthesis methods. Although molecular-identifier barcodes can detect mutations with a variant-allele frequency (VAF) as low as 0.1% using next-generation sequencing (NGS), sequencing depths of over 25,000× are required, thus hampering the detection of mutations at high sensitivity in patient samples and in most samples used in research. Here we show that low-frequency DNA variants can be detected via low-depth multiplexed NGS after their amplification, by a median of 300-fold, using polymerase chain reaction and rationally designed ‘blocker’ oligonucleotides that bind to the variants. Using an 80-plex NGS panel and a sequencing depth of 250×, we detected single nucleotide polymorphisms with a VAF of 0.019% and contamination in human cell lines at a VAF as low as 0.07%. With a 16-plex NGS panel covering 145 mutations across 9 genes involved in melanoma, we detected low-VAF mutations (0.2–5%) in 7 out of the 19 samples of freshly frozen tumour biopsies, suggesting that tumour heterogeneity could be notably higher than previously recognized.

High-throughput sequencing-by-synthesis, also known as next-generation sequencing (NGS), is one of the most powerful methods available today for highly multiplexed characterization of DNA at many genetic loci^{1,2}. Next-generation sequencing is used routinely to detect DNA sequence variants with an allele frequency of $\geq 5\%$. However, NGS struggles to report single nucleotide variants (SNVs) with low variant-allele frequencies (VAFs) because all current NGS platforms have an average intrinsic error of at least 0.2%^{3,4}. Depending on the size of the panel, and the expertise of the laboratory, the VAF limit of detection (LoD) for standard targeted NGS panels ranges between 1% and 5%. Error-correction methods based on unique molecular identifiers (UMIs) have been proposed and can circumvent the NGS intrinsic error limitation to achieve SNV LoDs of about 0.1%^{5–7}, but these generally require sequencing to extremely high depths of 25,000× or more. This results in a per-sample NGS cost of more than US\$1,000 for even moderate-size panels (for example, the 70-gene Guardant 360 panel^{8,9}), rendering sensitive detection of SNVs with a VAF of below 1% unaffordable for many researchers, clinicians and patients.

Simultaneously, there is growing research and clinical demand for high-sensitivity detection of DNA sequence variants at a low VAF. Examples for which low-VAF detection is needed include cell-free DNA (cfDNA) profiling for non-invasive cancer therapy guidance and post-treatment monitoring^{10,11}, rare-microbe detection for microbiome profiling^{12,13} and persistent bacterial subpopulations contributing to antibiotic resistance^{14,15}. Polymerase chain reaction (PCR)-based methods, such as digital PCR^{16,17} and allele-specific PCR¹⁸, can be used to detect and quantitate one or a few suspected DNA variants with a VAF of as low as 0.1% but cannot reasonably scale to fulfil the multiplexing needs of many research and clinical applications.

Here we present multiplex blocker displacement amplification (mBDA), a library preparation method that allows for robust NGS detection and quantitation of SNVs with a VAF as low as 0.019% using a sequencing depth of only 250×. Multiplex BDA functions by selectively enriching DNA sequence variants during a multiplex PCR target-enrichment step. Unlike other allele-enrichment methods¹⁹, mBDA can scale well to multiplex panels; we demonstrate a 300-fold median enrichment in an 80-plex panel. Because the post-mBDA NGS library exhibits VAFs far exceeding the NGS intrinsic error rate, low-depth sequencing is sufficient for variant detection and no error correction is needed. Furthermore, because the fold-enrichments of different variants are conserved across libraries and samples, it is possible to accurately quantitate the initial sample variants with 95% accuracy within a factor of two. By reducing the number of NGS reads needed by over 100-fold, mBDA uniquely enables the affordable sequencing of DNA variants with very low VAFs on lower-throughput NGS instruments such as the Illumina MiSeq.

Results

The mBDA method allows highly multiplex sequence-selective PCR amplification of single nucleotide polymorphism (SNP) alleles through the use of rationally designed ‘blocker’ oligonucleotides that perfectly bind to the intended SNP allele²⁰. The binding region of the blocker overlaps with that of the forward PCR primer, resulting in a competitive hybridization reaction in which the forward primer must displace the blocker to bind to the DNA template (Fig. 1a). The blocker sequence is designed to bind more strongly than the forward primer to the DNA templates bearing the intended SNP allele but less strongly to DNA templates bearing the variant SNP allele. Consequently, the variant templates are amplified with notably higher yield per PCR cycle than the wild-type templates.

¹Department of Bioengineering, Rice University, Houston, TX, USA. ²Systems, Synthetic, and Physical Biology, Rice University, Houston, TX, USA. ³NuProbe USA, Houston, TX, USA. ⁴Department of Therapeutic Radiology, Yale University, New Haven, CT, USA. ✉e-mail: dyz1@rice.edu

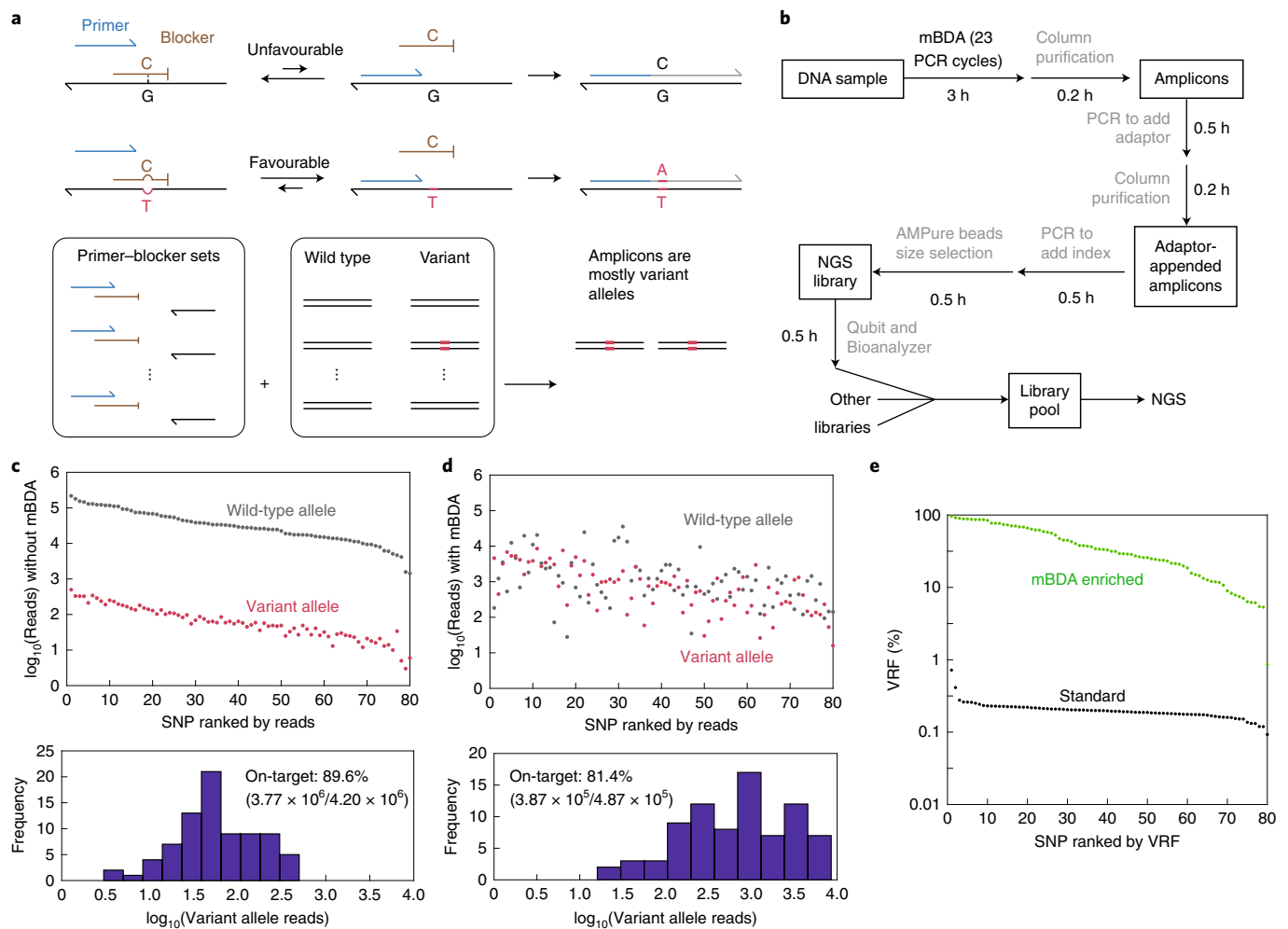


Fig. 1 | Allele enrichment with mBDA enables the detection of rare variants using low-depth sequencing. **a**, In mBDA, rationally designed blocker oligonucleotides compete with forward PCR primers in binding to the DNA templates at the loci of interest. The blocker is designed to be perfectly complementary to the wild-type DNA sequence. Any DNA sequence variant in the enrichment region (approximately 20 nt) results in mismatched blocker binding, leading to preferential PCR amplification of the variant. In highly multiplexed settings, a DNA sample may possess sequence variants in only a small fraction of the tested loci, resulting in amplicons being dominated by a small number of amplicons bearing the variant sequences. **b**, The entire library preparation workflow of mBDA NGS takes less than 6 h. **c**, Summary of NGS results on a library constructed using standard multiplex PCR target enrichment. Here we used a 300 ng sample of a 99.8%:0.2% mixture gDNA from the NA18537 and NA18562 cell lines. The panel targeted 80 loci on the human genome bearing SNPs in which NA18537 and NA18562 were homozygous for different alleles. Consequently, the VAFs of the NA18562-specific alleles were expected to be approximately 0.2% for all 80 amplicons. Bottom: histogram plot of the \log_{10} (Variant allele reads). **d**, Summary of the NGS results on a library constructed using mBDA on 300 ng of 99.8%:0.2% NA18537:NA18562. Compared with the library in **c**, ninefold fewer NGS reads were used but the variant alleles were sequenced to a higher depth. The SNPs were ranked as per **c**. Bottom: histogram plot of the \log_{10} (Variant allele reads). **e**, Summary of the VRF values for each SNP locus in the libraries described in **c** and **d**. The standard multiplex PCR NGS library showed a median VRF of approximately 0.2%, consistent with expectations.

Through the course of many PCR cycles, the variant allele can be preferentially amplified with a 1,000-fold higher efficiency than the wild-type allele.

