bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

More Signal, Less Noise: Single-primer Adenylation Method (SAM) for the Efficient Preparation of DNA for Transposon Sequencing (TnSeq)

Willcocks, S[I]; Stabler, R; Wren, B

[I]Address for Correspondence: The London School of Hygiene and Tropical Medicine, Keppel St, London, WC1E 7HT; email: sam.willcocks@lshtm.ac.uk; telephone: 02079272461

KEYWORDS: TnSeq; TraDIS; Next Generation Sequencing

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

Transposon Sequencing (TnSeq) is a cutting-edge tool that allows quantitative, genome-wide genetic analysis of a given population in an experimental condition. TnSeq utilises next generation sequencing that is designed to perform optimally when all of the input material is of interest, therefore TnSeq is inefficient since only a small portion of nucleic acid fragments contain the transposon target site for primer annealing. We describe an innovative modification to standard transposon library preparation protocols that provides a significant improvement to the quality of sequencing data recovered.

INTRODUCTION

The advent of next generation sequencing is revolutionising infectious disease research and holds great promise, particularly in elucidating the evolution and spread of organisms and in clinical diagnostics (Alfoldi and Lindblad-Toh 2013). When coupled with the creation of genome-wide transposon-mutant libraries, transposon-sequencing (TnSeq) has emerged as a powerful tool that allows the simultaneous assessment of every gene in a given genome, revealing essentiality in the chosen model - be this antibiotic resistance, virulence, or simply survival in optimal growth media (van Opijnen, Bodi et al. 2009; Gallagher, Shendure et al. 2011; Moule, Hemsley et al. 2014). TnSeq requires first the generation of a transposon-mediated mutant library that saturates the genome, so that every gene carries multiple insertions; individual mutants are grown on agarose plates and then collected and pooled prior to experimentation. Sequencing of the unknown region of genomic (g)DNA flanking the transposon reveals which gene has been mutated. Comparison of the genes represented in the ‘output’ versus the ‘input’ library reveals, in a quantitative manner, negative selection events *ie* mutation of genes that are required for survival in the test condition precludes

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

their detection in the output pool (Chaudhuri, Morgan et al. 2013). The applications of TnSeq are vast; as with all new technologies, however, there are technical hurdles.

RESULTS AND DISCUSSION

The widely used protocol for preparing DNA for TnSeq, as with Illumina related platforms, involves the fragmentation of extracted DNA, ligation of adaptor sequences, and the selective amplification of transposon-containing fragments (van Opijnen, Lazinski et al. 2014). What separates TnSeq from standard sequencing of gDNA or RNA-derived extracts, is that only a very small percentage of the genetic material contains the target of interest: the transposon (**Fig 1A**). This radically reduces the efficiency and statistical power of next generation sequencing. A common solution is to amplify, by means of polymerase chain reaction (PCR) with specific oligos, only fragments of DNA that contain the transposon. PCR is crucial to TraDIS since it is also used to introduce the sequence that allows the DNA fragments to anneal to the surface of the flow cell. While PCR improves the signal to noise ratio of target to background DNA, it is also the main step through which bias is introduced into the data (Kozarewa, Ning et al. 2009; Aird, Ross et al. 2011). Bias seems to be exacerbated in low-complexity (hundreds-scale pools) libraries, and somewhat masked in very large (hundreds of thousands scale) libraries.

Limiting amplification bias through primer design and changes to cycle parameters such as melting time and ramp speed may provide modest improvements to PCR efficiency. However, since the target site for primer annealing is conserved regardless of organism or location of the transposon, this is not a satisfactory candidate for explaining PCR amplification bias. More crucial are the characteristics of the nucleic acids outside the region of the transposon including sequence, fragment length, and GC content, that can all adversely affect melting temperature and primer annealing efficiency through secondary structure formation (Day, Speiser et al. 1996; Barnard, Futo et al. 1998; Ogino and Wilson 2002; Dohm, Lottaz et al. 2008). For this reason, TnSeq of organisms with unusually high (*Burkholderia spp*) or low (*Yersinia spp*) GC content, or even those with relatively

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

balanced GC content, but with interspersed regions of high GC content (*Escherichia coli*), can be problematic. Bias in amplification can result in apparent 'hotspots' that are not explained by the natural favouring by the transposon of particular conserved sequences. Worse, it can result in false conclusions of positivity or negativity.

