Genomics **62,** 500–507 (1999)
Article ID geno.1999.6048, available online at http://www.idealibrary.com on IDEAL®

Optimized Multiplex PCR: Efficiently Closing a Whole-Genome Shotgun Sequencing Project

Hervé Tettelin,*[,1] Diana Radune,* Simon Kasif,† Hoda Khouri,* and Steven L. Salzberg*[,‡]

*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850; †Department of Electrical Engineering and Computer Science, University of Illinois, Chicago, Illinois 60607; and ‡Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218

Received June 8, 1999; accepted October 26, 1999

A new method has been developed for rapidly closing a large number of gaps in a whole-genome shotgun sequencing project. The method employs multiplex PCR and a novel pooling strategy to minimize the number of laboratory procedures required to sequence the unknown DNA that falls in between contiguous sequences. Multiplex sequencing, a novel procedure in which multiple PCR primers are used in a single sequencing reaction, is used to interpret the multiplex PCR results. Two protocols are presented, one that minimizes pipetting and another that minimizes the number of reactions. The pipette optimized multiplex PCR method has been employed in the final phases of closing the *Streptococcus pneumoniae* genome sequence, with excellent results. © 1999 Academic Press

INTRODUCTION

In the late stages of a whole-genome shotgun sequencing project, most DNA sequences will be assembled into large contiguous blocks, or contigs (Fraser and Fleischmann, 1997). As the project nears completion, the number of contigs grows smaller as the contigs themselves grow larger. Due to nonrandomness in the library and unclonable sequences, some regions of the genome are not represented in the contigs, resulting in gaps. Other gaps result from extremely GC-rich or GC-poor regions and large repeat sequences. Significant effort is needed to close these gaps to finish the project. Constructing the initial shotgun clone library with plasmid vectors allows double-strand sequencing, using universal forward and reverse primers, producing sequence data from both ends of most clones. In some cases, gaps between contigs will be spanned by clones whose forward sequencing read is located at the extreme end of one contig and whose reverse read ("clone mate") is located at the end of another contig. Such gaps are called sequence gaps, and they can be "walked" by synthetic primers using the shotgun clone as a template (e.g., see gap closure methods in Fleischmann *et al.* (1995)). However, many contig ends will remain unlinked, especially when no physical map of the genome is available. Therefore the order of these contigs and the size of the gaps in between them are unknown.

Some of these physical ends can be extended by primer walking directly on genomic DNA (Heiner *et al.*, 1998). The efficiency of this approach is highly dependent on the purity and integrity of the genomic DNA, but it can be useful in linking more sequences or contigs to the contig's end. Genomic primer walking becomes tedious if the gap is larger than a few hundred basepairs, and any contigs linked this way still need to be checked by PCR to confirm their order in the overall genome. Walking on genomic DNA is possible only if the region of interest is unique in the genome, so that the walking primer will hybridize at only one location on the DNA and produce a unique sequence. Unfortunately, physical ends (and gaps) are frequently the result of repetitive sequences that cannot be resolved by sequence assembly algorithms. Because such repeats are usually longer than the average sequence read (else they would not have caused a problem for the assembler), walking using a primer located outside the repeat will not get across the repeat and therefore will not extend the physical end into the gap.

This problem can be circumvented by generating PCR products across each gap using unique primers located outside repeats. These PCR products can subsequently be walked using the product itself as a template, where the repeats do not cause a problem because they are unique within the PCR product (except in the case of long tandem repeats). In addition, PCR products do not need to be cloned prior to sequencing, and therefore regions potentially toxic to the host (another cause of gaps in a shotgun sequencing project) will nonetheless be sequenced.

To cover all gaps with PCR products, each physical end must be tested by PCR against all of the other

[1] To whom correspondence should be addressed. Telephone: (301) 838-3542. Fax: (301) 838-0208. E-mail: tettelin@tigr.org.

0888-7543/99 $30.00

ends. This can be achieved by combinatorial PCR. When a project reaches the stage where, for example, there are 24 physical gaps remaining, one would synthesize 48 primers and use $\binom{48}{2} = 48 * 47/2 = 1128$ PCRs. Running this many reactions is tedious and painstaking. Our methods present much more efficient alternatives.

