



Methods paper

Reducing chimera formation during PCR amplification to ensure accurate genotyping

R.P. Smyth^{a,b}, T.E. Schlub^d, A. Grimm^d, V. Venturi^d, A. Chopra^e, S. Mallal^e, M.P. Davenport^{d,*}, J. Mak^{a,b,c,*}^a Centre for Virology, Burnet Institute, 85 Commercial Road, Melbourne, Victoria, Australia^b Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria, Australia^c Department of Microbiology, Monash University, Clayton, Victoria, Australia^d Complex Systems in Biology Group, Centre for Vascular Research, University of New South Wales, Sydney, New South Wales, Australia^e Centre for Clinical Immunology and Biomedical Statistics, Royal Perth Hospital and Murdoch University, Perth, Australia

ARTICLE INFO

Article history:

Accepted 30 August 2010

Available online 15 September 2010

Received by J.V. Moran

Keywords:

PCR artifact

Recombination

Genotyping

Population genetics

454

ABSTRACT

Measurements of population diversity are fundamental to the reconstruction of the evolutionary and epidemiological history of organisms. Commonly used protocols to measure population diversity using the polymerase chain reaction (PCR) are prone to the introduction of artificial chimeras. These are often difficult to detect and can confound the correct interpretation of results due to the false generation of recombinants when the underlying DNA sample contains multiple distinct templates. This study presents a standardised procedure to suppress the formation of artificial chimeras during PCR amplification. The solution is based on the accurate determination of the efficiency and end point of the log-linear phase of a PCR. This procedure will facilitate the generation of data sets that more accurately reflect the underlying population diversity rather than artifacts introduced by the process itself.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Accurate measurements of population diversity are critical for the biological classification and the reconstruction of the evolutionary and epidemiological history of organisms. These measurements are widely used in studies of microbial diversity, pathogen transmission, the emergence of drug resistance, and genotyping, and thus have important implications for human health. The most common method to measure diversity is to PCR amplify the region of interest followed by cloning and sequencing of the PCR products. This can be problematic when sequences are amplified from a multi-template population (e.g. a pool of 16S sequences from a microbiome) as artificial chimeras can be generated. Although chimeras can be formed by a number of mechanisms, such as stochastic dissociation of the polymerase due to low polymerase processivity, misincorporation of nucleotides, and secondary structure inducing pausing and dissociation, the majority of chimeras are thought to arise due to incomplete

primer extension during the extension phase of the PCR. During subsequent cycles of PCR, partially extended products can bind to heterologous templates whereby they act as primers that can be extended to form artificial chimeras (Thompson et al., 2002; Meyerhans et al., 1990; Judo et al., 1998; Odelberg et al., 1995).

The formation of chimeras by PCR leads to the genetic mixing of sequences which can generate new diversity that is not present in the initial sample. This is a common problem for population genetic and molecular evolution analyses as artificial chimeras can alter the shape of phylogenetic trees by reducing the distances between the sequences, causing the trees to take on a star-like topology (Schierup and Hein, 2000). They can also inflate measurements of diversity, lead to the misidentification of haplotypes during genotyping, and can imply the presence of non-existent microorganisms when analysing microbial diversity (von Wintzingerode et al., 1997; Acinas et al., 2004; Lenz and Becker, 2008). The true impact of these artifacts to evolutionary interpretation remains unknown due to the inherent difficulty in distinguishing real sequences from chimeras. However, they have been reported in a variety of experimental systems and several studies indicate that it is a widespread problem. It has been estimated that up to 30% of PCR products from a standard PCR are artificial chimeras (Lenz and Becker, 2008; Wang and Wang, 1997, 1996; Cronn et al., 2002). The scale of the problem is highlighted by three studies that analysed the publically accessible NCBI 16S rRNA database for chimeric sequences – almost all the libraries (22/25 clone libraries or 11/19 phyla) tested contained obvious chimeras. (Hugenholtz and Huber, 2003; Ashelford et al., 2005, 2006). This problem is by no means restricted to the 16S rRNA database as

Abbreviations: PCR, polymerase chain reaction; DNA, deoxyribonucleic acids; HIV, human immunodeficiency virus; dNTP, deoxyribonucleotide triphosphates; qPCR, quantitative polymerase chain reaction; REP, recombination rates per nucleotide; Gag, group specific antigen; Pol, polymerase.

* Corresponding authors. M.P. Davenport is to be contacted at Tel.: +61 2 9385 2762; fax: +61 2 9385 1389. J. Mak, Centre for Virology, Burnet Institute, 85 Commercial Road, Melbourne, Victoria, Australia. Tel.: +61 3 9282 2271; fax: +61 3 9282 2100.

E-mail addresses: m.davenport@unsw.edu.au (M.P. Davenport), mak@burnet.edu.au (J. Mak).

chimeras from other experimental studies have been identified in GenBank (Ziegler et al., 2000; Kennedy et al., 2002).

