Whole Genome Amplification of Single Cells: Mathematical Analysis of PEP and Tagged PCR

Fengzhu Sun[1,*], Norman Arnheim[2] and Michael S. Waterman[1,2]

[1]Departments of Mathematics and [2]Molecular Biology,
University of Southern California, Los Angeles, California 90089-1113
[*]To whom correspondence should be addressed

Abstract

We construct a mathematical model for two whole genome amplification strategies, primer extension preamplification (PEP) and tagged polymerase chain reaction (Tagged PCR). An explicit formula for the expected target yield of PEP is obtained. The distribution of the target yield and the coverage properties of these two strategies are studied by simulations. From our studies we find that polymerase with high processivity may increase the efficiency of PEP and tagged PCR.

INTRODUCTION

Whole genome amplification can be contrasted with PCR in that the aim of the former is to amplify all DNA sequences in a sample whereas in the later only one specific genomic sequence is the target. Whole genome amplification methods can be used to select those genomic sequences that bind specific proteins (1), to prepare DNA probes for FISH (2, 3) and library screening and to permit multiple PCR analysis on very small samples such as single cells (4, 5) or molecules (6). RNA from a single neuron cell has also been amplified by a whole genome amplification method (7).

Whole genome amplification has two goals. The first is to increase the total amount of DNA sequences significantly (yield). The second is to insure that the amplification is not biased. Ideally all of the sequences in a sample should be amplified to the same extent (coverage).

The whole genome amplification method known as PEP (primer extension preamplification; (4)) has been evaluated for both yield and coverage when applied to single cell analysis. PEP involves multiple rounds of primer annealing followed by primer extension using a mixture (10^9 different sequences) of random 15 base long oligonucleotide primers. Starting with a single haploid cell, 50 primer extension cycles produce an estimated average of 60 copies and at least 78% of the genome is represented at least 30 times.

Given the number of cycles required, PEP is extraordinarily inefficient compared to PCR. We have developed a mathematical model of the original PEP procedure and a recent modification (8) to try and determine what factors could be altered to increase yields without lowering coverage. Our results are applicable to other whole genome amplification methods that use partially degenerate primers.

RESULTS

The Model

In PEP a collection of random primers 15 bases long are annealed to genomic DNA. We assume that they anneal and are extended with density λ; that is, the probability that a (genomic) base is at the 5′ end of an annealed primer and that it is extended is λ. After annealing, the annealed primers are extended. If λ is too small, too few Taq polymerase extension products will be made and little of the genome will be amplified. On the other hand if λ is too large, then extension from one primer will destroy downstream primers and primer extension products due to Taq polymerase's 5′ to 3′ exonuclease activity. Given the lowered processivity of the enzyme as it exhibits this exonuclease activity, an abundance of small Taq extension products will be produced. Another parameter is L, the length of a Taq extension product in nucleotides.

Consider a gene or target of length T in nucleotides. Our interest is in how many intact targets are found after n PEP cycles. Consider a single chromosome containing the target. We refer to this as a *generation 0* target or molecule. Suppose that in some PEP cycle two random primers anneal as shown in Figure 1. One primer (P_1) anneals 3′ of the target in an interval of length L-T so that its Taq extension product will contain the target. Primers in the interval (A, B) will destroy downstream primers (P_1) and their Taq extension products by the 5′ to 3′ exonuclease activity of Taq polymerase. It is possible to have a primer (P_2) annealed in the next interval of length L-T (at the 3′ end of the generation 0 molecule) since it's extension product will shorten the P_1 *generation 1* product but not destroy any of the target (Figure 1). The generation 1 product as shown in Figure 1 can, on another PEP cycle, have a primer anneal 3′ of the target and produce a *generation 2* product as shown.

The mathematical model for PCR is a branching process. If p is the probability of extending a primer so that it becomes a template for the next cycle, then after n PCR cycles the expected number of products is $(1+p)^n$, achieving an exponential growth of PCR products. The situation with PEP is related, but the branching process is not the straightforward type as in PCR. The reason for this is evident: while in the model of PEP λ is independent of generation number, the probability of producing a $(k+1)$-st generation product from a k-th generation product becomes smaller as k increases. The expected size of the population of target products after n PEP cycles is not obvious. We will now describe some mathematical results derived from the above model. Proofs of our results are given in a more mathematical treatment (9). For ease of description we will discuss the amplification of a single strand of DNA.