The mBDA NGS library preparation workflow takes less than 6 h from DNA to library; the process is summarized in Fig. 1b. To demonstrate allele enrichment with mBDA, we constructed an 80-plex mBDA NGS panel targeting 80 common human SNPs. These 80 SNPs were selected such that the NA18537 and NA18562 human genomic DNA (gDNA) samples were homozygous for different alleles^{21,22}. A 99.8%:0.2% mixture of NA18537 and NA18562 gDNA was thus an easily formulated reference sample with 0.2% VAF in all 80 SNP loci, when considering the NA18537 alleles as the wild type. Figure 1c shows the Illumina MiSeq sequencing results on the 0.2% VAF sample using a standard multiplex PCR target-enrichment

workflow. The number of reads aligned to the variant allele for each SNP locus was roughly 500-fold lower than the intended allele, consistent with expectations. Note that not all of these variants can be confidently called from these data; we will discuss the VAF LoD in the next section.

The mBDA NGS results on a different aliquot of the 99.8%:0.2% NA18537/NA18562 sample are shown in Fig. 1d for comparison. Despite using ninefold fewer total reads than the standard amplicon NGS library, the mBDA NGS library exhibited higher variant-allele read depth because the number of intended allele reads were dramatically reduced. Figure 1e displays the variant-read fraction (VRF) for each SNP locus, with the median VRF increasing from 0.2% in the standard NGS library to about 30% in the mBDA NGS library. Thus, the variant SNP alleles here are enriched by a median of 150-fold.

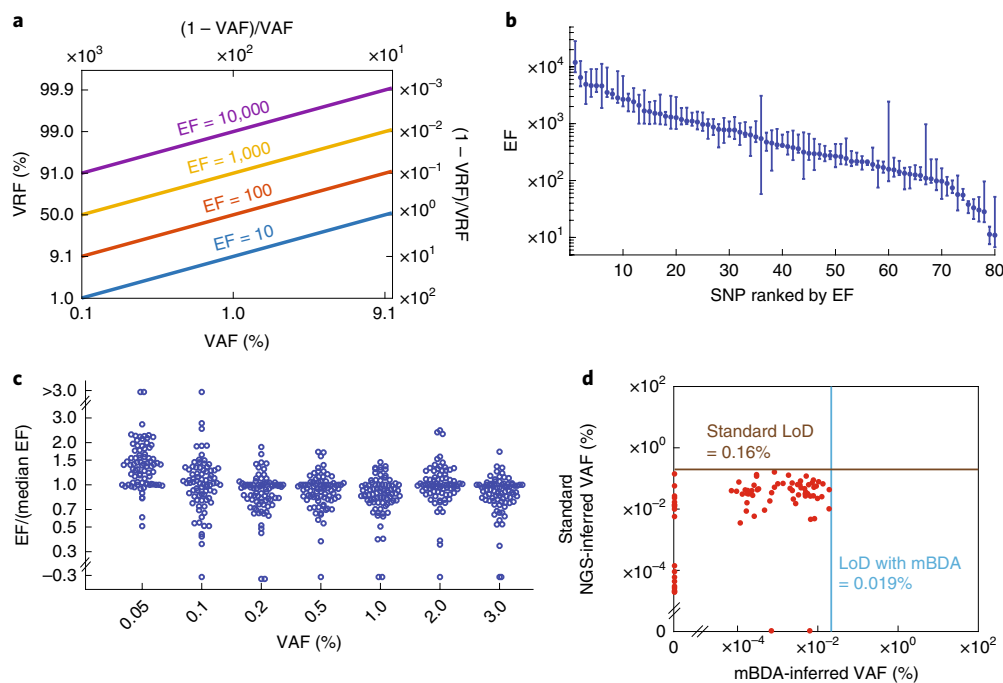


Fig. 2 | Quantitation of variant VAFs based on observed VRF values from mBDA libraries. **a**, Theoretical relationship between VAF and VRF for different allele EF values. The relationship between $\frac{(1 - VRF)}{VRF}$ and $\frac{(1 - VAF)}{VAF}$ is expected to be linear, with a slope of one and an intercept of $\log_{10}(EF)$. The EF values are expected to vary for different SNPs but are conserved across different experiments and VAFs for the same SNP allele. **b**, Summary of the inferred EF for each of the $n = 80$ variant SNP alleles based on a set of $n = 7$ calibration mBDA NGS libraries using samples with a VAF of 0.05, 0.1, 0.2, 0.5, 1, 2 and 3%. The dots show the median values and the error bars show the maximum and minimum EF values. The 0.05% VAF sample used 200 ng gDNA input; all other samples used 50 ng gDNA input. **c**, Relative EF values for different VAF inputs. The EF/(median EF) ratio can also be interpreted as the quantitation error for the calibration samples with known VAFs. More than 95% of the inferred VAFs are quantitated accurately to within a factor of two when $VAF \geq 0.1\%$. **d**, Comparison of the VAF LoD for standard amplicon NGS and mBDA NGS based on triplicate NGS libraries for each method. Here we define the LoD as the maximum inferred VAF from a pure wild-type (0% VAF) sample; the red dots show the inferred VAFs for each of the 80 SNPs. For the mBDA NGS, the VAF is calculated based on the median EF from the seven EF values summarized in **c** and **d**. For standard amplicon NGS, the VAF is calculated as the VRF. The VAF LoD was improved more than eightfold by mBDA relative to standard amplicon NGS.

See Supplementary Section 1 for the SNP loci and mBDA design details, and Supplementary Section 2 for the NGS protocol optimization experiments. We used a custom bioinformatics pipeline to filter the NGS reads from FASTQ files, make variant calls and determine the corresponding VRF values (Supplementary Section 3). Standard bioinformatics tools such as Picard (<http://broadinstitute.github.io/picard/>) and GATK (<https://gatk.broadinstitute.org/hc/en-us>) are optimized for hybrid-capture NGS panels, and we and others have observed these software to exhibit both false-positive and false-negative errors when making variant calls in amplicon sequencing libraries.

Quantitation of the VAF from mBDA NGS results. The relationship between the initial VAF in the sample and the observed VRF in the library can be described mathematically, based on the variant-allele enrichment fold (EF):

$$VRF = \frac{EF \times VAF}{EF \times VAF + (1 - VAF)} \quad (1)$$

Note that VRF is not linear with EF; VRF values close to 100% imply much greater EF values. For example, assuming $VAF = 0.1\%$, $VRF \approx 10\%$ when $EF = 100$, $VRF \approx 50\%$ when $EF = 1,000$ and $VRF \approx 90\%$ when $EF = 10,000$ (Fig. 2a).

The EF values will vary based on the sequence identity of the variant and intended alleles as well as the neighbouring context sequence. To calculate the EF for each of the 80 SNP loci, we ran a

series of different NGS libraries using samples with a known VAF (mixing NA18537 and NA18562). Rearranging equation (1) from above, we obtain:

$$EF = \frac{VRF \times (1 - VAF)}{VAF \times (1 - VRF)} \quad (2)$$

$$\log(EF) = \log \left[\frac{VRF}{(1 - VRF)} \right] - \log \left[\frac{VAF}{(1 - VAF)} \right] \quad (3)$$

The transformed relationship between EF, VAF and VRF is plotted in Fig. 2a. Figure 2b shows the maximum, minimum and median EF values calculated for each SNP locus from seven different mBDA NGS libraries with VAFs of 0.05–3%. Among the 80 SNP loci, 100% of the median EF values were greater than 10, 96% of the median EF values were greater than 30 and 85% of the median EF values were greater than 100. See Supplementary Section 4 for the EF calibration experiments.

The variation in EF values means that different LoDs are achievable for different mutations. Given the standard NGS LoD of 2% VAF, an EF value of ten corresponds to an LoD of about 0.2% VAF and an EF value of 30 corresponds to an LoD of 0.067% VAF. Thus, for the great majority of variants, the LoD for mBDA is not bottlenecked by the varying EF values but rather by the quantity of sample input and/or DNA damage.

Figure 2c shows the relative values of EF at each VAF compared with the median EF values.

The median values of EF/(median EF) are close to one for mutations with a VAF $\geq 0.1\%$, indicating that there is limited systematic bias in VAF quantitation. The systematic upward bias of $\frac{EF}{\text{median EF}}$ at 0.05% VAF may reflect false positives due to DNA polymerase misincorporation errors.

To determine the LoD of both standard amplicon and mBDA NGS, we ran triplicate NGS libraries on pure NA18537 DNA samples that were nominally 0% VAF for all 80 SNP loci. The VAF LoD was determined to be 0.16% for the standard amplicon NGS, based on the maximum observed VRF (y axis in Fig. 2d). For the mBDA NGS, the VAF LoD was determined to be 0.019%, based on the maximum VAF values calculated from the VRF and median EF values (x axis in Fig. 2d). Thus, the mBDA NGS method improved the VAF LoD by roughly eightfold compared with standard amplicon NGS because it eliminated the false-positive variant calls from the NGS intrinsic error. The remaining false-positive errors were suspected to be primarily due to the misincorporation of incorrect nucleotides by the DNA polymerase during the early cycles of the PCR²³.

Low-VAF variant calls from the mBDA NGS results. We next applied the mBDA NGS method to DNA samples with low fractions (0.07–0.25%) of conspecific contaminants (Fig. 3a). Because of the notable reduction in NGS reads required by the mBDA NGS libraries, we were able to run 21 different DNA samples on a single Illumina MiSeq chip while maintaining sensitivity to very low VAFs. In each of the libraries, the inferred VAFs for heterozygous alleles (regions shaded in green) and homozygous variant alleles (orange regions) are generally consistent with expectations, with $>80\%$ of the VAFs within a factor of two of the median for each group.

By contrast, with standard amplicon NGS, only two NGS libraries could be run on a MiSeq chip (Fig. 3b; MiSeq-2 and MiSeq-2 repeat) and roughly 20% of the homozygous intended alleles had inferred VAFs of more than 0.02%. When comparing the results from mBDA NGS and amplicon NGS on the same sample (roughly 0.18% HeLa in NA18537), we saw that the standard NGS method had somewhat tighter distributions of called VAFs than mBDA NGS. However, there was one heterozygous allele outlier with an expected VAF of 0.09% that produced a variant call of 0.58% VAF.