The strategy proposed here is to use multiplex PCR (Burgart *et al.*, 1992) combined with a pooling algorithm. Like combinatorial PCR, multiplex PCR can test all possible pairs of primers. Unlike combinatorial PCR, multiplex PCR does not need a separate reaction for each pair. Instead, multiple primers are pooled in a single PCR, and if any product results, further experiments are conducted to determine which two primers are "mates." There are physical limits to how many primers can be included in a single test tube; under our experimental conditions (see Materials and Methods), we have had success using up to 30 primers.

The formal problem associated with gap closure can be stated succinctly as follows: given a set containing N PCR primers and a limit K on the number of primers that can be included in a single PCR, what is the minimum number of reactions possible to check all the gaps in a genome? We also address a closely related question, which can be stated as: what is the minimum number of pipetting operations necessary to set up and run all the PCRs? This question is important because hand pipetting is a significant source of errors in the lab, and minimizing these operations reduces the likelihood of such errors. We call our method for answering the latter question pipette optimized multiplex PCR, or POMP. To answer the first question, we have a related method called reaction optimal multiplex PCR, or ROMP.

This problem is a variant on a class of well-known problems in graph theory, in the general area of block design. The ROMP solution presented below is a concrete application of one of the methods in this area. For details on the mathematics of block design, see Hall (1996).

The POMP method has been or is being used at The Institute for Genomic Research (TIGR) for the ongoing genome sequencing projects of *Streptococcus pneumoniae*, *Shewanella putrefaciens*, *Staphylococcus aureus*, and *Chlorobium tepidum*.

MATERIALS AND METHODS

Multiplex PCR. The conditions for the multiplex PCR were as follows: 20 mM Tris–HCl (pH 8.4), 50 mM KCl, 1.6 mM $MgCl_2$, 0.5 M betaine, 0.3 mM dNTP, 0.35 μM concentration of each primer, 0.7 μg of genomic DNA, and 5 units of Platinum *Taq* DNA polymerase (Gibco BRL Catalog No. 10966-018). The final reaction volume was 50 μl. Cycling conditions were 95°C for 5 min, 30 cycles of 45 s of melting at 94°C, 45 s of annealing at 60°C, 6 min of polymerization at 70°C; followed by a 10-min final extension at 72°C.

Long-range multiplex PCR. The conditions for the long-range multiplex PCR were as follows: 40 mM Tris–HCl (pH 9.3), 15 mM KOAc, 1.1 mM $Mg(OAc)_2$, 4.6 mM KCl, 1.5 μM EDTA, 0.3 mM dNTP, 15 μM dithiothreitol, 0.38 μg BSA, a 0.35 μM concentration of each primer, 0.7 μg of genomic DNA, 0.12 units of *Tth* DNA polymerase + 0.01 μg of TthStart antibody (Clontech Catalog No. K1906-1). The final reaction volume was 50 μl. Cycling conditions were 95°C for 1 min, 30 cycles of 15 s of melting at 95°C, 30 s of annealing at 60°C, 12 min of polymerization at 68°C; followed by a 12-min final extension at 68°C.

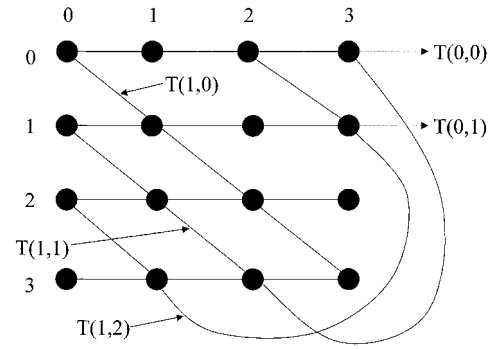

FIG. 1. Assigning pools to tubes using the affine planes technique. Each line in the figure corresponds to a single tube containing four pools. For example, the line $T(1, 2)$ includes all the pools labeled by (x, y) where $y = x + 2 \bmod 4$.

