BIOLOGY
Methods & Protocols

Biology Methods and Protocols, 2018, 1–8

doi: 10.1093/biomethods/bpy001
Methods Manuscript

METHODS MANUSCRIPT

Homology-based enzymatic DNA fragment assembly-based illumina sequencing library preparation

Hiroshi Shinozuka,[1,*] Shimna Sudheesh,[1] Maiko Shinozuka[1] and Noel O.I. Cogan[1, 2]

[1]Agriculture Victoria, AgriBio, Centre for AgriBioscience, 5 Ring Road, La Trobe University, Bundoora, Victoria, 3083 and [2]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, 3086, Australia

*Correspondence address: E-mail – hiroshi.shinozuka@ecodev.vic.gov.au

Abstract

The current Illumina HiSeq and MiSeq platforms can generate paired-end reads of up to 2 x 250 bp and 2 x 300 bp in length, respectively. These read lengths may be substantially longer than genomic regions of interest when a DNA sequencing library is prepared through a target enrichment-based approach. A sequencing library preparation method has been developed based on the homology-based enzymatic DNA fragment assembly scheme to allow processing of multiple PCR products within a single read. Target sequences were amplified using locus-specific PCR primers with 8 bp tags, and using the tags, homology-based enzymatic DNA assembly was performed with DNA polymerase, T7 exonuclease and T4 DNA ligase. Short PCR amplicons can hence be assembled into a single molecule, along with sequencing adapters specific to the Illumina platforms. As a proof-of-concept experiment, short PCR amplicons (57–66 bp in length) derived from genomic DNA templates of field pea and containing variable nucleotide locations were assembled and sequenced on the MiSeq platform. The results were validated with other genotyping methods. When 5 PCR amplicons were assembled, 4.3 targeted sequences (single-nucleotide polymorphisms) on average were successfully identified within each read. The utility of this for sequencing of short fragments has consequently been demonstrated.

Keywords: Gibson assembly; synthetic biology; next-generation sequencing (NGS); target enrichment; single nucleotide polymorphisms (SNPs)

Introduction

Sequencing-by-synthesis technologies (originally by Solexa) have realised the promise of high-throughput sequencing [1, 2], and a further substantial progress has been achieved since the appearance of the Illumina DNA sequencing platforms based on this method (https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/). One of the major improvements has been in the length of sequencing reads. The read length of an early version of the Illumina (Solexa) platform was only 25–35 bp (2 × 25–35 bp paired-end), considerably shorter than that of earlier Sanger sequencing-based platforms (up to 800–1000 bp) [1–4]. However, the current Illumina HiSeq platform can generate up to 2 × 250 bp reads, while even longer reads, 2 × 300 bp, can be generated on the MiSeq platform (https://www.illumina.com/techniques/sequencing/dna-sequencing.html). Owing to considerable reductions in both reagent cost and processing time, short-read technologies have become used for not only whole genome shotgun

Received: 12 October 2017; **Revised:** 3 January 2018; **Editorial decision:** 5 January 2018; **Accepted:** 19 January 2018

(re-)sequencing, but also other purposes, including small RNA (sRNA) sequencing, molecular inversion probe (MIP)-based mutant detection and characterisation of environmental DNA (eDNA) [5–7]. Generation of up to several billion sequencing reads from a single run permits identification of low prevalence sRNAs, accurate determination of sRNA expression levels, detection of functional sequence variation and presence of eDNA. The DNA molecules prepared for such purposes, however, can be considerably shorter than the potential length of sequencing reads from the current high-throughput sequencing systems.

The homology-based enzymatic DNA fragment assembly (or more commonly termed 'Gibson assembly') is a simple DNA manipulation method, widely used in synthetic biology [8]. Double-stranded DNA (dsDNA) fragments can be integrated into a larger single molecule through the activities of T5 exonuclease, DNA polymerase and DNA ligase [9]. DNA fragments are partially digested with T5 exonuclease, which catalyses DNA degradation in the 5′ to 3′ direction, so generating dsDNA with cohesive ends on the 3′-termini. The DNA fragments are then annealed on the basis of sequence homology between the cohesive ends (typically 15–80 bp in length), and larger concatenated DNA molecules are generated through gap-filling and ligation with the DNA polymerase and *Taq* DNA ligase. This method, however, is not suitable for assembly of short DNA fragments (<250 bp in length), presumably due to the strong enzymatic activity of T5 exonuclease [10]. In the present study, an alternative short dsDNA fragment assembly method has been developed through the use of T7 exonuclease (T7 gene 6 exonuclease), which has a moderate 5′→3′ exodeoxyribonuclease activity (https://www.neb.com/tools-and-resources/selection-charts/properties-of-exonucleases-and-endonucleases). The phosphorothioate bond (S-bond) is a modification used for synthetic oligonucleotides, and PCR primers containing S-bonds show partial or complete tolerance to a range of nucleases, largely depending on the number of inserted S-bonds (https://www.neb.com/tools-and-resources/selection-charts/properties-of-exonucleases-and-endonucleases, https://sg.idtdna.com/site/Catalog/Modifications/Category/8). Short PCR amplicons were prepared from genomic DNA (gDNA) template from a diploid crop plant species, field pea (*Pisum sativum* L.; $2n = 2\times = 14$) [11, 12], using the S-bond containing PCR primers, and a sequencing library for the Illumima platforms was subsequently prepared from these amplicons.