We could calculate the false-positive and false-negative rates of the variant-allele calls for the 1,760 mBDA NGS sample-locus

combinations, using the VAF LoD of 0.019% established in Fig. 3e. The overall specificity of the variant calls was 98.42%, and the sensitivity was 95.12% for the homozygous variant alleles and 95.04% for the heterozygous variant alleles (Fig. 3d). A change in the VAF LoD threshold would result in a trade-off between the relative sensitivity and specificity (Fig. 3e); the area under the receiver operator characteristic was calculated to be 0.9968. For comparison, the standard amplicon NGS libraries yielded a specificity of 100% and a sensitivity of 69.23% for homozygous variant alleles, and a sensitivity of 9.52% for heterozygous variant alleles, based on the previously established VAF LoD of 0.019%.

Because mBDA notably enriches the variant alleles, low-depth sequencing is sufficient for the detection and quantitation of variants with a low VAF. Figure 3c shows the inferred VAF for each of the 80 SNP loci based on the read depth of each SNP. The inferred VAFs above the 0.019% LoD quickly converged to their final values, typically within 250 reads. Thus, in theory, an NGS library as small as $250 \times 80 = 20,000$ reads would be sufficient to achieve a detection of 0.1% VAF. In practice, imperfect amplicon depth uniformity across the 80 loci means that roughly 50,000 reads are needed to obtain accurate inferred VAF values (Fig. 3f,g and Supplementary Section 5).

The experiments shown in Fig. 3 used 300 ng of input DNA to minimize the effects of Poisson's distribution during sample preparation and do not represent a limitation of the mBDA NGS method. In Supplementary Section 6 we show the experimental results for mBDA NGS on 10 ng of input DNA (0.18% HeLa contaminant in NA18537).

The mBDA VAF LoD we achieve matches or exceeds most UMI-based NGS methods, with 50-fold fewer reads required, but is worse than the consensus accuracy achieved by Pacific Biosciences sequencing of circularized templates. However, PacBio is an expensive platform that few laboratories have access to and its expensive consumables (on a per-read basis) render it non-competitive in the clinical sequencing space. For most clinical applications, the PacBio-level error rates (1×10^{-5} – 1×10^{-7}) are unnecessary due to biological sample limitations (for example, 10 ng DNA from fine-needle aspirates). In addition, the mutation VAF LoDs for biological samples are frequently limited by chemical damage to DNA, such as cytosine deaminations and guanine oxidations; PacBio sequencing will not be able to resolve these from true mutations.

Fig. 3 | Detection and quantitation of variants with a low VAF using mBDA NGS. **a**, Results for 20 of 21 mBDA libraries on a single MiSeq-1 chip using 11.8×10^6 reads. Each library was generated from 300 ng NA18537 gDNA contaminated with 0.07–0.25% of a different human DNA sample. Consequently, the homozygous variant alleles had a VAF of 0.07–0.25% and heterozygous variant alleles had a VAF of 0.03–0.13%. The individual library sizes ranged from 3.94×10^5 to 5.99×10^5 reads, and the library on-target rates varied from 75.2% to 83.8%. Each graph shows the inferred VAF for all 80 SNPs, sorted by contaminant genotype and then by inferred VAF. The black horizontal lines show the median inferred VAF for the homozygous and heterozygous variants. **b**, Comparison libraries using 300 ng NA18537 gDNA contaminated with approximately 0.18% D9 (HeLa) gDNA. One of the 21 mBDA libraries on the MiSeq-1 chip (left). Comparison libraries using standard amplicon NGS (middle and right): two repeat libraries using a single MiSeq chip and a total of 16.1×10^6 reads (MiSeq-2 and MiSeq-2 repeat). The inferred VAF for homozygous wild-type (NA18537) alleles was notably higher than the inferred VAF from the libraries with blockers, and a heterozygous SNP outlier had a VAF of about 57%. **a,b**, The regions shaded in orange, green and grey indicate the homozygous variants, heterozygous variants and homozygous wild type, respectively. **c**, Inferred VAF as a function of read depth for each SNP allele. Traces are coloured according to the contaminant SNP genotype (purple, homozygous variant; green, heterozygous; grey, homozygous wild type). The VAFs converge to their final values at a read depth of about 250x; VAFs below our 0.019% LoD are indicated by the region shaded in grey. **d**, Summary of the variant-call accuracy using the 0.019% VAF LoD threshold described in Fig. 4e. All of the inferred VAFs from **a** are displayed in this beeswarm plot. There was a false-positive variant-call rate of 1.58% and a false-negative rate of 4.88% or 4.96%, depending on whether the variant allele was homozygous or heterozygous, respectively. **e**, Receiver-operator-characteristic plot for variant calls using the data in **c**. Setting the variant-call threshold at 0.04% VAF would increase specificity to 100% at the cost of reducing sensitivity to roughly 85%. The area under the receiver-operator-characteristic curve was 0.9968 for this set of samples and would probably be higher for samples with larger contaminant fractions. **f**, A random sample of 50,000 reads out of the mBDA 0.2% HeLa contaminant library (3.58×10^5 reads in total) gave essentially the same inferred VAFs as the full library. **g**, Analysis of the number reads needed for accurate VAF quantitation using mBDA libraries. Each point shows the mean coefficient of determination (R^2) values for the inferred VAFs of a read sample versus the full library at different read sample sizes ($n=15$ for each read sample size); the error bars show one standard deviation. At $\geq 40,000$ reads, the R^2 value reliably converged to above 0.99.

Contaminant identification based on NGS data. In the specific application to human cell-line contamination, the mBDA NGS data can be used to inform the specific identity of the contaminant. In this setting, a database of known possible contaminant genotypes is constructed and the mBDA NGS results are compared against the genotypes of each potential contaminant. For each database genotype j , we can compute the likelihood L_j of it generating the observed mBDA NGS libraries based on the following formula:

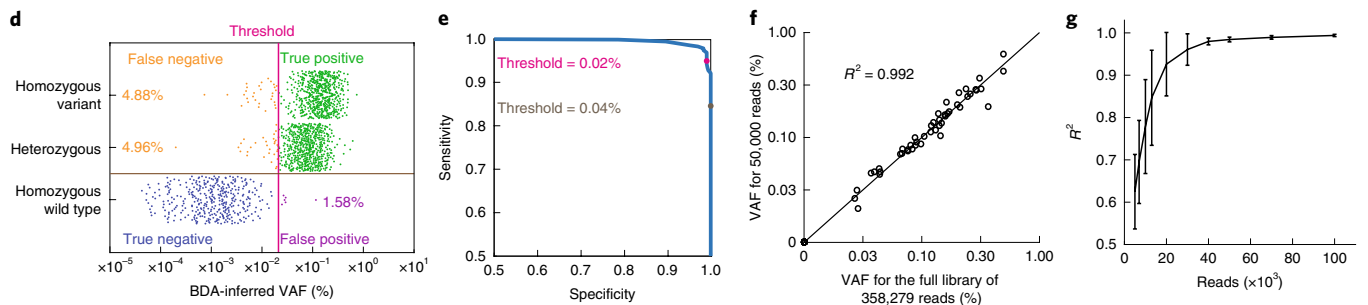
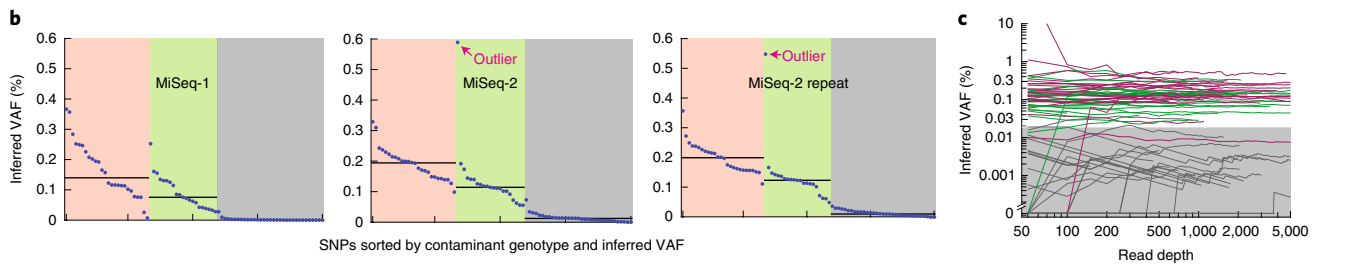
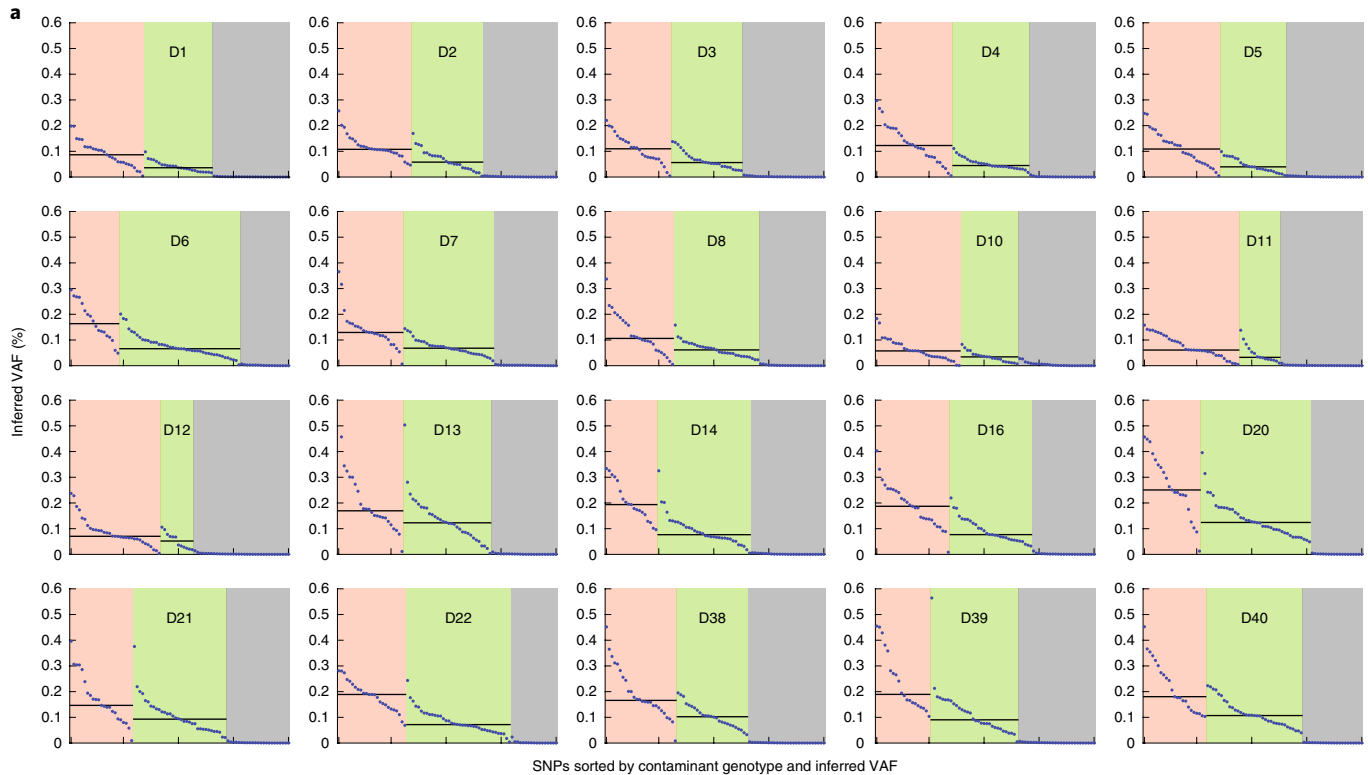
$$L_j = \prod_{i=1}^{80} p_{i,j} \quad (4)$$