Multiplex sequencing. The conditions for the multiplex sequencing reaction were as follows: 4 μl of BigDye Terminator mix from the Perkin–Elmer DNA sequencing kit (Catalog No. 4303154), a 0.6 μM concentration of each primer from the pool, 25 ng of PCR product cleaned using the Qiagen QIAquick PCR purification kit (Catalog No. 28106). The final reaction volume was 10 μl. Cycle sequencing reaction conditions were 96°C for 2 min, 25 cycles of 10 s at 96°C, 5 s at 55°C, and 4 min at 60°C. The products were precipitated with 70% ethanol at 0°C and loaded on sequencing gels.

Affine planes method for ROMP. Here we give the details of the affine planes technique, with $N = 48$ primers and the number of primers per reaction (tube) $K = 12$. We create 16 pools, $P_0, P_1, \ldots, P_{15}$, each of size 3. The pools are indexed (represented) using base 4 representation as follows: $P_0 = P(00)$, $P_1 = P(01)$, $P_2 = P(02)$, . . . , $P_{14} = P(32)$, $P_{15} = P(33)$. We then create 16 reaction tubes, each of which is also labeled using base 4 representation. In particular, tube $T(ab)$ contains the 4 pools $P(xy)$, where $y = (ax + b) \bmod 4$ and $0 \leq a < 4$, $0 \leq b < 4$. For example, tube $T(12)$ has $y = (1x + 2) \bmod 4$, so if $x = 0$, then $y = 2$. Pool $P(xy) = P(02) = P_2$, so pool P_2 is one of the 4 pools in tube $T(12)$. The other pools in $T(12)$ are numbers 7, 8, and 13.

Finally, we create four more tubes $T(i) = P(xy)$, where each tube contains the four pools defined by $x = i$. Specifically, $T(0) = (P_0, P_1, P_2, P_3)$, $T(1) = (P_4, P_5, P_6, P_7)$, $T(2) = (P_8, P_9, P_{10}, P_{11})$, and $T(3) = (P_{12}, P_{13}, P_{14}, P_{15})$. The complete allocation of pools to tubes is listed under Optimizing the Number of Reactions.

More generally, the affine planes technique used in ROMP uses P pools where P is a perfect square ($P = p^2$). If the total number of primers is N, then we group them into pools, using $N/P = S$ primers per pool. Thus, in the example with 48 primers, ROMP uses 16 pools of size 3 primers each, while with 400 primers, ROMP uses 100 pools of size 4 primers each. The affine planes method creates tubes by grouping pools. Pools are assigned to tubes by putting into tube $T(a, b)$ all pools P_{xy} where $y = ax + b \bmod p$. A geometric intuition for the assignment of pools to tubes is provided in Fig. 1.

The number of reactions (tubes) used by ROMP can be computed using a simple formula. Let $N = SP$, where $P = p^2$ and S is the size of each pool. Note that $K = Sp$. The number of reactions (tubes) required is therefore $R = P + \sqrt{P}$. For example, for $N = 48$ and $K = 12$ we will generate $S = 16$ pools of size $P = 3$ each, and the number of reactions (tubes) R is 20. The number of pipettings required by ROMP is $N + (R\sqrt{P})$. For the case of $N = 48$, we obtain 128 pipettings. Finally, we observe that the projective plane technique (Hall, 1996) might be also applicable to this problem for specific values of N and K.

