Published online 30 April 2010

Nucleic Acids Research, 2010, Vol. 38, No. 13 ***e137***
doi:10.1093/nar/gkq332

Titration-free massively parallel pyrosequencing using trace amounts of starting material

Zongli Zheng[1,*], Abdolreza Advani[2], Öjar Melefors[2,3], Steve Glavas[2], Henrik Nordström[2,3], Weimin Ye[1], Lars Engstrand[2,3] and Anders F. Andersson[2,4,*]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, [2]Swedish Institute for Infectious Disease Control, Solna, [3]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm and [4]Limnology/Department of Ecology and Evolution, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

Received November 5, 2009; Revised April 13, 2010; Accepted April 15, 2010

ABSTRACT

Continuous efforts have been made to improve next-generation sequencing methods for increased robustness and for applications on low amounts of starting material. We applied double-stranded library protocols for the Roche 454 platform to avoid the yield-reducing steps associated with single-stranded library preparation, and applied a highly sensitive Taqman MGB-probe-based quantitative polymerase chain reaction (qPCR) method. The MGB-probe qPCR, which can detect as low as 100 copies, was used to quantify the amount of effective library, i.e. molecules that form functional clones in emulsion PCR. We also demonstrate that the distribution of library molecules on capture beads follows a Poisson distribution. Combining the qPCR and Poisson statistics, the labour-intensive and costly titration can be eliminated and trace amounts of starting material such as precious clinical samples, transcriptomes of small tissue samples and metagenomics on low biomass environments is applicable.

INTRODUCTION

Recent developments in DNA sequencing techniques allow millions of DNA molecules to be sequenced in parallel in a short time (1–5). However, the requirement of large quantities of starting material (e.g. 500 ng DNA in the latest Roche 454 Rapid library protocol) limits the possibilities to sequence trace amounts of DNA, such as ancient DNA, precious clinical samples, EST sequencing of small tissue samples or metagenomic samples from low biomass environments. Even in applications not limited by input DNA amounts, library quantification is a substantial source for uncertainty in sequencing outcome: a too high DNA-to-bead ratio leads to a significant fraction of non-readable beads with multiple DNA templates (mixed beads), while a too low ratio will result in an inadequate amount of beads and cannot take advantage of the full sequencing capacity.

Standard methods for quantification of libraries such as UV spectrophotometry and fluorometry require DNA amounts hundreds to thousands times the amount needed for the actual sequencing and cannot distinguish amplifiable from non-amplifiable molecules, the latter stemming from, for example, inefficient ligation of adapters or DNA damage. Alternative approaches were recently suggested; including SYBR Green quantitative polymerase chain reaction (qPCR) (6), 5′ universal template Taqman qPCR or digital PCR (7). However, SYBR Green qPCR requires an accurate and difficult estimation of the DNA library size distribution, and 5′ universal template Taqman qPCR gives a relatively large variation in quantification values (7), while digital PCR is not accessible in many laboratories (8). The Roche 454 protocol recommends an optimal amount of library to be determined empirically using the sequencing-titration assay for FLX Standard libraries and, later, the emulsion-titration assay for Titanium libraries. Both methods require labour-intensive and costly emulsion-PCR and enrichment procedures, as well as additional sample material.

In this study, we have developed a qPCR method based on a highly sensitive and precise Taqman MGB-probe. The MGB probe was designed to be complementary to the adapters used to construct two different types of libraries for the Roche 454 Titanium sequencing platform. One library consisted of the traditional A and B adapters on either end of the DNA template (9), and the other consisted of the same 'Y' adapter on both ends (similar to the Illumina GA library (8); http://www.

*To whom correspondence should be addressed. Tel: +46 8524 82370; Fax: +46 8 31 49 75; Email: zhengzongli@gmail.com
Correspondence may also be addressed to Anders Andersson. Tel: +46 (0) 18 471 2725; Fax: +46 (0) 18 471 2700; Email: doubleanders@gmail.com

seqanswers.com). This qPCR setup quantifies only effective library, i.e. DNA that will be amplified to form a functional clone in the emulsion PCR. We also demonstrate that Poisson statistics, with its parameter λ equalling the qPCR-measured input library DNA-to-bead ratio, can be used to predict enrichment percentage, a key index for sequencing performance.

MATERIALS AND METHODS

Using Poisson distribution to predict enrichment percentage

Nine previously sequenced FLX Standard libraries in our sequencing core facility were re-quantified by a Taqman MGB-probe-based qPCR. These libraries had previously been sequenced using the sequencing-titration assay, as part of the FLX Standard protocol, in four concentrations (0.5, 2, 4 and 16 DNA-to-bead ratios) as measured by RNA6000 BioAnalyzer chip (Agilent). Because the sequencing was performed without enrichment, and generally almost all DNA-carrying beads passed the key filter, enrichment percentage (percentage of DNA-carrying beads) can be calculated using the sequencing result metrics (enrichment % = key passed wells/raw wells × 100%). Apart from these nine libraries, there were 23 additional libraries subjected to the sequencing titration assay in our core facility, but these samples were no longer available at the time of the current study. However, their sequencing titration assay results were useful for exploring how the sequencing outcome metrics (proportions of Single, Mix or Dots reads) change over different enrichment percentages. To plot the enrichment percentages on the *Y*-axis and input DNA-to-bead ratio on the *X*-axis, which however could not be re-quantified by qPCR, we derived the input DNA-to-bead ratio from Poisson-transformed enrichment percentage [–log (1 – enrichment fraction)]. This was appropriate since the purpose here was not to explore the effect of input DNA-to-bead ratio on enrichment percentage. We also carried out a prospective evaluation of Poisson prediction by an emulsion-titration using newly prepared AB and Y libraries.

By Poisson distribution, the probability that there are exactly k DNA molecules on one library capture bead ($k = 0, 1, 2, \ldots$) is equal to:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where λ is the input DNA-to-bead ratio. Therefore, the probability of having zero and one DNA molecule on one bead is $e^{-\lambda}$ and $\lambda \bullet e^{-\lambda}$, respectively, and the enrichment fraction is $1 - e^{-\lambda}$.