One obvious solution to the problem is to avoid PCR wherever possible. In reality, this is often impractical or impossible as PCR is invaluable when only small quantities of DNA are available for analysis. In some circumstances the formation of natural chimeras (recombinants) is expected to be rare, meaning that computational methods can be used to identify and remove artificial chimeras, as is often done before analysis of microbial 16S rRNA clone libraries (Acinas et al., 2004). The importance of these methods is underscored by the plethora of bioinformatic tools that are available for this purpose, including Chimeria_Check (Cole et al., 2005), Bellerophon (Hugenholtz and Huber, 2003), CCode (Gonzalez et al., 2005), Pintail (Ashelford et al., 2005) and Mallard (Ashelford et al., 2006). The application of multiple bioinformatics approaches will detect most obvious *in vitro* induced chimeras, but is likely to overlook more subtle naturally derived chimeras that occur between highly similar sequences. Unfortunately, highly similar sequences are the most likely to form chimeras during PCR, meaning the most commonly formed chimeras are the most difficult to computationally detect. Additionally, the application of purely bioinformatic techniques to remove chimeras is questionable, as it is increasingly becoming clear that most genetic material can undergo some level of natural recombination, which is indistinguishable from chimera formation during PCR (Kraytsberg et al., 2004; Awadalla et al., 1999). Certainly, high levels of recombination are a natural feature of the genomes of organisms, such as Archaea (Papke et al., 2004), and viruses, such as HIV (Schlub et al., 2010), so key questions will always remain regarding the validity of chimeras once they are detected.

Here, a standardised methodology for the optimization of PCR conditions that demonstrably enable the recovery of large amounts of DNA without forming large quantities of artificial chimeras is presented. It is reasoned that as long as the quantity of PCR product increases at a linear rate on a logarithmic scale with each cycle of PCR, the formation of chimeric DNA will remain close to zero. The solution is based on the accurate determination of the efficiency and end point of the log-linear phase of a PCR. This methodology is tested by designing a single PCR protocol for eight different primer sets. Using next generation 454 pyrosequencing it is demonstrated that the formation of artificial chimeras during PCR can be consistently suppressed to low levels.

2. Materials and methods

2.1. Molecular clones

The wild-type (WT) HIV-1 plasmid used was pDRNL(AD8) (Freed et al., 1995). The marker (MK) plasmid used was pDRNL(AD8)PolMarker50. pDRNL(AD8)PolMarker50 plasmid was created by introducing 35 genetic marker points by nucleotide substitution into *pol* of pDRNL(AD8) about 50 bp apart. This created 34 marker segments where the formation of chimera can be measured. Each marker point consisted of two silent changes in adjacent codons so that chimera formation could be distinguished from mutations introduced during the PCR or sequencing.

2.2. Primers

PCR Primers were designed so that the 34 marker segments were amplified as eight overlapping ~350 bp amplicons. PCR primers were designed against regions that were identical between WT and MK plasmid according to the manufacturer's directions (http://www.finnzymes.com/tm_determination.html) and are listed in table S1 (supplementary data).

2.3. PCR amplification conditions

The PCR reaction volume was 15 µl and consisted of 5 µl of template DNA, 1x HF buffer, 200 µM dNTP, 1 µM of each primer, 0.3 U of Phusion DNA polymerase. 0.5x SYBR Green I (Life Technologies) was also added to the PCR reaction in order to monitor the reaction using qPCR. PCR reactions were performed in quadruplicate to guard against PCR bias. PCR cycling conditions were 98 °C for 30 s followed by a variable number of cycles of 98 °C for 10 s and 1 min for 72 °C.

2.4. qPCR

qPCR was used to determine the efficiency of the PCR reaction and the linear range of the PCR. Standard curves were constructed by diluting wild-type pNL(AD8) plasmid according to the following formula: $m = n \times 1.096e^{-12}$ where *m* is mass of a single genome in nanograms, and *n* is the size of the plasmid in bp. For example, 10,000 copies of the 12,086 bp pNL(AD8) plasmid are present in $12,086 \times 1.096e^{-12} \times 10,000 = 1.324e^{-4}$ ng of DNA. Thresholds were adjusted manually and standard curves generally contained at least five points over a greater than one-log range. Standard curves were manually adjusted to remove points outside the linear range of the PCR and were accepted if the *R*² value was greater than 0.995. qPCR reactions were performed on the Stratagene Mx3000 quantitative PCR system (Agilent Technologies). Results were analysed using MxPro software [v4.10] (Agilent Technologies).

2.5. Determination of the PCR log-linear end point

To determine the log-linear end point of the PCR, eight PCR reactions containing 2500 copies of plasmid standard were setup for each primer set. One reaction was kept on ice as a control to determine the base line fluorescence. The remaining reactions were subjected to the PCR cycling conditions, as above, using a BioRad MyCycler. The number of PCR cycles was varied by manually removing the PCR tube at the end of cycle numbers 26 through to 32. The total amount of PCR product was measured using SYBR green I dye. To ensure the total fluorescence is within the linear range of the machine, and to ensure that the concentration of SYBR green I dye is not limiting, two-fold dilutions of the final PCR product were performed in 2x SYBR green I. Total fluorescence was measured using the plate-read function of the Stratagene Mx3000 at 72 °C. Base line fluorescence, calculated from the control PCR reaction, was subtracted from the total fluorescence of the variable cycle PCR reactions and plotted against cycle number for each dilution to determine the dilution at which SYBR green dye I is not limiting. The saturation point of the PCR was established from the first dilution that the SYBR green I dye was not limiting.