In Figure 2 we show the 5′ and 3′ configuration for k and k+1 generation PEP products. Y_k^3 denotes the length from the target to the 3′ end (in bases) while Y_k^5 denotes the length from the target to the 5′ end. Notice that $Y_k^5 = Y_{k+1}^3$ while usually $Y_k^3 > Y_{k+1}^5$. It is possible to derive the probability density function of (Y_k^3, Y_k^5). Surprisingly it depends only on the sum of the lengths for $k \geq 2$. This is the basis of our derivation of the closed form results we report next. In the discussion, we also consider the effect of primer-primer annealing on the model.

The Number of Extension Products Containing the Target

Let X_k^n be the total number of k-th generation target DNAs after n PEP cycles. The expected value of X_k^n is

$$E(X_k^n) = \binom{n}{k} p_k,$$

where

$$p_1 = e^{-\lambda T} - e^{-\lambda L},$$

$$p_k = \frac{e^{-\lambda T}}{(k-2)!} \int_0^{\lambda(L-T)} z^{k-2} e^{-z} \left(e^{-z} - e^{-\lambda(L-T)} \right) dz, k \geq 2$$

Figure 3 shows the expected number of second generation products as a function of primer density and number of PEP cycles when L=1000, T=250.

The standard deviation of X_k^n is of size $n^{k-1/2}$. So the mean and standard deviation are large, and since the standard deviation is smaller by a factor of $1/\sqrt{n}$ it is possible to prove a central limit theorem. For cycle numbers n of usual size (20 - 100), the variation is huge and a central limit theorem is not of practical value.

Since each generation has a successively smaller probability of having the whole target amplified, the question of which generation has the most expected target product is interesting. It turns out that the generation of "maximum size" is about the $\sqrt{n\lambda(L-T)}$-th generation. For n=50, L=1000, T=250, λ = .0015, this is about the $\sqrt{50} \approx 7$-th generation.

Above we gave the expected size of the k-th generation products after n-cycles, X_k^n. The total number T_n of product molecules is the sum over all generations.

$$T_n = \sum_{k=0}^{n} X_k^n$$

and

$$E(T_n) \approx e^{2\sqrt{n\lambda(L-T)}},$$

so the growth is neither polynomial nor exponential. Figure 4 shows the expected total number of products after 20 and 50 cycles as a function of primer density when L=1000, T=250.

Coverage Properties of PEP

Above we gave an explicit formula for the expected number of k-th generation target DNAs. Next we want to study the fraction of the genome that is amplified a certain fixed number of times, say M. We refer to this fraction as *coverage*. From ergodic theory in mathematics, we know that the expected coverage equals the probability that a specific target is amplified M times. Because we can prove a central limit theorem for X_k^n (Sun and Waterman 1994), we can approximate the probability that a target of length T is covered by at least M_k k-th generation target DNAs in the following way. Using a recursive formula (Sun and Waterman 1994), we can calculate $Var(X_k^n)$. Then by the central limit theorem we have

$$P\{X_k^n \geq M_k\} = P\left\{ \frac{X_k^n - EX_k^n}{\sqrt{Var(X_k^n)}} \geq \frac{M_k - \binom{n}{k} P_k}{\sqrt{Var(X_k^n)}} \right\}$$

$$\approx 1-\phi\left(\frac{M_k-\binom{n}{k}P_k}{\sqrt{Var\left(X_k^n\right)}}\right)$$

where ϕ is the distribution function of a standard normal with mean 0 and variance 1. This approximation is good only when k is small.