Deviation from Poisson distribution was tested using paired *t*-test comparing the difference, with 95% confidence intervals (CIs), between enrichment percentage observed from the titration sequencing and the enrichment percentage predicted by measured library concentration and Poisson distribution, after linearity transformation, ln(1 – enrichment percentage). Pearson's correlation coefficient between the observed and predicted enrichment percentage was calculated. When enrichment percentage is close to 100%, subtle enrichment differences will lead to large differences in DNA-to-bead ratio, we therefore excluded data when enrichment was > 95% (which is also far from the optimal DNA-to-bead ratio, see 'Discussion' section). This resulted in exclusion of 3 out of 36 data points. All data were analysed and plotted using the R software (10) (http://www.R-project.org.).

Construction of qPCR standards

We constructed a specific PCR product to be used as qPCR standards (available from the authors on request). One 1:1000 diluted stock FLX Standard library (1 µl) from our sequencing core facility was used as template for amplification with 5 pmol of each emPCR primer, 1 × PCR buffer, 1.5 mM $MgCl_2$, 1 U Taq DNA polymerase (Invitrogen) and 200 µM dNTP each. The PCR product was cloned with the TOPO TA kit (Invitrogen). Several colonies were selected, transferred to 50 µl of water, and boiled at 95°C for 10 min. The lysate (1 µl) was used as template for colony PCR amplification using primers (Supplementary Table S1, pCR4-TOPO vector) targeting the flanking region of the pCR4-TOPO vector ligation site. PCR products were visualized on 1% agarose gel and the cell lysate that generated an amplicon of about 360 bp (corresponding to an about 200 bp insert) was selected for Sanger sequencing using M13 forward primer. The result showed that the sequence was 202 bp long and contained one copy of adapter B and, therefore, one copy of qPCR probe complementary sequence. This cell lysate (1 µl) was then amplified using emPCR primers and purified with MinElute kit (Qiagen). The purified product was quantified using Qubit fluorescence quantification system (Invitrogen) and the number of molecules was calculated ($= \text{ng} \times 9.17 \times 10^{11}/202$). This DNA was 10-fold serially diluted and the concentrations from 10^7 to 10^2 copies were used as qPCR standards. The qPCR reaction mixture (20 µl) contained 1 × Taqman Fast Universal PCR Master Mix (Applied Biosystems), 900 nM of each emPCR primer (Invitrogen), 200 nM Taqman MGB-probe (Applied Biosystems) and 2 µl of standards or test samples. The qPCR was performed on the ABI 7900HT Fast System with the following conditions: fast cycling mode; 95°C for 20 s; 40 cycles including 95°C for 1 s and 60°C for 60 s. Fluorescence was detected during the extension step. Both standard and test samples were run in triplicates. It should be noted that, if quantifying a single-stranded library with this method, the readout must be multiplied by two to account for the double-stranded qPCR standards.

Library construction for FLX Titanium sequencing

We implemented a simplified library preparation protocol (8) that generates double-stranded DNA library using FLX Standard A and B adapters, but without biotin labelling on the B adapter. We designed a Taqman MGB-probe targeting the B adapter. We also implemented the concept of Illumina GA library and designed a Y adapter such that it is compatible with the Roche 454 Titanium platform and contains a complementary

sequence at its 'B' branch (Figure 1 and Supplementary Table S1). Genomic DNA was extracted from *Helicobacter pylori* strain HPAG1 (11) using DNeasy kit (Qiagen) according to the manufacturer's instructions. DNA was nebulized and selected for sizes between 300 and 800 bp by Solid Phase Reversible Immobilization (Agencourt) according to the Roche 454 protocol. One nanogram of DNA measured by Qubit Fluorometers (Invitrogen) was used for the downstream procedure. Polyallomer centrifuge tubes (Beckman Coulter) were used in all the experiments to minimize tube wall adsorption and denaturation of DNA (http://frstrauss.free.fr).

AB library. End polishing was conducted in 50 µl reaction volume with 1 ng of DNA, 1× End repair buffer, 5 µl end repair mix containing T4 DNA Polymerase and T4 PNK, 1 µl of 1 µM dNTPs (Enzymatics, MA, USA), and was incubated at 22°C for 30 min. After purification by AMPure beads (88 µl, 175% volume), DNA was eluted in 20 µl TE. Adapter ligation was conducted in a 25 µl reaction with 1× slow ligation buffer, 0.4 pmol of each of the A4 and B adapter (Invitrogen, USA) and T4 DNA Ligase 180 U (Enzymatics, MA, USA). We used the FLX Titanium adapter A4 (GS multiplex identifier number 4) because this adapter had performed well in previous experiments in our lab. The ligation reaction was incubated at 22°C over night, purified by AMPure beads (17.5 µl, 70% volume) and eluted in 25 µl TE. The eluted DNA was treated with 8 U Bst DNA polymerase Large Fragment (New England Biolabs), to fill in the 3′-junction nick between the adapter and sample DNA in a 30 µl reaction with 1× fill-in buffer, 30 µM dNTP and incubated at 37°C for 20 min. The double-stranded DNA library was purified with AMPure beads (21 µl, 70% volume) and eluted in 30 µl TE.