It has been noted by other researchers, that lowering the threshold of PCR to no more than 10 cycles allows enrichment of the transposon-containing fragments, while limiting the bias to an acceptable level. However, this exacerbates the problem of low signal to background with TnSeq. The user is required to decide the optimal DNA concentration to apply to the flow cell, which requires a narrow range for optimal results. One either normalises DNA concentration to the amount of total DNA, in which case the percentage of clusters that contain transposon sequence is very small, or one normalises to the concentration of transposon-only DNA, in which case the non-transposon 'background' DNA is so high that it prevents capture of good quality images by the sequencing machine software. This desultory element in terms of choosing PCR cycle number or DNA concentration is prohibitive; it is expensive in time and money and more importantly in precious samples that may not readily be re-attainable.

One solution to the issue of high background is to attempt to recover or enrich the transposon-positive fragments post-PCR. We have investigated methods including the enzymatic digest of non-PCR DNA, theoretically leaving only the transposon-positive amplicons, as well as the use of biotinylated primers and nano-bead capture technology to physically separate the target DNA of interest. However, with both of these methods, we found a loss of yield and no significant enhancement of signal to background as assessed by quantitative PCR.

We presently report an innovation that allows for the exclusive binding of transposon-containing DNA fragments to the flow cell, resulting in clean, high quality data that is highly reproducible. The use of the Single-primer Adenylation Method (SAM) for TnSeq obviates the need to find an elusive balance between having enough cycles to amplify the transposons whilst keeping bias to a

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

minimum. We are able to routinely use just 10 cycles of PCR, even on low complexity libraries to prepare a library with very limited bias, with negligible background from non-transposon DNA. Essentially, more signal and less noise.

In order for any DNA to bind to the Illumina MiSeq flow cell surface during a sequencing run, it must contain at least one of two defined sequences, ‘P5’ and ‘P7’; these are complementary to a lawn of primers that are pre-coated onto the flow cell. The P5 and P7 sequences are introduced by PCR that uses adapter sequences as a template for reverse strand synthesis, to introduce P7, and the transposon serves as the template for the forward primer, to introduce P5. Since PCR amplification with this primer pair only occurs with transposon-containing fragments, one may expect that only transposon-containing fragments have the P5 and P7 sequence introduced and therefore only transposon fragment are able to bind to the flow cell.

However, non-transposon DNA fragments are present in much higher amounts than transposon-containing fragments. It is therefore possible that, although no amplification takes place, a single extension of complementary DNA that carries the P7 sequence on the reverse primer would occur with the large proportion of non-transposon DNA that would subsequently be capable of binding to the flow cell. This portion of DNA would not be detected by qPCR methods that use either P5 or transposon sequence for the forward primer.

The use of adapters is essential, since it serves as a template for the reverse primer during amplification of transposon-containing fragments. The ligation process depends on the indiscriminate adenylation of the 3’ end of all DNA fragments using an enzyme such as Klenow fragment, exo-. Alternatively, by enzymatic digest of the transposon to leave an overhang, the efficiency of which depends on the frequency of the chosen restriction sequence in the target genome and usually, sequence modification of the transposon itself (van Opijnen, Lazinski et al. 2014). This is the key step that we have revised, so that only transposon-containing fragments receive the A-tail, forbidding the generation of P7-tagged non-transposon fragments and with no

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

need for restriction digest. In our revised protocol: genomic DNA extraction, fragmentation and end repair still occur as per standard preparation protocols. The critical step is to conduct a PCR using a single transposon-specific primer to introduce a 3′ A-overhang using *Taq* polymerase, instead of adenylating all fragments using Klenow fragment, exo-. The forward primer ideally is identical to the transposon-specific primer carrying the P5 tag that is used in the subsequent amplification PCR (**Fig 1B**).