It is difficult to obtain a limit distribution for T_n, the total number of target molecules after n cycles. Simulations showed that the variance of T_n is very large compared to its expectation. So a central limit theorem can not hold for T_n. To obtain the probability that a target is amplified at least M times, we can resort to simulations. In all the simulations described below, we replicated n cycles of PEP 5000 times and, as an example, we let L=1000 and T=250. Preliminary simulations showed that for n=20 PEP cycles, λ=0.002 gave the largest 5% quantile, the value of M that 95% of the simulation values exceeded. In the first set of simulations, we fixed the primer density at 0.002 and studied the effect of the number of PEP cycles. Figures 5 a and b give the histogram for the yield of target DNAs after 20 and 50 cycles respectively. For n=20, the simulation showed that in most experiments the target was amplified around 150-200 times with mean 330 and standard deviation 226. In 95% of the simulations the target was amplified at least 63 times. Extrapolating this simulation of one target to all the targets in the genome from ergodic theory allows us to conclude that 95% of the genome was represented in at least 63 copies. For n=50, in most experiments the target was amplified around 50000 times with mean 94895 and variance 69667, and 95% of the genome was amplified at least 15586 times. The improvement is enormous. It is important to note here that the yield does not center around its mean. The mode and median are much smaller.

Figure 4 b shows that for 50 PEP cycles, the primer density 0.002 is not optimum from the view of expected target yield. In another simulation we chose n=50 and a primer density λ = 0.01 which is close to the optimum. Out of the 5000 replications, there were 1282 times that the target yield exceeded 5×10^5. In order to compare the coverage with that when using primer density 0.002 (Figure 5 b), we draw the histogram on the same scale as for λ =0.002 (Figure 5 c). We also see that in most of the simulations (around 2000) we obtained less than 10000 copies of the target. The yield varied enormously. With probability 95% the target was amplified at least 23 times or 95% of the genome was represented at least 23 times, in contrast to the fact that 95% of the genome was represented at least 15586 times when λ = 0.002.

The above discussions show that in the design of experiments, not only do we need to consider the expected yield, but also need to consider the density function (shape), since under some conditions the yield of PEP varies enormously. Various experimental conditions should be carefully designed to obtain both good yield and coverage.

Whole Genome Amplification with T-PCR

In the previous sections we described the expected yield and the coverage properties of PEP. Experiments show that PEP does not amplify the DNA from a single cell up to amounts that can be detected on ethidium bromide stained gels after 50 PEP cycles (4). The tagged random primer method (8) attempts to combine the coverage properties of PEP with exponential amplification by PCR to give higher yields. Primers are designed with a random 3′ tail that can bind to arbitrary DNA sequences, and a constant 5′ head (the tag), for the subsequent amplification of the primer extension products. In the first step, $n \geq 2$ PEP cycles are carried out

using these tagged random primers. In the first PEP cycle, the 3′ tails of the tagged random primers anneal to the single-stranded sequences and *Taq* polymerase extends the primers by a constant length L. Note the first generation sequences are only 5′ end tagged. A second generation sequence is tagged at both ends if and only if the 5′ end of its first generation ancestor is tagged. Because we suppose the length of Taq extension is constant, third or higher generation sequences are always tagged at both ends (Tag-sequences) (Figure 6). After n PEP cycles, unbound primers are physically removed. In the second step, PCR is applied using primers complementary to the constant region of the tagged random primers. During this step molecules containing tag primers at both ends are amplified exponentially.

Because a tagged sequence will be amplified exponentially in the PCR step, we only need to consider coverage by Tag-sequences. Under our model we have a recursive formula for the probability c_n that a target of length T is covered by Tag-sequences after n PEP cycles (9).

$$1-c_{n+1}=(1-c_n)(1-h_n),$$

where $c_1=0$ and h_n is given by

$$h_n=\exp(-\lambda T)\left\{\iint_{0<x+y<L-T}\left[1-\prod_{i=1}^{n-1}\left[1-\left(1-e^{-\lambda y}\right)\left(1-e^{-\lambda ix}\right)\right]\right]\lambda^2 e^{-\lambda(x+y)}dxdy+\right.$$

$$\left.+e^{-\lambda(L-T)}\left(\left(\lambda(L-T)-\frac{1}{n}\right)+\frac{1}{n}e^{-\lambda n(L-T)}\right)\right\}.$$

c_n is the expected fraction of the genome that is represented by Tag-sequences. We refer to c_n as the *coverage* of T-PCR. Figure7 shows the coverage of target length 250 after 2, 3, 4, and 5 PEP cycles when L=1000. Next let us only consider n=5 PEP cycles. From Figure7 we see that if the primer density is low, the coverage is also low. When the primer density is 0.002, the coverage reaches its maximum of 58%. It is impossible to compare T-PCR theory directly with T-PCR experimental results since, for the experimental determination of coverage used in the T-PCR paper, the fraction of cosmids that hybridized to the T-PCR products is not the same as the coverage defined here.