where $p_{i,j}$ is the probability of generating the observed variant call at locus i based on a true variant state j . In this framework,

we consider only the presence or absence of a variant, rather than the VAF quantity, so $p_{i,j}$ is equal to the true-positive rate when the inferred VAF is above the 0.019% VAF LoD and comparison genotype i is either a homozygous variant or heterozygous. Thus, equation (4) can be rewritten as:

$$L_j = \Pr(\text{TP})^{\text{TP}} \times \Pr(\text{TN})^{\text{TN}} \times \Pr(\text{FP})^{\text{FP}} \times \Pr(\text{FN})^{\text{FN}} \quad (5)$$

where $\Pr(\text{TP})$ is the probability of a true positive based on Fig. 3d and TP is the number of true positives for genotype j , TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. Figure 4a shows graphical examples of the likelihood computation for two different genotypes; the correct comparison genotype has a much higher likelihood L than the incorrect genotype.



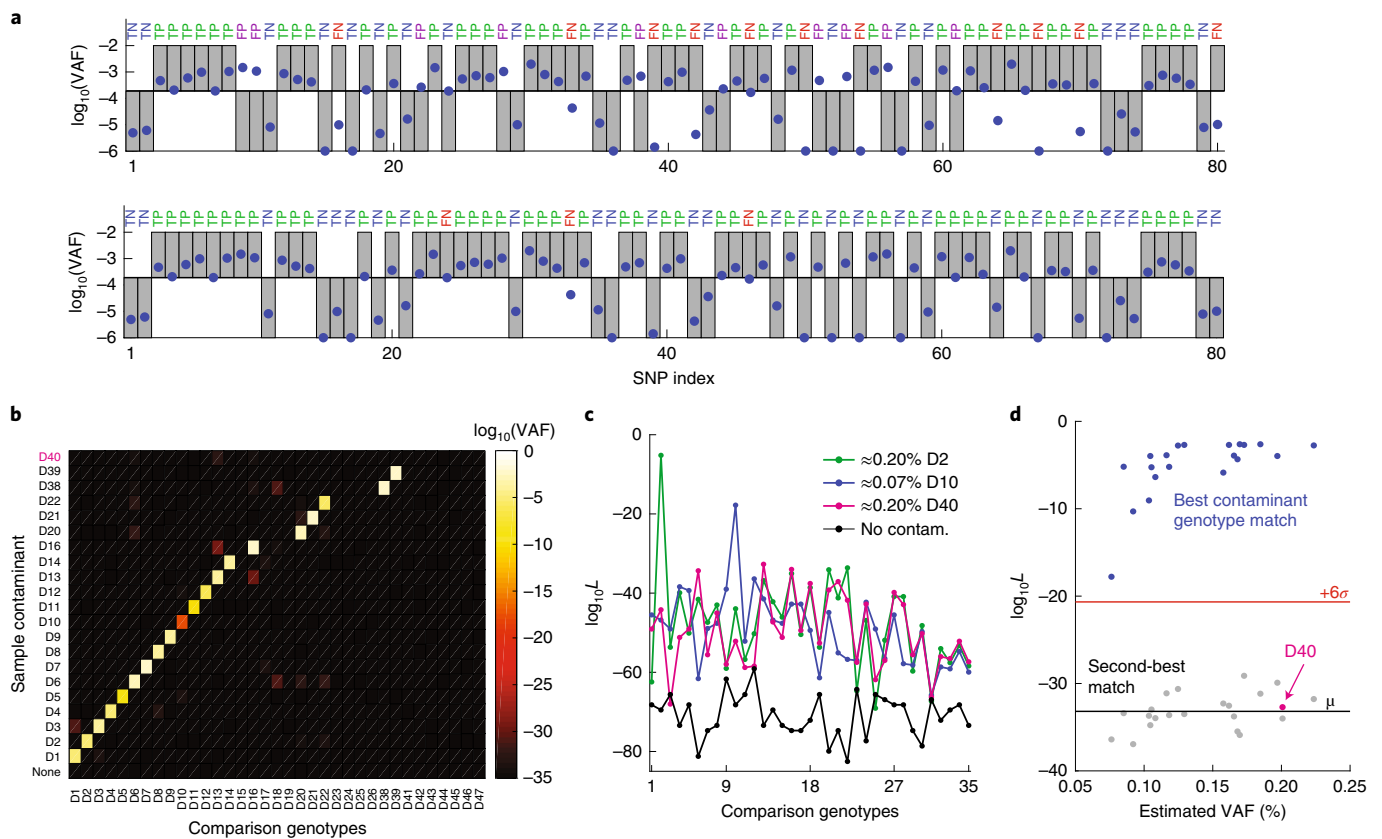


Fig. 4 | Determination of contaminant identity based on mBDA NGS data. a, Calculation of the contaminant likelihood. The sample contaminant is D1 and the comparison genotypes are D3 (top) and D1 (bottom). For each SNP, we decided a positive- or negative-variant-allele call based on whether the inferred VAF (blue dots) was greater than the threshold of 0.019%. For each potential contaminant, based on whether a variant allele existed for each SNP (grey boxes), we determined whether our mBDA-inferred variant would be a true positive (TP), true negative (TN), false positive (FP) or false negative (FN) relative to the potential contaminant. The overall likelihood L for the potential contaminant was calculated as the product of the likelihoods of all SNPs, with the probabilities of TP, TN, FP and FN set to 95.08, 98.42, 1.58 and 4.92%, respectively, based on Fig. 2d. The correct contaminant was expected to have a much higher value of L than incorrect contaminants. We calculated the $\log_{10}L$ values to be -33.4 and -5.2 for D3 and D1, respectively. **b**, Heatmap plot of the \log_{10} values for all pairwise combinations of all of the 21 samples tested (with a contamination fraction of 0.07–0.22%) versus 35 database genotypes. **c**, Representative distributions of $\log_{10}L$ for four different samples. D2, typical sample with a contaminant genotype that exists in the database. D10, sample with the weakest maximum $\log_{10}L$ value with a contaminant genotype that exists in the database. D40, sample with a contaminant genotype that is not included in the database. No contamin., sample of pure NA18537. **d**, Highest and second-highest $\log_{10}L$ values against the contamination fraction. The value of $\log_{10}L$ decreases for lower contamination fractions, consistent with expectations. However, even the D10 sample with 0.07% contamination is confidently identified at over six standard deviations above the mean second-highest $\log_{10}L$ value. The D40 sample, whose genotype was intentionally omitted from the database, can also be confidently identified as an unknown contaminant based on its highest $\log_{10}L$ value.

Figure 4b displays the \log -transformed value of L for every pairwise combination of the 21 mBDA NGS libraries in Fig. 3a and 35 different comparison genotypes. The genotype of contaminant D40 was intentionally withheld from the database to observe the effects of an unknown contaminant. In all cases other than D40, the correct genotype yielded the highest L value. In the case of contamination by D40, no comparison genotype produced a very high L but all L values were notably higher than the uncontaminated DNA sample (Fig. 4c). The L value of the correct contaminant genotype decreased for samples with very low fractions of contaminant (Fig. 4d); the D10 contaminant sample had the lowest contaminant fraction (0.07%) and corresponded to the lowest value of L for the correct genotype. However, based on the distribution of the second-highest values of L , even D10 could be confidently called with six standard deviations of confidence ($P < 10^{-9}$).

Cell-line contamination screening using qPCR. Quantitative PCR (qPCR) is by far the most commonly used method for detecting DNA markers due to its reliability, ease-of-use and short

turn-around time. Quantitative PCR can be effectively used to detect unique DNA sequences of a known contaminant. For example, the qPCR-based detection of mycoplasma contamination of human cell lines is well-established as a method for detecting inter-species contamination²⁴. However, there are many potential human cell-line contaminants. Studies report that the top-ten most common contaminants collectively only account for about 50% of all known contamination cases²⁵. To our knowledge, no qPCR methods have been reported that allow for the detection of arbitrary conspecific DNA contamination.