Materials and methods

T7 and T5 exonuclease activity assay

Locus-specific primers were designed for the field pea Psy_KP1_SNP_100000290 sequence [11, 12]. PCR primers with 1 and 6 S-bond modifications were designed and synthesised at Integrated DNA Technologies (Coralville, IA, USA) and GeneWorks (Thebarton, SA, Australia) in addition to those without modification (Supplementary Material 1). Short DNA fragments (62 bp or ~300 bp in length) were generated from gDNA templates through PCR using the Phusion polymerase kit (Thermo Fisher Scientific, Waltham, MA, USA), following manufacturer's instructions. The DNA fragments (~300 bp in length) were purified using the Agencourt AMPure XP kit (Beckman Coulter, Brea, CA, USA) and treated with T7 or T5 Exonuclease [5 U; New England Biolabs (NEB), Ipswich, MA, USA] in 1× NEBuffer 2 or 4 at room temperature (20–25°C) for 10 min. The assembly reaction can be performed owing to the activities of three enzymes (exonuclease, DNA polymerase and ligase), and all reactions should be performed in a single tube for a practical use. The NEBuffers 2 and 4 were, therefore, selected for the assay, as T7 and T5 Exonuclease show high enzymatic activity in the NEBuffer 4, and the NEBuffer 2 is relatively similar to a typical PCR buffer (https://www.neb.com/). T4 ligase shows high enzymatic activity in both NEBuffers 2 and 4 under the presence of ATP. For a time-course assay, the shorter DNA fragments were amplified using PCR primers with a single S-bond modification, and, then, treated with T7 or T5 exonuclease without DNA purification. An enzyme mixture (1 µl 10× NEBuffer 4, 1 µl T5 or T7 Exonuclease, 3 µl water) was prepared and added to 5 µl PCR products. The treatment was performed at room temperature for 3–10 min. For a control experiment, molecular biology grade water was used, instead of the enzymes. The treated DNA fragments were visualised using the 2200 TapeStation instrument and D1000 kit (Agilent Technologies, Santa Clara, CA, USA).

Gibson assembly-based sequencing library preparation

Locus-specific primers were designed to amplify short DNA fragments (57–66 bp in length) that included variable nucleotide positions (single-nucleotide polymorphisms; SNPs) of field pea (Table 1) [11, 12]. An 8 bp tag was attached to the 5′-terminus of each primer for the sequence homology-based assembly, and an S-bond was introduced between the ninth and tenth bases from the 5′-terminus to generate PCR fragments with cohesive ends (8 bp) after treating with the exonuclease. Sequencing adapters with 8 bp tags were also designed, which were protected from nuclease activity by S-bond modification at both the 5′- and 3′-termini. The tags were designed to assemble five separate PCR fragments and the two sequencing adapters into a single molecule. Two sets of PCR primers were prepared and designated PsySNP_Set1 and 2. Genotypes derived from two cultivars (Kaspa and PBA Oura), 7 genotypes (Psy_RIL99, Psy_RIL195, Psy_RIL268, Psy_RIL614, Psy_RIL656, Psy_RIL677 and Psy_RIL678) from recombinant inbred line (RIL) populations obtained by crossing genotypes from parental cultivars [12] and an unknown genotype (Psy_GenoX) of field pea were selected for the SNP genotyping assay. Using the ISOLATE Rapid Plant Buffer (beta version product; BIOLINE, London, UK), gDNA was extracted from leaves of the field pea genotypes and extracted DNA was purified and concentrated using the AMPure bead kit. The target DNA fragments, including the SNP sites, were amplified through uniplex PCR with the Phusion polymerase kit, and the same volume of PCR solution was pooled for enzymatic assembly. All products from 10 PCR reactions were pooled and designated PsySNP_All sample. An enzyme mixture [1 µl 10× NEBuffer 4, 1 µl 10 mM ATP (NEB), 1 µl T7 Exonuclease, and 0.5 µl T4 DNA ligase (400 000 units/ml; NEB)] was prepared and added to 5 µl PCR product mixture, which contained residual Phusion DNA polymerase and dNTPs. A sequencing adapter mixture (1.5 µl; 1.7 µM each) was then added (Table 1) [13], and the assembly mixture (total volume: 10 µl) was incubated at room temperature (20–25°C) for 15 min. After incubation, 10 µl EDTA (100 mM) were added to the mixture. The resulting DNA was purified using 16 µl AMPure bead solution, following manufacturer's instructions, and then eluted in 10 µl Tris–HCl buffer (10 mM). The assembled DNA was amplified in a 25 µl reaction volume using in-house indexing primers for Illumina sequencing systems and the Phusion DNA polymerase kit [13]. The sequencing libraries were visualised using the Agilent 2200 TapeStation platform, and pooled for multiplexed sequencing. The pooled libraries were subjected to two cycles of

Table 1: PCR primers for homology-based enzymatic DNA fragment assembly-based library preparation

Primer set	Locus (adapter)	Primer name	Primer sequence (5′→3′)
PsySNP_Set 1	Psy_KP1_SNP_100000290	Forward	AATCTCGTA*CCTCCGGTGCTATATAAGC
		Reverse	AGTGGCAAA*AATCGACTGTTGGAACTCC
	Psy_KP1_SNP_100000228	Forward	TTGCCACTC*CAGGGAATATGTTAGGGAAAC
		Reverse	CGTTGAAGG*GAACCAGTAATAATGCATCCA
	Psy_KP2_SNP_100000360	Forward	CTTCAACGG*TAGATCACAACCCGTATTC
		Reverse	AGTACGCTT*GCTCGTGACGCCTTGTAGA
	Psy_KP3_SNP_100000258	Forward	AGCGTACTT*TCTCGCACGCCTTACACTT
		Reverse	TCACCGTAT*GTCGTTTATTGGTACTCAG
	Psy_KP4_SNP_100000576	Forward	TACGGTGAG*CCAGAACCATCTGTAGCTATT
		Reverse	GGACAGTTC*TACTTTTGGACATATTCTGTC
PsySNP_Set 2	Psy_KP4_SNP_100000076	Forward	AATCTCGTT*CCGAAGAGGATTACCCCTA
		Reverse	AGTGGCAAC*ATCTACCATCAATAGCACG
	Psy_KP4_SNP_100000577	Forward	TTGCCACTG*GTCCTTGACCATACATAAA
		Reverse	CGTTGAAGC*CTGGTGCTCCAGGTCTTGG
	Psy_KP4_SNP_100000137	Forward	CTTCAACGT*GAGGCTGAAAACTTCTC
		Reverse	AGTACGCTA*AAGCTATTGAAGACCTCAA
	Psy_KP4_SNP_100000293	Forward	AGCGTACTG*ATGTAACACAGACACTCCG
		Reverse	TCACCGTAG*CGAAGAGTATAAAGGATAC
	Psy_KP4_SNP_100000267	Forward	TACGGTGAT*GCTACAGAGGGATCAGGCA
		Reverse	GGACAGTTC*CCCGACAGGGTAAACCATC
Sequencing library adapter	Positive strand of adapter	mpxPE1(+).adpPsytag1	A*C*A*CTTTCCCTACACGACGCTCTTCCGATCTAATCTCG*T
	Positive strand of adapter	mpxPE1(+).adpPsytag2	A*C*A*CTTTCCCTACACGACGCTCTTCCGATCTGGACAGT*T
	Negative strand of adapter	mpxPE2(-).GA-adp	A*G*ATCGGAAGAGCACACGTCTGAACTCCA*G*T*C

