https://doi.org/10.1093/nargab/lqab125

An improved molecular inversion probe based targeted sequencing approach for low variant allele frequency

Tamir Biezuner[1,†], Yardena Brilon[1,†], Asaf Ben Arye[2], Barak Oron[1], Aditee Kadam[1], Adi Danin[1], Nili Furer[1], Mark D. Minden[3], Dennis Dong Hwan Kim[3], Shiran Shapira[4], Nadir Arber[4], John Dick[5], Paaladinesh Thavendiranathan[6], Yoni Moskovitz[1], Nathali Kaushansky[1], Noa Chapal-Ilani[1] and Liran I. Shlush[1,7,8,*]

[1]Department of Immunology, Weizmann Institute of Science, Rehovot 761001, Israel, [2]Department of Statistics and Operations Research, Tel Aviv University, Ramat Aviv, Israel, [3]Princess Margaret Cancer Centre, University Health Network (UHN), Department of Medical Oncology & Hematology, Toronto, ON, Canada, [4]Sourasky Medical Center Tel Aviv, Israel, [5]Princess Margaret Cancer Centre, University Health Network (UHN), Department of Molecular Genetics, Toronto, ON, Canada, [6]Department of Medicine, Division of Cardiology, Ted Rogers Program in Cardiotoxicity Prevention, Peter Munk Cardiac Center, Toronto General Hospital, University Health Network, University of Toronto, Toronto, ON, Canada, [7]Division of Hematology, Rambam Healthcare Campus, Haifa, Israel and [8]Molecular Hematology Clinic Maccabi Healthcare Services, Tel Aviv, Israel

Received October 03, 2021; Revised November 25, 2021; Editorial Decision December 07, 2021; Accepted January 25, 2022

ABSTRACT

Deep targeted sequencing technologies are still not widely used in clinical practice due to the complexity of the methods and their cost. The Molecular Inversion Probes (MIP) technology is cost effective and scalable in the number of targets, however, suffers from low overall performance especially in GC rich regions. In order to improve the MIP performance, we sequenced a large cohort of healthy individuals ($n = 4417$), with a panel of 616 MIPs, at high depth in duplicates. To improve the previous state-of-the-art statistical model for low variant allele frequency, we selected 4635 potentially positive variants and validated them using amplicon sequencing. Using machine learning prediction tools, we significantly improved precision of 10–56.25% ($P < 0.0004$) to detect variants with VAF > 0.005. We further developed biochemically modified MIP protocol and improved its turn-around-time to ∼4 h. Our new biochemistry significantly improved uniformity, GC-Rich regions coverage, and enabled 95% on target reads in a large MIP panel of 8349 genomic targets. Overall, we demonstrate an enhancement of the MIP targeted sequencing approach in both detection of low frequency variants and in other key parameters, paving its way to become an ultrafast cost-effective research and clinical diagnostic tool.

INTRODUCTION

The development of next-generation sequencing (NGS) approaches has revolutionized molecular biology research as they can generate large volumes of sequencing data per run, however it has yet to be widely implemented into clinical practice. While complete omics approaches (whole genome/transcriptome/epigenome) provide opportunity for novel discoveries, they are still not cost-effective and therefore are not routinely used as diagnostic tools. To democratize NGS to a large number of samples and applications in a cost- and time-effective manner, several targeted enrichment approaches have been developed. Furthermore, deep sequencing aimed at identifying low variant allele frequency (VAF) mutations is usually based on targeted sequencing approaches.

With the growing demand for high performance and cost-effective targeted sequencing technologies it is generally required to choose between scalability (both for number of samples and number of targets) and cost. Currently, there are no targeted sequencing approaches that are both scalable, cost effective, simple and fast. Hybrid capture has high-performance but is still costly and time consuming (1). On the other hand, amplicon sequencing is simple and

[*] To whom correspondence should be addressed. Tel: +972 8 934 4441; Email: liran.shlush@weizmann.ac.il

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

cost effective but is not scalable for large number of targets. In a previous study (2), we analyzed several hundreds of DNA samples for low allele frequency mutations in age-related clonal hematopoiesis (ARCH) related positions using a probe capture approach. Since probe capture is effective but costly, we sought to scale up this type of analysis to a larger cohort at a cost-effective approach.

To that end, we have implemented the Molecular Inversion Probe (MIP) (3,4) technology which enables targeting multiple genomic regions and generating a sequencing library in economical, one pot reaction. Although MIP technology can potentially be fully automated and scalable, its main downsides are its low performance (i.e. uniformity (1,4), reduced coverage at GC-rich (5) regions). Another drawback of the MIP technology is the lack of an accurate noise model, an essential tool for low VAF analysis.

The library preparation step of any targeted sequencing approach has a unique issue of background error signatures which correlate with the specific chemistry and various steps of the protocol (6). Therefore, one needs to comprehensively understand the intrinsic background noise of the technology and to generate a noise model to determine if suspected variants are real (7). The state-of-the-art in MIP low VAF analysis is the algorithm published by Acuna-Hidalgo *et al.* (8). In their extensive study, ∼2000 samples were analyzed across ∼230 MIP targets. While this study introduces a new statistical approach to call low VAF variants based on a Poisson noise model it has several caveats such as the lack of extensive validation and the use of technical duplicates which were separated at the final step of the MIP protocol while true technical duplicates were not used. These drawbacks leave background noise model of the MIP protocol without a cross platform validation, and uncertainty regarding its accuracy.

In this study, we address the MIP drawbacks by (A) analyzing and modeling the MIP protocol noise in depth. This was accomplished by processing 4417 samples in duplicate using the MIP protocol, validating our findings with amplicon sequencing, and using machine learning algorithms for MIP low VAF calling; and (B) modifying the current MIP biochemistry to improve poor performance and noise properties, and testing the novel improved MIP chemistry (iMIP) on a large panel of 8349 MIPs.

MATERIALS AND METHODS

Biological resources

DNA samples were received from the Princess Margaret cancer center, University Health Network (UHN), Canada) (under UHN IRB protocol 01-0573); the WizeAging project (Weizmann Institute, Israel) (under WIS IRB protocol 283-1); the Cancer prevention clinic (Sorasky Medical Center, Israel), Denver. All the relevant ethical regulations were followed. Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and protocols approved by the relevant ethics committees. Sample donors are considered healthy without known ARCH defining mutations in their clinical records. Per reaction a total DNA of 50–500 ng/ul was used.