Here we show that mBDA is compatible with qPCR and can be used to build rapid and easy-to-use assays for detecting arbitrary conspecific cell or DNA contamination. In our mBDA qPCR implementation, we designed an mBDA primer-blocker set that specifically suppresses a set of SNP alleles on which the desired cell line is homozygous. When the set of selected SNPs is sufficiently large, the genotype of any contaminating DNA is likely to differ from the desired cell line in at least one of the SNPs in the panel. A double-stranded intercalating dye such as SybrGreen can be used

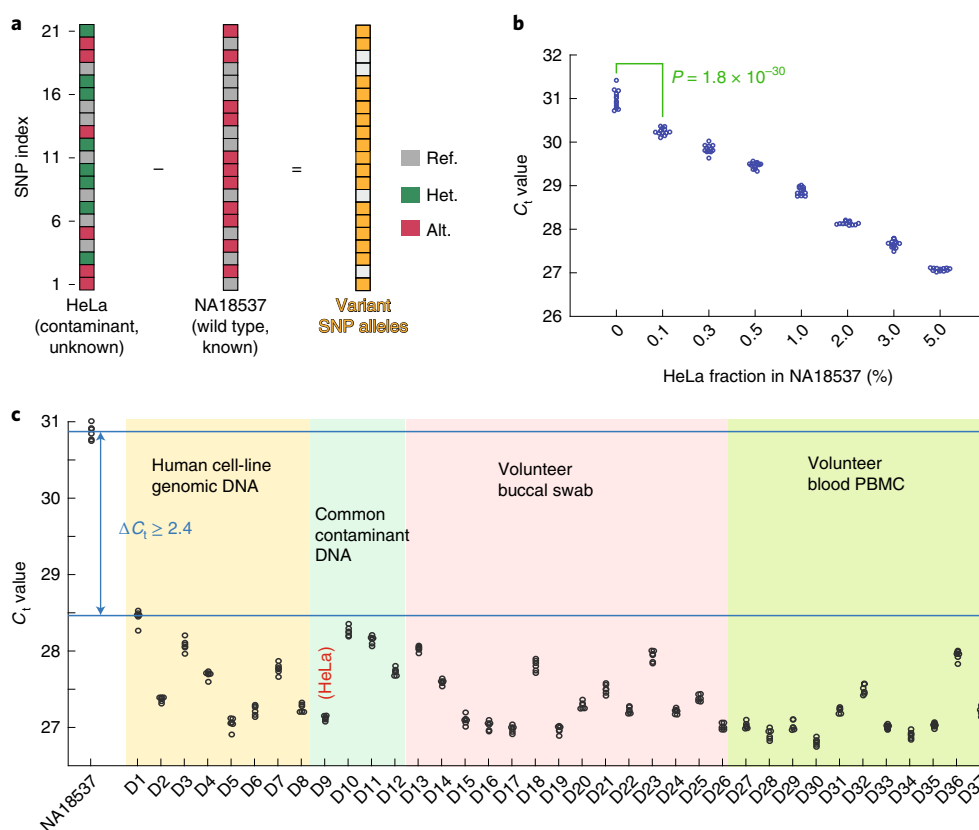


Fig. 5 | Detection of cell-line contamination using mBDA with qPCR. **a**, Given an intended human cell line (here, NA18537), our mBDA qPCR assay will detect potential contamination by any other human cells. The unknown contaminant (here, HeLa) is likely to differ in genotype in at least one SNP from the intended cell line. Variant SNP alleles from the contaminant will be preferentially amplified, resulting in lower C_i values when using a double-stranded DNA intercalating dye such as SybrGreen I. Here we selected a 21-plex subset of the 80-plex mBDA panel used earlier. Ref., reference; het., heterozygous; alt., alternate. **b**, Experimental qPCR C_i values of the 21-plex mBDA for 20 ng NA18537 with varying fractions of HeLa contamination. The beeswarm plot shows the observed C_i values for 12 independent reactions for each sample. Even the 0.1% HeLa contaminant could be confidently distinguished from the pure NA18537 gDNA sample. **c**, Summary of the qPCR results for the detection of 5% contaminant in NA18537 with 37 different human DNA contaminants (six replicates per sample, 20 ng DNA each). Because the different contaminants had different numbers and sets of variant alleles, there was variation in the C_i values of the contaminated samples. Based on the observed ΔC_i values at 5% contaminant, we extrapolated that a contaminant fraction of 1% should be detectable for all human DNA contaminants. PBMC, peripheral blood mononuclear cells.

to report the total quantity of PCR amplicons. In an ideal mBDA qPCR implementation with an infinite EE, an uncontaminated sample of cell-line DNA would give no amplification, and even an infinitesimal contamination could be detected and quantitated via the amplification cycle threshold (C_i).

Although the closed-tube nature of qPCR reactions minimizes the contamination risk, it precludes solid-phase reversible immobilization bead-based size selection steps that mitigate primer dimers and non-specific genomic amplification. Thus, we found that the full 80-plex panel developed for NGS exhibited notable primer dimers that generated a small qPCR C_i value even in the absence of DNA templates. We found that a reduction in the mBDA panel size to a 40-plex notably improved our qPCR LoD for detecting contaminants, and a reduction to a 21-plex resulted in an additional marginal improvement (Supplementary Section 7).

Figure 5a,b shows the qPCR detection of DNA samples with various fractions of HeLa contaminant in NA18537 using a 21-plex SNP assay. HeLa was selected as the contaminant because it is the most frequent source of human cell and DNA contamination, accounting for roughly 25% of all reported cases²⁵. Of the 21 SNP loci in the qPCR assay with NA18537 as the intended genotypes, HeLa had 17 variant alleles (Fig. 5a). The mean qPCR C_i values observed for uncontaminated NA18537 versus NA18537 with

0.1% HeLa were dramatically statistically different, with a P value of 1.8×10^{-30} .

We expected that the 21-plex qPCR assay would be able to detect arbitrary human mutations due to the low probability that a contaminant would have an identical genotype to the intended cell line in the 21 SNP loci. To experimentally support this hypothesis, we next ran the 21-plex mBDA qPCR assay on NA18537 with 5% contaminant from 37 different human DNA samples (Fig. 2c). The contaminants included DNA from common cell lines as well as DNA from volunteer individuals. All 5% contaminant samples could be confidently differentiated from the uncontaminated sample and the large ΔC_i difference suggested that all of the contaminants could be detected at 1% using qPCR.

Our simulations suggested that any arbitrary cell line would be homozygous in at least 21 of the 80 SNP loci that we built our mBDA NGS panel for (Supplementary Section 8). Of a 21-plex subset mBDA panel, our analysis and simulations suggested that there would be at least eight variant alleles out of the 21 SNP loci for an arbitrary human-cell contaminant, based on the population frequencies of each SNP allele reported in the 1000 Genomes database²². To help readers who work with cell lines construct their own qPCR panels to detect contamination in their cell lines, we have provided mBDA designs for both alleles of the 80 SNPs (Supplementary

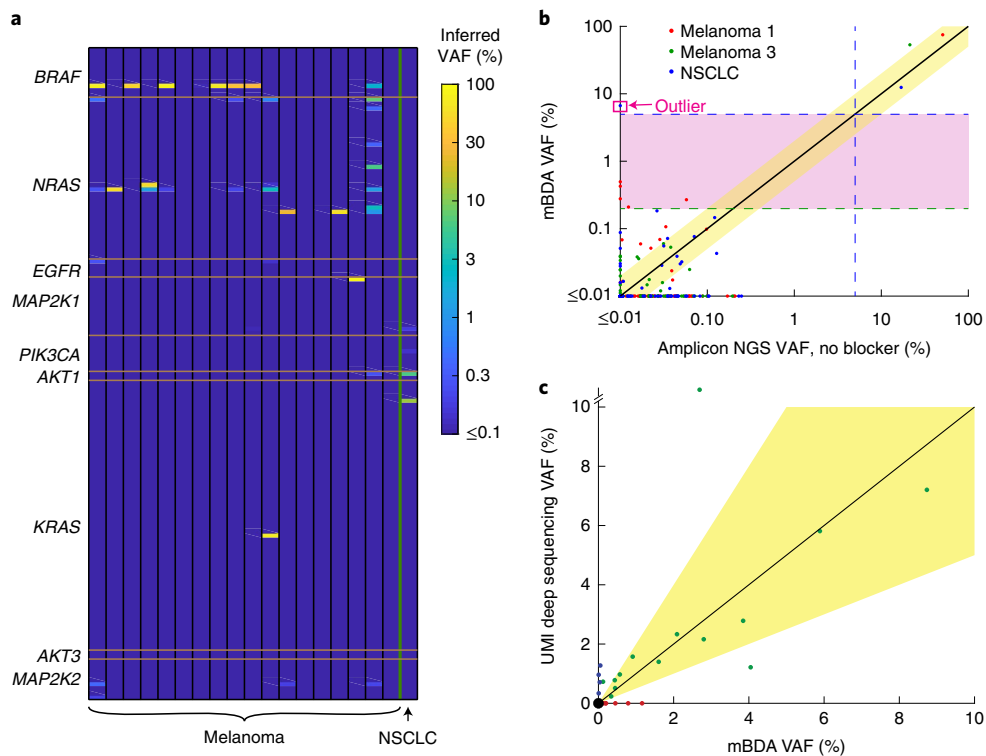


Fig. 6 | Validation of mBDA panels on clinical samples. **a**, Inferred VAFs for 18 melanoma tumour samples (fresh/frozen) and one NSCLC tumour sample ($n=19$ clinical tumour samples). This melanoma mBDA NGS panel covers nine genes and 145 mutations in the COSMIC database; the panel comprises a 16-plex mBDA reaction (see Supplementary Section 10 for contents). The samples were de-identified; 50 ng extracted DNA was used as input for each library. The number of NGS reads for each sample ranged from 45,844 to 89,718. **b**, Comparison of the inferred VAFs for mBDA and standard amplicon NGS. The blue dashed lines show the VAF LoD of 5% achievable by NGS without UMIs. The horizontal green dashed line shows 0.2% VAF LoD. The region shaded in pink shows the eight mutations identified by the mBDA panel at a VAF of between 0.2% and 5%. The boxed outlier corresponds to an *AKT1* mutation in the NSCLC sample with extremely low sequencing depth in the amplicon NGS panel (one mapped read). The sequencing depth for the *AKT1* amplicon in the melanoma samples was 3,058 and 4,788, indicating that this outlier may be specific to the NSCLC sample. **c**, Comparison of the inferred VAFs for mBDA and deep sequencing with UMIs for cfDNA samples from blood plasma from patients with Stage IV NSCLC. This mBDA panel comprises 31 amplicons across 14 genes; the quantities of input cfDNA ranged from 6.0 to 18.9 ng. Aliquots of these cfDNA samples were also analysed for tumour mutations through deep sequencing with UMIs following a previously published method⁴⁹. The region shaded in yellow indicates VAF agreement to within a factor of two. There is one *EGFR* exon-19 ELREA deletion outlier in which mBDA called 2.7% VAF and deep sequencing called 20.2% VAF. In another sample, the same *EGFR* exon-19 ELREA deletion was called as 8.7% and 7.2% VAF by mBDA and deep sequencing, respectively. One of the tested clinical cfDNA samples was excluded from this analysis because of a strong suspicion of sample mislabelling (sample R in Supplementary Data). Samples are colour-coded as per the legend in **b**.

Dataset). We envision that the genotype of the intended cell line, if unavailable in public databases, can be quickly and inexpensively obtained through Sanger sequencing, a small microarray or a small NGS library (without mBDA).