Y library. End polishing and 3′ dA extension was conducted in 22 µl reaction volume with 1 ng of DNA, 1× slow ligation buffer, 2.5 µl end repair mix containing T4 DNA polymerase and T4 PNK, 1 µl of 1 mM dNTPs (Enzymatics), 0.5 µl Klenow Fragment exo$^-$ (New England Biolabs), 1× Taq polymerase buffer (Mg^{2+} free) and 0.5 µl Taq polymerase (Invitrogen). It was incubated at 12°C for 10 min, 37°C for 10 min, 72°C for 20 min and held at 4°C. Adapter ligation was conducted directly by adding 1 µl of 1 µM Y adapter (Supplementary Table S1,

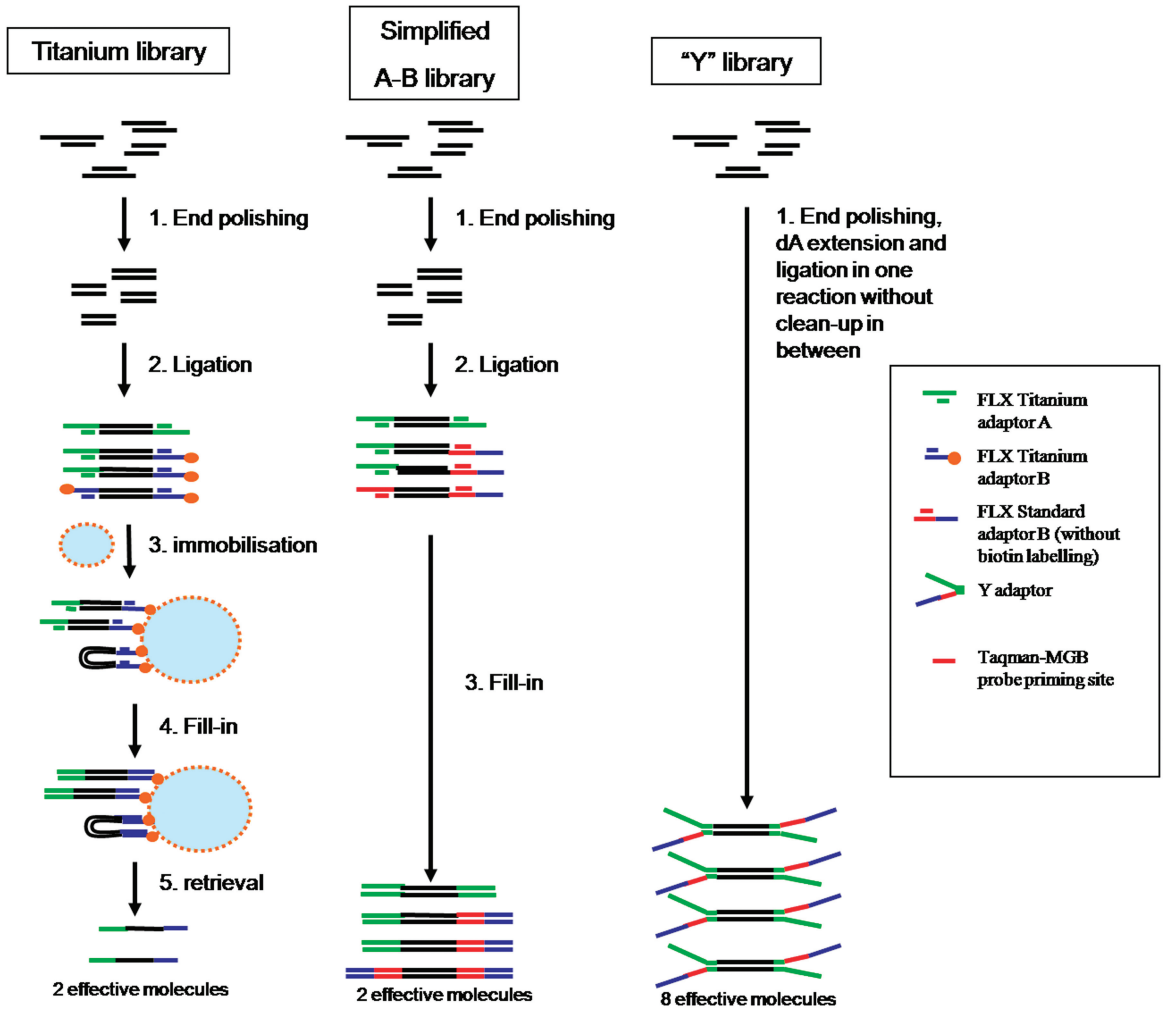

Figure 1. Schematic description of three types of library construction.

Integrated DNA Technologies), T4 DNA Ligase 180 U (Enzymatic) and was incubated at 12°C over night. The product was purified by AMPure XP beads (18 µl, 72% volume), and eluted in 30 µl TE. Immediately, not long, before library capture, the library was denatured into single-stranded, at 94°C for 2 min and held at 4°C, to avoid sequencing from both directions of a double-stranded template, which will result in mix reads.

It should be noted that this MGB probe is compatible with Roche FLX Standard library, not the FLX Titanium library nor the RL library, and that our AB and Y libraries are compatible with the FLX Titanium emPCR and sequencing kits.

RESULTS

Poisson behaviour of emulsion PCR

Optimizing the DNA-to-bead ratio in the emulsion PCR is a trade-off between minimizing the proportion of beads with multiple DNA templates and getting a sufficient amount of beads with template. The probability of getting zero, single or multiple DNA molecules per bead using different input DNA-to-bead ratios (λ) can be modelled using Poisson distribution. Since only DNA-carrying beads will be captured in the enrichment procedure, enrichment percentage corresponds to 1—probability of zero molecule per bead. Figure 2 shows theoretical enrichment percentage and probabilities of single and multiple templates per bead as functions of λ. When the DNA-to-bead ratio is small and covers the optimal range (0.08; see 'Discussion' section) it approaches a linear relationship with enrichment percentage (Figure 2B). In this range a 2-fold over- or underestimation of library concentration will all give satisfactory fractions of single-copy beads among enriched (98, 96 and 92% for ratios of 0.04, 0.08 and 0.16, respectively; Figure 2B). In contrast, if one would aim for a DNA-to-bead ratio of 1 (which should be avoided in practice) the same level of inaccuracy would result in 77, 58 and 31% single-copy beads for DNA-to-bead ratios of 0.50, 1.0 and 2.0, respectively.