The need for high fidelity is clear, and there are numerous products available commercially that contain a blend of a high fidelity polymerase with a standard, A-tailing *Taq* polymerase and the PCR parameters are as per manufacturer's recommendations for the chosen polymerase. Since there is no amplification during single-primer adenlyation, multiple cycles may be used without concern of bias introduction. Additives to assist with GC melting and an extended melting cycle time can also be used effectively at the discretion of the user, depending on the requirements of the source material. Additionally, there is the potential for size selection during single-primer adenylation: by controlling the extension time, one can dictate the size of fragments that are A-tailed. We recommend either a slow ramp-speed or a final slow cool-down step to assist correct annealing of the A-tailed strand with its complementary template.

After single-primer adenylation, the result is that the only location for the adapters to ligate to, are the ends of the transposon-positive fragments. Our data suggest that at least 50% of transposon-positive fragments are A-tailed this way.

The samples are then ready for amplification PCR as per standard TnSeq protocols, using the P5-containing transposon-specific forward primer, and the P7-containing adapter-specific reverse primer. Splitting the input material into several independent reaction tubes to be pooled post-PCR is recommended to reduce PCR amplification bias, as is a maximum of 10 cycles. Adapter-ligated transposon-containing fragments may be quantified using P5 and P7 specific primers for quantitative PCR prior to sequencing using Illumina MiSeq.

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SAM for TnSeq in contrast with existing methodology for DNA preparation is summarised in Fig1:

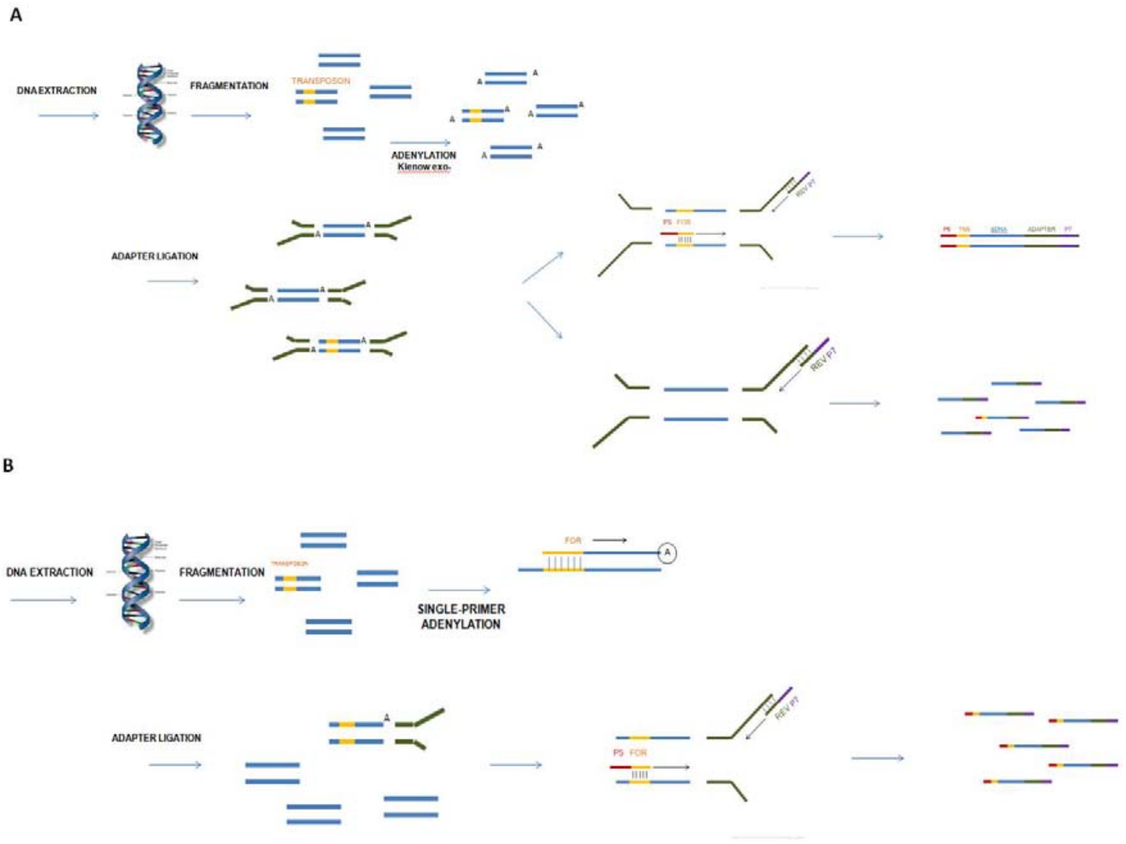

Fig 1 Single-primer Adenylation Method (SAM) for TnSeq. In the standard DNA preparation protocol for TnSeq (**A**), DNA fragments (blue) are indiscriminately adenylated, resulting in all fragments ligating with adapters (green) regardless of whether they contain the transposon (yellow). Even in the absence of transposon DNA, the reverse primer that introduces the P7 flow-cell binding sequence (purple) may still be introduced in the majority of fragments, although they are not subsequently amplified by PCR. This reduces the signal:background ratio and reduces efficiency of TnSeq. Using the Single-primer Adenylation Method (SAM) (**B**), a transposon-specific primer is used to selectively introduce an A-tail only to transposon-containing fragments. Subsequently, only these fragments are able to ligate with adapters, which may then be amplified by PCR in the usual manner.

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Only transposon-positive fragments are amplified and able to bind to the flow cell surface, virtually eradicating background signal from the majority of DNA that does not contain transposon DNA. It is possible that some clusters are still generated that seemingly do not yield useful transposon location information. This may be due to non-specific annealing of fragments to the flow-cell primers, which by definition are able to form clusters since the 'Read One' primer used by Illumina is related to the flow cell primer sequence. Alternatively, clusters may indeed contain transposon-containing sequences that are missed from post-sequencing analysis algorithms due to mismatches, and this will depend on the stringency demands set by the user.

The efficiency of SAM for TnSeq is enhanced over standard preparation protocols in two respects. First, only transposon-positive fragments are able to be ligated to the adapter sequence to serve as a template for amplification PCR. Second, only these fragments are able to receive the P5 sequence required for maximal binding to the flow cell, as this is included in the forward primer during the amplification step. With our method, non-transposon DNA fragments possess neither P7 nor P5 sequences, whereas in standard protocols, they contain both. This results in less total DNA binding to the flow-cell, and less cluster formation, but significantly more of these clusters map to the transposon-containing regions of DNA. Measurement of total DNA by methods such as Bioanalyzer or Nanodrop is of limited value in SAM for TnSeq, since this quantifies DNA that is not capable of binding to the flow-cell. Therefore qPCR using transposon-specific primers is essential to adjust the DNA to the appropriate concentration. Using SAM for TnSeq, quantification that uses a P7-specific primer is more accurate than the standard protocol, since P7 is only carried by transposon-containing fragments.

In conclusion, our method greatly enhances the efficiency of TnSeq, by virtue of a simple modification to an early step in the preparation of DNA for sequencing. The revised method is applicable for all TraDIS experiments, and is of particular value for researchers with challenging, GC-polarised genomic material, and those with small, low complexity libraries.