The sequence corresponding to assembly tag is shown underlined. An asterisk (*) indicates presence of S-bond modification between the nucleotides. The PCR primers were synthesised at Integrated DNA Technologies and GeneWorks.

size-selection with AMPure bead solution (0.8×). Further details of the sequencing library preparation method are provided in Supplemental Material 2.

Sequencing analysis and GBS data analysis

The pooled library was titrated using the TapeStation and Qubit instruments (Thermo Fisher Scientific). Sequencing analysis was performed with the MiSeq platform and MiSeq Reagent Nano Kit v2 (500 cycles). The library was denatured and loaded on the reagent cartridge, following manufacturer's instructions. Paired-end 260 bp reads were generated, and output data were analysed using the Sequencher 5.0 (Gene Codes, Ann Arbor, MI, USA) and Excel 2010 (Microsoft, Redmond, WA, USA) software packages (Supplemental Material 3).

Kompetitive allele-specific PCR (KASP) assay

Fluorescence-based genotyping was performed using KASP Master mix (LGC Genomics, Middlesex, UK), based on the modified protocol [12]. PCR primers were designed and synthesised at LGC Genomics (Supplemental Material 4), and the target sequence was amplified in a 10 μl PCR mixture. The PCR products were analysed on the FLUOstar Omega microplate reader (BMG LABTECH, Ortenberg, Germany), and the resulting data were visualised using KlusterCaller^{TM} software (LGC Genomics).

PCR–RFLP assay

A PCR-restriction fragment-length polymorphism (RFLP) assay was designed for three SNP-containing sites, designated Psy_KP1_SNP_100000290, Psy_KP4_SNP_100000076 and Psy_KP4_SNP_100000293. For the SNP_100000290 and SNP_100000076 sites, the PCR primers designed for sequencing library preparation were used, and PCR fragments were digested with the *Hpa*II and *Hpy*CH4III restriction enzymes (NEB), respectively. A reverse primer with a substituted base was designed for the Psy_KP4_SNP_100000293 site, and PCR fragments were digested with the *Bst*EII restriction enzyme (NEB). The DNA fragments were visualised using the 2200 TapeStation instrument. Further details can be found in Supplemental Material 5.

Results and discussions

In the exonuclease activity assay, DNA fragments were generated using PCR primers with or without S-bond modification. The DNA fragments (∼300 bp in length) generated using unmodified PCR primers were almost completely degraded by T7 exonuclease, while the fragments generated using PCR primers with 1 and 6 inserted S-bonds showed some tolerance to the activity (Fig. 1a, Supplementary Material 6). T5 exonuclease showed a higher enzymatic activity, and the fragments generated using PCR primers with S-bond modification were completely digested. As PCR primers with 6 S-bond modifications showed reduced PCR amplification (Supplementary Material 7), PCR primers with a single S-bond were preferred for the subsequent protocol development process. The DNA fragments prepared with primers with a single S-bond modification was subject to a further assay, to find that the DNA fragments were almost completely degraded within 3 min when T5 exonuclease was used (Fig. 1b). Moderate catalysis of T7 exonuclease was observed throughout 10 min. This result indicates that T7 exonuclease is more practically suitable for the controlled degradation of short DNA fragments.

Sequencing libraries were prepared following the described method (Fig. 2a). Short DNA fragments were generated from the field pea gDNA templates, using the PCR primers with a single S-bond modification (Table1, Fig. 2b, Supplementary Material 8). For assembly, 5 PCR products (*Psy*SNP_Set 1 and 2) or 10 PCR

