Preparation of high quality next generation sequencing libraries from picogram quantities of target DNA

Nicholas J Parkinson, Siarhei Maslau, Ben Ferneyhough, et al.

*Genome Res.*, published online November 16, 2011
Access the most recent version at doi:10.1101/gr.124016.111

Supplemental Material

http://genome.cshlp.org/content/suppl/2011/11/09/gr.124016.111.DC1.html

Published online November 16, 2011 in advance of the print journal.

This manuscript is Open Access.

Peer-reviewed and accepted for publication but not copyedited or typeset; preprint is likely to differ from the final, published version.

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Copyright © 2011, Cold Spring Harbor Laboratory Press
Title:

Preparation of high quality next generation sequencing libraries from picogram quantities of target DNA

Authors and Affiliations:

Nicholas J. Parkinson¹*, Siarhei Maslau¹,²*, Ben Ferneyhough¹, Gang Zhang¹, Lorna Gregory³, David Buck³, Jiannis Ragoussis³, Chris P. Ponting², Michael D. Fischer¹,⁴.

1 Systems Biology Laboratory UK, 127 Milton Park, Abingdon, Oxfordshire, OX14 4SA
2 MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, OX1 3QX
3 Genomics Laboratory, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK
4 Division of Clinical Sciences, Infection and Immunity Research Centre, St. Georges University of London, Cranmer Terrace, London, UK

* These authors have contributed equally to this work.

Correspondence:

Nicholas J. Parkinson
Systems Biology Laboratory UK,
127 Milton Park,
Abingdon,
Oxfordshire,
OX14 4SA
Email: nickp@sbl-uk.org
Work: 01235 827401
Fax: 01235 834965

Running Title: Next Generation libraries from picograms of DNA

Keywords: Next generation sequencing, library preparation, picogram input,
Abstract

New sequencing technologies can address diverse biomedical questions but are limited by a minimum required DNA input of typically 1μg. We describe how sequencing libraries can be reproducibly created from 20pg input DNA using a modified transpososome-mediated fragmentation technique. Resulting libraries incorporate in-line barcoding which facilitates sample multiplexes that can be sequenced using Illumina platforms with the manufacturer’s sequencing primer. We demonstrate this technique by providing deep coverage sequence of the *E. coli* K-12 genome that shows equivalent target coverage to a 1μg input library prepared using standard Illumina methods. Reducing template quantity does, however, increase the proportion of duplicate reads and enriches coverage in low GC regions. This finding was confirmed with exhaustive re-sequencing of a mouse library constructed from 20pg gDNA input (approximately 7 haploid genomes) resulting in approximately 0.4-fold statistical coverage of uniquely mapped fragments. This implies that a near-complete coverage of the mouse genome is obtainable with this approach using 20 genomes as input. Application of this new method now allows genomic studies from low mass samples and routine preparation of sequencing libraries from enrichment procedures.
Next Generation Sequencing, NGS, methods produce millions of short reads that are either subsequently compared to a reference genome \textit{in silico} or reassembled to provide \textit{de novo} target sequence data (Metzker. 2010). In addition to creating large datasets from an individual target sequence, methods exist to pool multiple, uniquely identifiable, sample libraries that can be de-multiplexed \textit{in silico} following sequencing thereby lowering overall costs of acquiring datasets for low complexity samples.

A common starting point for template preparation for NGS platforms is random fragmentation of target DNA and addition of platform-specific adapter sequences to flanking ends. Protocols typically use sonication to shear input DNA, coupled with several rounds of enzymatic modification to produce a sequencer-ready product. In addition to being labour intensive and difficult to automate, manufacturer’s protocols commonly need starting DNA quantities in the microgram range, most of which is lost during preparation with only a small fraction being present in the final library. The requirement for large quantities of DNA to prepare NGS libraries makes the sequencing of many limited material protocols, such as forensic and ChIP-seq samples, and single cell studies such as genotyping, sequencing or RNA-seq, challenging.

Recently, an alternative to the standard methods of fragmentation and adapter ligation has become available (Syed et al. 2009a; Syed et al. 2009b). Recombinant transposon-derived dsDNA integration complexes, transpososomes, have been produced with pre-adsorbed 19bp core transpososome recognition motif (TRM) containing oligonucleotides. Upon transpososome integration, target DNA is simultaneously fragmented and TRM oligos ligated to the 5’ end of each double stranded target (Syed et al. 2009a; Syed et al. 2009b). A series of platform-specific PCR amplifications are then required to produce sequencer-ready libraries. This technique allows NGS library synthesis from 50ng using the manufacturer’s standard protocols. Further titration of this method has been reported to produce un-validated libraries from as little as 10pg of template (Adey et al. 2010). Although a major advance, an important limitation of this technique is the incompatibility of tagmented libraries with standard platform specific sequencing primers.

Here we describe a modified tagmentation technique that permits picogram quantities of target DNA to be sequenced reproducibly on Illumina platforms with the standard Illumina paired end
(PE) sequencing primer. Our modified adapter sequences also incorporate an ‘in-line’ barcoding system that allows sample multiplexing without the need for additional index-specific reads. To ensure accurate input of picogram levels of DNA we have developed a high-sensitivity fluorescence based quantitation system that reproducibly reports sample DNA concentrations in the femtogram range. To validate our approach we used this technique to deeply sequence an *E. coli* K-12 genome from as little as 20pg of input genomic DNA. Parallel datasets obtained using sonication based Illumina library preparation methods produced from 1µg genomic DNA provide comparative information on target GC bias, levels of library diversity and coverage. Further, we show that reducing input quantities from 1µg to 2ng, 200pg or even 20pg results in libraries with comprehensive sequence coverage and high degrees of fragment diversity. These findings were confirmed with low coverage re-sequencing of 20pg of mouse genomic DNA.
Results

1. Quantitation of target input

The manufacturer’s standard tagmentation procedure requires the use of a pre-determined transpososome-to-DNA ratio. To maintain this relationship when titrating down levels of target it becomes necessary to accurately measure very low concentrations of DNA. We have developed a highly reproducible DNA sample quantitation method utilising a fluorescent DNA reporter dye, background signal quencher and highly sensitive optical plate reader (see Methods). We are able to reproducibly measure 500fgµL⁻¹ in the final 20µL reaction, equivalent to 10pgµl⁻¹ in the initial sample when using our standard 1:20 dilution (Methods and Supplemental Figure 1).