MIP targeted sequencing probe design

Molecular inversion probes (MIP) capture probes were designed using MIPgen (3) to capture ARCH related targets (Supplementary Table S1) (9,10) or a genotyping panel (Supplementary Table S2). Backbone, and primer sequences were adopted from previous studies (4). MIPs were ordered either as single strand MIPs (prepared as in Hiatt *et al.* (4)) or as an oligo mix (LCsciences, prepared as in Shen *et al.* (11))

Multiplex MIP capture protocol

One μl DNA template was added to a hybridization mix together with a MIP pool (final concentration of 0.05 pM per probe) in 1x Ampligase buffer (Epicentre). Mix was incubated in a thermal cycler at 98°C for 3 min, followed by 85°C for 30 min, 60°C for 60 min and 56°C for one or two overnight incubation. Product was mixed with (final concentration in brackets): dNTPs (Larova, 15 pM), Betaine (375 mM, Sigma-Aldrich,), NAD+ (1 mM, New England Biolabs), additional Ampligase buffer (0.5×), Ampligase (total of 1.25 U, Epicentre) and Phusion HF (0.16 U, New England Biolabs). Mixture was incubated at 56°C for 60 min followed by 72°C for 20 min. Enzymatic digestion of linear probes was performed by adding Exonuclease I (4 U, New England Biolabs) and Exonuclease III (25 U, New England Biolabs). Mixture was incubated at 37°C for 2 h, followed by 80°C for 20 min. Final product was amplified using iProof HF Master Mix (Biorad). Samples were pooled and concentrated using AMPure XP beads (Beckman Coulter) at 1.3× volumetric concentration, size-selected (190–370 bp) using Blue Pippin (Sage scientific), and sequenced in either NextSeq or Novaseq6000 (Illumina) 2 × 151 bp paired-end run using custom primers as described in the past (4). In total, we sequenced 4417 healthy individual DNA samples and processed and sequenced every DNA sample twice as true technical duplicate using the above MIP protocol.

Improved MIP (iMIP) capture protocol

One microliter DNA template was added to a hybridization mix together with a MIP pool (final concentration of 0.04 pM per probe) in 0.85× Ampligase buffer. Mix was incubated in a thermal cycler at 98°C for 3 min, followed by 85°C for 30 min, 60°C for 60 min and 56°C for 60 min. Product was mixed with (final concentration in brackets): dNTPs (14 pM), Betaine (375 mM), NAD+ (1 mM), additional Ampligase buffer (0.5×), Ampligase (total of 1.25U) and Q5 High-Fidelity DNA Polymerase (0.4 U, New England Biolabs). All the product of the hybridization was incubated at 56°C for 5 min followed by 72°C for 5 min. Enzymatic digestion of linear probes was performed by adding Exonuclease I (8U) and Exonuclease III (50 U). Mixture was incubated at 37°C for 10 min, followed by 80°C for 20 min. Final product was amplified using NEBNext Ultra II Q5 Master Mix (New England Biolabs). Samples were pooled and concentrated using AMPure XP beads at 0.75× volumetric concentration and sequenced as abovementioned described.

Amplicon sequencing for suspected variants detected in MIP protocol

Selected MIP probes were ordered as amplicon primers to enable target amplification using two-step amplicon sequencing. After collecting all potential variants (see below), the amplifying MIPs were sorted by the number of mutations in the cohort they will capture (highest first). MIPs were then converted to corresponding amplicons: to this end, the ligation arm was converted by 'reverse complement'. 5′ tail addition and index primers were as previously described (12). All selected amplicon primers were applied to all DNA samples in the experiment, generating a majority of sequencing data with no expected mutations at any sampled genomic region. This further allowed for per position true/false positive statistical validation. Selected primers were mixed in pools of ≤6 primer pairs/mix at a concentration of 2.5 uM per primer. First PCR reaction was performed by mixing NEBNext Ultra II Q5 Master Mix, 1 ul DNA template, and primer mix (0.5 uM). PCR program: 98°C activation for 30 s, followed by five steps of: denaturation at 98°C, annealing at 60°C and extension at 65°C, than 25 steps of: denaturation at 98°C, annealing and extension at 65°C. Final extension was at 65°C for 5 min. Reaction was diluted 1:1000 and second PCR (barcoding PCR) was at the same composition and protocol as the first PCR besides the reduction of the two steps from 25 to 12 cycles. Reactions were pooled at equal volumes and purified by AMPure XP beads at 0.7× volumetric concentration, size-selected (265–400 bp) using Blue Pippin and sequenced in Novaseq6000 2 × 151 bp paired-end run.

Data preprocessing and variant calling

Paired-end 2 × 151 bp sequencing data were converted to fastq format. Reads were merged using BBmerge v38.62 (13) with default parameters, followed by trimming of the ligation and extension arm using Cutadapt v2.10 (14). Unique Molecular Identifiers (UMI) were trimmed and assigned to each read header. Processed reads were aligned using BWA-MEM (15) to a custom reference genome, comprised of the MIP ARCH panel sequences ±150 bp extracted from broad HG19 [https://gatk.broadinstitute.org/hc/en-us/articles/360035890711-GRCh37-hg19-b37-humanG1Kv37-Human-Reference-Discrepancies#b37]. Aligned files were sorted, converted to BAM (SAMTools V1.9 (16), followed by Indel realignment using AddOrReplaceReadGroups (Picard tools) and later IndelRealigner (GATK v.3.7 (17)). Variant calling was done using mpileup for the single nucleotide variant (SNVs), and Varscan2 v2.3.9 (18) and Platypus v0.8.1 (19) for indels. Variants were annotated using ANNOVAR (20).

Statistical analysis of SNVs for MIPs and amplicon

The depth for reference calls and all possible variants of all positions was retrieved from the mpileup files. Only positions with depth >100 were included. To estimate background error rate at each position first we calculated the total read depth across all samples (DEPTH_SUM) and the alternate supporting reads (ALT_READS_SUM) (Supplementary Tables S3A, S3B). Next, the number of alternate reads in a sample (n) and the total depth for the sample in that position (N) were analyzed followed by the calculate of $m = ALT\_READS\_SUM - n$ and $M = DEPTH\_SUM - N$. For MIPs this was done separately on each technical duplicate. To test whether a specific VAF is significantly different from the background error rate we approximated the distribution of the variant using Poisson distribution and used Poisson exact test on each variant estimation (stats R package), and corrected for multiple hypothesis testing with Benjamini–Hochberg (BH) (21) test per *P*-value to get a BH score (Supplementary Table S4).

Calculating expected number of duplicate and duplication ratio

To utilize the information from the large number of samples we sequenced with the MIP panel ($N = 4417$), and that fact they were all had technical duplicates we added another layer of data dealing with the duplicate's reproducibility. Accordingly, mplieup files of the technical duplicates were merged to define consensus positions that have depth >100 in both duplicates. Each variant was defined as singleton if identified in one of the technical duplicates or as a duplicate if found in both. Next mplieup files of all sample IDs were merged and the number of singleton (single_n) and duplicates (dup_n) in the entire dataset was calculated. The same counting was also performed only on variants with VAF ≥0.006 to define single_cutoff and dup_n_cutoff. The expected number duplicates for each variant was calculated $exp\_dups = \frac{single\_n\_cutoff^2}{total\_sample\_ids}$ and the same for the duplicate ratio (dup_ratio) $dup\_ratio = \frac{dup\_n\_cutoff}{exp\_dups}$ (Supplementary Table S5).

Amplicon sequencing validation

In order to understand the MIP noise model we compared MIP sequencing to amplicon sequencing. The targets for amplicon sequencing were chosen based on putative true variants identified by the Poisson exact test. We focused on variants known to play a role in ARCH (Supplementary Table S2) and selected variants with BH1 and BH2 <0.002 to be validated by amplicon sequencing (Supplementary Table S6). To build the noise model of the amplicon sequencing approach we extended this experiment by targeting all samples in the experiment with all participating primers (see methods). We performed this validation in two iterations: The first iteration was composed of 84 DNA templates, and 48 amplicons covering 7930 bp. The second iteration was composed of 125 DNA templates, and 48 amplicons covering 7114 bp (Supplementary Table S6).