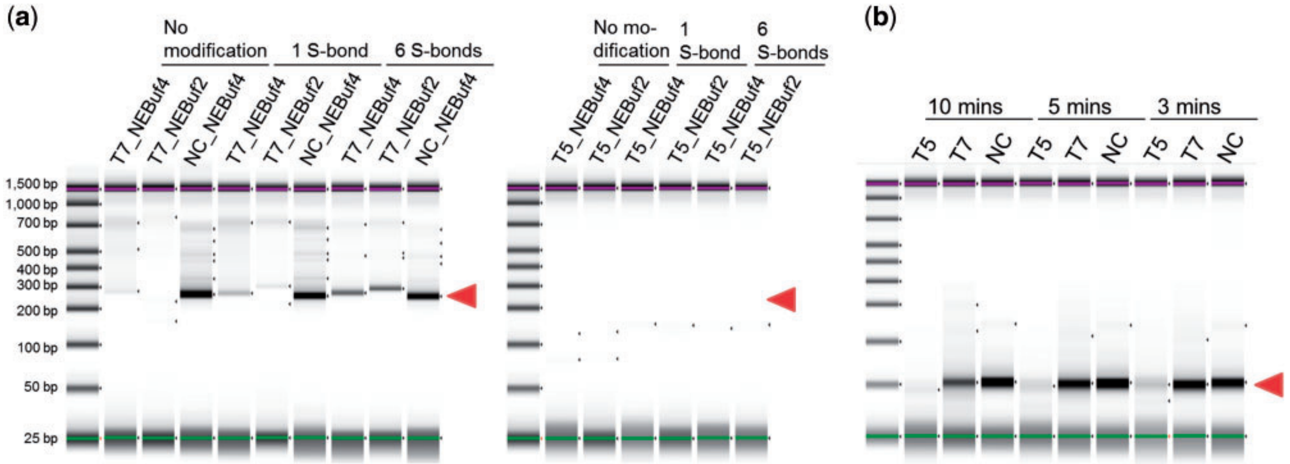

Figure 1: Visualised DNA fragments following the T7 and T5 exonuclease activity assay. (**a**) Effect of S-bond modification on the exonuclease activity. The size of the DNA ladder is shown on the right side of the T7 exonuclease activity assay image. (**b**) Time course assay of the exonuclease. The purple and green lines show the positions of the upper and lower markers of the Agilent D1000 kit, respectively. 'T7' and 'T5' stand for T7 and T5 exonuclease, respectively, and 'NC' stands for 'no-enzyme control', in which molecular biology grade water was used, instead of exonuclease. 'NEBuf2' and 'NEBuf4' denote that the reaction was performed in the NEBuffer 2 and NEBuffer 4, respectively. The position of the target DNA fragments is indicated with a red arrow. Although a slight DNA size difference can be observed between DNA fragments, the difference is within the size resolution of the instrument and kit (15% for DNA fragments between 35 and 300 bp).

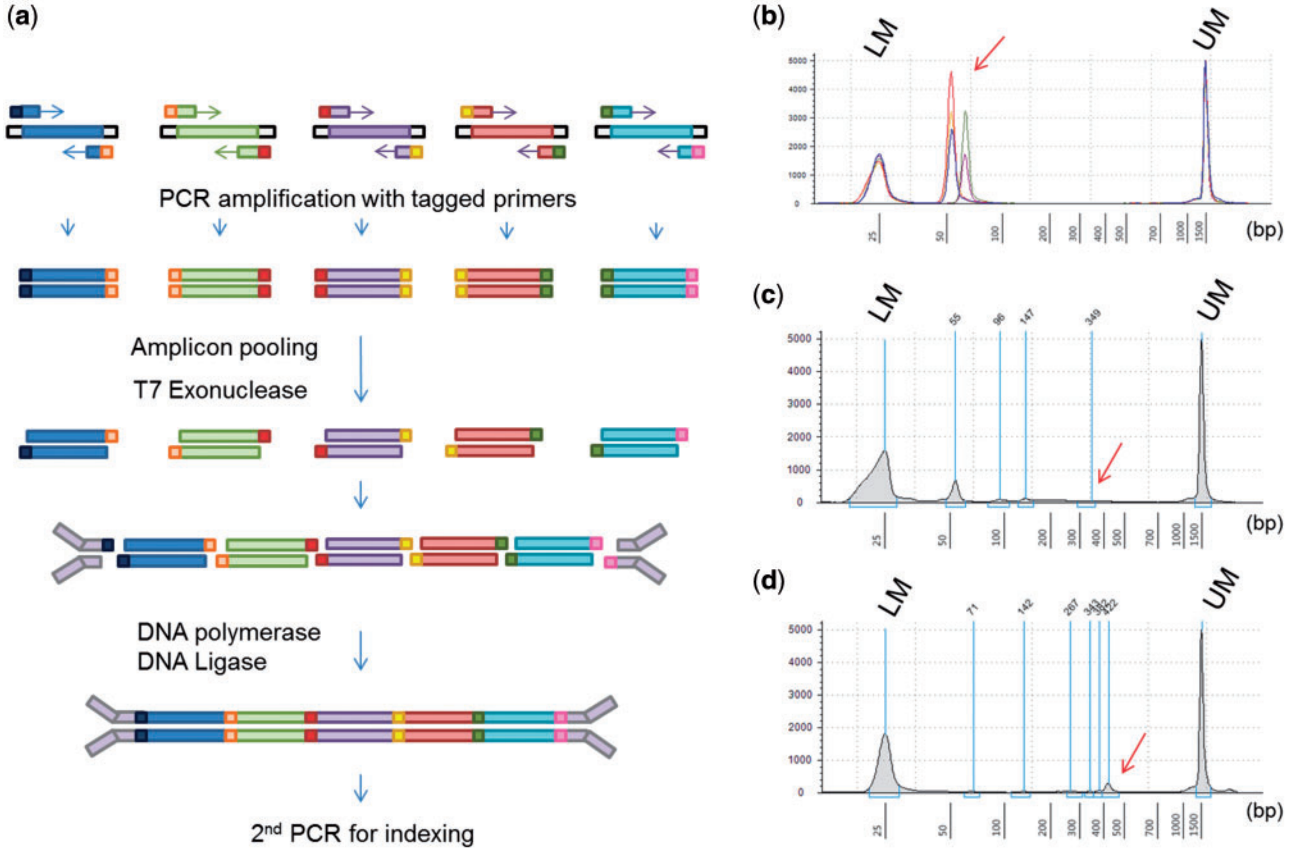

Figure 2: Short DNA fragment assembly-based Illumina library preparation method. (**a**) Procedure of the library preparation method. The target sequences (five regions, which are shown with blue, light green, purple, brown and aqua lines) are amplified using locus-specific primers with assembly tags (dark blue, orange, red, yellow, dark green and pink boxes) from gDNA templates. Following the PCR, partial DNA digestion is performed with T7 exonuclease. The S-bond modification in the PCR primers reduced the nucleotide catalysis in order to protect DNA fragments from excess digestion. Using DNA polymerase and ligase, the DNA fragments and sequencing adapters (grey boxes) are assembled into a single molecule, and the assembled DNA is used for the second PCR. (**b**) PCR amplicons generated for preparation of the *PsySNP*_Set 1 library. The signal peaks between the 50 and 100 bp positions show the PCR amplicons from the Kaspa genotype (**c**) DNA molecules after the assembly of the five PCR amplicons. (**d**) DNA molecules after the second PCR. The y and x axes denote the fluorescence intensity and DNA fragment size, respectively (b, c and d). The desired DNA molecules are indicated with a red arrow. LM and UM indicate the lower and upper markers of the Agilent D1000 kit, respectively.