2. In-line barcoding and read quality

The alteration to the manufacturer’s tagmentation protocol described here uses oligos complementary to the 19bp TRM sequence with an additional recognition site for a remote cutting type IIG restriction endonuclease. These enzymes are subsequently used to remove the majority of the 29bp flanking sequence, including the TRM, to leave a mandatory 2bp 3’ TG overhang at both ends of all library fragments which is subsequently used to ligate adapters in a highly efficient reaction (Figure 1). The chosen restriction endonucleases have short recognition sequences that also occur in the input DNA. Using such enzymes in the preparation of NGS library fragments will, in addition to trimming away transposon core sequences, also lead to cleavage of a number of fragments which will be lost from the library leading to reduced coverage of target DNA around endogenous enzyme recognition sites. To overcome this, we employed different type IIG restriction enzyme and tailed primer combinations (AcuI, Bpml, Bsgi, and BpuEl) to produce library fragments with identical 2bp 3’ GT overhangs in the flanking core oligo sequence. Endogenous cleavage sites for each enzyme are largely non-overlapping. Hence, sequencing libraries created by pooling sub-libraries from separate enzyme/primer combinations should minimise coverage reduction at endogenous recognition motifs.
Separate *E. coli* K12 genomic DNA libraries were prepared in parallel using either sonication based Illumina compatible techniques (1µg input DNA, see Methods) or our altered tagmentation method (with 1ng, 100pg and 10pg levels of input DNA, see Methods). For the tagmentation method, libraries were created with separate restriction enzyme/primer combinations and additional libraries made by pooling equimolar amounts of either two (*BpmI* and *BsgI*) or four (*AcuI*, *BpmI*, *BsgI*, and *BpuEI*) independent sub-libraries prior to adapter ligation. Each library used specific staggered length in-line barcoded adapters that allow indexing of both reads of a read-pair whilst simultaneously offsetting a common three base sequence found at the start of all tagmentation reads (Supplemental Table 2). Libraries were multiplexed, sequenced on an Illumina GAII 51bp paired end (PE) flowcell, and data retrieved. Resultant data files were converted to fastq format, de-multiplexed into constituent libraries and filtered for reads where all constituent base calls exceeded a *phred* quality of 20 (Methods).

At the 10pg level, multiplexed libraries for one enzyme (10pg input), two enzymes (20pg total input) and four enzymes (40pg total input) yielded 3.1x10^7 paired 51bp reads from the flowcell lane. 1.3x10^7 (42%) of these pairs contained constituent nucleotides that fell below the arbitrary *phred* 20 threshold (Methods and Table 1) and were discarded. Of the remaining data, 3.7x10^6 pairs (1.2% of the total) had chimeric or un-recognised barcodes. 98.9% of all high quality paired reads were thus successfully de-multiplexed using our in-line barcoding method resulting in sub-libraries with an average of 4.0x10^6 paired reads (~3.5x10^8 bp of sequence data) per library, equal to ~75-fold statistical coverage of the *E. coli* 4.6Mb genome (Table 1). To prevent alignment biases due to unequal read lengths we removed 7bp, corresponding to the longest barcode, from the 5’ end of all reads for all libraries.

The yields and quality of tagmented libraries outlined here are equivalent to data obtained from non-barcoded ‘standard’ Illumina GAII 51bp PE libraries run by our laboratory (Supplemental Table 1).
3. Read alignment quality and library diversity

Each 44bp PE dataset was aligned separately to an *E. coli* K-12 reference genome (Methods). Resultant datasets were quality filtered for uniquely mapping read pairs where both reads exceeded a mapping *phred* of 150 (Table 1).

PCR is used to increase the available material for sequencing in both standard Illumina and our modified tagmentation library preparation methods. This leads to the amplification of library fragments, and a consequent increase in DNA mass with a reduction in fragment diversity due to amplification bias. For this reason we sought to compare the relative diversity, and hence information content, of the aligned datasets. PCR duplicates within each filtered dataset were identified and excluded (Methods and Table 1). The 1µg Illumina libraries were found to contain approximately 1% library fragment redundancy whereas duplication levels for the 10pg tagmentation library set varied between 49% and 64%. Repeating the tagmentation process with higher DNA input amounts (100pg and 1ng levels) produced libraries with lower degrees of redundancy, 16-23% and 12-18%, respectively (Table 1 and Supplemental Table 3).

To allow direct comparison between preparation methods we randomly selected subsets of 1x10^6 (1M) non-redundant, uniquely mapping 44bp paired reads for all further analyses (see Methods).

4. Relative coverage

We first considered whether specific biases exist in statistical coverage between the two library preparation methods. Comparison of each library shows that the tagmentation method generates larger variation in coverage across the target genome when compared to the standard Illumina method (Figure 2a). As expected, tagmentation libraries produced using a single restriction endonuclease cover large numbers of genomic regions at zero (Figure 2a, Table 2) or low (Figure 2b and Table 2) statistical coverage. However, also as predicted, blended libraries from two or four independent tagmented reactions digested with separate enzymes substantially reduce the frequency of low coverage regions to levels similar to those observed for the 1µg Illumina library (Figure 2a, Figure 2b and Table 2). Equivalent results
were observed with 1ng and 100pg level tagmentation library datasets (Supplemental Table 4).

5. Relative GC content bias

We next considered whether coverage from the amplified libraries exhibited a bias in GC content. To compare relative and absolute sequence biases between Illumina and tagmented libraries we compared datasets to an unbiased *in silico* library of 1M randomly sampled uniquely mappable, non-redundant, *E. coli* K-12 genome fragments of equivalent fragment insert lengths to the test libraries (Figure 2c and Supplemental Figure 2). Statistical coverage levels between the two experimental datasets were most similar in genomic regions exceeding 50% GC content. Here both libraries showed under-representations of expected coverage levels compared to the simulated unbiased set. Overall, coverage for both libraries was biased towards AT-rich sequences with the tagmentation dataset showing greatest deviation.

6. Effect of enzymatic digest on local coverage

To quantitate the effect of using single enzymatic digests to produce tagmented libraries we analysed sequence coverage in the vicinity of endogenous recognition sites for the endonuclease used in the library preparation. Our data show that, as predicted, library fragments were reduced to 7% of normal coverage levels, spanning a few base pairs across the enzyme binding site (Figure 2d). Full coverage depth was restored within one insert length immediately flanking the recognition motif. Analysis of libraries made from two independent enzymatic digests showed minimum coverage levels are restored to 70% of normal levels at any individual enzyme site (Figure 2d). Blending 4 separately digested libraries increases the coverage at endogenous sites still further to 83% of median levels.

7. Sequence preference for transpososome insertion

The use of an enzymatic reaction to fragment target DNA as an alternative to sonication immediately raises the question of whether preferred transpososome sequence binding motifs exist and, if so, how this may introduce further bias in local sequence coverage. Consequently,
we next sought enriched sequence motifs at transpososome integration sites. Analysis of transpososome integration sites in our 10pg, 100pg and 1ng level tagmented library sets provided evidence for a weak ~13bp motif centered at the point of fragmentation (Figure 3) consistent with a preference reported independently (Adey et al. 2010). Analysis of the Illumina library yielded no evidence of sequence enrichments at sites of template fragmentation.