Calculating background error rates

For the calculation of background error rate, the mplieup files were filtered for variants with VAF <0.05, depth >100. Background errors were calculated as the number of alternate reads over all sequenced bases in the same position across the entire panel. We evaluated error rates for MIP amplicon and iMIP.

Refining low VAF detection in MIP sequencing

As the background noise of MIP was significantly higher than amplicon we used amplicon calling as true positives. We defined true variants in the amplicon sequencing based on the Poisson exact test ($P = 0$, depth > 100, VAF > 0.005), which identified $N = 42$ true variants. We than called SNVs in the MIP data by calculating Poisson exact test P values for both duplicates. We transformed the data to fit machine learning prediction algorithms (Supplementary Table S7). Next we applied various machine learning algorithms and chose to continue with SVM and the vanilladot Kernel (caret library R 4.0.4) to calculate sensitivity, specific and precision of the SVM predictions (Supplementary Figure S1).

Comparing MIP and iMIP performance

To be able compare the MIP and iMIP protocols we selected samples that had similar depth distributions in the original FATSQ files based on Kolomogorov–Smirnov p value (Supplementary Figure S2), MIP $N = 535$ and iMIP $N = 905$ samples. To evaluate the number of MIPs that were covered sufficiently across samples we compared the amount of targets which received above 100 reads in at least one sample, these MIPs were defined as working MIPs. Uniformity was calculated by the $\frac{\%\ \text{MIPs with depth} > (0.2{*}\text{mean depth})}{panel\ size}$. On-target rate was measured by the $\frac{\%\ \text{Mapped reads}}{total\ reads}$.

Defining GC rich targets

The MIP target sequence was retrieved, and the GC content was evaluated using gc5Base table from UCSC table browser. We defined GC rich regions as regions with GC content >55%. From all working MIPs we identified GC rich MIPs and grouped by genes.

Genotyping panel

To test the ability of iMIP to capture large number of targeted sequences we have used MIPgen to design a large panel of 8349 probes which capture SNPs. Such panel can be used for demultiplexing human samples from pools of samples. Once we discovered that a small fraction of our MIPs captured large proportion of reads, and that many MIPs did not perform optimally we have chosen from the original panel (Supplementary Table S2) a set of 4409 MIPs and sequenced with it 104 samples with minimum depth of 10e6 reads (Supplementary Table S2).

RESULTS

In the current study we aimed to improve the performance of the MIP based targeted sequencing approach both in calling low VAF variants and its uniformity and coverage. Using the MIP protocol we sequenced 4417 samples in duplicates using the ARCH panel. This panel is composed of 707 MIP probes targeting 70 134 genomic bases, of which, 616 probes were used for the analysis (working MIPs see methods).

Improving MIP noise model

The current noise model used for low VAF calling after MIP targeted sequencing is generally based on a Poisson exact test and correction for multiple hypothesis (8). Furthermore, previous methods for error correction were applied for UMI deduplication to minimize noise; however, we could not use UMI collapsing as the majority of read families in the current study have a size of <5 reads per family/group (which is the standard cutoff for consensus sequence) (22). The reason for the low number of families with >5 reads per family was the low total number of reads we allocated each sample in the current study. We aimed at detecting low VAF variants in a cost effective manner thus we intentionally had lower coverage than needed for the use of UMIs. Altogether, we concluded that new methods for error correction under the MIP targeted sequencing protocol without necessarily taking UMIs into account are needed.

To this, we compared the background error rate of amplicon and MIP sequencing. Amplicon sequencing yielded significantly reduced error rate in all possible single nucleotide variants (SNV) alternations (Figure 1A). We noticed a bimodal noise distribution in C > A in the MIP protocol in all MIP experiments (Supplementary Table S3A, S3B, Supplementary Figure S3), ruling out the chance for a batch effect. This could be explained by DNA damage introduced during the library preparation process as was suggested in the past (23). The high background error rate produced by the MIP protocol suggests that the current state of the art statistical noise reduction tools for MIP might produce substantial false positive rates. Furthermore, the lower background error rate of the amplicon protocol suggests that the statistical noise detection could be improved by training a model on variants with higher probability of being true as they were validated by amplicon sequencing. Accordingly, we defined true variants using strict statistical cutoff on the amplicon sequencing data and identified 42 true variants (Supplementary Table S6).

To evaluate the performance of the current state of the art statistical noise reduction algorithm, we applied it on our MIP data and compared it to the true variants extracted from the amplicon sequencing (Supplementary Table S7). The outcome of this calculation yielded a specificity ($\frac{TN}{TN+FP}$) of was 99.74%, sensitivity ($\frac{TP}{TP+FN}$) of 80.95%, and precision ($\frac{TP}{TP+FP}$) of 10% (Figure 1B). To improve the precision of the current method we used machine learning algorithms which took into account only the parameters used in the past (VAF, Depth and Poisson exact test P values of the duplicates). Although this approach improved precision (50%) sensitivity was significantly lower (16.67%) ($P = 0.004$). Next we tested the hypothesis that adding information on the number of samples sequenced, duplicate ratio and other parameters extracted from the large dataset we created might improve our prediction model (Supplementary Table S7). We eventually used an SVM model which yielded the following results: specificity of 99.98%, sensitivity of 81.81%, and significantly higher precision of 56.25% ($P = 1.4\text{E}{-}5$) (Figure 1B). Altogether, the model we developed significantly reduced the number of false positive variants (Figure 1B).