We plotted the sequencing outcome metrics (single template and mixed template) as functions of enrichment percentage using the data from sequencing-titration assays of 32 FLX Standard libraries (four different DNA-to-bead ratios each) previously sequenced at our core facility. In accordance with the Poisson model, single template beads peaked at the range when the enrichment percentage was about 40–60; thereafter undesired beads quickly started dominating (Figure 2C).

Taqman MGB-probe-based qPCR for library quantification

In order to facilitate quantification of minute amounts of sequencing library, and to quantify only those library molecules that are effective, we designed a Taqman-MGB probe targeting the B adapter and the 'B' branch of the Y adapter (Figure 1 and Supplementary Table S1). To estimate the precision and reproducibility of this Taqman-MGB probe setup, we re-quantified nine of the previously sequenced FLX Standard libraries (the remaining 23 libraries were not available). The libraries were 1:100 diluted and quantified with eight replicates each. The estimated concentrations of the diluted libraries were within the range of 10^5–10^8 molecules per microlitre. The mean (standard error) of the coefficient variation was 9.5% (1.6%). In a repeated run, these libraries were further diluted 1:10, spanning the range of working concentrations of 10^4–10^7, and quantified with triplicates; the mean (standard error) of the coefficient variation was 5.3% (1.5%).

Evaluating qPCR quantification and Poisson prediction retrospectively

We compared the qPCR quantification with previous quantifications done with BioAnalyzer. Large variations in enrichment percentages were observed for the same input DNA-to-bead ratios as measured by BioAnalyzer (Figure 3A), and the enrichment percentages were not predictable using Poisson (P value for the paired t-test comparing the observed and the predicted enrichment percentage was 0.0005, mean difference −42.2% (linearity transformation 3.77, 95% CIs 1.77–5.76). In contrast, when the libraries were measured with the qPCR method (Figure 3B), enrichment percentages were predictable using Poisson statistics: there was no significant difference between observed and predicted enrichment percentage (paired t-test $P = 0.9157$, mean difference 2.0% (linearity transformation −0.020, 95% CIs 0.365 to −0.500). Likewise, there was an improvement in the correlation between observed and predicted enrichment percentages using qPCR (Pearson's correlation coefficient = 0.68) as compared with BioAnalyzer (0.46).

Evaluating qPCR quantification and Poisson prediction prospectively

To further evaluate how well MGB-probe qPCR and Poisson distribution can predict enrichment percentage, and also to demonstrate its applicability on the double-stranded libraries (AB library and Y library; Figure 1 and 'Materials and Methods' section), we performed emulsion-titration assays using newly prepared AB and Y libraries.

From 1 ng of fragmented DNA (mean size 500 bp), 1.15 million emPCR amplifiable AB library molecules and 53.6 million single-stranded Y library molecules were generated, respectively, as measured by qPCR (Figure 4A and C).

The qPCR amplification products were subsequently analysed by gel electrophoresis to make sure that the libraries had expected size distributions and to ensure that no adapter dimers had been carried over (Figure 4B and D). The putative dimers (84 bp for AB adapters and 79 bp for Y adapter) could not be observed among the amplified products. In addition, to test whether or not A–A and B–B tagged library molecules can be amplified, qPCR reactions containing the AB library templates and only one of the emPCR primers (either A or B) were performed. No detectable fluorescence (Figure 4A) and no