8. Tagmentation with 20pg of mouse genomic input

To investigate the utility of this technique with complex animal genomes, seven separate C57BL/6J mouse liver genomic DNA libraries were prepared using either sonicated (1µg input) or our modified tagmentation method (20pg, 1ng or 4ng input). Tagmented libraries with inputs at 1ng and 4ng were multiplexed and run on a single Illumina GAII lane. The library produced using the sonication based Illumina protocol was run on a separate lane. Both lanes were run as 51bp PE sequences on the same GAII flowcell. The 20pg input tagmented library was run using two lanes of an Illumina HiSeq 2000 platform at 100bp PE. As before, resultant data files were converted to fastq format, de-multiplexed into constituent libraries, trimmed to 44bp residual length and filtered for reads where all constituent nucleotides exceeded a base call phred score of 20 (Methods). Approximately 3.2x10^7 51bp paired reads were recovered from the tagmentation multiplex lane. 8.5x10^6 (26.5%) of these fell below the quality threshold and were discarded. Of the remaining dataset 6.5x10^5 paired reads (2.0% of the total) had chimeric or un-recognised barcodes. 2.2x10^7 paired reads (68.8% of the lane) contained identifiable barcodes, passed the phred 20 filter, were successfully demultiplexed into constituent libraries (average 4.4x10^6 ± 1.3x10^6 paired reads per library) and were trimmed to remove the barcode sequence and produce 44bp residual length reads. Each individual library therefore represents 3.9x10^8 bp of sequence data. 2.9x10^7 51bp paired reads were recovered from the mouse genomic DNA library lane prepared using the sonication based Illumina method. 2.1x10^7 PE reads (72.4% of lane) contained all constituent base calls exceeding the base call phred 20 threshold, representing 1.8x10^9 bp of sequence data. A total of 2.3x10^8 100bp PE reads were produced from the 20pg input mouse genomic DNA tagmented library. 1.1x10^8 PE reads (47.5%) contained identifiable barcodes, passed the phred 20 filter and were trimmed to 44bp representing an equivalent of 9.7x10^9 bp of sequence data.
As with the *E. coli* analyses, each 44bp PE dataset was aligned separately to the C57BL/6J reference assembly (Methods). Resultant datasets were quality filtered for uniquely mapping paired ends with alignment scores above our arbitrary threshold (Table 3). PCR duplicates within each filtered dataset were identified based on mapping coordinates and $1 \times 10^6$ (1M) non-redundant, uniquely mapping PE reads from each library were selected for further analyses. Finally, gross target genome coverage and constituent fragment diversity were derived by computing unique nucleotide coverage and comparing to the maximum predicted unique nucleotide coverage for 1M 44bp PE reads (Table 3).

No gross differences were observed in the levels of library quality, fragment redundancy, target coverage or fragment diversity between the $1\mu$g input sonication based Illumina and 1ng input tagmentation libraries (Table 3). PCR duplication rates of up to 39-fold those of the other libraries were, however, seen in the 20pg library and strongly suggest that at these input levels, diversity of the starting material becomes a limiting factor.
Discussion

Standard protocols for next generation library synthesis typically require ~1µg of input DNA. Some sequencing centres have successfully generated libraries from ~2ng (Paolo Piazza, Wellcome Trust Centre for Human Genetics. Pers Comm.). Tagmentation methods allow reduction to 50ng but require sample specific sequencing primers (Syed et al. 2009a; Syed et al. 2009b). Tagmentation has been reported to recover sequence from 10pg genomic DNA, however methods for accurately quantitating the DNA input and the quality of the resultant libraries are un-reported (Adey et al. 2010). We have modified tagmentation technology to allow routine preparation of fully platform compatible NGS libraries from picogram quantities of DNA. Our study carried out detailed analysis of low input sequencing datasets and showed that biases are equivalent to standard Illumina 1µg input libraries.

The production of picogram level libraries requires the accurate quantification of target DNA. We have therefore developed a method for reliably quantifying sample concentrations down to the femtogram/µL range.

Our protocol uses more PCR than standard sonication based Illumina methods resulting in increased duplicate sequences and an AT rich sequence bias. Bacterial or mouse libraries produced from 1ng of material displayed a maximum of 1.2x more duplicates when compared to the 1µg sonication based Illumina preparation. Interestingly, at the 10pg level there is a greater discord between the relative duplicate levels observed between the E.coli, 2.9x, and mouse, 38.9x, libraries compared to the 1µg sonication based Illumina preparation. This apparent reduction in the final information content of the libraries can largely be attributed to the limiting input material. 20pg of gDNA represents a theoretical maximum of ~6.7x10^7 300bp average insert fragments. Diversity of both mouse and E. coli genomes are sufficient that false duplicates due to fragmentation at identical genomic coordinates should not be a significant issue at this level of input material. In our E.coli libraries we recovered 1.5x10^6 of the available input fragments, representing a capture of 2.2% of the theoretical maximum available starting material. Our mouse library recovered 1.8x10^6 of such fragments, approximately 2.7% of the maximum available starting material. Approximately 17-fold greater high quality uniquely mapped reads were sequenced for the mouse 20pg input library compared to the E.coli counterpart. Thus, the 13-fold increase in duplicate rates observed between these libraries.
suggests that the majority of available fragment diversity in the mouse library as fixed before PCR amplification has been sequenced to exhaustion with the large number of additional sequencing reads increasing the apparent duplicate rate without adding to overall library information.

