Hi-Plex2: a simple and robust approach to targeted sequencing-based genetic screening

Fleur Hammel1,4, Khalid Mahmood2,3, Thomas R Green2, Tu Nguyen-Dumont1,4, Melissa C Southey1,4,5, Daniel D Buchanan2, Andrew Lonie3, Katherine L Nathanson6, Fergus J Couch7, Bernard J Pope2,3 & Daniel J Park*,1,3

ABSTRACT
We have previously reported Hi-Plex, a multiplex PCR methodology for building targeted DNA sequencing libraries that offers a low-cost protocol compatible with high-throughput processing. Here, we detail an improved protocol, Hi-Plex2, that more effectively enables the robust construction of small-to-medium panel-size libraries while maintaining low cost, simplicity and accuracy benefits of the Hi-Plex platform. Hi-Plex2 was applied to three panels, comprising 291, 740 and 1193 amplicons, targeting genes associated with risk for breast and/or colon cancer. We show substantial reduction of off-target amplification to enable library construction for small-to-medium sized design panels not possible using the previous Hi-Plex chemistry.

METHOD SUMMARY
Hi-Plex2 is an enhanced PCR amplicon-based method for the construction of small-to-medium-sized targeted DNA sequencing libraries that incorporates simple improvements to minimise off-target effects while retaining the benefits of its predecessor, Hi-Plex.

KEYWORDS
genetic screening • Hi-Plex • multiplex PCR • targeted DNA sequencing • variant detection

Previously, we have reported the development and application of Hi-Plex for cost-effective and streamlined targeted sequence-screening. Hi-Plex is a highly multiplex PCR-based approach to sequencing library preparation. It applies amplicon-specific primer-based ‘seeding’ and universal primer-based ‘harvesting’ in a single reaction vessel at the start of the protocol. As this step includes sample indexing, products from many different samples can be pooled at an early stage, preventing the need for parallel handling through multiple processing steps. The division of overall target-space into uniform-sized ‘tiles’ allows relatively stringent selection of products of the expected size. Hi-Plex uses inexpensive off-the-shelf reagents, involves few processing steps and is compatible with high-fidelity DNA polymerases.

Alternative targeted sequencing approaches exist. Molecular inversion probe-based targeting offers one general approach (e.g., Roche Heat-Seq), but requires relatively expensive reagents and involves a substantial number of enzymatic and purification steps in parallel. This can be costly and labour-intensive. Agilent SureMASTR offers another exemplar approach. This method is closely related to Hi-Plex, but does not prevent parallel processing across multiple specimens.

Hi-Plex has proved effective for numerous panel designs, including target sizes ranging from 16 to 1000 amplicons [1–10]. However, more extensive application of the method across a number of different target designs led to observations of excessive off-target amplification, manifesting as ‘smearing’ under agarose gel electrophoretic analysis, with no discernable library bands – such products are not suitable for subsequent sequencing.

To mitigate this, we sought to develop a standardized protocol that would reduce the off-target influence on library preparation, while maintaining a streamlined, high-throughput workflow. We present Hi-Plex2, an upgraded protocol that delivers robustness and improved targeting accuracy while maintaining the benefits of its predecessor (Figure 1).

A major advancement in Hi-Plex2 is the separation of the seeding and harvesting stages. This offers an approach to restricting off-target product persistence while still enabling the funneling of large sample numbers (i.e., 96-well plate) to single-tube format very early in the procedure. Accurate size-selection of pooled target constructs is facilitated using spiked-in DNA size standards. This markedly improves robustness across designs and yields purified, sequencing-ready library construct in a single-day workflow.