2.6. Measuring chimera formation

PCR products for sequencing were created by pooling quadruplicate PCR reactions. These were purified using the High Pure PCR product purification kit (Roche, Basel, Switzerland) and unique 6 nucleotide identifiers (barcodes) were attached by following a parallel tagged sequencing method to allow multiplexing on the same sequencing run (Meyer et al., 2007). Single-stranded sequencing libraries were constructed from 1 µg of initial starting material using the 454 library preparation kit (Roche) according to the manufacturer's directions. Libraries were quantified and stored at −20 °C until further use. Emulsion PCR and sequencing were performed according to standard GS FLX titanium procedures. Sequencing data was processed to remove the 6 nucleotide barcodes and assigned to a sample only upon a perfect barcode match. All sequencing analysis was performed using software custom written in BioRuby (www.bioruby.com). Details are available from the authors upon request. The conversion of raw rates of chimera

formation into recombination rates per nucleotide (REPN) and statistical comparisons between recombination rates were performed using the methods in Schlub et al. (Schlub et al., 2010). Briefly, recombination is detected by monitoring the linking of marker points from separate genomes into a single genome. However, multiple recombination events can occur between two marker points. Consequently, all odd numbers of recombination events between two markers (1,3,5,7,9,...etc.) will be counted as a single recombination event, while even numbers of recombination events between two markers (2,4,6,8,10,...etc.) will go undetected. The observed recombination rate is a function of the true recombination rate and the distances between two marker points. Our mathematical procedure accounts for these variations to estimate the true recombination rate within the experimental system (Schlub et al., 2010).

2.7. Analysis of hot spots

95% confidence intervals, which assess whether chimera formation is uniformly distributed along the genome, were created using the 'average' optimal rate of chimera formation that was measured along the gene. From this rate the probability of observing a chimera in each interval between markers was calculated (Schlub et al., 2010). Using the binomial distribution we then calculated the 99.847% confidence interval for the expected number of chimeras to be observed in each interval. A 99.847% confidence interval is used so that after 33 comparisons (1 comparison for each interval between markers), any significant deviation will be significant at the 5% level ($0.99847^{33} \approx 0.95$). The 99.847% confidence interval for the frequency of chimeras in each interval is back transformed into a rate of chimera formation (Schlub et al., 2010). When taken as a single 95% confidence interval over the gene, any deviation outside the confidence interval is significant evidence (at the 5% level) of higher or lower rates of chimera formation.

3. Results

3.1. Experimental design

The fact that PCR can generate artificial chimeras has been known for some time and the PCR conditions which promote or suppress their formation are well understood (Meyerhans et al., 1990; Judo et al., 1998; Bradley and Hillis, 1997). They generally form during the latter plateau stages of PCR when dNTPs and primer are exhausted and PCR inhibitors have accumulated. Therefore, it is generally recommended that a minimum number of cycles to obtain a usable PCR product are carried out. On the other hand, as the product of a PCR reaction is often utilised in further downstream enzymatic processes, the temptation remains to increase the number of PCR cycles to obtain a higher yield of final product at the risk of increasing the formation of DNA chimeras. In

this study a mixture containing equimolar amounts of two highly similar plasmids was amplified using eight different primer sets (Fig. 1A). First, the efficiency and log-linear end point of the PCR were determined using an optimized PCR protocol (Figs. 1B and C). This enabled the effect of stopping the PCR during log-linear (optimal) and non-log linear (non-optimal) phases of the PCR to be tested on the formation of artificial chimeras using 454 pyrosequencing (Fig. 1D).

3.2. PCR amplification conditions

PCR reagents were selected using a rational design approach. The major cause of DNA chimera formation is thought to be incomplete extension of primers during a cycle of PCR which can happen if the polymerase dissociates before completing synthesis of the entire template. Therefore, Phusion polymerase (Finnzymes) was selected due to the widely held view that it has a low error rate and high processivity, which are characteristics that are preferred by researchers concerned with chimera formation. A high processivity increases the number of nucleotides per binding event and decreases the number of binding events required for elongation. This polymerase also has the advantage of having an extremely low error rate (4.4×10^{-7} per bp per cycle) and a fast extension speed (Wang et al., 2004). A high concentration of primer (1 μ M) was also used so that any partially extended primers are outcompeted for template for binding sites by un-extended primers in subsequent cycles. Usually the addition of excess primer is not recommended as it can decrease the efficiency of the PCR and increases the likelihood of primer dimer formation. However, quantitative PCR (qPCR) can be used to ensure that amplification is robust and consistent with large quantities of primers, and the primer concentration can always be decreased if the PCR does not perform optimally at a higher concentration.

PCR cycling conditions were also determined rationally making use of reports from the literature. A two-step PCR protocol (where annealing and extension steps are carried out at the same temperature) was used to ensure that primer binding occurs at the optimal temperature for primer extension. In keeping with previous reports, the manufacturer's recommended extension time was increased (Meyerhans et al., 1990; Judo et al., 1998). The final extension step was also eliminated as it is designed to fully extend partial PCR products, which is something to be avoided.