Another factor that may affect T-PCR is the different rates of amplification of Tag-sequences during the PCR step. It has been observed that short sequences are more efficiently amplified than long sequences in PCR. Therefore in the final T-PCR products, short sequences will dominate. For whole genome amplification to be effective, any aliquot should contain roughly equal numbers of amplified sequences from any parts of the genome. If long sequences are rare relative to short ones in the final T-PCR products, then the fraction of the genome that are represented at least a certain number of times, such as M, can be affected.

As an example, let us assume that the efficiency of PCR is $r(t)$, where t is the length of the target sequence. We assume that the efficiency decreases with the length of the target. Then the fraction of the genome that is covered by Tag-sequences of length at most $T+l$ ($T+l\le L$ nucleotides) after n PEP cycles is represented at least $(1+r(T+l))^m$ times after T-PCR, with m equals the number of PCR cycles. This follows because sequences longer than $T+l$ are amplified less efficiently than $r(T+l)$. Using our model, we can estimate this fraction. We do not have an explicit formula for this quantity and resort again to simulations. We use n=5, L=1000 and the optimum primer density λ= 0.002 (Figure7). Table 1 gives the fraction of the genome that is covered by Tag-sequences of length at most $T+l$. Thus as $T+l$ decreases, the coverage is reduced

from 0.554 when $T+l=1000$ to as little as 0.172 when $T+l=500$. We emphasize that Table 1 is given in the form of a cumulative probability distribution; that is, 0.406 = fraction covered by sequences of length $T+l = 750$ or less. Therefore the fraction covered by DNA sequences of length between 600 and 750 equals 0.406 - 0.274 = 0.132. The coverage 0.554 for $T+l = 1000$ is a little less than the predicted coverage 0.58 due to the fluctuations in our simulations.

Results When the Initial Number of Molecules is Greater Than One

Suppose we have m double-stranded molecules at first and we amplify them using PEP or T-PCR. The expected target yield will be $2m$ times the expected target yield from a single stranded sequence. The effect on coverage is not so simple. For PEP, let $c_n(M)$ be the fraction of the genome that is amplified at least M times after n PEP cycles from amplifying a single stranded sequence. Then the fraction of the genome that is amplified at least M times will have a lower bound $1-(1-c_n(M))^{2m}$. For T-PCR, let c_n be the fraction of the genome that is covered by Tag-sequences from amplifying a single stranded sequence. Then the coverage will be $1-(1-c_n)^{2m}$ if we amplify m double stranded sequences.

DISCUSSION

It was reported (4) that in a series of 50 cycle PEP experiments, the average yield of a specific fixed target was estimated to be 62 Taq extension products, and that about 78% of the genome was amplified at least 30 times. Using our model we can estimate the primer density used in those experiments. We adjusted the primer density in our simulations to obtain an expected target yield of about 60. There are two solutions because of the shape of the target yield as a function of the primer density (Figure 4 b). The primer densities giving an expected target yield of 60 are around 0.00025 or 0.035. First we chose $\lambda = .035$ and found that the variance of the number of copies of the target is huge (data not shown). In about 4900 out of 5000 replications, the target was not represented in the final PEP products, *i.e.* 4900/5000=98% of the genome was not amplified. This fact contradicts the experimental results. Next we chose $\lambda = .00025$. Figure 8 shows the histogram of the yield of target DNAs for $\lambda = .00025$ and n=50. From this figure we see that in most of our simulations the target was amplified around 40 times with mean 55 and standard deviation 29, and that 79% of the genome was amplified at least 30 times, a result very close to that observed experimentally. We estimate from our model therefore that the primer density in those experiments was around 0.00025 using 40 μM PEP primers. If we start from a diploid cell, i.e, two double-stranded DNA sequences, at least $1-(1-.79)^{2\times2}$= 99.8% of the genome will be amplified at least 30 times from the above formula.