MATERIALS & METHODS
Primer design was conducted as described previously [1,5], with unpublished modifications to implement a nearest-neighbor thermodynamic model to estimate melting temperatures (Tm) [11] and a custom algorithm related to the Smith-Waterman string alignment algorithm [12] to iteratively avoid predicted primer-dimer conflicts. This new gaps-aware primer-dimer conflict-scoring algorithm takes a coarse-grained dynamic programming approach to identifying relevant potential dimerization structures by considering ‘tile’ binding elements (https://github.com/djpark1974/primer_dimer_tile_scoring). One strategy to consider is the use of a tool such as PrimerPooler (http://people.ds.cam.ac.uk/sssb22/pooler/), which separates primer sets into pools in an attempt to avoid primer-dimer events, although without appar-
Reports

Reports

quantified using the Qubit 2.0 Fluorimeter (Hilden, Germany) and fluorimetrically line pellet via QIAamp DNA Blood kit (Qiagen, Genomic DNA, extracted from an Epstein–

Panel 1: 1 μM (IDT), IL, USA). Aggregate concentrations: primer (ASP; Integrated DNA Technologies μ and 0.0017 (ThermoFisher Scientific, MA, USA; no. F540)

The target Tm was set to 64°C for the amplicon-specific portion of primer, the target amplicon insert length was 100 bp.

A Hi-Plex2 PCR mix was prepared on ice, containing final concentrations of: 1x Phusion HF buffer, 2.5 mM MgCl₂, 1.6 mM dNTPs (dATP, dTTP, dGTP, dCTP at 400 μM each), 1 U Thermo Scientific Phusion Hot Start High-Fidelity DNA polymerase (ThermoFisher Scientific, MA, USA; no. F540) and 0.0017 μM of each amplicon-specific primer (ASP; Integrated DNA Technologies [IDT], IL, USA). Aggregate concentrations: Panel 1: 1 μM; Panel 2: 2.5 μM; Panel 3: 4 μM. Genomic DNA, extracted from an Epstein–Barr virus-transformed lymphoblastoid cell line pellet via QIAamp DNA Blood kit (Qiagen, Hilden, Germany) and fluorimetrically quantified using the Qubit2.0 Fluorimeter (ThermoFisher Scientific), was included in the mix to achieve 50 ng per 25 μl reaction. On ice, the mix was dispensed into wells of a 96-well PCR plate prior to direct transfer into a Veriti™ thermocycler (ThermoFisher Scientific) at 98°C. Thermocycling conditions applied immediately were: 98°C – 1 min, four cycles of 98°C – 30 s, 60°C – 2 min, 64°C – 2 min, 72°C – 1 min. The thermocycler block temperature was maintained at 72°C for addition of indexed adapter primers (Integrated DNA Technologies; Supplementary Table 1) to 1 μM final concentration, as described previously [13]. Thermocycling conditions applied were: four cycles 98°C – 30 s, 66°C – 2 min, 72°C – 1 min, followed by a 72°C – 7 min extension, and reaction termination by addition of 2.8 μl 100 mM EDTA. 10 μl from each well per plate (total 960 μl per panel) was pooled into a single tube (Figure 1) with addition of 1.5 μg Invitrogen 50 bp ladder (ThermoFisher Scientific, No. 10416014). Two serial one-sided size-selections were then performed using JetSeq™ Clean (Bioline, London, UK, No. 68030) paramagnetic beads employing bead:reaction ratios of 0.91. The resultant DNA was eluted in 100 μl lowTE and 15 μl Trackit Cyan/Orange loading dye (ThermoFisher Scientific, No. 10482028) was added. Eluate was loaded into a preparative (25 mm × 1.5 mm) well of 1.75% (w/v) Invitrogen UltraPure™ Agarose-1000 (Thermo-Fisher Scientific, No.16550100)/ 1x TBE gel containing 0.5x GelRed nucleic acid stain (Biotium, Freemont, CA, USA, No.41003).

Electrophoresis was conducted in 1x TBE buffer at 100V for 70 min. The gel region bounded by, and including, the 250 bp and 300 bp ladder bands, was excised from the gel. DNA was extracted from agarose using the QIAquick Gel Extraction Kit (Qiagen, No. 28706) according to the manufacturer’s instructions. Prior to elution, columns were centrifuged at 16,000 x g and incubated for 10 min at 50°C to completely remove all traces of ethanol. DNA was eluted from the column using 50 μl 10 mM Tris pH 8.0.