As input DNA quantity in NGS library synthesis is reduced to low levels, two process-related factors may reduce diversity in the resultant dataset: firstly, only a fraction of the initial sample contributes directly to the final library; and, secondly, increased loss of effective template necessitates the use of more PCR, ultimately raising the duplication frequency in the final dataset. The input template quantity contributing to the final library is set by both the DNA fraction fragmented within a selected size range and by the amount of template which becomes successfully ligated and amplified. Our tagmentation protocol uses a single tube reaction to fragment input DNA, ligate TRM containing oligos and amplify fragments. Restriction endonucleases are then used to create a ligatable 2bp overhang as a more efficient alternative to the analogous ‘A-tailing’ Illumina stages. Hence, in our method, process-induced reduction in diversity mainly arises from the efficiency of enzymatic fragmentation and controlled rejection of material at size selection. In current Illumina protocols, substantial additional losses are incurred during the sonication process, through inefficiencies incurred at several enzymatic manipulations and through multiple sample purifications. Our 20pg input library data suggest that 2-3% of the starting material is captured with our method. An analogous figure has not been reported for the Illumina method. Thus, when our modified tagmentation method is used to synthesize libraries from 1ng input material, acceptable levels of diversity are observed in the final dataset when sequenced at this depth. However, our data suggest that further reducing input material to the 100pg and 10pg level has a limiting effect on final library diversity. Despite the loss of ~95% of the initial material in our library prep, it is worthy of note that our 20pg mouse libraries recovered a ~0.4x coverage of unique mapped fragments of the non-repetitive mouse genome from only 7x genome equivalents of starting material. This is a significant recovery of mouse genomic data from genome level inputs and has not currently been achieved by any other technique. Extrapolation from this result suggests that near complete coverage of mapped fragments for the mouse genome should be possible with ~50pg of input. This figure would be closer to 60pg if coverage is measured as sequenced bases using 100bp PE reads. Further improvements in diversity should be
achievable, relative to those reported here, by increasing the size range of library fragments selected, and by titrating down the use of PCR. These alterations should decrease the amount of starting material required to achieve a particular coverage. However, it should be noted that it will never be possible to recover a full genome coverage for samples where the quantity of starting material is close to or less than a single genome equivalent. As a complement to whole genome studies, this technique is also likely to have widespread utility in the analysis of samples derived from enrichment procedures that typically result in low DNA yields but where coverage over target regions is very high.

Techniques are currently available for amplifying low input samples prior to NGS library preparation (Tang et al. 2009). However, these approaches are likely to introduce sample bias and amplification artifacts that are impossible to distinguish in the final library. Our technique avoids amplification prior to fragmentation and uses an in silico paired-end-read duplicate filter to exclude gross artifacts providing a more accurate representation of the relative relationship between input molecules. This is an important consideration when attempting to capture relative data from very low quantity starting procedures such as single cell transcription profiling which is currently only possible with target pre-amplification (Tang et al. 2009).

The use of type IIG endonucleases in our library synthesis was shown to cause predictable, highly localised coverage loss at endogenous recognition sites. Blending sub-libraries created with separate restriction enzymes resolves this issue but requires parallel production of multiple samples. We have shown that blending two libraries is sufficient to increase target coverage to levels similar to the Illumina preparation. Hence, our laboratory uses a two restriction enzyme method as standard. The use of alternative type IIG restriction enzymes with less frequent endogenous sequence motifs may be explored so that a single preparation technique giving acceptable coverage might be used.

Our modification to the standard Illumina PE adapter sequence results in the repeated sequencing of a mandatory CAG motif at the start of all reads. This may result in a failure of some Illumina base-calling software. To avoid this we use a variable length in-line barcoding region in our adapter to offset these constant sequences and simultaneously allow library multiplexing. Our in-line barcoding system indexes both reads of a paired-end fragment
allowing important quality checks such as the identification of inter-library fragment chimeras which are not possible using standard indexing systems.

This research has focused on the application of this technology to Illumina library synthesis. However, the basic fragmentation, ligation and digestion protocol described here provides a universal entry point to the creation of libraries for any current NGS or third generation platform that requires the ligation of an adapter to the ends of fragmented target DNAs during library preparation.

Overall, we feel that the modified tagmentation method presented here realises the true potential of transpososome mediated NGS library preparation technology. It allows reproducible sequencing of picogram quantities of target DNA in a fully platform compatible method. Application of this technique will finally make practicable many studies where amplification prior to fragmentation is undesirable or where limited genetic materials are available such as genomic analysis of un-culturable bacteria, low mass forensic or museum samples, low input ChIP-seq and genotyping, sequencing or RNA-seq transcription analysis of single cells.
Methods

Input sample quantitation

A 2x serial dilution of lambda genomic DNA (Invitrogen) in 1xTE (Promega) was produced to give final concentrations from 100pg/µl to 390fg/µl in 1xTE. 10µL aliquots of each standard were added to a 384 well Optiplate-F (Perkin-Elmer) in triplicate. 10µL aliquots of 1:10 sample dilutions were added to separate plate wells. 10µl picogreen working solution, 1:100 Quant-iT Picogreen dsDNA reagent (Invitrogen), 1:10 AccuBlue High Sensitivity Enhancer 1000x (Biotium) in 1xTE (Promega), was added to each sample well. The plate was centrifuged briefly at 1000xg and placed into a BMG PHERAstar. Optimal focal height and gain were set, samples shaken for 30 seconds then fluorescence values were read with 485nm excitation and 520nm emission filters. Standard dilutions (final plate standard dilutions 50pg/µl to 195fg/µl) were plotted against blank normalised relative fluorescence using MARS Data Analysis Software (BMG LABTECH) and concentrations of unknown samples extrapolated.