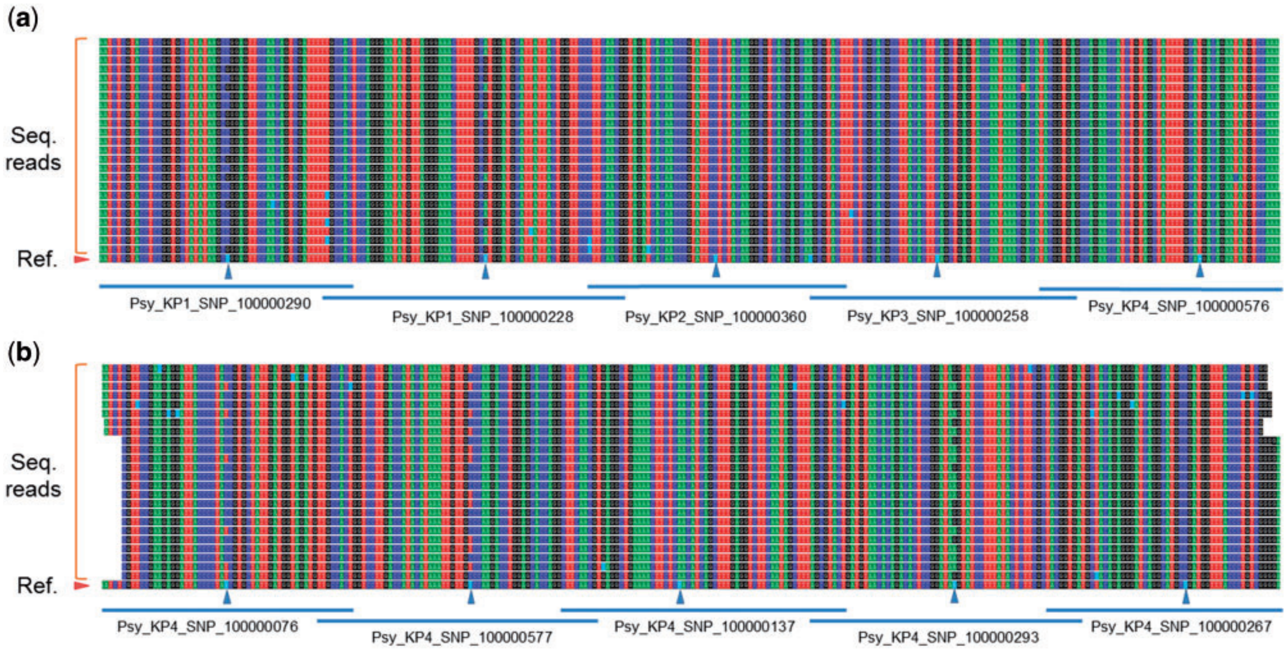

Figure 3: Alignment of sequencing reads to the reference sequences. Sequencing reads (Seq. reads) from the *PsySNP*_Set 1 library of the Psy_RIL677 genotype (**a**) and *PsySNP*_Set 2 library of the PsyRIL99 genotype (**b**) were visualised on the Sequencher software. The reference sequences (Ref.) is shown at the bottom of the alignment. The A, C, G and T bases are shown in green, dark blue, black and red, respectively. A gap (:) is shown in light blue. The position of target SNP is indicated with the blue arrow, and ambiguity codes (light blue) are used to show candidate nucleotides at the SNP sites of the reference sequences. Under the reference sequence, the corresponding region for each PCR fragments including the 8-bp tags is shown with a blue line.

products (*PsySNP*_All) were pooled. The PCR fragments and library prep adapters were assembled into larger single molecules. The expected size of assembled DNA from *PsySNP*_Set 1 samples was 334 bp in length, and a signal peak around the 330–360 bp position was detected using the TapeStation instrument, suggesting the presence of the desired molecules (Fig. 2c). After size-selection, the target DNA was amplified using indexing primers that were designed in-house. Using the TapeStation instrument, a peak at around the 420 bp position, corresponding to the target DNA fragments, was observed along with some signal peaks of shorter DNA molecules (Fig. 2d). Similar results were observed when the *PsySNP*_Set 2 and *PsySNP*_All libraries were analysed on the TapeStation instrument (Supplementary Material 9). The sequencing libraries were pooled for multiplexed sequencing, and the desired short DNA fragments were purified through two cycles of size selection using the AMPure bead kit.

The pooled library was sequenced using a portion of a single Illumina MiSeq run. The raw reads were filtered for the downstream analysis. Totals of 4014–29 005 filtered reads were obtained from each library (Fig. 3). The data were analysed by counting the number of reads containing the target SNP sequence. From the *PsySNP*_Set 1 and 2 libraries, over 4000 counts were obtained for each SNP site (Table 2). From the *PsySNP*_All libraries, the counts varied between 590 and 12 484 depending on the SNP sites, and the lowest count (590) was obtained from the PBA Oura genotype, from which the fewest (4014) filtered reads were obtained (Supplementary Material 10). Among the total of 100 sequenced SNP-containing sites (10 locations × 10 genotypes), 82 and 18 were classified into the homozygote and heterozygote categories, respectively. However, for the 82 homozygous sites, substantial differences in the ratio between the major (positive) and minor (negative) alleles were observed, the prevalence of the minor alleles varying between 0% and 7%, suggesting the possibility of a low level of cross-contamination during the library preparation procedure. Of the 18 heterozygous sites, the corresponding read count ratio was relatively close to 1: 1, except for the Psy_KP4_SNP_100000076 site of the Psy_RIL99 genotype, of which the read count ratio was 72% (C) and 28% (T). As a similar ratio was obtained from both *PsySNP*_Set 2 and *PsySNP*_All libraries, this divergence from the 1: 1 ratio could be attributed to amplification bias during the initial PCR, rather than cross-contamination. Consistent genotyping data were obtained from the *PsySNP*_All, and *PsySNP*_Set 1 and 2 libraries.