Next we consider the sensitivity of the estimated primer density with respect to the length of *Taq* extension. If the target length T is very small compared to L, we can take T=0 and both the expected target yield and coverage are functions of λL from our formulas. If λL is kept constant, the expected target yield and coverage will be constant too. That is, decreasing (or increasing) L by a number of times is equivalent to increasing (or decreasing) λ by the same number of times. For any fixed $T > 0$, we do not have an explicit formula relating λ and L to give the same expected target yield. We use simulation again in this case. For T=250 and L=500, using the above method, we estimated the primer density used in the experiments (4) was around 0.0008. Using simulations, we found that for T=250 and any $L > 500$, a rough estimate of the primer density used in the experiment can be estimated by a formula $\lambda = .19/(L-250)$.

Our model assumes that primers and Taq polymerase are not limiting. This assumption could be invalid for two reasons. First, PEP generates new templates for extension each cycle

while the number of Taq molecules remains constant. It is possible that, before 50 cycles are completed, not enough Taq polymerase molecules to extend all the annealed primers are available. Under the experimental conditions used by PEP (4), it is estimated that there are approximately 2×10^{15} primers and 1.5×10^{11} Taq polymerase molecules. Under the assumption that primers anneal to the single stranded templates according to a Poisson process with constant rate (and there is no primer-primer annealing; see below), we found that when primer density is less than a critical value of $\lambda_c(L)$, primers and enzymes are not limiting. We calculated that λ_c (500) = 0.0009 and λ_c (1000)=0.0011 (The exact calculation can be obtained from the first author). The primer density $\lambda = 0.00025$, estimated to be associated with extension length of 1000bps from the original PEP data (4), does not lead to limiting primers or enzymes before 50 cycles as λ = 0.00025 is smaller than the corresponding critical primer density λ_c (1000) = 0.0009.

A second possibility is that Taq could be limiting due to primer-primer interactions which can serve as extension templates. If one assumes that primer-primer annealing is as likely as primer-target annealing, then the Taq enzyme should be proportionally distributed among the the two possible template forms. Since the number of base pairs represented in the primers $\left(15\times2\times10^{15}\right)$ is greater than the DNA in a single cell $\left(6\times10^{9}\right)$ by a factor of 5×10^{6}, only $\left(1.5\times10^{11}\right)/\left(5\times10^{6}\right)=3\times10^{4}$ of the $\left(1.5\times10^{11}\right)$ Taq molecules may be available in the first round of PEP to extend primers on genomic DNA. Therefore the primer density could not be greater than 0.000005 $\left(=\left(3\times10^{4}\right)/\left(6\times10^{9}\right)\right)$. In later rounds, fewer Taq molecules will be available to extend primers annealed to the templates not arising from primer-primer interactions since the same number of enzyme molecules must also be partitioned among the 1st, 2nd and later generation extension products. Therefore the experimental results would be worse than our model predicts using $\lambda = 0.000005$. In fact our simulations show that, with this primer density, it is impossible to achieve the yield and coverage reported in the PEP experiment (4).

Finally, if primer-primer interactions had a major influence on PEP efficiency then we might expect that lowering the primer concentration would improve the efficiency. To the contrary, extensive optimization of random primer concentration showed that lowering the concentration below 40 uM, gave no improvement of PEP efficiency (Zhang et al. 1992 and unpublished data, L. Zhang and N. Arnheim).

The experimental results are far below what our model predicts under optimal conditions. There are several possible reasons for this. The primer density might be too low or the length of Taq extension is shorter than L=500bp. Using a higher primer density may be difficult owing to the already very high levels which border on inhibiting the reaction. Since yield and coverage are approximately a function of λL for $L \geq 1000$ from our analysis, increasing the extension length L is equivalent to increasing the primer density λ. Therefore a polymerase with higher processivity leading to extension length larger than 1000 might improve the experimental results.

For T-PCR, if the primer density is 0.00025 which is the primer density we estimated in the PEP experiment where the concentration of primers is 40μM, the coverage is only 11% for L=1000, T=250 and n=5 (Figure7), which is much smaller than the coverage obtained by PEP. If we start from a diploid cell, using our formula, we see that the coverage is $1-(1-0.11)^4=0.37$, which is still very low from a practical point of view.