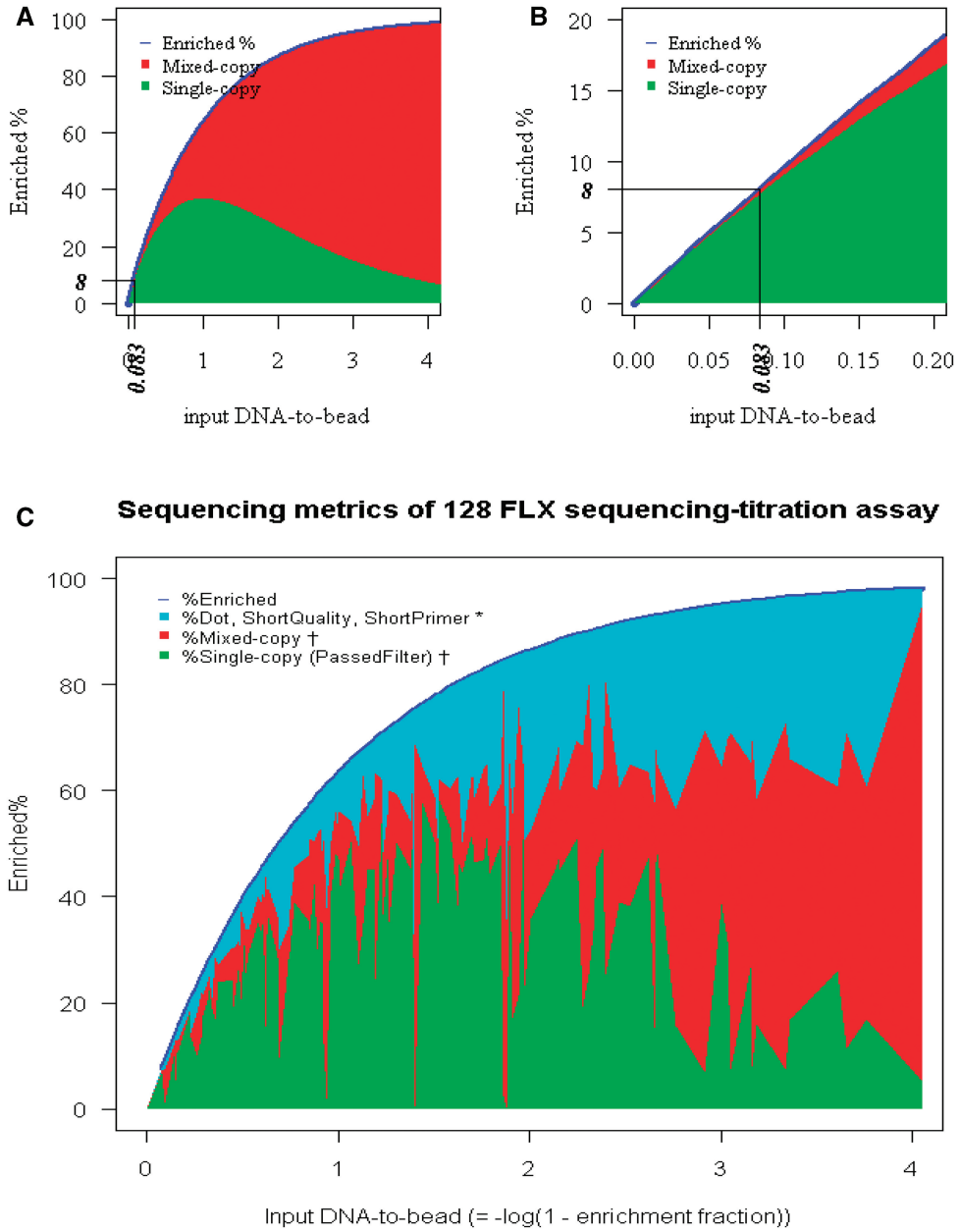

Figure 2. (**A**) Enrichment percentage, single copy and mixed copy DNAs in an emulsion drop/bead according to Poisson theory. In a subset (**B**), enriched percentage as a function of input DNA-to-bead ratio approaches linearity when DNA-to-bead ratio decreases. The 8% enrichment percentage recommended by the Roche 454 protocol corresponds to a 0.083 DNA-to-bead ratio. (**C**) Sequencing outcome metrics as functions of the percentage of DNA-bearing beads (called enrichment percentage if enrichment were performed). The 40–60% enrichment percentage zone harbored a critical turning point, after which undesired beads becoming dominant quickly. As indicated by asterisks the Roche 454 Dot filter filters away those reads having too many negative flows due to poor incorporations or interruptions, the ShortQuality filter filters those reads failed the length test because of quality trimming and the ShortPrimer filter filters those failed the length test because of primer sequence trimming. As indicated by dagger the Roche 454 pipeline uses a positive flow percentage of 70% as a cut-off to distinguish Single- from Mixed-copy of templates.

PCR products visible on the gel were obtained (Figure 4B), indicating that the qPCR measurement corresponds to the amount of A-B tagged library, not B-B tagged library. The latter contains the probe complementary site and, if amplified, would interfere with the quantification.

For the emulsion-titration assay, we used input DNA-to-bead ratios of 0.0796 for the AB library (see 'Sequencing outcome' below) and 0.1, 0.2, 0.6 and 2.0 for the Y library. According to Poisson prediction, the enrichment percentages should be 7.7, 9.51, 18.1, 45.1 and 86.5%. The actual enrichment percentages observed from the emulsion-titration assay for the five ratios (with duplicates) were 8.1 and 6.5%, 13.3 and 9.79%, 19.8 and 18.4%, 39.2 and 40.7%, and 81.3 and 81.0% (Figure 5). Hence, for these newly prepared libraries, observed

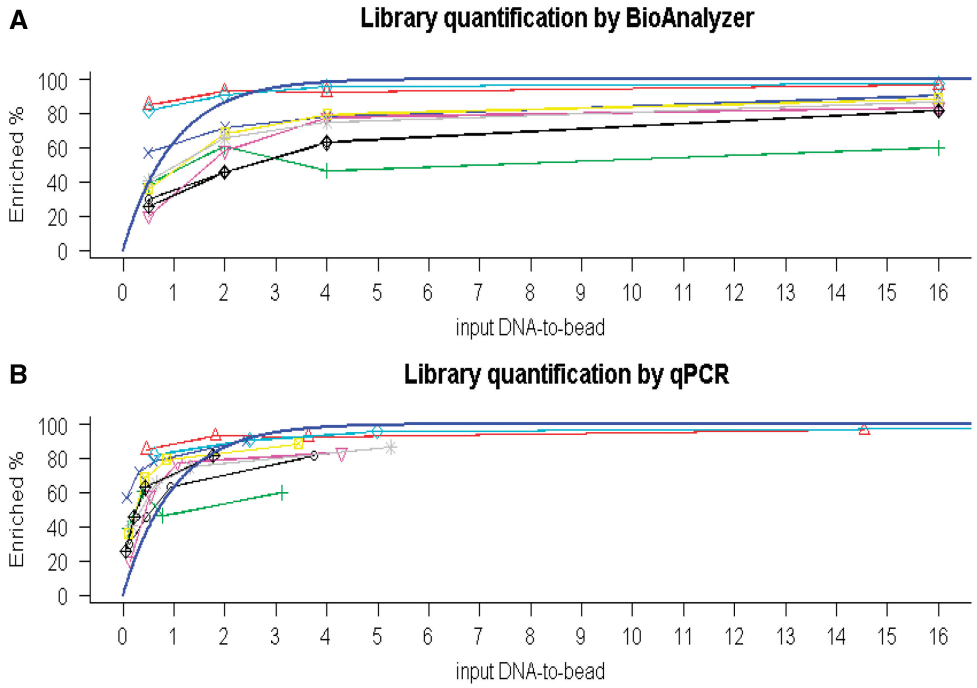

Figure 3. Nine stock libraries were previously quantified by BioAnalyzer as described in the Roche 454 Manual (**A**) and re-quantified by qPCR (**B**). Each colour indicates one library with four titration points. The blue curve indicates Poisson prediction of enrichment percentage. There was a deviation from Poisson prediction in (A) ($P = 0.0005$) and no deviation in (B) ($P = 0.9157$).

enrichment fits Poisson prediction well (Pearson's correlation coefficient = 0.998 between predicted and observed enrichment).