For validation of the GBS-based genotyping results, KASP and PCR-RFLP assays were performed. The KASP assay was performed for nine field pea individuals, and the results were consistent with those from GBS-based method, except for those at the Psy_KP4_SNP_100000076 and Psy_KP4_SNP_100000293 sites (Table 2). The GBS-based method suggested a heterozygous allele combination (C/T) at Psy_KP4_SNP_100000076 of the Psy_RIL614 genotype, as compared to a homozygous (T/T) allelic status when using the KASP-based method. In addition, although the GBS-based method categorised 7 individuals into the heterozygous (A/G) class at the Psy_KP4_SNP_100000293 site, those individuals were identified as homozygous (G/G) with the KASP assay. This disagreement may be attributed to use of different PCR primer sets between the two assays. A PCR–RFLP assay for the SNP site was, therefore, performed, using equivalent primers to those for the GBS assay. The results from the PCR–RFLP assay were consistent with the GBS results, suggesting that the use of different PCR primer sets caused the discrepancy (Table 1, Supplementary Materials 4 and 5). A PCR–RFLP assay was also designed for the SNP_100000290 site with the PCR primers used for the GBS assay, and the resulting genotyping data for this SNP site were consistent with those from both GBS and KASP-based methods.

Table 2: Genotyping results from the PsySNP_Set 1, PsySNP_Set 2 sequencing libraries and KASP-based methods

PsySNP_Set 1	Psy_KP1_SNP_100000290				Psy_KP1_SNP_100000228				Psy_KP2_SNP_100000360				Psy_KP3_SNP_100000258				Psy_KP4_SNP_100000576			
	GBS			KASP	GBS			KASP	GBS			KASP	GBS			KASP	GBS			KASP
	C	G	Alleles		A	G	Alleles		C	T	Alleles		C	G	Alleles		A	T	Alleles	
Kaspa	**6414 (99%)**	32 (1%)	C/C	C/C	51 (1%)	**6547 (99%)**	G/G	G/G	**6386 (99%)**	33 (1%)	C/C	C/C	**5391 (100%)**	11 (0%)	C/C	C/C	21 (0%)	**5255 (100%)**	T/T	T/T
PBAOura	238 (2%)	**12 174 (98%)**	G/G	G/G	**12 675 (99%)**	108 (1%)	A/A	A/A	151 (1%)	**12 804 (99%)**	T/T	T/T	135 (1%)	**18 669 (99%)**	G/G	G/G	53 (0%)	**20 230 (100%)**	T/T	T/T
Psy_GenoX	**11 366 (98%)**	235 (2%)	C/C	NA	**11 867 (99%)**	136 (1%)	A/A	NA	150 (1%)	**11 930 (99%)**	T/T	N.A.	208 (1%)	**20 975 (99%)**	G/G	N.A.	**24 285 (99%)**	217 (1%)	A/A	N.A.
Psy_RIL99	**12 777 (99%)**	116 (1%)	C/C	C/C	73 (1%)	**13 462 (99%)**	G/G	G/G	**14 004 (99%)**	85 (1%)	C/C	C/C	**15 000 (100%)**	13 (0%)	C/C	C/C	**7625 (50%)**	**7596 (50%)**	A/T	A/T
Psy_RIL195	247 (2%)	**13 010 (98%)**	G/G	G/G	**13 870 (99%)**	118 (1%)	A/A	A/A	192 (1%)	**14 082 (99%)**	T/T	T/T	**21 636 (100%)**	61 (0%)	C/C	C/C	90 (0%)	**21 955 (99%)**	T/T	T/T
Psy_RIL268	76 (1%)	**11 207 (99%)**	G/G	G/G	**11 927 (99%)**	54 (0%)	A/A	A/A	**11 972 (100%)**	38 (0%)	C/C	C/C	**12 913 (100%)**	28 (0%)	C/C	C/C	33 (0%)	**13 061 (100%)**	T/T	T/T
Psy_RIL614	**13 825 (100%)**	54 (0%)	C/C	C/C	56 (0%)	**14 541 (100%)**	G/G	G/G	**15 081 (100%)**	46 (0%)	C/C	C/C	**16 374 (100%)**	23 (0%)	C/C	C/C	**16 505 (99%)**	120 (1%)	A/A	A/A
Psy_RIL656	**16 607 (100%)**	54 (0%)	C/C	C/C	63 (0%)	**17 157 (100%)**	G/G	G/G	**17 945 (100%)**	43 (0%)	C/C	C/C	**10 307 (54%)**	**8661 (46%)**	C/G	C/G	56 (0%)	**19 880 (100%)**	T/T	T/T
Psy_RIL677	**5749 (50%)**	**5773 (50%)**	C/G	C/G	**5744 (48%)**	**6100 (51%)**	A/G	A/G	**12 109 (100%)**	6 (0%)	C/C	C/C	**12, 670 (100%)**	3 (0%)	C/C	C/C	15 (0%)	**12 729 (100%)**	T/T	T/T
Psy_RIL678	**6656 (51%)**	**6406 (49%)**	C/G	C/G	**6962 (52%)**	**6436 (48%)**	A/G	A/G	**13 609 (100%)**	22 (0%)	C/C	C/C	**13, 998 (100%)**	20 (0%)	C/C	C/C	15 (0%)	**14 118 (100%)**	T/T	T/T
PsySNP_Set 2	Psy_KP4_SNP_100000076				Psy_KP4_SNP_100000577				Psy_KP4_SNP_100000137				Psy_KP4_SNP_100000293				Psy_KP4_SNP_100000267			
	GBS			KASP	GBS			KASP	GBS			KASP	GBS			KASP	>GBS			KASP
	C	T	Alleles		C	T	Alleles		A	C	Alleles		A	G	Alleles		C	G	Alleles	
Kaspa	**6915 (100%)**	5 (0%)	C/C	C/C	8 (0%)	**6912 (100%)**	T/T	T/T	35 (1%)	**6938 (100%)**	C/C	C/C	**6924 (100%)**	21 (0%)	A/A	A/A	8 (0%)	**5965 (100%)**	G/G	G/G
PBAOura	**6815 (100%)**	5 (0%)	C/C	C/C	18 (0%)	**6695 (100%)**	T/T	T/T	**6860 (100%)**	6 (0%)	A/A	A/A	**3383 (50%)**	**3415 (50%)**	A/G*	G/G	**6280 (99%)**	24 (0%)	C/C	C/C
Psy_GenoX	**2791 (49%)**	**2859 (51%)**	C/T	NA	**4971 (100%)**	11 (0%)	C/C	N.A.	**5072 (100%)**	3 (0%)	A/A	N.A.	10 (0%)	**5041 (100%)**	G/G	N.A.	**4628 (100%)**	8 (0%)	C/C	N.A.
Psy_RIL99	**4510 (72%)**	**1790 (28%)**	C/T	C/T	**3203 (50%)**	**3257 (50%)**	C/T	C/T	**6542 (100%)**	3 (0%)	A/A	A/A	**3209 (49%)**	**3282 (51%)**	A/G*	G/G	**6134 (100%)**	7 (0%)	C/C	C/C
Psy_RIL195	**6351 (100%)**	3 (0%)	C/C	C/C	13 (0%)	**6605 (100%)**	T/T	T/T	23 (0%)	**6647 (100%)**	C/C	C/C	**6616 (100%)**	14 (0%)	A/A	A/A	16 (0%)	**5786 (100%)**	G/G	G/G
Psy_RIL268	**6038 (100%)**	6 (0%)	C/C	C/C	11 (0%)	**6207 (100%)**	T/T	T/T	**6300 (100%)**	7 (0%)	A/A	A/A	**3326 (53%)**	**2961 (47%)**	A/G*	G/G	**2453 (48%)**	**2638 (52%)**	C/G	C/G
Psy_RIL614	**2988 (46%)**	**3480 (54%)**	C/T*	T/T	**6577 (100%)**	14 (0%)	C/C	C/C	**6659 (100%)**	6 (0%)	A/A	A/A	**3427 (51%)**	**3230 (49%)**	A/G*	G/G	**6414 (100%)**	7 (0%)	C/C	C/C
Psy_RIL656	**7092 (100%)**	4 (0%)	C/C	C/C	10 (0%)	**7140 (100%)**	T/T	T/T	**7221 (100%)**	3 (0%)	A/A	A/A	**3853 (54%)**	**3286 (46%)**	A/G*	G/G	6 (0%)	**6420 (100%)**	G/G	G/G
Psy_RIL677	**8555 (100%)**	3 (0%)	C/C	C/C	13 (0%)	**8721 (100%)**	T/T	T/T	**8827 (100%)**	8 (0%)	A/A	A/A	**4597 (52%)**	**4226 (48%)**	A/G*	G/G	6 (0%)	**6769 (100%)**	G/G	G/G
Psy_RIL678	**9268 (100%)**	2 (0%)	C/C	C/C	9 (0%)	**9443 (100%)**	T/T	T/T	**9519 (100%)**	3 (0%)	A/A	A/A	**4670 (49%)**	**4784 (51%)**	A/G*	G/G	5 (0%)	**7915 (100%)**	G/G	G/G