While our theoretical calculations suggest that T-PCR may be less efficient than PEP for single cell analysis, a direct experimental comparison cannot be made between them. T-PCR coverage has been defined experimentally by hybridizing radioactive T-PCR product made from DNA isolated from a pulse field gel purified single yeast chromosome preparation to a yeast cosmid library. Since even a T-PCR product of a few hundred base pairs in length can yield a

hybridization signal on a cosmid with a 40,000 bp insert, the fraction of the target which is actually amplified cannot be known exactly.

Another whole genome amplification method called DOP (3) uses a primer containing 5′ and 3′ segments with a specific sequence interrupted by a region where any one of the 4 bases may be present (of the general form 5′*******NNNNNN*****3′). Unlike T-PCR, repeated cycles of amplification are carried out using only this primer. Primer extension products carrying primer sequences at both ends (at least second generation products) may further interact with the primer in a quasi-random fashion (PEP-like) or specifically at the end (PCR-like) depending upon the experimental conditions. Since it has been used only for clone preparation and FISH, the coverage for DOP cannot be compared to that of either PEP or T-PCR for single cell analysis.

Only PEP has been evaluated for both yield and coverage for single cell analysis. However a final evaluation of these techniques must await a direct comparison of coverage and yield among the methods that are defined for the specific application desired.

REFERENCES

1. Kinzler, K.W. & Vogelstein, B. (1989) *Nucl. Acids Res.* **17**, 3645-3653.
2. Ludecke, H., Senger, G., Claussen, U. & Horsthemke, B. (1989) *Nature* **338**, 348-350.
3. Telenius, H., Carter, N.P., Bebb, C.E., Nordenskjold, M., Ponder, B.A. & Tunnacliffe, A. (1992) *Genomics* **13**, 718-725.
4. Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W. & Arnheim, N. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 5847-5851.
5. Kristleifur, K., Chong, S.S., Van den Veyver, I.B., Subramanian, S., Snabes, M.C. & Hughes, M. R. (1994) *Nature Genet.* **6**, 19-23.
6. Dear, P.H. & Cook, P.R. (1993) *Nucl. Acids Res.* **21**, 13-20.
7. Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M. & Coleman P. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 3010-3014.
8. Grothues, D., Cantor, C.R. & Smith, C.L. (1993) *Nucl. Acids Res.* **21**, 1321-1322.
9. Sun, F. & Waterman, M. (1995) Submitted to *Advances in Applied Probability.*

NOTE TO TYPESETTER: TABLE 1

$T+l$	500	550	600	650	700	750	800	850	900	950	1000
c	.172	.222	.274	.327	.365	.406	.437	.469	.499	.528	.554

NOTE TO TYPESETTER: TABLE CAPTIONS

Table 1: Fraction c of the genome that is covered by Tag-sequences of length at most $T+l$ after 5 cycles with primer density λ=0.002.

NOTE TO TYPESETTER: FIGURE CAPTIONS

Figure Legends

Figure 1: PEP with target length=T, Taq extension product length=L, ||| = the primer annealing sites. The box denotes the target.
Figure 2: Generations k and k+1 with ||| = the primer annealing sites. The box denotes the target.
Figure 3: Number of second generation products as a function of primer density and number of PEP cycles. L=1000, T=250.

Figure 4: Total number of products after (a). 20 and (b). 50 cycles as a function of primer density. L=1000, T=250.
Figure 5 : Histogram of the target yield after (a). 20 and (b). 50 cycles with primer density 0.002. (c). 50 cycles with primer density 0.01. M = number of amplification products and the frequency refers the number of occurrences out of 5000 replications. All instances of M geq 500,000 are represented by the bar at 500,000.
Figure 6: The mechanism of T-PCR. The sequences with boxes at both ends are Tag-sequences. Third or higher generation sequences are always Tag-sequences.
Figure 7: T-PCR coverage after 2,
3, 4, and 5 cycles with L=1000 and target length 250.
Figure 8: Histogram of the target yield after 50 cycles with primer density 0.00025 (5000 replications).