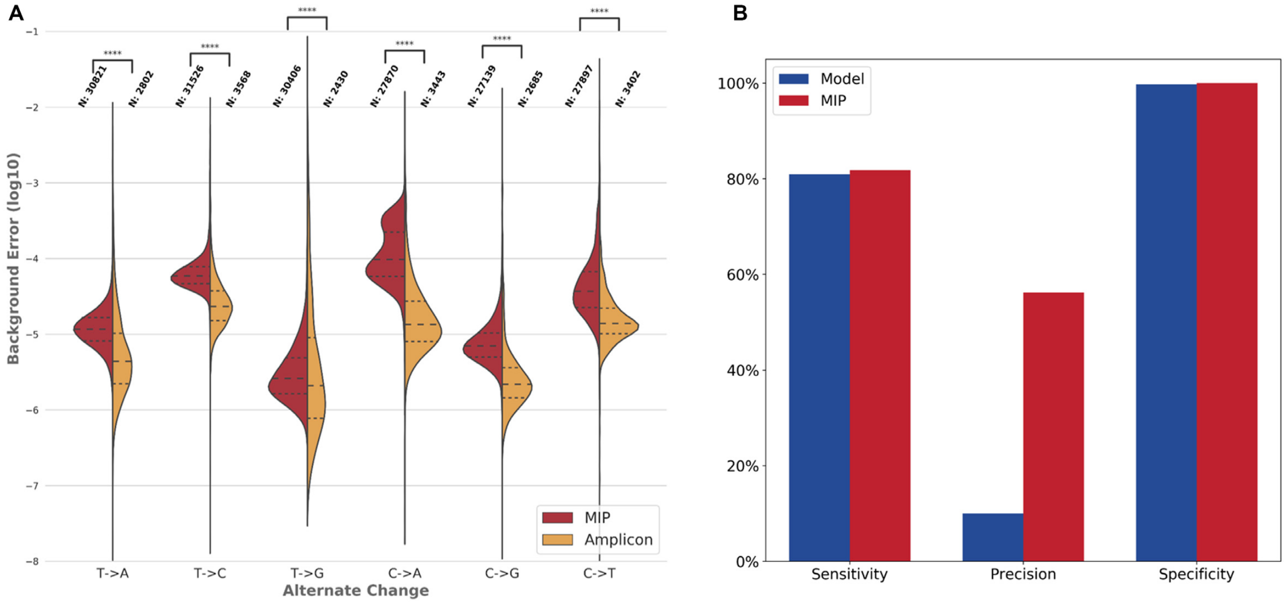

Figure 1. Increased background error rate in the MIP protocol results in high false positive rate which can be improved by machine learning algorithms. (**A**) Distribution of per base background error rate (log 10) of each possible alteration comparing the molecular inversion probe (MIP) protocol (red) and Amplicon sequencing protocol (yellow). Mann–Whitney–Wilcoxon test two-sided with Bonferroni correction ns: $0.05 < P \leq 1$, *: $0.01 < P \leq 0.05$, **: $0.001 < P \leq 0.01$, ***: $0.0001 < P \leq 0.001$, ****: $P \leq 0.0001$. (**B**) Variants from the MIP protocol were validated by amplicon sequencing and true positives were defined based on the results of the amplicon sequencing. Performance (sensitivity precision and specificity) was calculated for the state of the art Poisson distribution error suppression method (blue) and for a machine learning variant caller (MIP) trained on our entire MIP dataset (red). The precision of the machine learning variant caller (MIP) to detect variants with variant allele frequency (VAF) >0.005 was significantly better, Fischer exact test $P < 0.0001$.

Refining the biochemistry of the MIP protocol to improve performance and to reduce noise

While we were able to reduce the false positive rate of the MIP protocol (Figure 1), further improvements of known caveats of this method are still needed, namely poor performance properties such as on-target rate and uniformity (see methods). To this end, we tested several enzymes in each of the MIP protocol steps to recalibrate the protocol. Furthermore, with the possibility of potential use of this technology for large-scale screening studies for early cancer detection in clinical laboratories, we fine-tuned the protocol's timing to under 4 h (end to end). We then analyzed new 1569 samples using the same MIP ARCH panel mentioned above using the improved MIP protocol (termed: iMIP, see methods). Our results demonstrated significantly lower background error rate in the iMIP protocol versus the previous MIP protocol for all possible alterations, except for T > C (Figure 2A). Furthermore, the iMIP protocol had a significant lower background error rate compared to amplicon sequencing in T > G and C > A transversions, while in other alterations amplicon sequencing was still superior (Figure 2A). Of note the iMIP protocol had less small families (<5) and more large families (>5) (Supplementary Figure S4A). Same background calculation was performed on indels using the variant calling algorithms Varscan and Platypus (Supplementary Figure S5).

To study the effect of our iMIP protocol on the panel performance we compared the median number of MIPs that work (see methods) for both the MIP and iMIP protocols and demonstrated a significant increase in the median MIPs that work in the iMIP protocol (609 versus 558 respectively $P < 0.00001$) (Figure 2B). The iMIP protocol further demonstrated a significant improvement in the on-target rate (Figure 2C) and in the panel uniformity (Figure 2D). Another downside of the MIP protocol we noticed was the relatively low correlation of VAFs between the technical duplicates in the MIP protocol ($R^2 = 0.68$). The introduction of the iMIP protocol significantly improved the correlation between duplicates ($R^2 = 0.71$, $P < 0.00001$) (Supplementary Figure S4B).

Next, we aimed to improve the uniformity and on target rate specifically in the GC-rich regions, as it was reported in the past that MIP protocols perform poorly in such regions. In line with that our data demonstrates significantly lower uniformity and median depth in GC rich regions (Supplementary Figure S6B). As expected, many of the MIPs that got poor coverage in the MIP protocol and better coverage in the iMIP protocol exhibited high GC rich content (Supplementary Figure S6A). Furthermore we had barely any coverage in important GC rich regions such as the gene CEBPA and others. To resolve these issues, we specifically modified the MIP protocol and created the iMIP. Indeed when we compared the coverage across GC rich regions it was significantly higher in the iMIP protocol for all regions beside MIPs in the gene *SETBP1* (Figure 3A). Overall uniformity was also significantly higher in the iMIP protocol (Figure 3B). Specifically in the GC rich region of CEBPA which is known to be a challenging region across various NGS technologies (24) we significantly improved the coverage (Figure 3C).