Illumina library preparation

1µg non-methylated *E. coli* K-12 MG1651 gDNA (NEB) was placed in a DNA LoBind microcentrifuge tube (Eppendorf) and diluted to a total volume of 40µL in 1xTE. Sonication was performed in a Misonix 4000 sonicator (Misonix) using a cup horn with circulating ice cold water (Amplitude 100, 4x cycles of 60secs sonication with samples left to cool on wet ice for 60secs between cycles). Following sonication, 1µL of neat sample was run on a Bioanalyser 2100 (Agilent) using a HS DNA chip (Agilent) to monitor for optimal sample fragmentation. 10µL of 6x gel loading buffer (Maniatis et al. 1982) was added to the remaining sample. This was then mixed and split into two equal aliquots which were loaded on adjacent wells of a 3% Nusieve agarose gel (Lonza) pre-stained with 0.5x GelRed (Biotium) and subjected to 35mins of electrophoresis at 100V in 1xTAE. 100ng of 1kb ladder (Promega) was run in a parallel lane. Following electrophoresis the gel was imaged in a UVP bioimaging system adapted for use with hyper-bright green 528nm LEDs (RS) and visualised using a 617nm±73nm filter (Semrock). A 250-350bp region of fragmented target sample was excised and DNA recovered using Zymoclean Gel DNA recovery kit (Zymo Research). Target DNA was then end repaired
in a 100µl total volume reaction (1x NEB phosphorylation buffer, 0.5mM NEB dNTPs, 5µL NEB T4 DNA polymerase E6003S, 1µL NEB Klenow fragment E6004S, 5µL NEB T4 PNK E6005S) at 20°C for 60mins. The reaction mix was then cleaned using a DNA Clean and Concentrator Kit (Zymo Research) and eluted twice with 10µL 1xTE. Pooled eluates were then subject to a 50µL total volume A-Tailing reaction (1x NEB NEBuffer 2, 0.2mM NEB dATP, 3µL NEB Klenow 3’-5’ exo- E6006S) at 37°C for 30mins. The completed reaction mix was then cleaned using a DNA Clean and Concentrator Kit (Zymo Research) and eluted twice with 10µL 1xTE, pooled and 1µL of eluate subject to high sensitivity DNA quantitation in a final volume of 20µL. Samples were ligated to standard Illumina Paired End adapters (Illumina) in a 15µL reaction (1x NEB T4 DNA ligase buffer with 1mM final concentration ATP, 10x molar excess Illumina PE adapter, 2,000 units NEB T4 DNA ligase) at 20°C for 30mins. Following incubation, 2µL of neat ligation reaction was used as a template in a 25µL PCR enrichment reaction (1x Finnzymes Phusion HF master mix, 1µM Illumina PE PCR primer 1, 1µM Illumina PE PCR primer 2). PCR was performed in a Piko Thermal cycler (Finnzymes) with the following cycling parameters (1x 98°C 30s, 14x [98°C 10s, 65°C 30s, 72°C 30s], 1x 72°C 5min). Following amplification, 1µL of a 1:100 dilution of the neat PCR sample was run on a Bioanalyser 2100 (Agilent) using a HS DNA chip (Agilent) to check for appropriate amplification. Following confirmation of enrichment the remaining sample was mixed with 5µL 6x gel loading buffer (Maniatis et al. 1982), loaded on a 3% Nusieve agarose gel (Lonza) pre-stained with 0.5x GelRed (Biotium) and subjected to 35mins electrophoresis at 100V in 1xTAE. 50ng of 1kb ladder (Promega) was run in a parallel lane. Following electrophoresis the gel was imaged in a UVP bioimaging system (UVP) adapted for use with hyper-bright green 528nm LEDs (RS) and visualised using a 617nm±73nm filter (Semrock). Amplified product, corresponding to the target library, was excised and DNA recovered in 10µL 1xTE using Zymoclean Gel DNA recovery kit (Zymo Research). A 1µL aliquot of the final library was subject to high sensitivity DNA quantitation in a final volume of 20µL. Library dilutions were adjusted to 10nmol and used for cluster generation and sequence analysis on Illumina GAII Genome Analyser or HiSeq 2000 platforms.
Tagmentation library preparation

250pgµl⁻¹, 25pgµl⁻¹ or 2.5pgµl⁻¹ dilutions of non-methylated E. coli K-12 gDNA (NEB) were produced by serial dilution in 1xTE and concentrations confirmed with high sensitivity DNA quantitation in a final volume of 20µL. 4µl aliquots of input samples were added to the appropriate wells of a 24 well PCR plate (Finnzymes). 1µL aliquots of a 5x reaction master mix (5x Epibio Nextera LMW Reaction buffer, 0.2µl EpiBio Roche 454-FLX compatible Nextera Enzyme mix at dilutions 1/10, 1/100 and 1/1000 for 10ng, 100pg and 10pg input dilutions respectively) were added, samples sealed with cap strips, combined by brief centrifugation and incubated at 55°C for 5mins in a Piko thermal cycler (Finnzymes) followed by an immediate hold step at 4°C. 20µL aliquots of a PCR master mix were added to give a final concentration of 1x Phusion HF reaction master mix (Finnzymes) and 2µM tailed primer in a 25µL final reaction volume. Tailed primers are specific for the endonuclease to be used in the preparation (AcuI tailed oligo GCGCGCCTGAAGATGTGTATAAGAGACAG, BsgI tailed oligo GCGCGCGTGCAGATGTGTATAAGAGACAG, BpmI tailed oligo GCGCGCCTGGAGATGTGTATAAGAGACAG, BpuEI tailed oligo GCGCGCCTTGAGATGTGTATAAGAGACAG) and were synthesised and PAGE purified (IDT). Reactions were immediately cycled at 1x 72°C 5mins, 1x 98°C 30s, 10x [98°C 10s, 65°C 30s, 72°C 45s], 1x 72°C 5min in a Piko thermal cycler (Finnzymes). The number of cycles of PCR used was varied dependent on library input quantity such that 1ng = 10x cycles, 100pg = 12x cycles and 10pg = 14x cycles. Following amplification 5µL of 6x gel loading buffer (Maniatis et al. 1982) was added to the sample, mixed and loaded on a 3% Nusieve agarose gel (Lonza) pre-stained with 0.5x GelRed (Biotium) and subjected to 35mins electrophoresis at 100V in 1xTAE. 50ng of 1kb ladder (Promega) was run in a parallel lane. Following electrophoresis the gel was imaged in a UVP bioimaging system (UVP) adapted for use with hyper-bright green 528nm LEDs (RS) and visualised using a 617nm±73nm filter (Semrock). A 250-350bp region of tagmented target sample was excised using 4mm Genecatcher disposable gel excision pipette tips (Gel Company) and DNA recovered using Zymoclean Gel DNA recovery kit (Zymo Research) with a final elution of 2 aliquots of 10µL 1xTE. Recovered DNA was then digested with 5 units of the appropriate class III endonuclease (AcuI, BpmI, BsgI, BpuEI, NEB) using the manufacturer’s recommended conditions in a 30µL final volume.
at 37°C for 30mins. Digested samples were cleaned using a DNA Clean and Concentrator Kit (Zymo Research) and eluted with 10µL 1xTE. 1µL of eluate was subject to high sensitivity DNA quantitation in a final volume of 20µL. 20µL ligation reactions were set up using all remaining target template (final concentration 1x NEB T4 DNA ligase buffer, 2,000 units NEB T4 DNA ligase, 10x molar excess of Illumina compatible adapter) at 20°C for 30mins. Following incubation, 4x 5µL aliquots of neat ligation reaction were used as a template in replicate 25µL final volume PCR enrichment reactions (1x Finnzymes Phusion HF master mix, 1µM Illumina PE PCR primer 1, 1µM Illumina PE PCR primer 2). PCR was preformed in a Piko Thermal cycler (Finnzymes) with the following cycling parameters (1x 98°C 30s, 8x [98°C 10s, 65°C 30s, 72°C 30s], 1x 72°C 5min). Following amplification, 1µL of a 1:100 dilution of the neat PCR sample was run on a Bioanalyser 2100 (Agilent) using a HS DNA chip (Agilent) to check for appropriate amplification. Following confirmation of enrichment the remaining samples were mixed with 5µL 6x gel loading buffer (Maniatis et al. 1982), loaded on a 3% Nusieve agarose gel (Lonza) pre-stained with 0.5x GelRed (Biotium) and subjected to 35mins electrophoresis at 100V in 1xTAE. 50ng of 1kb ladder (Promega) was run in a parallel lane. Following electrophoresis the gel was imaged in a UVP bioimaging system (UVP) adapted for use with hyperbright green 528nm LEDs (RS) and visualised using a 617nm±73nm filter (Semrock). Amplified sample, corresponding to the target library was excised using 4mm Genecatcher disposable gel excision pipette tips (Gel Company), gel slices from amplifications of the same target library pooled and DNA recovered into 10µL 1xTE using a Zymoclean Gel DNA recovery kit (Zymo Research). 1µL of eluate was subject to high sensitivity DNA quantitation in a final volume of 20µL. Library dilutions were adjusted to 10nmol and used for cluster generation and sequence analysis on an Illumina GAII Genome Analyser and delivered to our local NGS service provider who sequenced the library using standard manufacturer’s procedures.