Sequencing outcome

We set out to sequence the AB library using two M/S lanes (2 × 1/8 plate). The Roche 454 emPCR Kit supplies 4.8 million library capture beads for one M/S lane. We therefore used 10 µl of the AB library, corresponding to 0.382 million DNA molecules and hence a DNA-to-bead ratio of 0.0796. This was expected to generate 367195 (= $0.0765 \times 4.8 \times 10^6$, where 0.0765 is the Poisson prediction of enrichment fraction $1 - e^{-0.0796}$) enriched beads. Assuming a 10% bead loss during the laboratory procedure, the remaining 90%, i.e. 330 475 enriched beads, would be nearly the amount needed according to the Roche 454 standard protocol (340 000). We processed two identical 10 µl of the library samples in parallel and thus generated two M/S lanes on a sequencing plate.

Following the enrichment process, the bead counter showed 389000 and 310000 beads for the two emPCR replicates. We loaded the enriched beads onto the sequencing plate and ran the sequencing according to the standard protocol. There were 104 091 and 96 151 filter-passed reads for the two lanes, falling well into the range of 80 000 to 120 000 reads per M/S lane of a successful sequencing run according to the Roche 454 protocol. Percentage of the filter-passed was 59.3% and 63.0%, median reads length 358 and 419 bp, and total bases 33.4 millions and 35.0 millions, respectively. Nearly all (99.3%) of the reads could be aligned to the *H. pylori* (strain HPAG1) genome and its plasmid using BLASTN with at least 30 bases matched and a significance value of 1E-6 (12), indicating that the contamination during library preparation, if any, was negligible. *De novo* assembly using Roche gsAssembler software resulted in a 1 584 532 bp genome, which was 99.3% the size of the reference genome HPAG1 (11).

DISCUSSION

With the Taqman-MGB probe-based qPCR and Poisson distribution, we were able to avoid the costly and labour-intensive titration assay. Using only one nanogram of fragmented DNA, we prepared enough library (1.15 million amplifiable AB library molecules and 53.6 millions Y library molecules, which is sufficient for 10 Titanium runs (~4 million enrichment beads to yield ~1 million high quality reads per run) for Roche 454 Titanium sequencing without the need for template pre-amplification by various means of whole genome amplification (13,14). However, it should be acknowledged that some factors causing sample loss remained, such as fragmentation of genomic DNA, low efficiency of ligation and a limited recovery in the enzymatic reaction clean-up (15). This was clearly shown by the much higher yield of Y library (sticky-end ligation) than AB library (blunt-end ligation and two additional reaction clean-up steps).

Library quantification by qPCR has earlier been proposed to overcome the lower detection limits of conventional methods (6,7). We used a Taqman MGB-probe to take advantage of its significantly higher sensitivity, specificity and reproducibility than conventional Taqman probes, at the same time only a shorter priming site is needed (16). This MGB probe rendered a precision (coefficient variation of 9.5%) comparable with digital