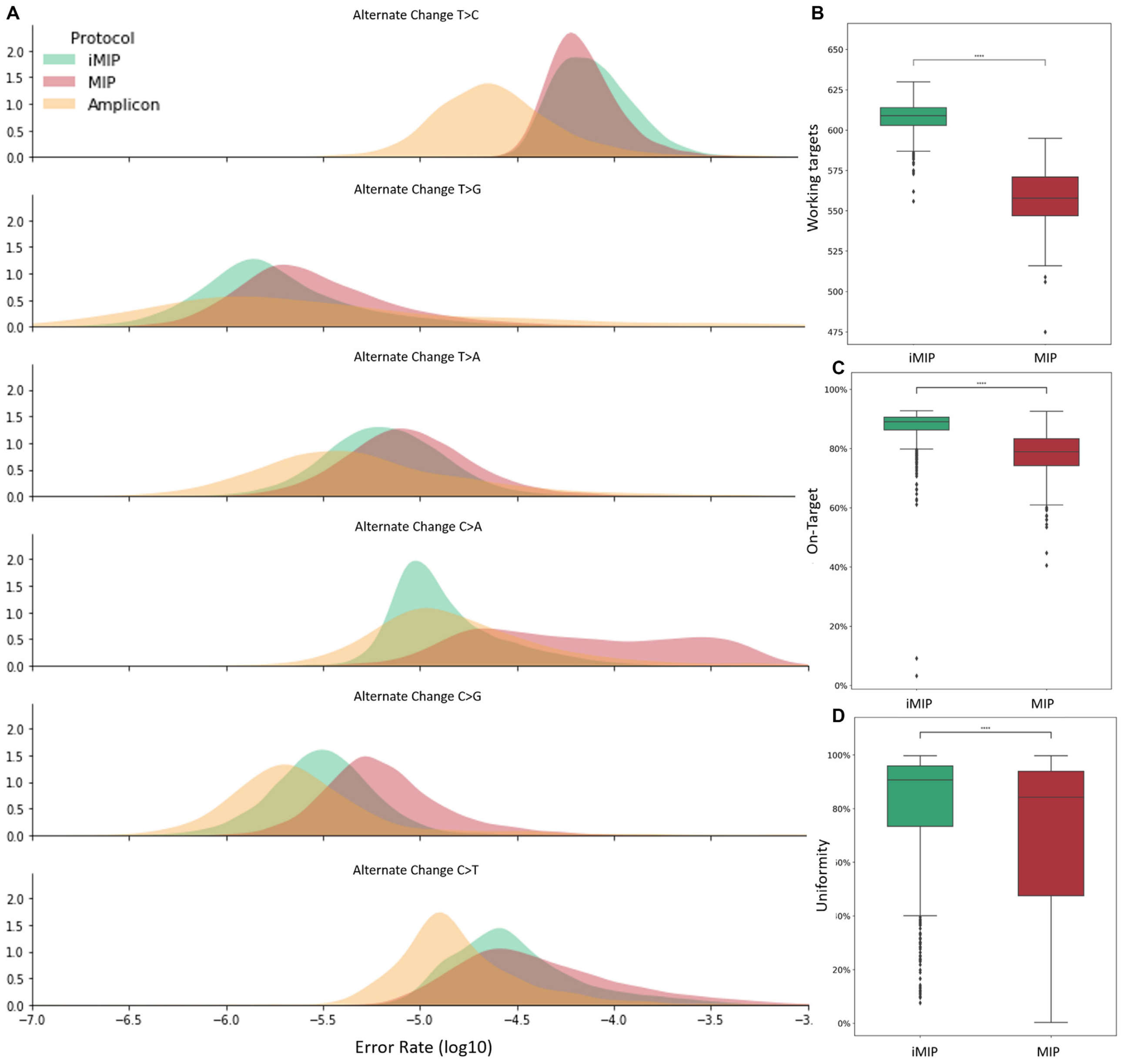

Figure 2. A novel improved MIP (iMIP) protocol has reduced background error rate and improved sequencing quality attributes. (**A**) Background error rate was calculated for iMIP (green) MIP (red) and amplicon (yellow). Comparisons of error background rates are presented in supplementary table S3B. (**B**) Number of MIP targets that worked across the selected samples between MIP and iMIP, Mann−Whitney−Wilcoxon test two-sided with Bonferroni correction *P*_val < 10∧-217(****). (**C**) On target rate across the selected samples, Mann−Whitney−Wilcoxon test two-sided with Bonferroni correction *P*_val < 10∧-131(****). (**D**) Uniformity of MIP and iMIP across the selected samples, Mann–Whitney–Wilcoxon test two-sided with Bonferroni correction *P*_val < 10∧-11(****).

Performance of the iMIP protocol on a large panel of 8349 targets

In order to further extend our understanding of the iMIP performance for larger MIP panels we have tested the same protocol with a genotyping panel we designed (Supplementary Table S2) which contains 8349 MIPs. Our initial results demonstrated that the majority of the samples with more than one million reads had on average 95% reads on target (Supplementary Figure S7A). However, when we compared the uniformity of the genotyping panel to that of the ARCH panel we faced a significantly lower uniformity (Figure 4C, first two boxplots). In order to better understand this low uniformity we analyzed the MIP properties of the mapped data and observed that a significant low number of MIPs took over a large proportion of the mapped reads (Supplementary Figure S7B). By back tracing the origin of these MIPs we noticed that some of these MIPs' ligation or extension arms have higher number of copies in the reference genome (Figure 4A). Notably, there are two problematic groups of MIPs with high copy number: the first one is when both ligation and extension arms have between 1 and 100 copies (Figure 4A, green), and the second one when both

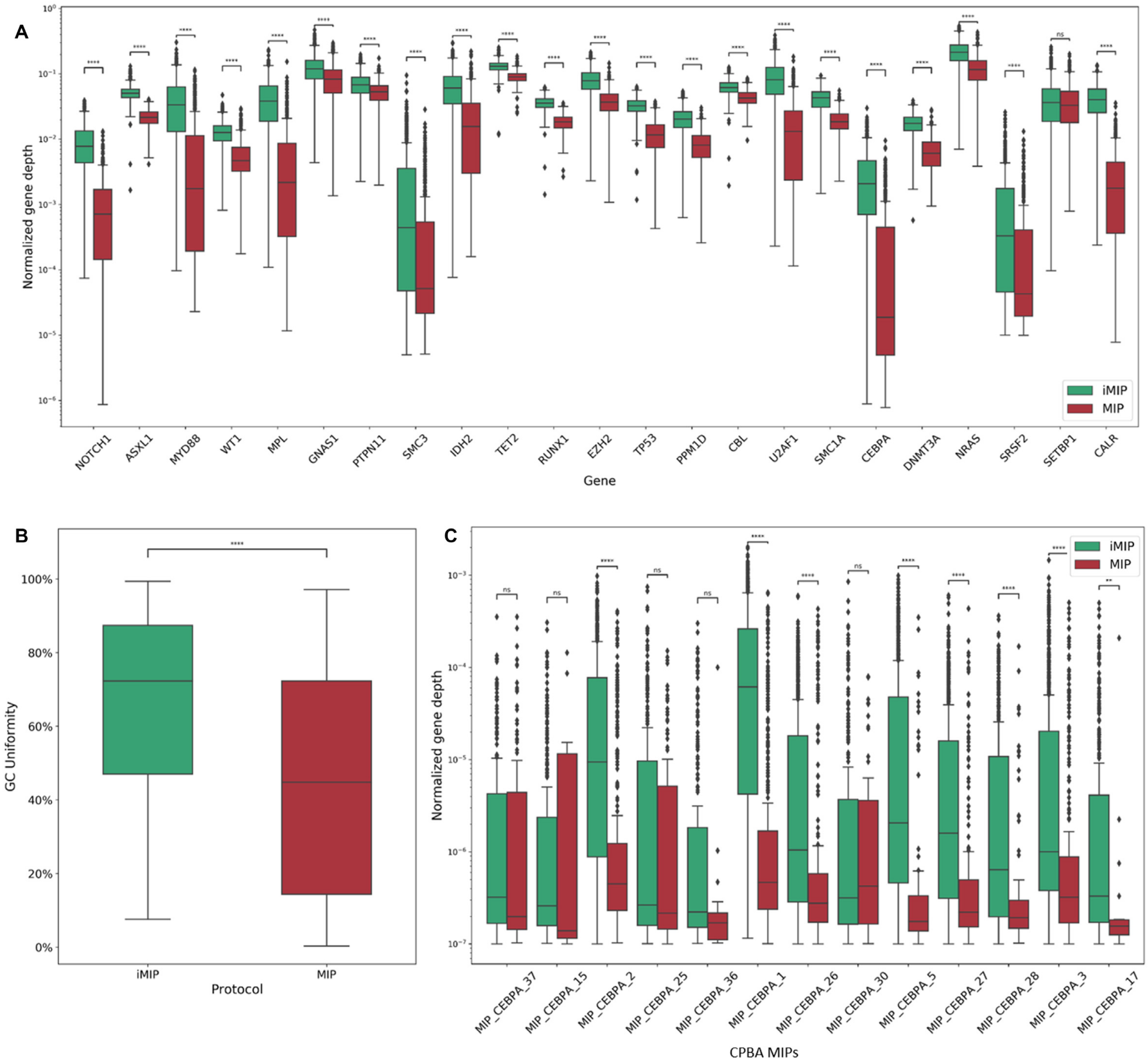

Figure 3. The iMIP protocol has better coverage and uniformity across GC-rich regions. (**A**) MIP ($n = 535$ samples) vs iMIP ($n = 905$ samples) comparison of GC-rich genes coverage (GC-rich targets have higher than 55% GC content). Targets which are part of each gene were included, the data is normalized by: sum (targets depth)/ number of targets/original fastq reads *100. Other than SETBP1 all p values were significant (****: $P \leq 0.001$). Mann–Whitney–Wilcoxon test two-sided with Bonferroni correction. Note: the values are in log scale, and for visualization, zero values were omitted. (**B**) Uniformity between MIP and iMIP across GC-rich targets P_val < 10∧-15(****) Mann–Whitney–Wilcoxon test two-sided with Bonferroni correction. (**C**) Improved coverage of the iMIP protocol across CEBPA (known to be difficult to capture), depth was normalized in the same way as (A). Note: the values are in log scale, and for visualization, zero values were omitted.