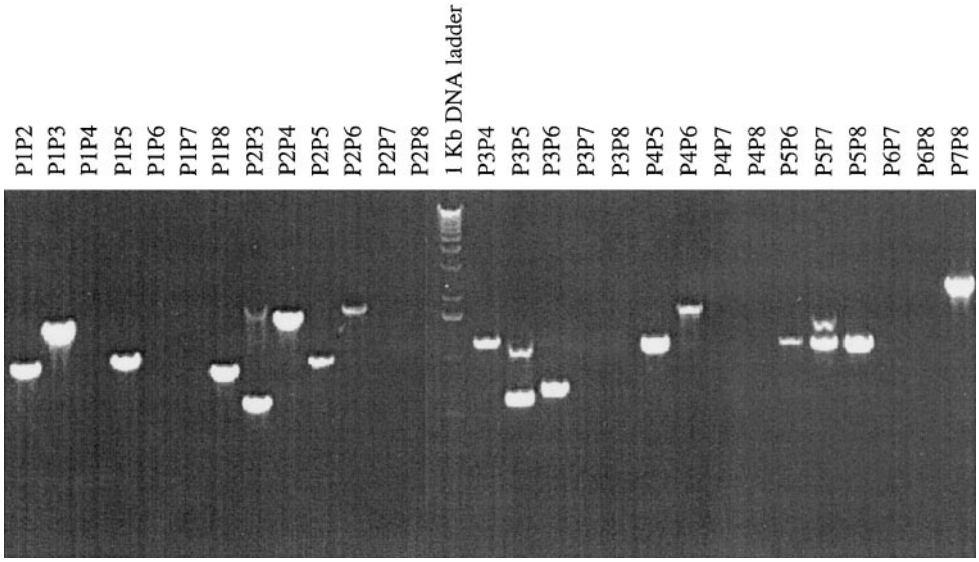

FIG. 2. Primary PCR gel results for 48 primers in pools of 6, with 2 pools per reaction. Each lane shows the PCR results from a reaction using two pools; e.g., pools P_1 and P_2 are in the first lane.

GAP CLOSURE OF *S. pneumoniae*

We will illustrate the POMP protocol with an example using $N = 48$ and $K = 12$, matching the initial number of primers we used when we started our multiplex PCR experiments for the *S. pneumoniae* genome sequencing project under way at TIGR (http://www.tigr.org/tdb/mdb/mdb.html).

For $N = 48$, combinatorial PCR would require 1128 separate reactions. Each reaction would require two primers to be pipetted into a test tube; thus the number of pipetting operations would be 2256 (we do not include in this count the one additional pipetting per reaction needed to add the reaction mix). The POMP protocol we used required just 28 reactions and 104 pipetting operations.

The pool size is $K/2$, so with 12 primers per tube, the pool size is 6. Dividing the 48 primers into pools of 6 gives us 8 pools, labeled $P_1, P_2, \ldots, P_8$.

Now simply choose all pairs from this set of 8 pools; e.g., P_1 is paired with P_2, P_3, through P_8; P_2 is then paired with P_3, P_4, through P_8. (See Fig. 2). Each of these pairs is placed together in a reaction tube. The total number of PCRs is $\binom{8}{2} = 28$.

This simple protocol guarantees that all primers have been paired with all other primers at least once. Note that each reaction tests $\binom{12}{2} = 66$ pairs of primers, and the entire protocol tests $66 \times 28 = 1848$ pairs. Since only 1128 distinct pairs can be made from 48 primers, there is considerable redundancy in the protocol; i.e., some pairs of primers appear in multiple reactions. This redundancy can be used to advantage, as explained below.

The number of pipetting operations is as follows: 48 pipettings are needed to create the 8 pools. Two additional pipettings are necessary per reaction (one from each pool), for a total of $48 + (2 \times 28) = 104$ pipetting operations.

Interpreting POMP Results

Once the initial 28 reactions are run, the action to take depends on the gel results (shown in Fig. 2). Protocols for determining which primers are mates of other primers in a reaction are described next, using examples from our experiments on *S. pneumoniae.*

The PCRs in the POMP design may contain at most half as many products as there are primers in a pool. Here we describe how to interpret each possible outcome and further experiments to disentangle the results from the initial round.

Case 1

No product is observed. When there are no products from the multiplex PCR containing pools P_i and P_j, no further experiments are needed. The interpretation is that none of the K primers in the two pools together spans a physical gap, unless the gap is too large to be detected by PCR.

Case 2

One product is observed. The most commonly observed result from the initial 28 reactions was a single product. As explained below, this agrees with statistical expectations. There are two subcases to consider: (2a) the product appears when P_i is mixed with P_j but not in other reactions involving P_i and (2b) the product appears in all reactions involving P_i.