3.3. Determination of the PCR efficiency

The end point of a PCR is caused by the build-up of final product, which exhausts the supply of reagents such as nucleotides, and is inhibitory to DNA polymerases. Consequently, the rate at which the final product builds up is likely to be a critical component of when the PCR reaction reaches saturation point. This is called the efficiency of the PCR and is typically measured using qPCR. One well-established

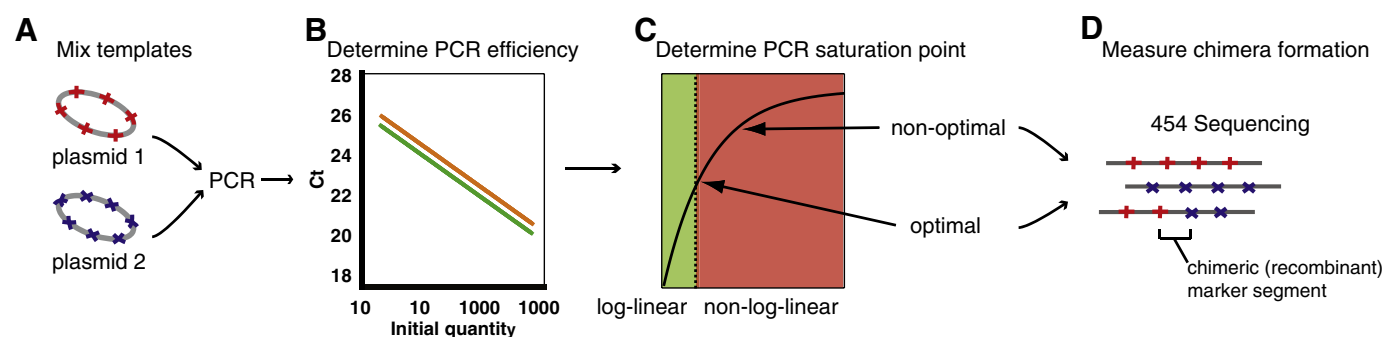


Fig. 1. Outline of methodology to exclude the formation of artificial chimeras during PCR. (A) two similar plasmids that differ at defined positions were mixed and subjected to PCR. (B) primer efficiencies were tested to ensure all primer pairs behave in the same way. (C) the optimal point to stop the PCR was found by plotting a PCR product saturation curve for each primer pair. (D) optimal and non-optimal PCR conditions were verified for chimera formation by 454 sequencing.

qPCR method is based on the addition of a fluorescent dye (commonly SYBR green I) to a conventional PCR reaction that becomes fluorescent upon binding to double stranded DNA (Rengarajan et al., 2002; Higuchi et al., 1992). A PCR efficiency of 100% indicates that for every PCR cycle the quantity of product doubles. Accordingly, a PCR efficiency of 80% indicates that each cycle the quantity of product increases by factor of 1.8 (an inefficient PCR), whereas a PCR efficiency of 120% indicates that each cycle the quantity of product increases by factor of 2.2 (signifying non-specific amplification). As small differences in the efficiency of the primers will have a large effect on the amount of final product, it is important to determine (1) that all primers amplify with similar efficiencies (2) over what range of input template the PCR remains log-linear – too little input template increases the risk of primer dimer formation whereas too much input template can inhibit the PCR. Eight sets of primers against HIV-1 *pol* were assessed using plasmid standards (range 100–10,000) and shown to amplify with near 100% efficiency (range 98%–106.4%) over a broad range of input template copy number (range 100–10,000) (Table 1). The specificity of the PCR was determined using melt curve analysis and gel electrophoresis and shown to be specific within the log-linear range of the standard curve (see supplementary data).

3.4. Determination of the PCR log-linear end point

To avoid excessive chimera formation, the key parameter that needs to be determined is the end of the log-linear phase of the PCR. The concentration of SYBR dye typically used in qPCR is too low to determine the log-linear end point of the PCR based on qPCR amplification curves alone. Consequently, the saturation in total fluorescence observed during qPCR usually reflects binding of all available SYBR green I dye rather than saturation of the PCR. Nevertheless, SYBR Green I dye can be used to quantify DNA as long as it is present in excess (Rengarajan et al., 2002). To determine the log-linear end point of the PCR, a second set of quantitative PCR reactions were carried out with 2500 copies of template for varying numbers of cycles (range 26–32 cycles). The relative amount of DNA was then quantified by measuring total SYBR green I fluorescence using the plate-read function on a Stratagene MX3000. To ensure that the SYBR green I dye was not saturating, several two-fold serial dilutions (in x2 SYBR green I) of the final PCR products were performed (Fig. 2A). The dilutions of the final PCR reactions (cycles 26–32) were plotted as log total fluorescence against number of PCR cycles to determine whether all samples were within the dynamic range of the assay. Fig. 2B shows an example plot for the first set of primers (P1). Clearly, the PCR reactions from cycle 32 and 31 are not in the dynamic range of the assay at dilutions one and two as the total fluorescence intensity does not decrease when the sample is diluted. These samples are expected to contain the most DNA indicating that the concentration of SYBR green I dye is still limiting at these dilutions. However, the total fluorescent of all the samples decreases

in a log-linear fashion after the third serial dilution, demonstrating that after this point all the samples fall within the range of the assay.