arms have above 100 copies (Figure 4A, light red). These two groups comprise of a low fraction in the panel (Figure 4A, left bar plot), but capture a significant fraction of the reads (Figure 4A, right bar plot and Figure 4B). This data points on the importance of the arm copy number parameter on the MIP panel performance. As the recommendations regarding arm copy number filtering are not clear, we analyzed the uniformity across the different copy number groups and found that the best uniformity was achieved when choosing MIPs with copy number of one in at least one of the arms. To validate this hypothesis and to improve the performance of our genotyping panel, we have generated a reduced genotyping panel, that contained only MIPs with copy number of one in at least one of the arms. MIPs that demonstrated low coverage were also omitted from the reduced genotyping panel. We sequenced 104 samples with the reduced genotyping panel and reached a median uniformity of 80.3% (Figure 4C) and median 50× coverage of 89.6% (Figure 4D). Our results, demonstrate the ability of the iMIP protocol to target thousands of genomic targets and extended our knowledge as for how to design better panels.

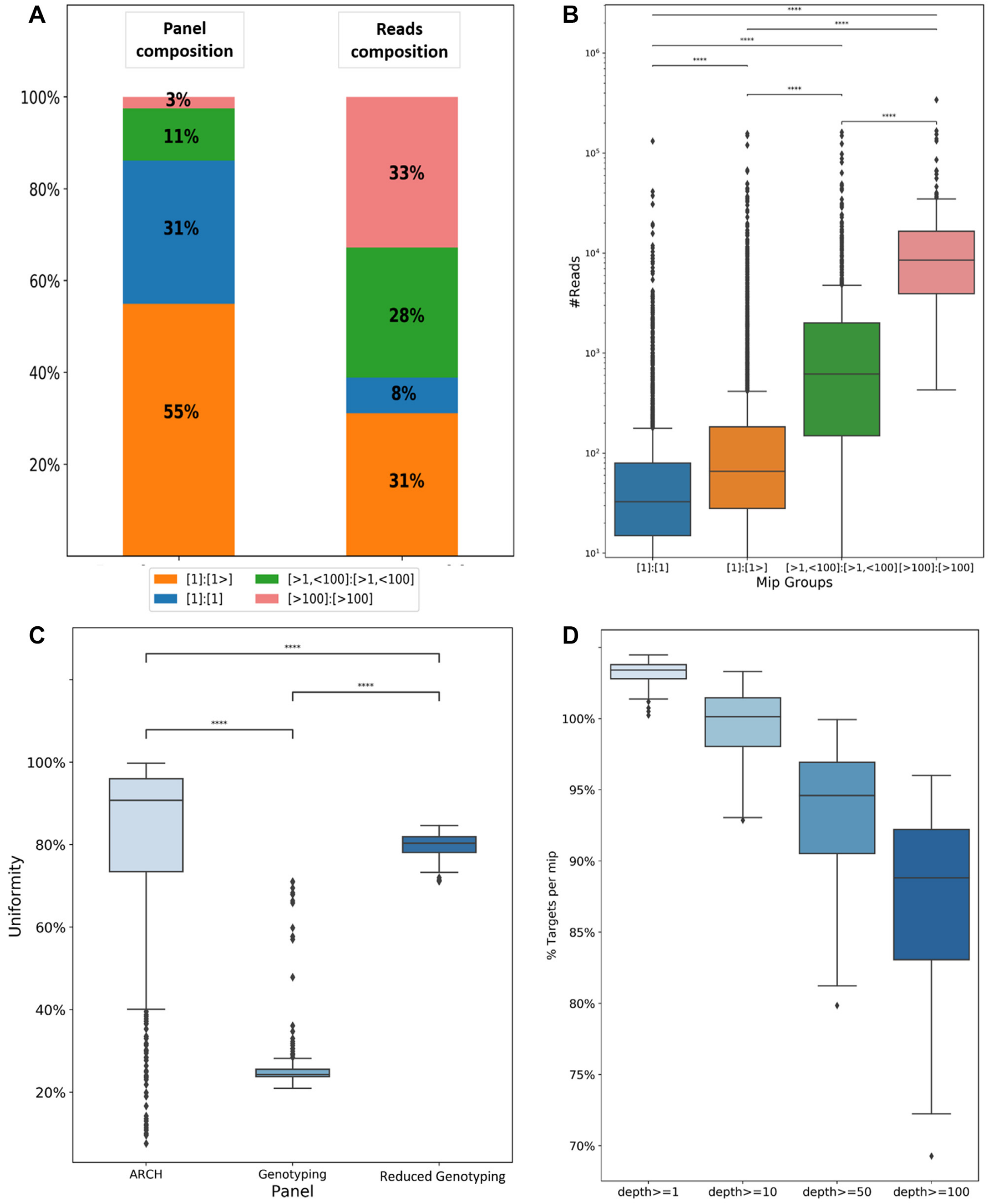

Figure 4. The iMIP protocol can successfully capture a genotyping panel of 8349 targets. (**A**) MIPs were divided into groups: [1]:[1]—ligation and extension arms have one copy in the genome, [1]:[1>] – one of the arms (either ligation or extension) has one copy in the genome, [>1 and <100]: [>1 and <100]—both ligation and extension arms have between 1 and 100 copies, [>100]:[>100]—both arms have above 100 copies. Left bar—percentage of MIPs in each of the groups, out of the total panel. Right bar—percentage of the read across all data. (**B**) Median depth compared between the target groups based on arms copy number. (**C**) Uniformity of the different panels. The uniformity of the genotyping panel was significantly lower than the ARCH panel Mann–Whitney–Wilcoxon test. However, the improved genotyping panel was significantly higher than the original genotyping panel. (**D**) Performance of the improved genotyping panel (without MIPs with copy number). Boxplots were calculated including 104 samples and the values that are presented are %targets with depth of at least one read, 10 reads, 50 reads and 100 reads.