The number of reads corresponding to each allele is shown for the GBS-based method, of which major (positive) allele(s) are shown in bold. An asterisk (*) denotes that the GBS-based genotyping result is not consistent with that from the KASP-based method, but the PCR-RFLP-based method supported the GBS-based result. NA stands for 'not analysed'.

Target enrichment procedures are used to select genomic regions of interest, which generally represent only 1% or less of the whole genome, in order to reduce duration and cost of the sequencing process [14–16]. The efficiencies of target enrichment methods have been evaluated as the proportion of desired sequence among total reads which passed the quality filtering. This ratio may be effectively increased to over 0.7 using PCR- or hybridization-based enrichment methods [14]. Following this definition, the efficiency of sequencing-based SNP genotyping would be represented as the average number of SNP sites sequenced per read. When a complex multi-allelic gene (locus) is targeted, it may be possible to sequence DNA fragments including multiple SNP sites, even with the short read-sequencing instruments. Bi-allelic SNP sites in simpler sequence contexts are, however, more commonly used for whole-genome genotyping experiments and/or for marker-assisted selection schemes, and it has therefore been unusual to capture more than a single SNP site in each sequencing read. The current study has demonstrated a simple sequencing library preparation method, based on assembly of short PCR fragments into larger DNA molecules. The average number of sequenced SNP sites per read from each field pea-derived library was between 2.8 and 4.7, and the average across all libraries was 4.3 with a standard deviation of 0.48 (Table 3). As multiple short fragments can be sequenced in a single read, this library preparation method may substantially reduce the sequencing cost for SNP-based GBS, as well as sRNA sequencing, MIP-based mutant detection and eDNA characterised.

Table 3: Enrichment efficiency of the short fragment assembly-based library preparation method

Genotype		PsySNP_ Set 1	PsySNP_ Set 2	PsySNP_ All	Total	
Kaspa		4.04	4.60	4.52		
PBAOura		3.38	4.57	4.01		
Psy_GenoX		2.81	4.21	4.30		
Psy_RIL99		4.01	4.70	4.55		
Psy_RIL195		3.12	4.61	4.17		
Psy_RIL268		3.96	4.41	4.48		
Psy_RIL614		3.71	4.73	4.53		
Psy_RIL656		3.78	4.62	4.62		
Psy_RIL677		4.52	4.56	4.39		
Psy_RIL678		4.49	4.63	4.56		
	Total	3.78	4.56	4.41	Average	4.25
	SD	0.55	0.15	0.20	SD	0.48

The average number of sequenced SNP sites in each read is shown from each sequencing library. 'SD' stands for standard deviation.

With PCR-based enrichment methods, Illumina sequencing libraries are commonly prepared through two rounds of PCR [7]. Target sequences are initially amplified from gDNA templates through single or low-plex PCRs, using PCR primers containing locus-specific sequences and a part of the sequencing adapter, and the rest of the adapter sequences are then attached to both termini of the target fragments through second-round PCR. Although this procedure is simple and effective, as DNA ligase or transposase is not required, a high level of multiplexing in the first PCR may not be feasible for Illumina sequencing instruments, due to the length of the tagged PCR primers. The present method uses short tags, and the sequences of interest could be effectively amplified through a multiplexed PCR. Through multi(penta)-plexed PCR with the primer sets of PsySNP_Set 1, the short DNA fragments were simultaneously amplified from gDNA of the Kaspa, PBA Oura and Psy_GenoX genotypes. Additionally, short DNA assembly was successfully performed using the products from the penta-plexed PCR (Fig. 4), suggesting that this approach may be suitable for a highly multiplexed PCR-based target enrichment method.