**Sequencing adapters**

Modified Illumina adapters were syntheised to allow ligation of the GT sticky end left by the class III endonuclease digests. Five pairs of adapters (consisting of partially complementary oligos 1 and 2) were produced with staggered barcode sequences as (Supplemental Table...
2). Adapter primers were synthesised using PAGE purification (IDT technologies) and diluted to 100µM in 1xTE. A 50µL adapter annealing reaction was carried out (final concentration 20µM adapter 1, 20µM adapter 2, 1x T4 DNA Ligase buffer) on a Pico thermocycler (Finnzymes) using the following cycling parameters (1x 95°C 2mins, 130x [85°C -0.5°C] followed by a 4°C hold. Final annealed adapter concentrations were 20µM per adapter.

**Sample de-multiplexing, trimming and quality control**

Raw Illumina format PE datasets were simultaneously filtered for reads containing base calls with *phred* scores less than 20, demultiplexed in to constituent sub-libraries based on in-line barcodes, trimmed to remove barcoding nucleotides and converted into fastq format using our own software. Overall library quality before and after de-multiplexing, filtering and trimming were monitored using the FastQC software package (Andrews. 2010).

**Genome alignment.**

Individual sequenced datasets were aligned to the reference genomes of the bacterium *Escherichia coli* K-12 MG1655 (NCBI. 2007a) and repeat-masked mouse *Mus musculus* C57BL/6J assembly m37 (NCBI. 2007b). Genome alignment was carried out with Novoalign v.2.05.33 (Novocraft Technologies. 2011), currently the most accurate algorithm for short-read alignment to reference genomes (Li and Homer. 2010; Lunter and Goodson. 2010). Novoalign was run in the paired-end mode and with default program settings.

**Duplicate filtering.**

Duplicate paired-end reads within the same library were identified based on alignment coordinates of both reads within a pair using our SAM file duplicate filtering utility (Maslau. 2011) to leave a subset of non-degenerate read-pairs. This utility has been made available to the community under the GNU General Public License.

**Data analysis.**

Successfully de-multiplexed, trimmed and quality thresholded datasets were further filtered. Read-pairs were kept only when both reads aligned to unique locations in the reference
genome, had mapping *phred* scores >150 and possessed a fragment size within the central 98% distribution of apparent mapped library fragment insert size.

Genome coverage statistics were calculated using the sequenced bases of the forward and reverse reads mapped to the reference genome. The GC content dependent bias in the genome coverage was calculated using the complete inferred genomic fragments. GC content was assigned to the middle base of each 250 bp window of the genome, with 1 bp relative offset. The statistical dispersion in genome coverage at each position was described using median values and inner-quartile ranges (25th and 75th percentile).

For the analysis of coverage at class IIG endonuclease sites the full complement of enzyme recognition sites for the reference genome were initially identified *in silico*. Genome coverage at each nucleotide across a 1kb region centered at each enzyme cut site (coordinate +1) was calculated and data for all cut sites overlaid. The statistical dispersion in genome coverage at each position was described with median values and inner-quartile ranges (25th and 75th percentile).

Putative consensus sequence motifs at fragmentation sites were investigated within 50bp regions centered at each of 1000 randomly selected fragmentation sites. Sequence conservation was displayed using the Weblogo program (Crooks et al. 2004) and annotated with the Gnuplot program (Janert. 2010). Three replications resulted in the identification of essentially identical sequence motifs.
Data Access

Sequencing datasets used in this study are available through the EMBL-EBI European Nucleotide Archive (http://www.ebi.ac.uk/ena/). Accession numbers of alignment files for each dataset are as follows; *E. coli* genomic DNA 1µg Illumina library (Acc. ERS066939), *E. coli* genomic DNA 1ng *AcuI* Tagmentation library (Acc. ERS066940), *E. coli* genomic DNA 1ng *BsgI* Tagmentation library (Acc. ERS066941), *E. coli* genomic DNA 1ng *BpuEl* Tagmentation library (Acc. ERS066942), *E. coli* genomic DNA 1ng all enzymes Tagmentation library (Acc. ERS066944), *E. coli* genomic DNA 100pg *AcuI* Tagmentation library (Acc. ERS066945), *E. coli* genomic DNA 100pg *BsgI* Tagmentation library (Acc. ERS066946), *E. coli* genomic DNA 100pg all enzymes Tagmentation library (Acc. ERS066947), *E. coli* genomic DNA 10pg *AcuI* Tagmentation library (Acc. ERS066948), *E. coli* genomic DNA 10pg *BsgI/BpmI* Tagmentation library (Acc. ERS066949), *E. coli* genomic DNA 10pg all enzymes Tagmentation library (Acc. ERS066950), Mouse genomic DNA 1µg Illumina library (Acc. ERS066931), Mouse genomic DNA 1ng *AcuI* Tagmentation library (Acc. ERS066932), Mouse genomic DNA 1ng *BsgI* Tagmentation library (Acc. ERS066933), Mouse genomic DNA 1ng *BpuEl* Tagmentation library (Acc. ERS066935), Mouse genomic DNA 1ng *BpmI* Tagmentation library (Acc. ERS066934), Mouse genomic DNA 1ng all enzymes Tagmentation library (Acc. ERS066936), Mouse genomic DNA 10pg *BsgI/BpmI* Tagmentation library lane 1 (Acc. ERS066937), Mouse genomic DNA 10pg *BsgI/BpmI* Tagmentation library lane 2 (Acc. ERS066938).
Acknowledgments

This work was made possible with Alamy grants administered by the Fischer Family Trust (M.D.F, N.J.P, G.Z, B.F and S.M) and Medical Research Council, UK funding (C.P.P.). We would like to acknowledge Christoffer Nellåker and Andreas Heger at MRC Functional Genomics Unit, University of Oxford and Nick Carruccio at EpiBio for useful discussion and technical information.
Figure Legends