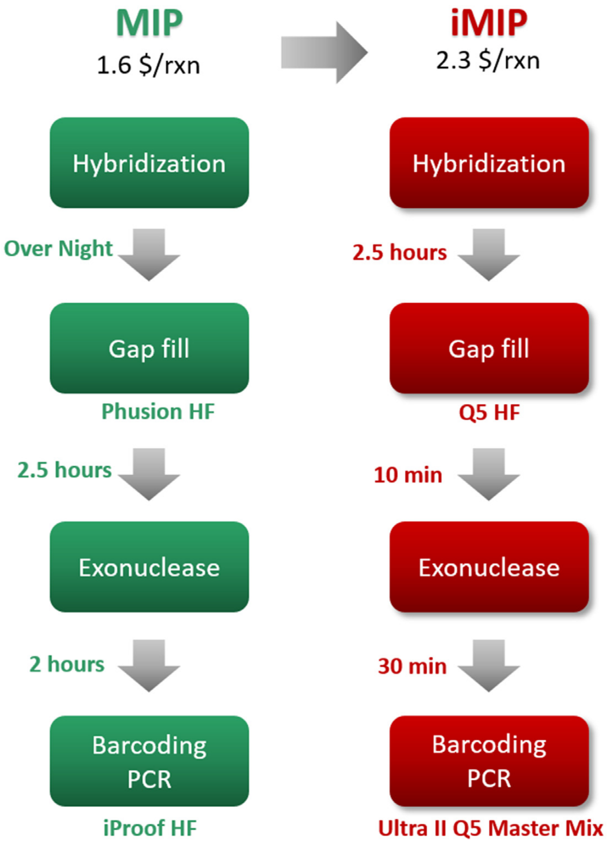

Figure 5. Main modifications of iMIP protocol from previous MIP protocols: (1) shorter hybridization incubation of 2.5 h (instead of overnight). (2) gap filing using Q5 High-Fidelity (HF) DNA Polymerase which takes ∼10 min (instead of 2.5 h). (3) Enzymatic digestion of linear probes is performed by adding Exonuclease I and Exonuclease III followed by 30 min incubation (instead of 2 h). (4) Amplification of final product using Ultra II Q5 Master Mix.

DISCUSSION

Here, we have devised a two directional (i.e. statistical and biochemical) approach for the improvement of the MIP technology, a previously low performance but highly scalable and economical technology. To achieve this goal we have studied the noise pattern of the technology in large dataset and created a benchmark amplicon based sequencing strategy to validate our candidate variants. This further improved the state-of-the-art algorithm for MIP noise reduction and generated a high precision low VAF machine learning calling model. We were also able to reduce noise by changing the protocol timing and enzymes (Figure 5 summarizes the main changes from previous MIP protocols). The improved iMIP protocol aided in the reduction of overall SNV noise of all possible alternations and eliminated the bimodal noise of C > A alternations (Figure 1). This may be explained by DNA damage due to longer exposure to oxidative stress during the library prep procedure in the MIP protocol (overnight hybridization, longer gap filling) previously observed for the MIP and other library prep protocols (4,8,11,25,26). Therefore, we suggest that shorter protocols could result in lower background error rate however this still needs formal evidence. The short 4-h protocol we devised should be attractive for both clinical laboratories and large scale screening efforts.

Calling low VAF using the MIP protocol could be further improved by utilizing unique molecular identifiers (UMI)/molecular tags (25). Although our MIP structure is composed of UMIs (seven nucleotides), we chose not to use it. This is mainly because UMI utilization for low VAF requires higher depth per target that allows large number of families with size >5 (22). In the current study we chose to allocate each sample ∼2 million reads and accordingly the vast majority of the families in our study had a size <5 (Supplementary Table S8). Nevertheless, it was also shown in the past that using a correct statistical model in hybrid capture protocols, enables correct VAF calling without the need for UMI correction (2), and we have provided similar evidence here for the MIP protocol. Our model is therefore suitable for detection of variants with VAF as low as 0.5% with sensitivity of 80% and significantly higher precision, but can be further improved. If lower VAFs or higher sensitivity are needed we suggest that deeper sequencing will be used with the addition of UMI collapsing. However, in many instances, this is not needed and our current protocol can answer the need for a cost effective low VAF protocol. Our model and protocol can be generalized for every MIP panel and can be combined with UMI error correction, however for much deeper sequencing (which might be needed for minimal residual disease detection) the number nucleotides in the UMI should be increased correlatively to depth and VAF thresholds. While, deep targeted sequencing has its own needs in the early diagnosis of cancer and other applications, the vast majority of targeted sequencing applications do not require low VAF detection and still suffer from high costs, long and complicated protocols. In the current study we present a four hour single tube fully automated protocol which is now ready for clinical use as we significantly improved its performance.

The MIP protocol notoriously suffered from low: on-target%, uniformity and GC content coverage which were all significantly improved in the current study (Figures 2 and 3). However, other caveats of the MIP protocol while improved could be further modified. The relatively low correlation between the duplicates which was significantly improved by the iMIP protocol remains an obstacle specifically for applications requiring accurate VAF such as, disease burden or somatic copy number variation detection. Based on our analysis of germline heterozygous mutations we identified that the average of the duplicates yielded better results. Future protocols could further optimized this problem possibly by reducing the number amplification cycles and calibrating the amount of DNA input to probe ratio. Clearly deeper sequencing and UMI correction will also improve duplicate correlations, however as mentioned above deep sequencing is not needed for many NGS applications. Another obstacle to the wide use of the MIP protocol is the availability of few MIP design tools which fit previous protocols (chemistry). Our data from the genotyping panel (Figure 4) suggests that refinement of MIPgen copy number thresholds and MIPgen predictions regarding MIP performance could improve the overall performance of the panel. Furthermore the fact that we introduced a new MIP protocol supports the need for a refined design tool that will take into account parameters that our current dataset harbors (e.g. effect of specific motifs like GC content and others in the MIP arms and in the target itself).

In recent years, molecular inversion probes were used to target and sequence a variety of genomic and transcriptomic targets, e.g. the exome (11,27), short tandem repeats (28), disease related targets (29–32), methylation patterns (33) and RNA expression (34). We foresee the improvements laid in this work as a stepping stone towards advancing MIP library prep not just to the clinic, mainly due to the ease of use short turnaround time, but also to other targeted sequencing applications due to improved performance specifically in GC rich of small and medium size panels.

DATA AVAILABILITY

The dataset generated and analyzed during the current study are available in the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under accession number PRJEB49813. Code is available on GitHub under https://github.com/ShlushLab.

SUPPLEMENTARY DATA

Supplementary data are available at NARGAB online.

ACKNOWLEDGEMENTS

Biological samples from Mobilized PBSC autologous transplant products were generously provided by Dr Mark Minden. L.S. is the incumbent of The Ruth and Louis Leland career development chair.

FUNDING

EU horizon 2020 grant project MAMLE [714731]; LLS and rising tide foundation [RTF6005-19]; ISF-NSFC [2427/18]; ISF-IPMP-Israel Precision Medicine Program [3165/19, BIRAX 713023]; Ernest and Bonnie Beutler Research Program of Excellence in Genomic Medicine (to L.I.S.); L.I.S. is an incumbent of the Ruth and Louis Leland career development chair; N.K. is an incumbent of the Applebaum Foundation Research Fellow Chair; Sagol Institute for Longevity Research; Barry and Eleanore Reznik Family Cancer Research Fund; Steven B. Rubenstein Research Fund for Leukemia and Other Blood Disorders; Rising Tide Foundation; Applebaum Foundation; Princess Margaret Cancer Foundation (to D.K., in part).

Conflict of interest statement. Tamir Biezuner is co-founder of Sequentify LTD. Liran I. Shlush is a consultant in Sequentify LTD. Yardena Brilon and Adi Danin are employees in Sequentify LTD.