The current study has demonstrated an efficient assembly procedure for short dsDNA fragments. Although bacterial genomes (600 kb–1 Mb) have been effectively synthesised using the original homology-based enzymatic DNA fragment assembly (Gibson assembly) scheme, this approach is not suitable for assembly of short DNA fragments [9, 10]. One of the obstacles to short fragments assembly has been a requirement for relatively long overlapping sequences. The present method performed the assembly with 8 bp homologous sequences (tags) which are shorter than the 15–80 bp sequences used in the original protocol, and the method may be useful for rapid synthesis of DNA fragments up to several kilobases in length, such as synthetic virus genomes for the purpose of vaccine production [17].

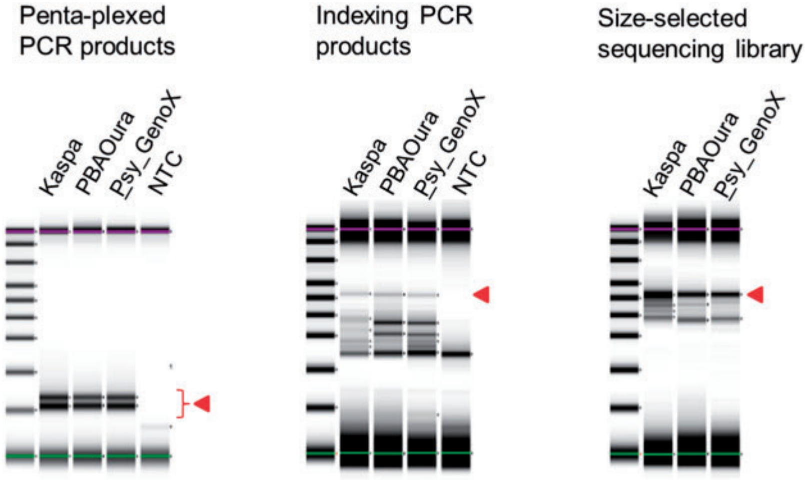

Figure 4: Sequencing libraries prepared through penta-plexed PCR. Products and sequencing libraries after PCR-enrichment and size-selection were visualised on the 2200 TapeStation instrument using the D1000 Kit. The target DNA is indicated with a red arrow. NTC stands for no template control.

Supplementary data

Supplementary data is available at *Biology Methods and Protocols Journal* online.

Acknowledgements

The authors would like to thank Profs John Forster and Ben Cocks for technical advice and proof reading.

Funding

This work was supported by funding from the Victorian Department of Economic Development, Jobs, Transport and Resources.

Authors' contributions

H.S., S.S. and M.S. contributed to the experimental work and data analysis. N.O.I.C. provided overall project leadership. H.S. prepared the primary drafts of the article and contributed to finalisation of the text. N.O.I.C. and S.S. co-developed interim and final drafts of the article.

Conflict of interest statement. None declared.

References

1. Bentley DR, Balasubramanian S, Swerdlow HP *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;**456**:53–9.
2. Chi KR. The year of sequencing. *Nat Methods* 2008;**5**:11–4.
3. Hillier LW, Marth GT, Quinlan AR *et al*. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* 2008;**5**:183–8.
4. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. *Genomics* 2016;**107**:1–8.
5. Nagy T, Kis A, Poliska S, Barta E *et al*. [Letter to the Editor] Comparison of small RNA next-generation sequencing with and without isolation of small RNA fraction. *BioTechniques* 2016;**60**:273–8.
6. Stefan CP, Koehler JW, Minogue TD. Targeted next-generation sequencing for the detection of ciprofloxacin resistance markers using molecular inversion probes. *Scientific Reports* 2016;**6**:25904.
7. Miya M, Sato Y, Fukunaga T *et al*. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Roy Soc Open Sci* 2015;**2**:150088.
8. Gibson DG. Enzymatic assembly of overlapping DNA fragments. *Meth Enzymol* 2011;**498**:349–61.
9. Gibson D. One-step enzymatic assembly of DNA molecules up to several hundred kilobases in size. *Protocol Exchange* 2009; doi: 10.1038/nprot.2009.77.
10. Hillson NJ. DNA assembly method standardization for synthetic biomolecular circuits and systems. In Koeppl H, Setti G, di Bernardo M and Densmore D (eds), *Design and Analysis of Biomolecular Circuits*, Springer, New York, NY, 2011, 295–314.
11. Javid M, Rosewarne GM, Sudheesh S *et al*. Validation of molecular markers associated with boron tolerance, powdery mildew resistance and salinity tolerance in field peas. *Front Plant Sci* 2015;**6**:917.
12. Sudheesh S, Rodda M, Kennedy P *et al*. Construction of an integrated linkage map and trait dissection for bacterial blight resistance in field pea (*Pisum sativum* L.). *Mol Breeding* 2015;**35**:185.
13. Shinozuka H, Forster JW. Use of the melting curve assay as a means for high-throughput quantification of Illumina sequencing libraries. *Peer J* 2016;**4**:e2281.
14. Mamanova L, Coffey AJ, Scott CE *et al*. Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010; **7**:111–18.
15. García-García G, Baux D, Faugère V *et al*. Assessment of the latest NGS enrichment capture methods in clinical context. *Sci Rep* 2016;**6**:20948.
16. Bodi K, Perera AG, Adams PS *et al*. Comparison of commercially available target enrichment methods for next-generation sequencing. *J Biomol Tech* 2013;**24**:73–86.
17. Dormitzer PR, Suphaphiphat P, Gibson DG *et al*. Synthetic generation of influenza vaccine viruses for rapid response to pandemics. *Sci Transl Med* 2013;**5**:185ra68.