The eluate, containing purified, size-selected seeding PCR pool was subjected to a single-reaction harvesting PCR comprising: 1x Phusion HF buffer, 2.5 mM MgCl₂, 400 μM dNTPs, 1 μM both P5 and P7 primers (5′–3′ sequence of AATGATACGGCGACACCAGGA and CAAGCGAAGCGGCCAGCATCGA, respectively), 1 U Thermo Scientific Phusion Hot Start High-Fidelity DNA polymerase in a total volume of 75 μl. PCR parameters were: 98°C 1 min, 12 cycles of 98°C – 30 s, 60°C – 1 min, 72°C – 1 min, followed by extension at 72°C – 5 min. The –270 bp expanded library was then purified by gel electrophoresis in a 2% (w/v) agarose/1xTBE gel using preparative well combs and extracted via the QIAquick Gel Extraction Kit (Qiagen, No. 20051). Libraries were quantitated using the QuantiFluorONE™ dsDNA System kit (Promega, WI, USA, No. E4870) and quality assessed on the 4200 TapeStation (Agilent Technologies, CA, USA) using a High-Sensitivity D1000 ScreenTape (Agilent Technologies, No. 5067–5584).

For sequencing, 2 nM libraries for Panels 1, 2 and 3 were pooled in the volume ratio 291:740:1193 (reflecting the number of amplicons in each design) to achieve a 2 nM final pooled library. Sequencing was performed via a NextSeq 550 (Illumina, CA, USA) High Output run using 150-bp paired-
end reads and v2 chemistry (the subject libraries were sequenced in conjunction with other libraries not relevant to this publication, requiring a High Output run for the total). Denatured library was loaded at 1.8 nM, HPLC-purified Read1, Read2, i5 and i7 (sequence details in Supplementary Table 1) custom sequencing primers were added to their respective sequencing cartridge positions to 0.3 μM each.

FASTQ format data were retrieved from the NextSeq and mapped to the reference human genome (hg19) using the BWA-MEM algorithm (v0.7.15) [14]. The mapped reads were then filtered to discard secondary and supplementary mappings. The primary mapped sequences were then used to calculate target coverage-related statistics using in-house tools. The complete set of sequence analysis tools and automated pipeline (hiplexpipe) used in this study are available from and detailed at https://github.com/khalidm/hiplexpipe.

RESULTS & DISCUSSION
To demonstrate that Hi-Plex2 provides a robust and consistent targeted sequencing method to build amplicon-based libraries, we applied the approach to three panels, designed using our updated primer design system. These panels targeted coding regions of genes that have been implicated in predisposition to breast and/or colon cancers. The panels, representing a variety of sizes and gene target sets, referred to herein as Panel 1, Panel 2 and Panel 3, contain 291, 740 and 1193 amplicons. Panel 1 design targets are a subset of genes targeted in Panel 3 – however, primer sequences are not necessarily in common as the panel design is influenced by different potential primer-dimer contexts. Using these designs, we tested both the previous Hi-Plex chemistry and Hi-Plex2 for their ability to generate clean library profiles, as indicated by agarose gel electrophoretic analysis. Whereas Hi-Plex2 yielded clearly resolved libraries across replicates for all three panels, the previous Hi-Plex chemistry only yielded clear library bands for Panel 1. In the cases of Panels 2 and 3, Hi-Plex resulted in indistinct smears unsuitable for further processing.

We then benchmarked Hi-Plex2 in the 96-well plate setting in the context of sequence-based analysis. To do this, for each panel, we prepared 96 uniquely barcoded reactions each templated with an aliquot from the same human genomic DNA preparation. Figure 2A shows an image of the post-seed size-selection gel for each panel pool. All DNA size markers upwards of 250 bp were clearly visible and absence of markers below 250 bp indicates all smaller PCR elements were successfully removed. Figure 2B depicts plate-level, post-harvesting PCR libraries for each of the three panels. In all cases, clearly resolved library bands were observed at the expected size with no background smearing.