The total log fluorescence against PCR cycle number was then plotted for the optimal dilution three for all primer sets (Fig. 2C). The curves were similar for all primer sets, which was to be expected as all primers amplified with similar efficiencies and yield products of similar size. When the data was averaged for all the primer sets it showed that the optimal point to stop a PCR starting with 2500 copies of input template was at cycle 29 as after cycle 30 the PCR quickly becomes non-log-linear (Fig. 2D). Similar results were obtained with dilution four (data not shown).

3.5. Quantification of chimera formation during amplification from plasmid

To directly compare the levels of chimera formation following an optimal and non-optimal PCR, the *pol* region of HIV-1 was amplified from an equimolar ratio of two different HIV-1 plasmids and the resulting PCR product was subjected to high-throughput 454 pyrosequencing. The two HIV-1 plasmids differ at defined positions within *pol*, but are otherwise highly similar. As sequence similarity drives the formation of chimeras during PCR, this ensures that the conditions are optimal for formation of chimeras. Each marker position consists of two silent base pair changes in consecutive codons, interspersed about 50 bp apart, which allowed mutations introduced during the PCR to be differentiated from the formation of chimeras. The PCR optimization curve suggests that starting from 2500 copies of HIV the PCR remains linear on a logarithmic scale after 29 cycles (Figs. 2C and D). This optimal condition was analysed and 33 chimeras from 89,804 marker segments were detected (Table 2). The PCR optimization curve indicated that the PCR begins to become non-log-linear after 30 cycles (Figs. 2C and D). To shift the end of the PCR into the non-linear phase the number of input templates was increased from 2500 to 12,500 copies. This is expected to be equivalent to stopping the PCR between cycle number 31 and 32, based on the 100% efficiency of the PCR. For this non-optimal condition, 132 chimeras within 41,955 marker segments were detected (Table 2). This gave an overall rate of chimera (recombinant) formation of 2.4×10^{-5} recombination events per nucleotide (REPN) and 18×10^{-5} REPN for 2500 copies and 12,500 copies, respectively (Table 2). This corresponded to approximately 1 chimera per 41,700 nucleotides from 2500 initial copies, and 1 chimera per 5600 nucleotides from 12,500 initial copies. Therefore, the rate of chimera formation increased by a factor of 7.5 when the number of templates was increased from 2500 to 12,500, which was significantly different ($p < 0.01$).

3.6. Identification of artificial hot spots

The formation of natural chimeras (recombinants) occurs in many biological systems and is generally non-uniformly distributed as recombination 'hot-spots' (Jeffreys et al., 2001; McVean et al., 2004). Artificial chimeras can potentially confound the understanding of natural patterns of variation in recombination rates. To assess whether the overall rate of chimera formation is non-uniformly distributed, the rate of chimera formation was plotted for each of the 34 marker segments for the optimal and non-optimal PCR conditions (Figs. 3A and B). For the optimized PCR condition, the rate of chimera formation was evenly distributed across the template (Fig. 3A). On the other hand, for the non-optimal PCR condition there are clear regions of the template that show higher than average rates of chimera formation (Fig. 3B) ($p < 0.05$), which could be incorrectly interpreted as 'hot-spots' for recombination.

4. Conclusion

PCR has proven to be an invaluable technique but is prone to the introduction of chimeras. This study tested a method to determine the

Table 1

Primer set efficiencies. Primer efficiencies and linear range of primer sets P1–P8 from two independent quantitative PCR reactions.

Primer set	R ²	Efficiency (%)	Range (copies)
P1	0.995	98	10,000–500
P2	0.999	106.4	10,000–500
P3	1	100.7	10,000–500
P4	1	99.5	10,000–500
P5	1	98.8	10,000–500
P6	0.999	98.2	10,000–500
P7	0.999	101.8	10,000–500
P8	0.999	105.3	10,000–100