Although we expect that the primary interest of cell-line researchers is to detect arbitrary contaminants in known cell lines, it is also possible to use mBDA qPCR assays to detect known contaminants without genotype information regarding the specific intended cell line. In this scenario, we selected SNP loci in which the alternate allele has an extremely low frequency in the population and the contaminant is heterozygous. As a demonstration, we designed a three-plex mBDA assay for HeLa-specific SNPs and experimentally demonstrated that we could detect 1% HeLa contaminant in eight different cell lines with unknown genotypes (Supplementary Section 9).

Melanoma mBDA NGS panel reveals frequent heterogeneity in tumour tissue samples. Most qPCR-, Sanger- and NGS-based assays and panels for oncology targeted therapy selection^{26–28} have mutation-VAF sensitivities of between 1% and 5%, and are thus unable to identify low-VAF drug-resistance mutations arising from

trace subclones due to tumour heterogeneity. However, under the selective pressures of targeted therapy, tumour subclones with drug-resistance mutations can rapidly expand and cause treatment failure or cancer recurrence^{29–32}. Thus, the reliable detection of low-VAF drug-resistance mutations can inform personalized treatment selection, including the use of combination therapies, to improve patient outcomes.

To address this challenge, we next constructed a 16-plex mBDA NGS panel covering 145 commonly observed melanoma mutations across nine genes (Fig. 6a). The BDA enrichment regions were designed to cover the most frequently observed mutations in the nine genes, based on the COSMIC database³³. Because this cancer panel covers a range of different mutations, we arbitrarily selected one of the two DNA strands to be the target strand for blocker binding, unlike the case of SNP detection where we can intentionally target the strand with a larger mismatch thermodynamic penalty. We ran calibration NGS experiments using human gDNA (NA18537) spiked in with different quantities of synthetic gene blocks with a length of 490–500 nucleotides (nt) to measure the VAF LoD and EF values for the cancer mutations (see Supplementary Section 10 for reference sample preparation and EF calibration details). To analytically

Table 1 | Comparison of the performance of different NGS methods in profiling somatic mutations

	mBDA (this work)	SafeSeqS/OncoMine/AmpliSeq	CAPP-Seq/Guardant 360/FoundationOne Liquid	Standard NGS
VAF LoD (%)	0.1	0.1	0.1	5
Depth required	250x	25,000x	60,000x	500x
UMI	No	Yes	Yes	No
Library preparation time (h)	6	6	24	24
Reads per sample for a 100 kb panel ($\times 10^6$)	1.2	25	36	0.3
MiniSeq NGS cost per sample (US\$)	90	1,875; impossible	2,700; impossible	22.50
MiniSeq sample no. for full loading	16	0.80	0.56	66
NovaSeq NGS cost per sample (US\$)	6.50	136	195	1.63
NovaSeq sample no. for full loading	2,800	132	92	11,000

Only mBDA and standard NGS are feasible on the lower-throughput NGS platforms, such as the MiniSeq, and only standard NGS is incapable of detecting VAFs that are notably <5%. In contrast, mBDA is unsuitable for high-throughput NGS platforms, such as the NovaSeq, due to the large number of samples that must be pooled to leverage the economies of scale of the instruments. The NGS cost for MiniSeq assumes US\$1,500 for 20×10^6 reads on a 2×150 nt flow cell; the NGS cost for NovaSeq assumes US\$18,000 for 3.3×10^7 reads on a 52.2×150 nt flow cell. For mBDA, an enrichment region of 20 nt per 90 nt amplicon is assumed. SafeSeqS, OncoMine and AmpliSeq assume 140 nt amplicons on average to generate 100 nt of usable sequence. CAPP-Seq, Guardant 360 and FoundationOne Liquid assume 160 nt DNA input.

validate the accuracy of our mBDA NGS panel in the detection and quantitation of mutations with a low VAF, we performed comparison experiments on spike-in samples against the BioRad QX200 droplet digital PCR (Supplementary Section 11).

We applied the melanoma mBDA NGS panel to a total of 19 clinical fresh/frozen tumour samples purchased de-identified from a commercial supplier; 18 of these were melanoma tissue samples and one was a non-small-cell-lung-cancer (NSCLC) tissue sample. All samples had a tumour fraction of at least 75%, based on histological analysis. The called variants are summarized in Fig. 6b and Supplementary Section 10. A total of 7 of the 19 samples (37%) had mutations with a low VAF of between 0.2% and 5%. The 95% confidence interval of fresh/frozen tumour samples with low-VAF mutations is 19–58%, based on binomial distribution analysis. To confirm our findings, we performed droplet digital PCR comparison experiments for these fresh/frozen tissues as well as four formalin-fixed paraffin-embedded (FFPE) tissue samples (Supplementary Section 11).

As part of our design and bioinformatics interpretation process, we designed primers and blockers to avoid non-pathogenic SNPs with population allele frequencies above 1% using the 1000 Genomes and Kaviar databases. The high-VAF oncogene mutations that we observed in the tumour samples, such as BRAF-V600E and NRAS-Q61K, are not reported to be present in the population at notable frequencies, presumably due to the strong selective pressure against individuals with these mutations. Consequently, we believe these are likely to be somatic tumour mutations, although we cannot be sure without matched normal tissue or blood samples.

Interestingly, all samples with low-VAF subclonal mutations also had at least one high-VAF mutation (Table 1). A χ^2 analysis weakly suggests that the presence of low-VAF and high-VAF mutations are not statistically independent ($P=0.046$). In other words, the presence of low-VAF subclonal drug-resistance mutations seems to be higher in tumour samples with high-VAF clonal mutations.

Validation on clinical cfDNA samples. Cell-free DNA in peripheral blood represents a promising class of biomarkers for non-invasive tumour profiling and can be useful in cancer management not only for therapy selection in cases where a tumour biopsy is not convenient but also longitudinally for post-treatment monitoring. Detection of actionable mutations in cfDNA is also well-suited for mBDA, given that the mutation VAFs in cfDNA can be quite low (<1%) because of the wild-type cfDNA derived from healthy dying cells as a normal part of homeostasis.

Here we validate the effectiveness of mBDA NGS panels on clinical cfDNA samples from patients with Stage IV NSCLC. The mBDA panel covers 31 hotspots in 14 genes (*AKT1*, *ALK*, *BRAF*, *DDR2*, *EGFR*, *ERBB2*, *KRAS*, *MAP2K1*, *MET*, *NRAS*, *PIK3CA*, *PTEN*, *ROS1* and *TP53*) and was commercially developed by Nuprobe as the VarMap NSCLC panel. The called mutations and their VAF were compared against the VAF called by deep sequencing using UMIs. Across a total of 528 loci (12 samples \times 44 loci), we observed a 78% positive-concordance rate (14/18) and a 98.6% negative-concordance rate (503/510). Qualitative and quantitative concordant calls were made for mutations with a VAF as low as 0.23% in clinical samples (Supplementary Dataset).

Discordant calls were observed for mutations with a VAF of up to 1.3% for samples with very low input DNA quantities (6 ng); 6.0 ng cfDNA corresponds to about 1,800 haploid genomic copies and 1.3% VAF corresponds to 23.4 expected mutant molecules. The conversion yield of an NGS panel is the fraction of original DNA molecules at a locus that are represented in the final NGS library. For commercial hybrid-capture NGS panels, the median conversion yield is roughly 20%, so 4.6 of the 23.4 mutant DNA molecules theoretically present in the sample are expected to be represented in the NGS library. The number of actual mutant molecules in the NGS library follows a Poisson distribution and there is an 83.7% probability that a random sample will have three or more molecules. Consequently, some discordance can be expected even at the 1.3% VAF level, due to sampling.

Discussion

The mBDA technology presented here notably scales the multiplexing of BDA to 80-plex and demonstrates integration with NGS to allow broad detection and quantitation of rare DNA mutations with a VAF LoD of $\leq 0.1\%$. Although allele-enrichment methods have been extensively researched (for example, ICE-COLD PCR¹⁹, Boreal Genomics³⁴, PNA/LNA blocker PCR³⁵, nuclease-assisted minor-allele enrichment³⁶ and CUT-PCR³⁷), we are not aware of any other allele-enrichment technologies that have been successfully scaled and integrated with NGS to allow low-depth sequencing analysis of low-VAF mutations in settings of more than tenplex. Generally speaking, other allele-enrichment methods are sensitive to operational conditions (temperature, time and enzyme/DNA purity). This renders multiplexed panels challenging to design and optimize because variant alleles from multiple loci may enrich optimally at different conditions.

In this work, we first demonstrated mBDA using an 80-plex amplicon panel that covered approximately 8,000 nt in total and performed allele enrichment on about 1,600 nt.

Given our observed on-target rates of more than 80%, we do not think that primer dimers and non-specific amplification are at present the bottleneck for scaling up mBDA. Primer dimers are typically much shorter than on-target amplicons, and non-specific genomic amplification results in amplicons that are typically much longer than on-target amplicons. Consequently, the vast majority of primer dimers and non-specific amplicons can and are removed through solid-phase reversible immobilization bead-size selection. Commercial amplicon target-enrichment NGS panels (for example, AmpliSeq) include up to 24,000-plex PCR primers, so we expect that scaling up mBDA to over 1,000-plex in a single tube should be possible with some experimental optimization.

In its present design, the mBDA technology is primarily directed to the detection and quantitation of mutations and small insertions/deletions in hotspot regions. Thus, mBDA is well-suited for the detection of actionable cancer mutations in clinical guidelines, such as those of the National Comprehensive Cancer Network, as well as to personalized mutations found from initial whole-exome sequencing or whole-genome sequencing analysis of tumour biopsies³⁸. For standard single-base-replacement mutations, we are able to achieve mBDA success rates of well over 95%, with failures primarily due to extremely high G/C content or context sequences with highly repetitive DNA sequences that cause PCR mispriming. We observed a slightly worse performance in enriching single-base insertions and deletions from homopolymeric repeats of over 6 nt, but these are sequences that all PCR and NGS struggle with.

However, mBDA is less suitable for the discovery of new mutations across the entire exon regions of many genes and is unsuitable for the discovery of new structural variations (for example, chromosomal translocations). To comprehensively detect mutations in tumour-suppressor genes such as *BRCA1/BRCA2*, we would need to pursue a split-tube strategy, wherein the exons are tiled by different amplicons split across multiple tubes. Tube splitting is needed for most amplicon sequencing approaches for comprehensive exon coverage because mutations in the regions covered by primers cannot be detected in amplicon sequencing; different primers in tube 2 are needed to detect potential mutations in loci covered by the primers in tube 1.