Figure 1. Schematic overview of modified Tagmentation procedure. Purified genomic DNA is tagmented using a specific ratio of enzyme and target (see Methods). Following tagmentation (addition of standard TRM oligo shown as white box) the reaction is quenched through the addition of premixed PCR reagents, subject to a brief extension step, and amplified with limited PCR cycles using a single tailed-oligo (black boxes) resulting in a library of fragments flanked by identical 29bp sequences that can be size selected by standard gel electrophoresis techniques. Tailed oligos contain a recognition site for a remote cutting type IIG restriction endonuclease that is used to remove the majority of the 29bp flanking sequence, including the core transposon motif, and leaves a mandatory 2bp 3’ TG overhang at both ends of all amplified library fragments. Modified Illumina sequencing adapters (grey boxes), incorporating variable length in-line barcoding sequences, are then ligated to this 2bp overhang in a highly efficient reaction. A second limited cycle PCR is performed directly on the ligation reaction and products run on an agarose gel and subject to a second round of size selection. The purified product is then sequenced on an Illumina GAII sequencer using the manufacturer’s standard methods.

Figure 2. Analysis of bias in 10pg level libraries. (a, b) 1x10^6 non-redundant, uniquely mapping, high quality, paired end reads were randomly selected for further analysis of the Illumina library (Red; 1µg input) and each tagmented library. Tagmented libraries used 10pg input (Blue; single enzyme, AcuI), 20pg input (Black; two enzyme, BsgI and BpmI) and 40pg input (Green; four enzyme, AcuI, BsgI, BpmI and BpuEI). Coverage depth across the genome and percentage of genome covered at increasing cumulative coverage depths for each library were compared. (c) Median coverage depth for genomic regions defined by GC content was also analysed for a single enzyme tagmentation library AcuI (Blue), an Illumina library (Red) and 1x10^6 size matched fragments randomly selected in silico from the reference genome to model coverage in a non-biased manner (Black). Shaded regions represent 25th and 75th inner quartile regions for each dataset. The region between vertical black lines represents 99% of the total reference genome. (d) Fold coverage across 1kb genomic regions containing endogenous cleavage sites for single enzyme (Green; AcuI, AcuI sites shown) and two enzyme (Black; BsgI and BpmI; BpmI sites shown) tagmented libraries. Approximately 3,000
genomic regions are represented in each analysis. Absolute coverage and percentage median coverage by base position are shown for each region.

**Figure 3. Sequence motif analyses at tagmentation and Illumina fragmentation sites.** Genomic sequences at 1000 randomly chosen sites of transpososome integration (top; single enzyme, *AcuI* 10pg input tagmentation library) or physical sonication induced shearing positions (bottom; Illumina library) were analysed for over-represented sequence motifs and plotted by nucleotide prevalence at each base position. Fragmentation site is shown at base +1. A weak sequence preference spanning a ~13bp region was found in the tagmented library. No sequence preference was found in the Illumina library.
Table 1. Quality Statistics for 1µg Illumina and 10pg level Tagmented libraries. Gross yield of paired-end reads containing all nucleotides with phred values >20 is shown following in silico de-multiplexing for tagmented 10pg input (single enzyme, Acul), 20pg input (two enzyme, BsgI and BpmI), 40pg input (four enzyme, Acul, BsgI, BpmI and BpuEI) and Illumina 1µg inputs. Gross numbers and percentage of initial paired-end reads are shown at each sequential filtering stage; number of pairs with unique alignment to reference genome, pairs where both reads have alignment phred scores greater <150, pairs that fall within middle 98th percentile of mapped library fragment size, non-redundant pairs where both reads have unique mapping coordinates with respect to other read pairs within the library. Library Diversity was calculated as the percentage of non-redundant read pairs in the sample of read pairs passing all quality control filters.
Table 2. Coverage statistics for subsets of $1\mu$g Illumina and 10pg level tagmented libraries. $1 \times 10^6$ non-redundant, uniquely mapping, high quality, paired end reads were randomly selected for onward analysis of the Illumina library ($1\mu$g) and each tagmented (10pg input single enzyme, \textit{Acul}, 20pg input two enzyme, \textit{BsgI} and \textit{BpmI}, 40pg input four enzyme, \textit{Acul}, \textit{BsgI}, \textit{BpmI} and \textit{BpuEl}) libraries. The percentage of reference genome covered at a depth greater than 1x, 5x and 10x, median genome coverage and coverage dispersion with values at 25th and 75th inner quartile ranges are shown for each dataset.

<table>
<thead>
<tr>
<th>Target DNA</th>
<th>E. coli gDNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Illumina</td>
</tr>
<tr>
<td>DNA Input</td>
<td>$1\mu$g</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10pg</td>
</tr>
<tr>
<td></td>
<td>4x 10pg</td>
</tr>
<tr>
<td>% Genome sequenced at coverage ≥ 1x</td>
<td>98 %</td>
</tr>
<tr>
<td>% Genome sequenced at coverage ≥ 5x</td>
<td>97 %</td>
</tr>
<tr>
<td>% Genome sequenced at coverage ≥ 10x</td>
<td>91 %</td>
</tr>
<tr>
<td>Median genome coverage</td>
<td>19x</td>
</tr>
<tr>
<td>Coverage dispersion (IQR 25th - 75th)</td>
<td>14x - 24x</td>
</tr>
<tr>
<td>Target DNA</td>
<td>Mouse Liver gDNA</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Method</td>
<td>Illumina</td>
</tr>
<tr>
<td>DNA Input</td>
<td>$1 \mu g$</td>
</tr>
<tr>
<td>Demultiplexed Reads &gt;phred 20</td>
<td>$21,289,016$ (100%)</td>
</tr>
<tr>
<td>Uniquely Mapped Autosomal read pairs</td>
<td>17,148,508 (80.6%)</td>
</tr>
<tr>
<td>Genome Alignment phred Q &gt;150-150</td>
<td>15,647,414 (73.5%)</td>
</tr>
<tr>
<td>Read pairs with 98th percentile of library fragment length</td>
<td>15,322,795 (72.0%)</td>
</tr>
<tr>
<td>Non-Redundant read pairs</td>
<td>14,291,998 (67.1%)</td>
</tr>
<tr>
<td>Library Diversity (% unique fragments)</td>
<td>93.3%</td>
</tr>
<tr>
<td>Total Target covered (bp)</td>
<td>85,540,188</td>
</tr>
<tr>
<td>Target genome covered</td>
<td>3.46 %</td>
</tr>
<tr>
<td>Percentage of maximum possible coverage</td>
<td>97.2 %</td>
</tr>
</tbody>
</table>
Table 3. Quality statistics and coverage for low coverage mouse libraries. Gross yield of paired end reads containing all nucleotides with *phred* values >20 is shown following *in silico* de-multiplexing for tagmented 20pg input (two enzyme *BsgI* and *Bpml*), 1ng input (single enzymes, *AcuI*, *BsgI*, *Bpml* or *BpuEI*), 4ng input (four enzyme, *AcuI*, *BsgI*, *Bpml* and *BpuEI*) and Illumina 1µg inputs. Gross numbers and percentage of initial paired-end reads are shown at each sequential filtering stage; number of pairs with unique alignment to the autosomal component of the reference genome, end pairs where both reads have alignment *phred* scores greater <150, pairs that fall within middle 98th percentile of mapped library fragment size, non-redundant pairs where both reads have unique mapping coordinates with respect to other read pairs within the library. Library diversity was calculated as the percentage of non-redundant read pairs in the sample of read pairs passing all quality control filters. 1x10⁶ non-redundant, uniquely mapping, high quality, paired end reads were randomly selected for onward analysis of each library. Cumulative total genome covered, either in unique nucleotides or as a percentage of the complete reference autosomal genome, and the percentage of maximum possible target coverage for 1x10⁶ paired-end reads as a measure of diversity are also given for each dataset.
Supplemental Figure and Table legends