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MATERIALS AND METHODS

Genomic DNA is extracted from the experimental organism followed by fragmentation by either sonic disruption or a frequently-cutting restriction enzyme and then purified. The ends of the DNA fragment are repaired to make blunt ends using (for a 75 µl sample): 10 µl phosphorylation buffer, 4 µl dNTPs, 1.5 µl Klenow DNA polymerase, 5 µl T4 PNK, 5 µl T4 DNA polymerase at room temperature (RT) for 30 mins. The reaction is purified and then subjected to single-primer adenlyation. A blended polymerase containing A-tailing *taq* and a proof-reading enzyme is recommended, such as Advantage Polymerase (Takara Clontech) and the PCR parameters are as per manufacturer's recommendations for the chosen polymerase. The forward primer ideally is identical to the transposon-specific primer carrying the P5 tag that is used in the subsequent amplification PCR (using a high fidelity enzyme such as NEBNext or Phusion). Since there is no amplification during single-primer adenlyation, multiple cycles may be used without concern of bias introduction. If desired, the extension time may be intentionally restricted to limit the size of fragments that are A-tailed. We recommend either a slow ramp-speed or a final slow cool-down step to assist correct annealing of the A-tailed strand with its complementary template. Following PCR purification, DNA fragments are ligated with annealed adapters according to standard TnSeq procedure and again purified. The samples are then ready for amplification PCR, using the P5-containing transposon-specific forward primer, and the P7-containing adapter-specific reverse primer. Splitting the input material into several independent reaction tubes to be pooled post-PCR is recommended to reduce PCR amplification bias, as is a maximum of 10 cycles. Adapter-ligated transposon-containing fragments may be quantified using P5 and P7 specific primers for quantitative PCR prior to sequencing using Illumina MiSeq.

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PCR parameters for Single-Primer Adenylation Using Advantage Polymerase:

94 °C for 3 mins

94 °C for 1 min

60 °C for 1 min   < 30 cycles

68 °C for 3 mins

95 °C for 2 mins

72 °C for 5 mins   Slow cool-down

50 °C for 5 mins

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

Aird, D., M. G. Ross, et al. (2011). "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries." Genome Biol 12(2): R18.

Alfoldi, J. and K. Lindblad-Toh (2013). "Comparative genomics as a tool to understand evolution and disease." Genome Res 23(7): 1063-1068.

Barnard, R., V. Futo, et al. (1998). "PCR bias toward the wild-type k-ras and p53 sequences: implications for PCR detection of mutations and cancer diagnosis." Biotechniques 25(4): 684-691.

Chaudhuri, R. R., E. Morgan, et al. (2013). "Comprehensive assignment of roles for Salmonella typhimurium genes in intestinal colonization of food-producing animals." PLoS Genet 9(4): e1003456.

Day, D. J., P. W. Speiser, et al. (1996). "Identification of non-amplifying CYP21 genes when using PCR-based diagnosis of 21-hydroxylase deficiency in congenital adrenal hyperplasia (CAH) affected pedigrees." Hum Mol Genet 5(12): 2039-2048.

Dohm, J. C., C. Lottaz, et al. (2008). "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing." Nucleic Acids Res 36(16): e105.

Gallagher, L. A., J. Shendure, et al. (2011). "Genome-scale identification of resistance functions in Pseudomonas aeruginosa using Tn-seq." MBio 2(1): e00315-00310.

Kozarewa, I., Z. Ning, et al. (2009). "Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes." Nat Methods 6(4): 291-295.

Moule, M. G., C. M. Hemsley, et al. (2014). "Genome-wide saturation mutagenesis of Burkholderia pseudomallei K96243 predicts essential genes and novel targets for antimicrobial development." MBio 5(1): e00926-00913.

Ogino, S. and R. B. Wilson (2002). "Quantification of PCR bias caused by a single nucleotide polymorphism in SMN gene dosage analysis." J Mol Diagn 4(4): 185-190.

bioRxiv preprint doi: https://doi.org/10.1101/012732; this version posted December 13, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

van Opijnen, T., K. L. Bodi, et al. (2009). "Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms." Nat Methods 6(10): 767-772.

van Opijnen, T., D. W. Lazinski, et al. (2014). "Genome-Wide Fitness and Genetic Interactions Determined by Tn-seq, a High-Throughput Massively Parallel Sequencing Method for Microorganisms." Curr Protoc Mol Biol 106: 7 16 11-17 16 24.