For SNVs with VAFs between 0.03% and 3%, we showed that results from mBDA NGS could accurately quantitate sample VAF through a mathematical transformation assuming conserved EF values, with up to 95% accuracy within a factor of two. For low-VAF somatic mutations, quantitation accuracy is frequently limited by biological variability. Even adjacent tissue sections can have different VAFs and a wide range of biological factors (for example, exercise, time of day and bacterial/viral infection) affect the tumour-mutation VAF in cfDNA. However, mBDA suffers from less accurate VAF quantitation when the initial VAF and/or EF is very high. For example, a 10% VAF sample with $EF = 10,000$ would result in an mBDA library VRF of 99.9%, which cannot be accurately quantitated for the same reason that a VRF of 0.1% cannot be accurately quantitated. When accurate quantitation of mutations with a high VAF is needed, we recommend constructing a standard amplicon NGS library with low sequencing depth (for example, 250 \times) in addition to the mBDA library. The mBDA library will accurately quantitate low VAFs of 0.03–3% and the standard amplicon library will accurately quantitate VAFs of 3–97%.

We demonstrated mBDA using a panel of non-pathogenic human SNPs and showed that it could be used for the detection of human cell-line contamination in both NGS and qPCR settings. Cell lines serve as critical models for biomedical research, allowing the use of effective and inexpensive experiments to understand the function of cells and organisms as well as to establish feasibility for

potential therapeutics^{39–41}. However, contamination and misidentification of cell lines have become widespread problems that limit the reproducibility of experimental results and threaten the validity of published conclusions^{42–44}. Recent estimates suggest that up to 35% of cell lines in use today suffer from contamination and the findings of over 30,000 research publications may be compromised⁴⁵. With our mBDA qPCR assays, researchers will be able to detect low-level contamination with less than 1% conspecific DNA, increasing the rigour of scientific research based on cell lines.

As a proof-of-concept for clinical applications of mBDA-based NGS panels, we constructed a 16-plex melanoma mBDA NGS panel and applied it to 19 tumour tissue samples. We found that 37% of the samples (95% confidence interval of 19–58%) contained detectable subclonal mutations, including known drug-resistance mutations, in our panel at a VAF of between 0.2% and 5%. Furthermore, we expect that the fraction of patient tumour samples with low-VAF mutations would be higher given a broader panel covering more genes and loci. The increase in drug-resistance mutations in patients undergoing targeted therapies has generally been considered as a combination of an increase in de novo mutation(s) during treatment and the expansion of pre-existing subclones with the resistance mutation(s). Our results here statistically indicate that the latter mechanism may be applicable to a notable fraction of patients. Alternatively, reports of genomic heterogeneity of somatic mutations in the skin of healthy ageing individuals⁴⁶ suggest that the correlation between low-VAF and high-VAF mutations may be due to more complex mechanisms not yet fully understood.

In this work, we applied the mBDA NGS panel to fresh/frozen tumour tissue, rather than FFPE tissue, because of well-documented deamination and oxidation damage to FFPE DNA^{47,48}. With fresh/frozen tissue samples we can be more confident that the low-VAF mutations called are truly indicative of subclonal mutations due to tumour heterogeneity rather than artifacts from DNA damage.

In our preliminary experiments applying mBDA panels to melanoma FFPE samples, we observed systematic cytosine deamination damage resulting in C > T and G > A false positives at a VAF of up to 0.6%. Because we typically set our LoD reporting cutoff at double the highest observed false-positive VAF, this means that for FFPE samples, we cannot report C > T and G > A variants at a VAF below 1.2%. Other mutations, such as the BRAF-V600E T > A mutation, can still be reported at a VAF as low as 0.1%. However, because the extent of FFPE damage is highly dependent on the age and the storage/handling of the sample, it is not generalizable to determine the VAF LoDs across different FFPE samples. Thus, for FFPE samples, we expect to determine the LoD for different types of mutations within the same sample by designing BDA to other portions of the genome that are not expected to be mutated (for example, housekeeping genes such as *GAPDH*).

To date, all NGS panels that achieve a VAF LoD of 0.1% use UMIs with ultra-deep sequencing of a depth of more than 25,000 \times . This renders rare-mutation profiling feasible only on instruments with the highest throughput NGS, such as the NovaSeq (Table 1). However, the high capital expense (roughly US\$1,000,000 for a NovaSeq) and the large number of samples that need to be pooled to obtain economies of scale render high-throughput NGS instruments out of reach for most hospitals and reference laboratories. By enabling accurate and quantitative sequencing of low-VAF mutations using a depth of only 250 \times , our technology allows NGS instruments with a lower throughput (for example, a US\$50,000 MiniSeq) to analyse clinical samples for low-VAF mutations across many genes. The reduced sequencing reads required by mBDA also manifests as a notable reduction in the bioinformatics analysis time and data storage. Some other aspects of NGS analysis, most notably DNA extraction kits, are not reduced by mBDA, but these are currently less than US\$20 per sample and are not the dominant cost consideration for tumour or cfDNA sequencing. We envision that

facilitating the decentralized NGS testing of clinical samples for low-VAF somatic mutations will accelerate the adoption of precision medicine and lead to improvements in patient outcomes.

Methods

Primer and blocker oligonucleotides. Primers and blockers were purchased from Integrated DNA Technologies. All DNA oligos were purchased with standard desalting and LabReady formulation (100 μM in Tris–EDTA pH 8.0; Integrated DNA Technologies). The DNA stock solutions were diluted to 5 μM using 1 \times TE buffer (purchased from Sigma Aldrich as a 100 \times stock solution, cat. no. 574793). All primers and blockers can be stored at 4 $^{\circ}\text{C}$.

Cell-line and human DNA samples. All experiments were conducted in compliance with all of the relevant ethical regulations. Human gDNA samples (NA18562, NA18537, NA19223, NA12815, NA20507, NA18545, NA18572 and NA20502) were purchased from Coriell Biorepository and stored at -20°C .

HeLa cell-line DNA was ordered from New England Biolabs (cat. no. N4006S). K562, T24 and PC3 cells were obtained from collaborators and DNA was extracted using standard methods. Fresh/frozen cancer tissue samples were ordered from OriGene Technologies Inc.

Volunteer DNA samples. This research is classified as IRB Exempt under NIH Exemption 4. Buccal swabs samples were collected from volunteers from the Houston, TX area. Swabs were placed in 400 μl 1 \times PBS buffer in 1.5 ml tubes and incubated in a Multi-Therm instrument (Genesee) at 37 $^{\circ}\text{C}$ for 10 min. The swabs were then removed and the remaining solution was centrifuged at 12,000 r.p.m. for 5 min. The supernatant was discarded and 200 μl PBS buffer was added to resuspend the cell pellet, while vortexing vigorously for 1 min. Next, 20 μl proteinase K and 4 μl RNase A solution (100 mg ml $^{-1}$) were added and the suspension was vortexed for 15 s. Finally, the solution was extracted using a QIAamp DNA blood mini kit (Qiagen, cat. no. 51104) following the manufacturer's instructions.

Blood samples were purchased from ZenBio, Inc. from consented de-identified volunteers from the Chapel Hill, NC area. The blood samples were centrifuged at 500 r.p.m. for 15 min to separate the plasma, red blood cells and buffy coat. DNA was extracted from the buffy coat using a QIAamp DNA blood mini kit following the manufacturer's instructions.

NGS library preparation protocol for mBDA. For each library, the DNA sample was first mixed with the appropriate concentrations of primers and blockers, and then underwent 23 cycles of PCR using Phusion hot start flex DNA polymerase (New England Biolabs, cat. no. M0535L). The reaction volume was 50 μl and the thermocycling protocol included an initial 98 $^{\circ}\text{C}$ denaturation for 30 s, followed by 23 cycles of 98 $^{\circ}\text{C}$ for 10 s, 63 $^{\circ}\text{C}$ for 5 min and 72 $^{\circ}\text{C}$ for 2 min. The amplicon products were then purified from the reaction mixture using a column-based DNA clean and concentrator kit (Zymo Research, cat. no. D4005).

Next, we appended sequencing adaptors to the BDA amplicons via two cycles of PCR using adaptor primers. The forward and reverse adaptor primers were each at a final reaction concentration of 15 nM in a reaction volume of 50 μl . The same PCR thermocycling protocol as described above was applied for two cycles, and the amplicons were subsequently re-purified using a DNA clean and concentrator kit. We then optionally quantitated the adaptor-appended amplicons using qPCR (BioRad iQaq SYBR Green supermix, cat. no. 1725122; 95 $^{\circ}\text{C}$ for 3 min, followed by 40 cycles of 95 $^{\circ}\text{C}$ for 10 s and 60 $^{\circ}\text{C}$ for 30 s). The observed C_i value was determined using the BioRad software.

We then perform index PCR using the Illumina Nextera sequences as index primers. Here, we used Phusion hot start flex DNA polymerase (New England Biolabs, cat. no. M0535L) and added 5 μl of the index primers into the final 50 μl reaction, and followed the following thermocycling protocol: 98 $^{\circ}\text{C}$ for 30 s, followed by 11 cycles of 98 $^{\circ}\text{C}$ for 10 s, 63 $^{\circ}\text{C}$ for 1 min and 72 $^{\circ}\text{C}$ for 1 min. If qPCR was used in the previous step, we would instead perform index PCR for $C_i + 4$ cycles.

Before the next step of amplicon size selection, we incubated AMPure beads at room temperature for 30 min. The prepared AMPure beads (Beckman Coulter, cat. no. A63881) were then used to size select the NGS library; ratios of 0.7 \times and 0.3 \times AMPure beads were used (that is, 0.7 \times means 50 μl library and 35 μl AMPure beads). After size selection, we used a Qubit dsDNA HS kit (Thermo Fisher, cat. no. Q32851) to determine the concentration of the library and a Bioanalyzer DNA 1000 kit (Agilent, cat. no. 5067-1504) was used to quantify the length of the amplicons to ensure the library quality. Different libraries were pooled to a final concentration of 4 nM and loaded onto an Illumina MiSeq instrument following standard protocols with 5–10% PhiX.