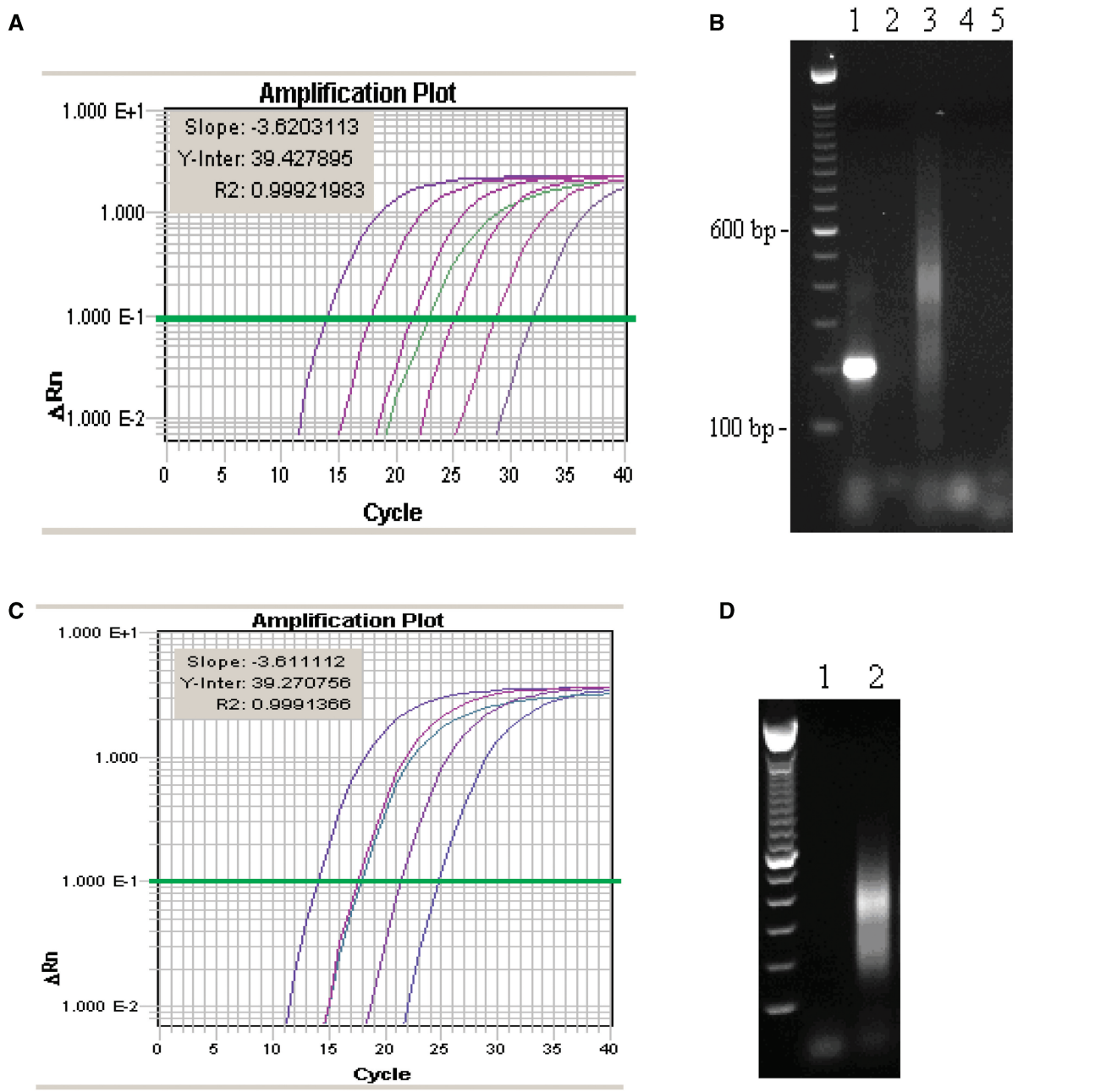

Figure 4. Library quantification by qPCR. All samples were run in triplicate and only the ones with median Ct values were plotted. (**A**) The six standards ranged from 10^7 to 10^2 copies. The tested AB library was 38 200 molecules/µl. The reactions containing the AB library template and one of the emPCR primers (A only or B only) generated no detectable fluoresces. (**B**) Gel electrophoresis of qPCR products. Lane 1: standard 10^4 copies; lane 2: no template control; lane 3: tested AB library; lane 4: the reaction containing the library template and primer A only; lane 5: the reaction containing the library template and primer B only. (**C**) The standards ranged from 10^7 to 10^4 copies. The tested Y library was 894 000 molecules/µl. (**D**) Gel electrophoresis of qPCR products. Lane 1: no template control; lane 2: Y library.

PCR (11.8%), but better than a 5′ universal template Taqman probe (21.2%) (7).

We demonstrated that the distribution of DNA on beads after emulsification follows a Poisson distribution, which enabled us to predict enrichment percentage, a key index for successful sequencing. In general, the lower DNA-to-bead input, the lower percentage of beads with mixed templates and hence the more desirable results, provided that there are enough beads for sequencing loading. As a trade-off, a DNA-to-bead ratio close to 0.08 should be aimed at for optimal results (Figure 2). This is consistent with the Roche 454 protocol recommendation of 8% enrichment percentage (corresponds to a 0.083 DNA-to-bead ratio, Figure 2B). The study using digital PCR by White *et al.* (7) used ratios in the range of 0.08 to 0.30, providing independent support for this DNA-to-bead ratio. Furthermore, the Poisson distribution demonstrates that imprecise library quantifications within 2-fold over- or under-estimations will all give satisfactory results when DNA-to-bead ratio is low, while

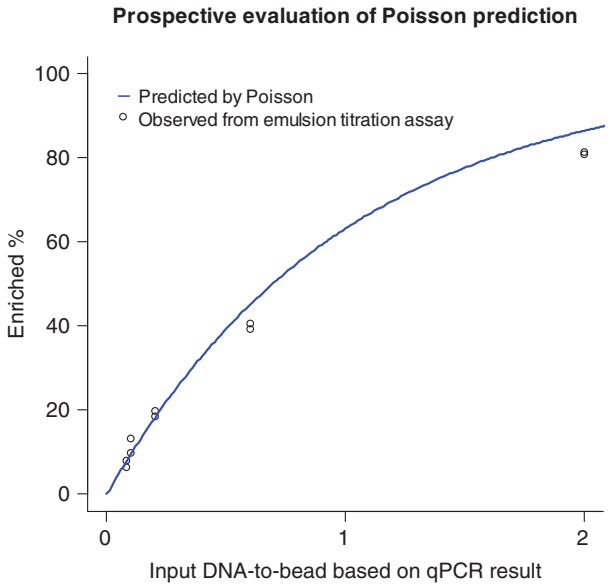

Figure 5. Enrichment percentage observed from emulsion-titration assay and predicted by Poisson distribution.

this level of inaccuracy will lead to undesired results when DNA-to-bead ratio is high. In a previous study, a linear regression model was used to correlate input DNA-to-bead ratio with enrichment percentage (17). The linear regression method may suffer from limitations of a positive intercept (meaning that there will be enriched beads even with no input library) and unlimited enrichment percentage (can be higher than 100% when DNA amount increases). However, when the input DNA-to-bead ratio is low, prediction from a linear regression will approach that from a Poisson distribution and is then acceptable.

It should be noted that the physical capture of libraries on beads is performed before emulsification for FLX Standard libraries, and during emPCR for FLX Titanium libraries. The emulsification of the Roche 454 FLX Titanium platform is performed under tightly controlled conditions, such that aqueous 'microreactors' containing no more than a single bead are generated (18). Moreover, the total volume of the aqueous phase (DNA library) added for emulsification is strictly controlled so that almost all of the aqueous droplets contain one bead, i.e., few droplets contain no bead. Only when both of these two conditions are held, the distribution of DNA molecules on the beads can be expected to follow Poisson distribution. This explains why a limited volume of library (1–10 µl for small-scale and <100 µl for large-scale emPCR, as described in the Roche 454 protocol) should be used for good results.

Our results confirmed a previous study that the AB library (a mixture of A–B, A–A and B–B tagged templates, without the need to clean out the latter two) are suitable for Roche 454 sequencing (9). The A–A and B–B tagged library templates could not be amplified in PCR. This is because hairpin structures, formed in the annealing process by the complementary sequences on either end of the single-stranded DNA, can prevent annealing of primers because: (i) the much higher local annealing temperature of the hairpin structures (>80°C for both the 40 bp 'A hairpin' and 44 bp 'B hairpin') than that of the PCR primers (60°C) and (ii) the complementary sequences in the ends of the same single-stranded DNA molecule are closer and more accessible for binding than surrounding primers at normal concentrations. This was confirmed by the observation that there was no detectable fluorescence and no visible PCR products generated in the qPCR or in a normal no-probe PCR reactions containing primer A only or primer B only. Interestingly, a recent study using transposons to generate shotgun libraries showed that A–A or B–B tagged templates generated ~1:1000 the signal of A–B tagged templates (19). The fact that their A–A or B–B templates were actually amplified (although weakly) could be due to a weaker amplification inhibitory effect associated with their shorter hairpins than with ours.