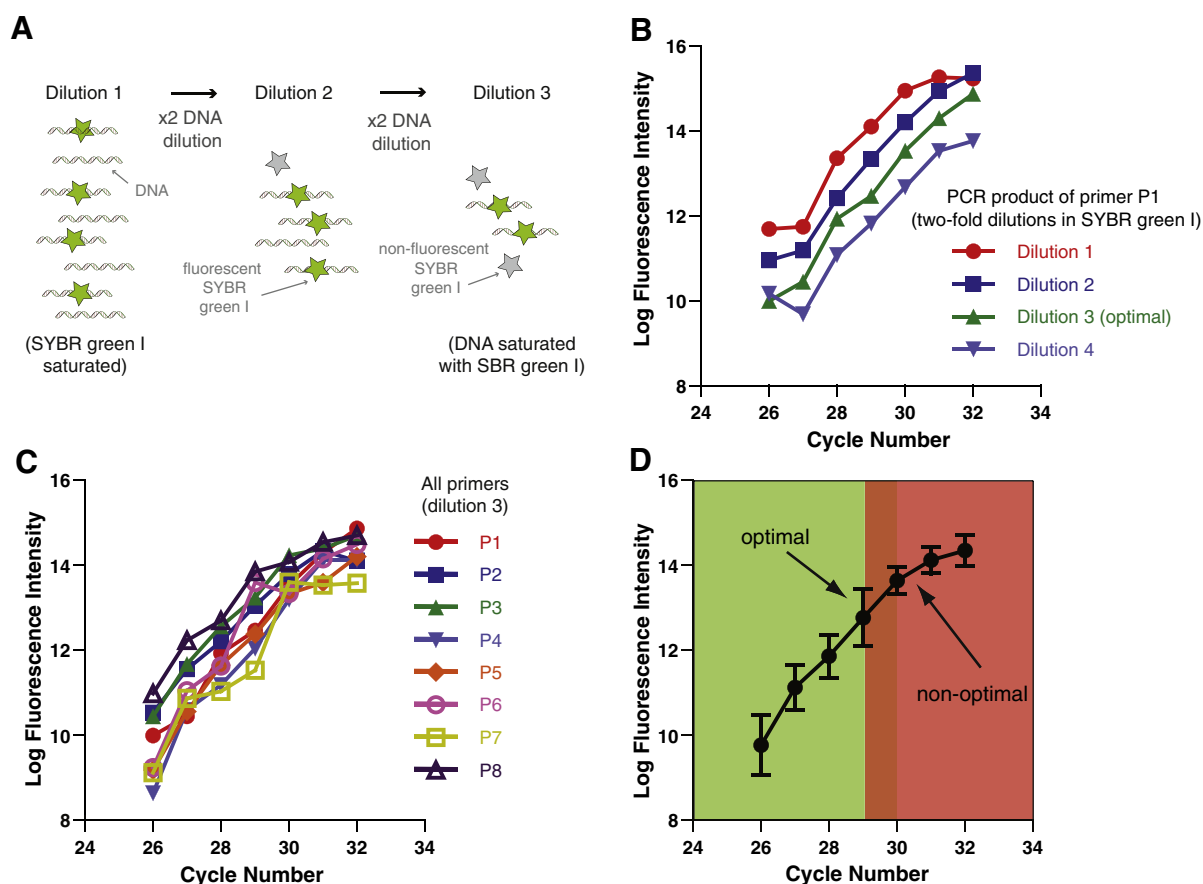


Fig. 2. Determination of the PCR saturation point. (A) dilution of the final PCR product in SYBR green dye I ensures that total fluorescence is not limited by the concentration of the dye (B) log total fluorescence (SYBR green I) for two-fold dilutions of PCR products from 26 to 32 cycles of PCR in SYBR green I. The log total fluorescence is limited by SYBR green I concentration at cycles 31 and 32 at dilutions 1 and 2. After dilution 3, the concentration of SYBR green dye is no longer limiting (C) log total fluorescence plotted for each cycle number for dilution 3 of all primer sets (D) the mean log total fluorescence for all primer sets. Error bars are 95% confidence intervals. Optimal point to stop the PCR to minimize formation of chimeras is cycle 29.

correct PCR cycling conditions for amplifying DNA without the generation of large numbers of artificial chimeras. The advantage of this approach is that multiple primers sets, with a defined amount of input template, can be run using a single PCR cycling condition. This ensures that consistency is preserved between experiments and potential biases are kept to a minimum. The approach is simple, and can be performed using only a quantitative PCR machine. This methodology should be general to any PCR polymerase, but the parameters will be specific to the individual amplification reaction and polymerase used. First, a standard curve was prepared for each primer set to determine the efficiency and linear range of the PCR. Small changes in efficiency have large effects on the yield of PCR product so it is important that all primer sets have similar efficiencies if they are to be assayed using the same PCR conditions. It is also

important to understand the linear range of the PCR to ensure that an appropriate quantity of input template is added to each PCR reaction as changing this variable can affect the outcome of the PCR reaction. Second, the identification of log-linear end point of the PCR was determined by measuring total fluorescence with SYBR green I. Other DNA quantification methods can be used, such as agarose gel electrophoresis followed by densitometry, but this method can be performed quickly and easily with a multichannel pipette and requires no further manipulation of the PCR product.

After optimal PCR conditions are determined, it is only necessary to quantify each biological sample to ensure that the appropriate amount of template is added to each PCR, which is simple to perform by routine qPCR. Another advantage of utilizing qPCR to analyse samples before sequencing is that abnormal quantitative PCR curves can indicate problems with the biological sample, such as the presence of PCR inhibitors common in biological lysates (Al-Soud and Radstrom, 2001). Careful examination of the slope of the qPCR can reveal the presence of inhibitors. Biological samples that produce abnormal curves upon quantification should be discarded or subjected to further purification as inhibition of the PCR will increase the likelihood of chimera formation. It is important to note that different enzymes and master mixes can give rise to different qPCR results. Therefore it is essential that the same polymerase mix used for qPCR and generation of saturation curves is used in the generation of PCR products for analysis. The most straightforward way of ensuring that qPCR results and saturation curves reflect the PCR master mix is to add 0.5x SYBR green I to the master mix, as is done here, as this concentration of SYBR is not thought to inhibit PCR (Mao et al., 2007).

Table 2

Rate of chimera formation. Equimolar mixtures of WT and MK plasmid were subjected to 'optimal' and 'non-optimal' PCR conditions. PCR products were barcoded and multiplex sequenced using 454 titanium pyrosequencing. Raw sequencing reads were assigned to 'optimal' and 'non-optimal' samples only upon a perfect barcode match. Sequences were analysed for recombination using a custom script written in BioRuby. The rate of chimera formation was quantified using the method outlined in Schlub et al. (2010).