NGS library preparation protocol for standard multiplex PCR amplicons without blockers. The protocol used was mostly the same as the 'NGS library preparation protocol for mBDA' protocol, except that the number of PCR cycles in the first step was reduced from 23 to 13 cycles. In addition, for historical reasons, during size selection we used 0.6 \times and 0.3 \times AMPure beads (instead of

0.7 \times and 0.3 \times); we do not expect this difference to make a notable difference in our conclusions.

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

Data availability

The main data supporting the results in this study are available within the paper and its Supplementary Information. All requests for raw and analysed data will be reviewed by the Legal Department of Rice University to verify whether the request is subject to any intellectual property or confidentiality constraints. Requests for patient-related data not included in the paper will not be considered. Data can be shared for non-commercial research purposes via a material transfer agreement.

Code availability

All requests for code will be reviewed by the Legal Department of Rice University to verify whether the request is subject to any intellectual property or confidentiality constraints. Custom code can be shared for non-commercial research purposes via a material transfer agreement.

Received: 18 September 2019; Accepted: 9 March 2021;
Published online: 03 May 2021

References

- Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
- Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **6**, 287–303 (2013).
- Schirmer, M. et al. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43**, e37 (2015).
- Schirmer, M., DaAmore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinform.* **17**, 125 (2016).
- Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **108**, 9530–9535 (2011).
- Newman, A. M. et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* **34**, 547–555 (2016).
- Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109**, 14508–14513 (2012).
- Schwaederle, M. C. et al. Utility of genomic assessment of blood-derived circulating tumor DNA (ctDNA) in patients with advanced lung adenocarcinoma. *Clin. Cancer Res.* **23**, 5101–5111 (2017).
- Ma, C. X. et al. Neratinib efficacy and circulating tumor DNA detection of *HER2* mutations in *HER2* nonamplified metastatic breast cancer. *Clin. Cancer Res.* **23**, 5687–5695 (2017).
- Hata, A. N. et al. Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat. Med.* **22**, 262–269 (2016).
- Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).
- Zitvogel, L., Ma, Y., Raoult, D., Kroemer, G. & Gajewski, T. F. The microbiome in cancer immunotherapy: diagnostic tools and therapeutic strategies. *Science* **359**, 1366–1370 (2018).
- Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
- Fisher, R. A., Gollan, B. & Helaine, S. Persistent bacterial infections and persister cells. *Nat. Rev. Microbiol.* **15**, 453–464 (2017).
- Merker, M. et al. Whole genome sequencing reveals complex evolution patterns of multidrug-resistant *Mycobacterium tuberculosis* Beijing strains in patients. *PLoS ONE* **8**, e82551 (2013).
- Hindson, B. J. et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83**, 8604–8610 (2011).
- Watanabe, M. et al. Ultra-sensitive detection of the pretreatment EGFR T790M mutation in non-small cell lung cancer patients with an EGFR-activating mutation using droplet digital PCR. *Clin. Cancer Res.* **21**, 3552–3560 (2015).
- Didelot, A. et al. Competitive allele specific TaqMan PCR for KRAS, BRAF and EGFR mutation detection in clinical formalin fixed paraffin embedded samples. *Exp. Mol. Pathol.* **92**, 275–280 (2012).
- Milbury, C. A., Li, J. & Makrigiorgos, G. M. Ice-COLD-PCR enables rapid amplification and robust enrichment for low-abundance unknown DNA mutations. *Nucleic Acids Res.* **39**, e2 (2010).

20. Wu, L. R., Chen, S. X., Wu, Y., Patel, A. A. & Zhang, D. Y. Multiplexed enrichment of rare DNA variants via sequence-selective and temperature-robust amplification. *Nat. Biomed. Eng.* **1**, 714–723 (2017).
21. *International HapMap Project* (National Human Genome Research Institute, 2012); <https://www.genome.gov/10001688/international-hapmap-project/>
22. *1000 Genomes Project* (The International Genome Sample Resource, 2021); <http://www.internationalgenome.org/>
23. Potapov, V. & Ong, J. L. Examining sources of error in PCR by single-molecule sequencing. *PLoS ONE* **12**, e0169774 (2017).
24. Volokhov, D. V., Graham, L. J., Brorson, K. A. & Chizhikov, V. E. Mycoplasma testing of cell substrates and biologics: review of alternative non-microbiological techniques. *Mol. Cell. Probes.* **25**, 69–77 (2011).
25. Capes-Davis, A. et al. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int. J. Cancer* **127**, 1–8 (2010).
26. Qu, K. et al. Detection of BRAF V600 mutations in metastatic melanoma: comparison of the Cobas 4800 and Sanger sequencing assays. *J. Mol. Diagn.* **15**, 790–795 (2013).
27. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
28. Beltran, H. et al. Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity. *Eur. Urol.* **63**, 920–926 (2013).
29. Russo, M. et al. Tumor heterogeneity and lesion-specific response to targeted therapy in colorectal cancer. *Cancer Discov.* **6**, 147–153 (2016).
30. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628 (2017).
31. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).
32. Romano, G. et al. A preexisting rare *PIK3CA*^{E545K} subpopulation confers clinical resistance to MEK plus CDK4/6 inhibition in *NRAS* melanoma and is dependent on S6K1 signaling. *Cancer Discov.* **8**, 556–567 (2018).
33. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2018).
34. Pel, J. et al. Nonlinear electrophoretic response yields a unique parameter for separation of biomolecules. *Proc. Natl Acad. Sci. USA* **106**, 14796–14801 (2009).
35. Kang, S. et al. Targeted sequencing with enrichment PCR: a novel diagnostic method for the detection of EGFR mutations. *Oncotarget* **6**, 13742–13749 (2015).
36. Song, C. et al. Elimination of unaltered DNA in mixed clinical samples via nuclease-assisted minor-allele enrichment. *Nucleic Acids Res.* **44**, e146 (2016).
37. Lee, S. H. et al. CUT-PCR: CRISPR-mediated, ultrasensitive detection of target DNA using PCR. *Oncogene* **36**, 6823–6829 (2017).
38. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
39. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
40. Sharma, S. V., Haber, D. A. & Settleman, J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat. Rev. Cancer* **10**, 241–253 (2010).
41. Kikuchi, T. et al. Human iPS cell-derived dopaminergic neurons function in a primate Parkinson's disease model. *Nature* **548**, 592–596 (2017).
42. Schweppe, R. E. et al. Deoxyribonucleic acid profiling analysis of 40 human thyroid cancer cell lines reveals cross-contamination resulting in cell line redundancy and misidentification. *J. Clin. Endocrinol. Metab.* **93**, 4331–4341 (2008).
43. Freedman, L. P. et al. Reproducibility: changing the policies and culture of cell line authentication. *Nat. Methods* **12**, 493–497 (2015).
44. Huang, Y., Liu, Y., Zheng, C. & Shen, C. Investigation of cross-contamination and misidentification of 278 widely used tumor cell lines. *PLoS ONE* **12**, e0170384 (2017).
45. Horbach, S. P. & Halfman, W. The ghosts of HeLa: how cell line misidentification contaminates the scientific literature. *PLoS ONE* **12**, e0186281 (2017).
46. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
47. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin. Chem.* **61**, 64–71 (2015).
48. Chen, G., Mosier, S., Gocke, C. D., Lin, M. T. & Eshleman, J. R. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Mol. Diagn. Ther.* **18**, 587–593 (2014).
49. Goldberg, S. B. et al. Early assessment of lung cancer immunotherapy response via circulating tumor DNA. *Clin. Cancer Res.* **24**, 1872–1880 (2018).

Acknowledgements

This work was supported by NIH grant nos R01CA203964 and R01CA233364, and CPRIT grant no. RP180147 to D.Y.Z. We thank J. Nie for proofreading assistance and G. Bao for providing access to his BioRad QX200 digital droplet PCR instrument. We thank Nuprobe for providing early access VarMap NSCLC kits for cfDNA testing.

Author contributions

P.S. and D.Y.Z. conceived the project. P.S., S.X.C. and Y.H.Y. performed the mBDA sequence design for the cell-line contamination panels. L.Y.C. and P.D. performed the mBDA sequence design for the cancer panels. A.A.P. provided clinical cfDNA samples and performed the comparison deep sequencing experiments. P.S. and Y.H.Y. performed experiments and analysed qPCR data. P.S. and A.P. performed NGS experiments. P.S., S.X.C. and D.Y.Z. analysed the NGS data. P.S. and D.Y.Z. wrote the manuscript with input from all authors.

Competing interests

There are patents pending on the BDA (patent number EP3146080B1) and mBDA (patent number WO2019164885A1) methods used in this work. P.S., S.X.C., L.Y.C. and P.D. declare competing interests in the form of consulting for Nuprobe USA. A.A.P. declares a competing interest in the form of consulting for Nuprobe USA as well as consulting for and equity ownership in Binary Genomics. D.Y.Z. declares a competing interest in the form of consulting for and equity ownership in Nuprobe and Torus Biosystems as well as consulting for Avenge Bio.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41551-021-00713-0>.

Correspondence and requests for materials should be addressed to D.Y.Z.

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

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

© The Author(s), under exclusive licence to Springer Nature Limited 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

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

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data were collected and analysed by using Illumina and Biorad commercial software for qPCR and NGS experiments.

Data analysis

Custom Python and Matlab scripts for analyzing NGS FASTQ files. The codes can be shared for non-commercial research purposes via a material transfer agreement.

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The main data supporting the results in this study are available within the paper and its Supplementary Information. All requests for raw and analysed data will be reviewed by the Legal Department of Rice University to verify whether the request is subject to any intellectual property or confidentiality constraints. Requests for patient-related data not included in the paper will not be considered. Data can be shared for non-commercial research purposes via a material transfer agreement.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size	No sample-size calculation was carried out during experiment design.
Data exclusions	One clinical cfDNA sample was excluded on strong suspicion of sample mislabelling.
Replication	Replicates of the of initial analytical-validation experiments were performed, as summarized in the paper and its Supplementary Information.
Randomization	No experiments were randomized.
Blinding	The experiments were not blinded.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	No inclusion criteria were established for the SNP studies. Melanoma tumour samples were purchased, with the inclusion criteria that fresh/frozen tissue slides must contain at least 50% of tumour.
Recruitment	Consented volunteers from the Houston area contributed de-identified blood samples for the SNP experiments. Other samples were purchased commercially or provided by clinical collaborators in de-identified format.
Ethics oversight	This research is classified as IRB-exempt under NIH Exemption 4.

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