REFERENCES

1. Chastain,E.C. (2015) In: Kulkarni,S. and Pfeifer,J. (eds). *Clinical Genomics*. Academic Press, Boston, pp. 37–55.
2. Abelson,S., Collord,G., Ng,S.W.K., Weissbrod,O., Cohen,N.M., Niemeyer,E., Barda,N., Zuzarte,P.C., Heisler,L., Sundaravadanam,Y. *et al.* (2018) Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature*, **559**, 400–404 .
3. Boyle,E.A., O'Roak,B.J., Martin,B.K., Kumar,A. and Shendure,J. (2014) MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics*, **30**, 2670–2672.
4. Hiatt,J.B., Pritchard,C.C., Salipante,S.J., O'Roak,B.J. and Shendure,J. (2013) Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.*, **23**, 843–854.
5. Almomani,R., Marchi,M., Sopacua,M., Lindsey,P., Salvi,E., Koning,B., Santoro,S., Magri,S., Smeets,H.J.M., Martinelli Boneschi,F. *et al.* (2020) Evaluation of molecular inversion probe versus truseq(R) custom methods for targeted next-generation sequencing. *PLoS One*, **15**, e0238467.
6. Park,G., Park,J.K., Shin,S.H., Jeon,H.J., Kim,N.K.D., Kim,Y.J., Shin,H.T., Lee,E., Lee,K.H., Son,D.S. *et al.* (2017) Characterization of background noise in capture-based targeted sequencing data. *Genome Biol.*, **18**, 136.
7. Ma,X., Shao,Y., Tian,L., Flasch,D.A., Mulder,H.L., Edmonson,M.N., Liu,Y., Chen,X., Newman,S., Nakitandwe,J. *et al.* (2019) Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.*, **20**, 50.
8. Acuna-Hidalgo,R., Sengul,H., Steehouwer,M., van de Vorst,M., Vermeulen,S.H., Kiemeney,L., Veltman,J.A., Gilissen,C. and Hoischen,A. (2017) Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *Am. J. Hum. Genet.*, **101**, 50–64.
9. Shlush,L.I. (2018) Age-related clonal hematopoiesis. *Blood*, **131**, 496–504.
10. Tuval,A. and Shlush,L.I. (2019) Evolutionary trajectory of leukemic clones and its clinical implications. *Haematologica*, **104**, 872–880.
11. Shen,P., Wang,W., Chi,A.K., Fan,Y., Davis,R.W. and Scharfe,C. (2013) Multiplex target capture with double-stranded DNA probes. *Genome Med.*, **5**, 50.
12. Biezuner,T., Spiro,A., Raz,O., Amir,S., Milo,L., Adar,R., Chapal-Ilani,N., Berman,V., Fried,Y., Ainbinder,E. *et al.* (2016) A generic, cost-effective, and scalable cell lineage analysis platform. *Genome Res.*, **26**, 1588–1599.
13. Bushnell,B., Rood,J. and Singer,E. (2017) BBMerge - accurate paired shotgun read merging via overlap. *PLoS One*, **12**, e0185056.
14. Martin,M.J.E.j. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads.*EMBnet*, **17**, 10–12.
15. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: https://arxiv.org/abs/1303.3997, 26 May 2013, preprint: not peer reviewed.
16. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
17. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
18. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
19. Rimmer,A., Phan,H., Mathieson,I., Iqbal,Z., Twigg,S.R.F. and WGS500 ConsortiumWGS500 Consortium, Wilkie,A.O.M., McVean,G. and Lunter,G. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, **46**, 912–918.
20. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
21. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.
22. Shugay,M., Zaretsky,A.R., Shagin,D.A., Shagina,I.A., Volchenkov,I.A., Shelenkov,A.A., Lebedin,M.Y., Bagaev,D.V., Lukyanov,S. and Chudakov,D.M. (2017) MAGERI: computational pipeline for molecular-barcoded targeted resequencing. *PLoS Comput. Biol.*, **13**, e1005480.

23. Chen,L., Liu,P., Evans,T.C. Jr and Ettwiller,L.M. (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752–756.
24. Behdad,A., Weigelin,H.C., Elenitoba-Johnson,K.S. and Betz,B.L. (2015) A clinical grade sequencing-based assay for CEBPA mutation testing: report of a large series of myeloid neoplasms. *J. Mol. Diagn.*, **17**, 76–84.
25. Waalkes,A., Penewit,K., Wood,B.L., Wu,D. and Salipante,S.J. (2017) Ultrasensitive detection of acute myeloid leukemia minimal residual disease using single molecule molecular inversion probes. *Haematologica*, **102**, 1549–1557.
26. Stefan,C.P., Hall,A.T. and Minogue,T.D. (2018) Detection of 16S rRNA and KPC genes from complex matrix utilizing a molecular inversion probe assay for next-generation sequencing. *Sci. Rep.*, **8**, 2028.
27. Turner,E.H., Lee,C., Ng,S.B., Nickerson,D.A. and Shendure,J. (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods*, **6**, 315–316.
28. Carlson,K.D., Sudmant,P.H., Press,M.O., Eichler,E.E., Shendure,J. and Queitsch,C. (2015) MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Res.*, **25**, 750–761.
29. Verkouteren,B.J.A., Wakkee,M., van Geel,M., van Doorn,R., Winnepenninckx,V.J., Korpershoek,E., Mooyaart,A.L., Reyners,A.K.L., Terra,J.B., Aarts,M.J.B. *et al.* (2019) Molecular testing in metastatic basal cell carcinoma. *J. Am. Acad. Dermatol.*, **85**, 1135–1142.
30. Kluck,V., van Deuren,R.C., Cavalli,G., Shaukat,A., Arts,P., Cleophas,M.C., Crisan,T.O., Tausche,A.K., Riches,P., Dalbeth,N. *et al.* (2020) Rare genetic variants in interleukin-37 link this anti-inflammatory cytokine to the pathogenesis and treatment of gout. *Ann. Rheum. Dis.*, **79**, 536–544.
31. Khan,M., Cornelis,S.S., Pozo-Valero,M.D., Whelan,L., Runhart,E.H., Mishra,K., Bults,F., AlSwaiti,Y., AlTalbishi,A., De Baere,E. *et al.* (2020) Resolving the dark matter of ABCA4 for 1054 stargardt disease probands through integrated genomics and transcriptomics. *Genet. Med.*, **22**, 1235–1246.
32. Pogoda,M., Hilke,F.J., Lohmann,E., Sturm,M., Lenz,F., Matthes,J., Muyas,F., Ossowski,S., Hoischen,A., Faust,U. *et al.* (2019) Single molecule molecular inversion probes for high throughput germline screenings in dystonia. *Front. Neurol.*, **10**, 1332.
33. Diep,D., Plongthongkum,N., Gore,A., Fung,H.L., Shoemaker,R. and Zhang,K. (2012) Library-free methylation sequencing with bisulfite padlock probes. *Nat. Methods*, **9**, 270–272.
34. Arts,P., van der Raadt,J., van Gestel,S.H.C., Steehouwer,M., Shendure,J., Hoischen,A. and Albers,C.A. (2017) Quantification of differential gene expression by multiplexed targeted resequencing of cDNA. *Nat. Commun.*, **8**, 15190.