Sequence analysis conducted across all samples for Panel 1, 2 and 3 revealed a median of 96.4, 98.8 and 91.3% reads mapped to the genome (hg19), with an average on-target coverage depth of 1851, 2706 and 1288 (Table 1). The median percentage of reads mapping to the target, of total sequenced reads, were 81.0% for Panel 1, 98.3% for Panel 2 and 55.8% for Panel 3. The percentages of amplicons for which at least one primary-mapped read spanned the mid-point of the amplicon for Panels 1, 2 and 3 were 100, 99.3 and 99.3%, respectively. Coverage uniformity was assessed

Figure 2. Agarose gel profiles of Hi-Plex2 library preparation. (A) Size selection gel. 96 individual sample Hi-Plex2 PCR reactions seeded with panel amplicon-specific primers (as indicated) followed by adapter primers, cleaned with JetSeq™ Clean, prior to agarose gel electrophoresis. Dashed box indicates the region of gel containing the design target library products. The within-lane 250 and 300 bp marker bands (arrowed) guide accurate size selection of library products. (B) Agarose gel representative library band profile after expansion PCR.

M: 50 bp MW marker; NTC: No template control (non-templated seeding PCR reaction processed through size selection and expansion PCR steps as for templated reactions).

Figure 3. Modelled percentage on-target. On-target rate metrics resulting from progressive removal amplicon read pairs ranked (top 1–50) for highest contribution to off-target reads.
by calculating the primary-mapping depth per amplicon compared to the median across the target per panel. 99.7, 96.9 and 96.4% of amplicons in Panels 1, 2 and 3, respectively, exhibited primary-mapping depth within 20-fold of the median across the target.

The sequencing metrics demonstrated good all-round performance across the three panels, with potential to further improve the on-target rate for Panel 3. Considering Panel 3 in particular, we next investigated possible strategies for further enhancing on-target rates. We reasoned that if small subsets of primers tend to contribute to disproportionately large off-target effects, and that these could be identified, it would enable rational approaches to iterative design optimization — either by simply removing small numbers of troublesome primers from a given panel or by attempting to redesign such primers. To investigate this, we applied custom software that enabled identification of specific primer sequences occurring within output reads to determine, for each primer sequence, the number of reads that mapped to the genome, mapped to specific target genomic regions or were unmapped. Indeed, we observed that for each panel the vast majority of off-target reads involved a very small proportion of the total off-target reads. For Panel 1, this primer pair contributed to 23% total off-target, whereas for Panel 3 the corresponding pair contributed only 0.9% total off-target reads.

A high level of off-target priming observed for a given amplicon’s primers does not necessarily detract from on-target coverage achieved. We observed that primers implicated in contributing the highest counts of off-target reads generally yielded acceptable levels of their intended on-target products. Contributions to off-target product production can be complex. A primer pair targeting the same genomic region in Panels 1 and 3, with very slight sequence variations, demonstrated large differences in contributions to total off-target reads. For Panel 1, this primer pair contributed to 23% total off-target, whereas for Panel 3 the corresponding pair contributed only 0.9% total off-target reads.

As we identified that very few primers contribute the large majority of off-target effects, we next sought to model likely performance achievable by simply removing these from the system. To do this, we iteratively removed FASTQ read-pairs that displayed evidence of high off-target-associated amplicon primer sequence at read-starts - this was performed step-wise in rank order of amplicon-level contribution to off-target signal. The filtered FASTQ files were reanalysed through hipepipe to obtain metrics of projected on-target rates for each progressive amplicon primer-pair subtraction (Figure 3). This modelling demonstrated that incremental removal of the highest off-target contributing amplicon primer pairs in each panel would likely result in significant increases in on-target rates. For example, for Panel 1, the total reads on-target rate increased from an observed 81.0 to 90.6% with the removal of just two amplicons (equating to the retention of 99.3% of target amplicons). Similarly, for Panel 3, by removing the worst-performing 0.5% of amplicons, the total reads on-target rate increased from an observed 56.3% to a projected 86.5% (with the retention of 99.5% of target amplicons). As such, fractional reductions in design targets by excluding primers causing the highest off-target effects should yield significant gains in sequencing efficiency. Alternatively, iterative redesign of the worst performing primers could be an effective strategy.