**Supplemental Figure 1. Standard curve for low quantity DNA assay.** Final concentration of standards (50pgµl⁻¹, 25pgµl⁻¹, 12.5 pgµl⁻¹, 6.25 pgµl⁻¹, 3.13pgµl⁻¹, 1.56pgµl⁻¹, 781fgµl⁻¹, 390fgµl⁻¹, 195fgµl⁻¹) in 20µL reaction volume are displayed on the X axis. Relative fluorescence units (+/- SD) normalised against a blank reading (1xTE) are displayed on the Y axis. Samples were assayed in triplicate. Reproducible range of assay is ~500fgµl⁻¹. Trend line is shown, \( R^2 = 0.9999 \). Inset shows magnification of lower region of graph values.

**Supplemental Figure 2. Sequence content bias for 20pg, 200pg and 2ng input libraries.** Median coverage depth at genomic regions defined by sequence content for \( 1 \times 10^6 \) randomly selected, uniquely mapping, high quality, paired end reads from two enzyme Bpml + BsgI (Blue) tagmentation libraries produced from three different input levels: 20pg total input (left panel), 200pg total input (centre panel) and 2ng total input (right panel). \( 1 \times 10^6 \) randomly selected, uniquely mapping, high quality, paired end reads from a 1µg input Illumina library (Red) and \( 1 \times 10^6 \) size matched fragments randomly selected \textit{in silico} from the reference genome to model non-bias coverage (Black) are shown for comparison in each panel. Shaded regions represent 25\(^{th}\) and 75\(^{th}\) inner quartile regions for each dataset. The region between vertical black lines represents 99% of the total reference genome.

**Supplemental Table 1. Comparison of yields and de-multiplex efficiency of Tagmented and Illumina Libraries.** The average gross paired-end read yields, gross number and percentage of initial library of reads containing all nucleotides with \textit{phred} values >20 and numbers of reads with successful \textit{in silico} de-multiplexing are compared between tagmented barcoded (n=6), Illumina barcoded (n=7) and Illumina non-barcoded (n=4) flowcell lanes.

**Supplemental Table 2. Oligonucleotide sequences used to make modified barcoded adapters.** Oligonucleotide sequence pairs used to produce tagmentation modified Illumina adapters incorporating an ‘in-line’ staggered barcode (Bold Text). /5Phos/ denotes a 5’ Phosphorylation. * denotes a 3’ Phosphorothioate bond modification.
Supplemental Table 3. Quality Statistics for 1µg Illumina, 1ng tagmented and 100pg level tagmented libraries. Gross yield of paired-end reads containing all nucleotides with *phred* values >20 is shown following *in silico* de-multiplexing for tagmented 100pg input (single enzyme, *AcuI*), 200pg input (two enzyme, *BsgI* and *BpmI*), 400pg input (four enzyme, *AcuI*, *BsgI*, *BpmI* and *BpuEI*), 1ng input (single enzyme, *AcuI*), 2ng input (two enzyme, *BsgI* and *BpmI*), 4ng input (four enzyme, *AcuI*, *BsgI*, *BpmI* and *BpuEI*) and Illumina 1µg inputs. Gross numbers and percentage of initial paired-end reads are shown at each sequential filtering stage; number of pairs with unique alignment to reference genome, pairs where both reads have alignment *phred* scores greater <150, pairs that fall within middle 98th percentile of mapped library fragment size, non-redundant pairs where both reads have unique mapping coordinates with respect to other read pairs within the library. *Note. 2ng *BpmI*/BsgI* dataset created *in silico* using randomly selection of 50% total read pairs from separate 1ng *BsgI* and 1ng *BpmI* input libraries.

Supplemental Table 4. Coverage Statistics for 1µg Illumina, 1ng tagmented and 100pg level tagmented libraries. 1x10^6 uniquely mapping, high quality, paired-end reads were randomly selected for onward analysis of the Illumina library (1µg) and each tagmented (100pg input single enzyme, *AcuI*; 200pg input two enzyme, *BsgI* and *BpmI*; 400pg input four enzyme, *AcuI*, *BsgI*, *BpmI* and *BpuEI*; 1ng input single enzyme, *AcuI*; 2ng input two enzyme, *BsgI* and *BpmI*; and, 4ng input four enzyme, *AcuI*, *BsgI*, *BpmI* and *BpuEI*) libraries. The percentage of reference genome covered at a depth greater than 1x, 5x and 10x, median genome coverage and coverage dispersion with values at 25th and 75th inner quartile ranges are shown for each dataset. *Note. The 2ng *BpmI*/BsgI* dataset was created *in silico* by blending 5x10^5 randomly selected, uniquely mapping, high quality, paired-end reads from separate 1ng *BsgI* and 1ng *BpmI* input libraries.

2. Andrews S. FastQC High Throughput Sequence QC Report. 0.7.0


Figure 1 - Nicholas Parkinson

Purified Genomic DNA input

Single tube Tagmentation reaction, polymerase elongation and PCR with tailed transpososome core oligos (~1hr)

Agarose gel electrophoresis and size selection (~30mins)

Gel purification, class IIg restriction digest and column clean up (~1hr)

Single tube Illumina compatible barcoded adapter ligation and enrichment PCR (~30mins)

Second agarose gel size selection and purification (~30mins)

Library QC and run on Illumina using standard PE sequencing primers
Figure 3: Tagmentation vs Illumina motifs.

- Tagmentation
- Illumina

Genome coordinate: -9 to +11

Motif bits: 0.0 to 0.5