Sample	Sequences analysed	Segments analysed	Chimeras observed	Rate of chimera formation (per bp $\times 10^{-5}$)
Optimal	17,516	89,804	33	2.4
Non-optimal	8165	41,955	132	18

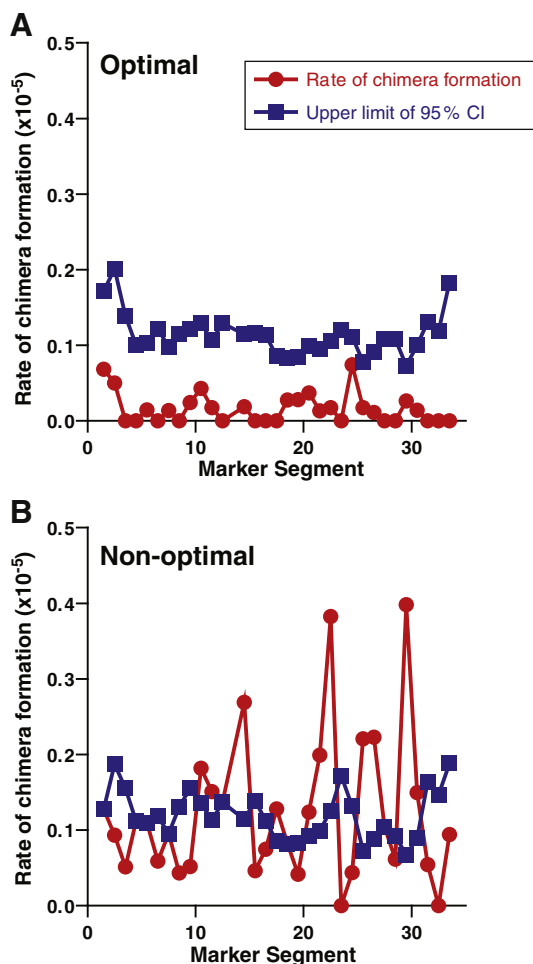


Fig. 3. Rate of chimera formation for optimal and non-optimal PCR conditions. The rate of chimera formation for each marker segment for (A) optimal and (B) non-optimal conditions. The rate of chimera formation is indicated by the red circles. The upper limit of the 95% confidence intervals (99.847% confidence interval which reduces to 95% after 33 'between marker interval comparisons', $0.99847^{33} \approx 0.95$) of the average rate of chimera formation is indicated by the blue squares. The lower limit of this confidence interval is uniformly zero. 'Hot-spots' for chimera formation are indicated by points that fall outside of 95% confidence intervals.

Several previous studies have indicated that high-fidelity polymerases, such as *Pfu* and *Vent*, generate more artificial chimeras than *Taq*, and recommended avoiding the use of proofreading polymerases (Judo et al., 1998; Lenz and Becker, 2008; Zylstra et al., 1998). This is in sharp contrast to the Phusion proofreading polymerase used in this study, which generated very few chimeras. This is probably due to high processivity of Phusion polymerase compared to standard proofreading enzymes. Phusion is an engineered polymerase containing a high-fidelity *Pyrococcus*-like polymerase domain fused to a DNA binding domain which results in a 10-fold greater processivity compared to *Pfu* and a 6-fold greater processivity compared to *Taq*, according to the manufacturer's product sheet. On the other hand, a recent study by Lahr and Katz showed that for the same initial starting template concentration and cycle number Phusion polymerase actually had a higher rate of chimera formation than *Vent*, an enzyme with lower processivity (Lahr and Katz, 2009). This counterintuitive result could be explained if the Phusion amplification was more efficient than *Vent*, resulting in an earlier saturation of the PCR and hence the formation of more chimeras for the equivalent number of PCR cycles. The Lahr and Katz study concluded that Phusion polymerase is efficient at suppressing the formation of artificial chimeras only when small amounts of initial template are amplified. This begs the question, what constitutes a small quantity of template?

The advantage of the approach outlined in this study is that optimal conditions for suppression of PCR chimeras can be exactly determined through accurate identification the log-linear end point of the PCR. Collectively, these data highlight the parameter specific nature of individual PCR reactions and the polymerase used. As a result the selection of the optimal PCR polymerase for this general procedure will need to be determined by identifying the optimal conditions for each enzyme, followed by sequencing analysis to measure the absolute rate of chimera formation for each of these PCR polymerases.

Artificial chimeras are generally indistinguishable from natural recombination, which is an important force that shapes genome evolution. It allows sites to evolve independently and also acts as a source of genetic variation. The correct interpretation of recombination rate variations relies on the assumption that the variations are due to biological recombination rather than artificial recombination introduced by the experimental procedure. This study showed that under non-optimal PCR conditions, artificial chimeras are non-uniformly distributed, which could lead to false identification of recombination hot-spots. These hot-spots could be a consequence of small differences in amplification efficiency between the primer sets, or alternatively may reflect an inherent bias in the template due to sequence specific characteristics. This result demonstrates that even small differences in initial template quantity (or cycle number) could potentially bias interpretation of data. The use of optimal PCR conditions ensures that not only are artificial chimeras formed at a low level, but that they do not contribute to variations in recombination rates.