We have demonstrated Hi-Plex2 to have robust performance characteristics across a variety of design sizes, including a design that had proved intractable by the parent method (data not shown). This is realized by a library construction protocol that physically separates reaction seeding and harvesting stages while maintaining the streamlined philosophy, such that the method is amenable to processing many specimens (e.g., 96-well plate format) in single-tube format immediately following initial seeding. The simplicity of the methodology allows for further upscaling in the same processing time and, moreover, we have identified potential for improved performance, optionally, by redesign or removal of a very small fraction of the total design space.

<table>
<thead>
<tr>
<th>Panel (amplicons)</th>
<th>Median target coverage</th>
<th>Total reads MtoG (%)</th>
<th>Target covered (%)</th>
<th>Target coverage ±20-fold median (%)</th>
<th>On-target rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel 1 (291)</td>
<td>1851</td>
<td>96.4</td>
<td>100</td>
<td>99.7</td>
<td>81.0</td>
</tr>
<tr>
<td>Panel 2 (740)</td>
<td>2706</td>
<td>98.8</td>
<td>99.3</td>
<td>96.9</td>
<td>98.3</td>
</tr>
<tr>
<td>Panel 3 (1193)</td>
<td>1288</td>
<td>91.3</td>
<td>99.3</td>
<td>96.4</td>
<td>55.8</td>
</tr>
</tbody>
</table>

†~99.5% target refers to projected data based on reanalysis having removed ~0.5% of the worst offending (with respect to off-target contribution) amplicon primer-containing reads (0.7% in the case of Panel 1). MtoG: Mapping to genome; On-target rate (%): The fraction of total reads mapping to the design target.
The methodology is applicable to a wide variety of settings in which targeted regions of genomes are of interest for sequencing across many specimens, including the contexts of medical diagnostics and epidemiological discovery efforts. To provide an illustration of the cost benefits of Hi-Plex2, the library-build component of a sequence screen of size 300 amplicons using Hi-Plex2 costs the order of US$2 per specimen, depending on the scale of the screen, essentially reducing the majority of the cost of screening to that associated with the sequencing chemistry.

FUTURE PERSPECTIVE

The cost of DNA sequencing will continue to decrease. However, considerations relating to sequence depth requirements, a desire for serial sampling over site and time, and data volume and processing requirements, will likely motivate the ongoing application and relevance of targeted sequencing approaches. Somatic mutation analysis in cancer provides one scenario in which all of these factors are relevant. Hi-Plex2 offers a niche solution for small-to-moderate target sizes that exhibits robust performance and that is simple and cost effective. The introduction of molecular indexing is one potentially powerful way to modify this technique in the future. Software design optimizations also offer numerous potential avenues for improvement. Repetitive and low-complexity sequence contexts, for example, remain challenging for targeted sequencing approaches, generally. Improved approaches to considering such regions and other factors that contribute to off-target effects as part of system design will be valuable.

SUPPLEMENTARY DATA

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com doi/suppl/10.2144/btn-2019-0026

AUTHOR CONTRIBUTIONS

FH contributed to experimental design, conduction of experiments and preparation of the manuscript. KM and TRG contributed to data analysis. TN-D, MCS, DDB, AL, BJP, KLN and FJC contributed to experimental design. DJP conceived the study and contributed to experimental design and data analysis.

ACKNOWLEDGMENTS

We thank Helen Tsimiklis and Catherine Chatfield for supplying the DNA used in this study, Derrick Theyes for operation of the NextSeq 550 and Jason Steen for processing sequencing data.

FINANCIAL & COMPETING INTERESTS DISCLOSURE

This work was supported by the Australian National Health and Medical Research Council (NHMRC) grants GNT1108179, GNT1074383 and GNT1125269, and national Breast Cancer Foundation (NBFC) grant NT-14–016. DD Buchanan is a NHMRC RD Wright Career Development Fellow, TN-D is a National Breast Cancer Foundation (Australia) Career Development Fellow and BJP is a Victorian Health and Medical Research Fellow. MC Southey is a NHMRC Senior Research Fellow (GNT106177). The authors have no other relevant affiliations or financial involvement with any organization or entity in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

OPEN ACCESS

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/

REFERENCES