The qPCR amplification slopes of the libraries were lower than those of the standards, indicating that portions of the libraries, despite being A–B tagged, were not well amplified. Similarly in the emPCR of the FLX platform, not every DNA in the emulsion droplet will be well amplified (18). Because short amplicons are generally amplified better than long ones, we used small sizes (202 bp) as qPCR standards to quantify effective library. In contrast, a standard of size similar to the library median (e.g. 500 bp) would lead to quantification of total library rather than effective library and, consequently, an overestimated DNA-to-bead ratio. Thus, if one aims for a 0.08 DNA-to-bead ratio, a quantification based on 500-bp standards would result in too few beads being enriched. In addition, to mimic the long extension time in the emPCR, the qPCR extension time was prolonged from the normal 20 seconds to 60 seconds. If the amplicons are not in the expected size range, as seen from the gel electrophoresis, a longer extension such as an addition of 68°C for 60 s in the qPCR cycling programme may be needed.

During the revision of this manuscript, Roche released a simplified library protocol, 'Rapid', in which a starting amount of DNA of at least 500 ng is recommended. This is an improvement compared to previous protocols, but for applications where limited amounts of material are available this may still be too much. The qPCR quantification and Poisson prediction methods presented here for the Roche 454 sequencer are expected to be useful also for the SOLiD, Solexa and Ion Torrent sequencing platforms. However, our methods have been validated on one bacterial species. For sequencing of more complex samples such as metazoan or plant genomes, further validation studies might be needed.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors acknowledge the suggestions from one of the reviewers that improved the manuscript. They also thank

Wilhelm Paulander for help with cloning; Hedvig Engstrom Jakobosson for help with qPCR; Lena Eriksson for sample preparation and Sten Linnarsson for discussion on handling of trace amounts of DNA.

FUNDING

Sixth Research Framework Programme of the European Union, project INCA (LSHC-CT-2005-018704). KID grant (June 2006), the Karolinska Institutet faculty funds for funding of postgraduate students to Z.Z. Grant from The Swedish Research Council Formas (FORMAS) to A.A. Funding for open access charge: Sixth Research Framework Programme of the European Union, project INCA (LSHC-CT-2005-018704).

Conflict of interest statement. None declared.

REFERENCES

1. Harris,T.D., Buzby,P.R., Babcock,H., Beer,E., Bowers,J., Braslavsky,I., Causey,M., Colonell,J., Dimeo,J., Efcavitch,J.W. *et al*. (2008) Single-molecule DNA sequencing of a viral genome. *Science*, **320**, 106–109.
2. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
3. Shendure,J., Porreca,G.J., Reppas,N.B., Lin,X., McCutcheon,J.P., Rosenbaum,A.M., Wang,M.D., Zhang,K., Mitra,R.D. and Church,G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
4. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
5. Clarke,J., Wu,H.C., Jayasinghe,L., Patel,A., Reid,S. and Bayley,H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, **4**, 265–270.
6. Meyer,M., Briggs,A.W., Maricic,T., Hober,B., Hoffner,B., Krause,J., Weihmann,A., Paabo,S. and Hofreiter,M. (2008) From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Res.*, **36**, e5.
7. White,R.A. 3rd, Blainey,P.C., Fan,H.C. and Quake,S.R. (2009) Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics*, **10**, 116.
8. Linnarsson,S. (2010) Recent advances in DNA sequencing methods—general principles of sample preparation. *Exp. Cell Res.*, doi:101016/j.physletb.2003.10.071.
9. Wiley,G., Macmil,S., Qu,C., Wang,P., Xing,Y., White,D., Li,J., White,J.D., Domingo,A. and Roe,B.A. (2009) Methods for generating shotgun and mixed shotgun/paired-end libraries for the 454 DNA sequencer. *Curr. Protoc. Hum. Genet.*, **Chapter 18**, Unit181.
10. R Development Core Team. (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
11. Oh,J.D., Kling-Backhed,H., Giannakis,M., Xu,J., Fulton,R.S., Fulton,L.A., Cordum,H.S., Wang,C., Elliott,G., Edwards,J. *et al*. (2006) The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. *Proc. Natl Acad. Sci. USA*, **103**, 9999–10004.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Blow,M.J., Zhang,T., Woyke,T., Speller,C.F., Krivoshapkin,A., Yang,D.Y., Derevianko,A. and Rubin,E.M. (2008) Identification of ancient remains through genomic sequencing. *Genome Res.*, **18**, 1347–1353.
14. Pinard,R., de Winter,A., Sarkis,G.J., Gerstein,M.B., Tartaro,K.R., Plant,R.N., Egholm,M., Rothberg,J.M. and Leamon,J.H. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, **7**, 216.
15. Maricic,T. and Paabo,S. (2009) Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands. *Biotechniques*, **46**, 51–52, 54–57.
16. Kutyavin,I.V., Afonina,I.A., Mills,A., Gorn,V.V., Lukhtanov,E.A., Belousov,E.S., Singer,M.J., Walburger,D.K., Lokhov,S.G., Gall,A.A. *et al*. (2000) 3′-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Res.*, **28**, 655–661.
17. Sandberg,J., Stahl,P.L., Ahmadian,A., Bjursell,M.K. and Lundeberg,J. (2009) Flow cytometry for enrichment and titration in massively parallel DNA sequencing. *Nucleic Acids Res.*, **37**, e63.
18. Roche Diagnostics GmbH. (2008) *GS FLX Titanium emPCR Method Manual*. Roche Diagnostics, Roche Applied Science, 68298 Mannheim, Germany.
19. Syed,F., Grunenwald,H. and Caruccio,N. (2009) Optimized library preparation method for next-generation sequencing [advertising feature]. *Nat. Methods*, **6**, i–ii.