The methodology outlined in this manuscript enables the determination of a single PCR condition for multiple primers which recovers the maximum yield of final product for downstream enzymatic processes without generating a large proportion of artificial chimeras. This ensures that measurements of diversity more accurately reflect the biological sample rather than artifacts generated by the PCR process. The routine application of this methodology will facilitate the correct interpretation of diversity measurements where PCR is used and thus shed light on genome evolution.

Acknowledgements

RPS and JM thank Dr. Alyssa Barry for critical reading of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at [doi:10.1016/j.gene.2010.08.009](https://doi.org/10.1016/j.gene.2010.08.009).

References

- Acinas, S.G., et al., 2004. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* 430 (6999), 551–554.
- Al-Soud, W.A., Radstrom, P., 2001. Purification and characterization of PCR-inhibitory components in blood cells. *J. Clin. Microbiol.* 39 (2), 485–493.
- Ashelford, K.E., et al., 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* 71 (12), 7724–7736.
- Ashelford, K.E., et al., 2006. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl. Environ. Microbiol.* 72 (9), 5734–5741.
- Awadalla, P., Eyre-Walker, A., Smith, J.M., 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science* 286 (5449), 2524–2525.
- Bradley, R.D., Hillis, D.M., 1997. Recombinant DNA sequences generated by PCR amplification. *Mol. Biol. Evol.* 14 (5), 592–593.
- Cole, J.R., et al., 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* 33 (Database issue), D294–D296.
- Cronn, R., et al., 2002. PCR-mediated recombination in amplification products derived from polyploid cotton. *Theor. Appl. Genet.* 104 (2–3), 482–489.

- Freed, E.O., Englund, G., Martin, M.A., 1995. Role of the basic domain of human immunodeficiency virus type 1 matrix in macrophage infection. *J. Virol.* 69 (6), 3949–3954.
- Gonzalez, J.M., Zimmermann, J., Saiz-Jimenez, C., 2005. Evaluating putative chimeric sequences from PCR-amplified products. *Bioinformatics* 21 (3), 333–337.
- Higuchi, R., et al., 1992. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology (N Y)* 10 (4), 413–417.
- Hugenholtz, P., Huber, T., 2003. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Syst. Evol. Microbiol.* 53 (Pt 1), 289–293.
- Jeffreys, A.J., Kauppi, L., Neumann, R., 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* 29 (2), 217–222.
- Judo, M.S., Wedel, A.B., Wilson, C., 1998. Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res.* 26 (7), 1819–1825.
- Kennedy, L.J., et al., 2002. Sequence analysis of MHC DRB alleles in domestic cats from the United Kingdom. *Immunogenetics* 54 (5), 348–352.
- Kraytberg, Y., et al., 2004. Recombination of human mitochondrial DNA. *Science* 304 (5673), 981.
- Lahr, D.J., Katz, L.A., 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* 47 (4), 857–866.
- Lenz, T.L., Becker, S., 2008. Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci—implications for evolutionary analysis. *Gene* 427 (1–2), 117–123.
- Mao, F., Leung, W.Y., Xin, X., 2007. Characterization of EvaGreen and the implication of its physicochemical properties for qPCR applications. *BMC Biotechnol.* 7, 76.
- McVean, G.A., et al., 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304 (5670), 581–584.
- Meyer, M., et al., 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res.* 35 (15), e97.
- Meyerhans, A., Vartanian, J.P., Wain-Hobson, S., 1990. DNA recombination during PCR. *Nucleic Acids Res.* 18 (7), 1687–1691.
- Odelberg, S.J., et al., 1995. Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Res.* 23 (11), 2049–2057.
- Papke, R.T., et al., 2004. Frequent recombination in a saltern population of *Halorubrum*. *Science* 306 (5703), 1928–1929.
- Rengarajan, K., et al., 2002. Quantifying DNA concentrations using fluorometry: a comparison of fluorophores. *Mol. Vis.* 8, 416–421.
- Schierup, M.H., Hein, J., 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156 (2), 879–891.
- Schlub, T.E., et al., 2010. Accurately measuring recombination between closely related HIV-1 genomes. *PLoS Comput. Biol.* 6 (4), e1000766.
- Thompson, J.R., Marcelino, L.A., Polz, M.F., 2002. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic Acids Res.* 30 (9), 2083–2088.
- von Wintzingerode, F., Gobel, U.B., Stackebrandt, E., 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.* 21 (3), 213–229.
- Wang, G.C., Wang, Y., 1996. The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* 142 (Pt 5), 1107–1114.
- Wang, G.C., Wang, Y., 1997. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl. Environ. Microbiol.* 63 (12), 4645–4650.
- Wang, Y., et al., 2004. A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro. *Nucleic Acids Res.* 32 (3), 1197–1207.
- Ziegler, A., et al., 2000. Polymorphisms in olfactory receptor genes: a cautionary note. *Hum. Immunol.* 61 (12), 1281–1284.
- Zylstra, P., et al., 1998. PCR amplification of murine immunoglobulin germline V genes: strategies for minimization of recombination artefacts. *Immunol. Cell Biol.* 76 (5), 395–405.