Case 2a. Case 2a is illustrated in our experiments on *S. pneumoniae* by a single product that appeared when pool P_2 was mixed with P_6. Since this product did not appear in all reactions containing P_2 or in all reactions containing P_6, this result indicated that one of the primers in P_2 was an interpool mate with one primer from P_6 (see Fig. 2). This is the easiest case to interpret. One strategy could be to run up to 12 additional

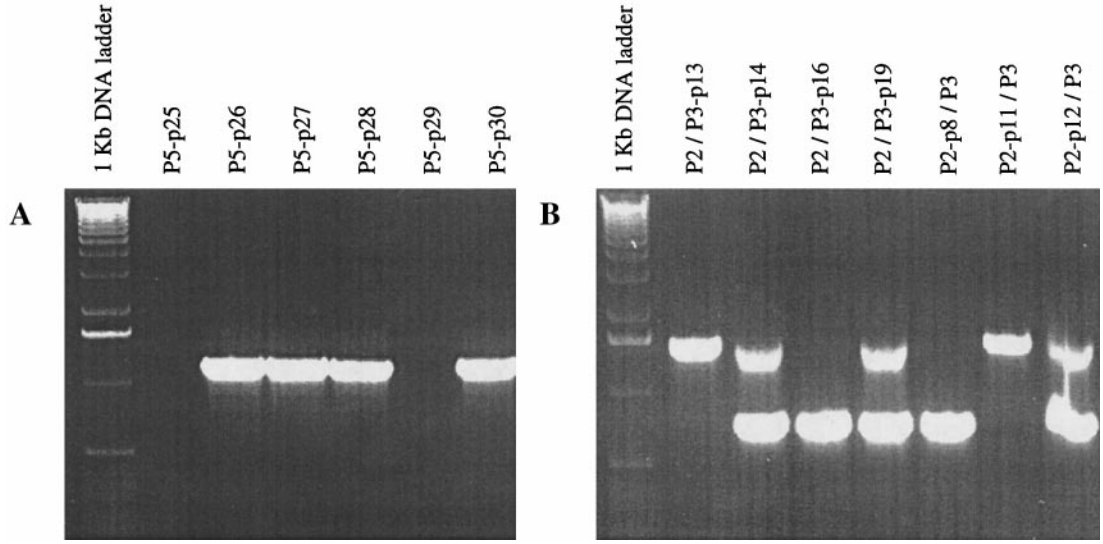

FIG. 3. PCR gel results for interpreting primary gels. (A) An intrapool PCR product in pool P_5, which has six primers. The product disappears when a primer involved in the PCR product is absent. (B) Two interpool products between pools P_2 and P_3, containing six primers each. Only seven reactions were necessary for interpretation; the other five primers were identified by multiplex sequencing. The smaller product is missing when primers p_{13} and p_{11} are absent from the pools. The larger product disappears when primers p_{16} and p_8 are absent.

reactions, eliminating the six primers from each pool, respectively, one at a time. A simpler strategy, which is what we implemented, is to sequence the product directly, using each pool in two separate sequencing reactions with the PCR product as a template. This multiplex sequencing method works because only one primer in the pool will hybridize to the PCR product, generating a unique sequence. Thus no further PCRs were necessary for interpretation in this case; the primers involved were identified by alignment of the sequence obtained to the contig ends.

Case 2b. Case 2b is illustrated by pool P_5, which showed the same product in every reaction in which it was used (see Fig. 2). This means that two of the six primers in P_5 are intrapool mates. To determine which of the two primers were mates, we ran six more reactions: in each reaction, we eliminated primers 25–30 one at a time from the pool and observed which of the reactions resulted in the disappearance of the product (see Fig. 3A). In this case, primers 25 and 29 were mates and were used to sequence the PCR product (note that in principle only five reactions are necessary; if only one mate was found after any five primers were eliminated, the last primer could be inferred to be the second mate).

Case 3

Two products are observed. This is the most common result after the pools with zero or one product are eliminated. If a reaction produces multiple products, then the direct multiplex sequencing solution does not work, because more than one primer in the pool will hybridize to the PCR product, resulting in a mixed sequence. Three subcases must be considered.

Case 3a. Both products result from intrapool mates. We can determine this directly by observing that for both P_i and P_j, at least one product will appear in all reactions involving these pools. The length of the product will be the same for all reactions containing pool P_i and likewise for P_j. To identify the primers involved, each pool must be interpreted separately as described in Case 2b. We had no examples in our experiments where this occurred.

Case 3b. The two products represent one interpool product and one intrapool product. This was the case with P_3–P_5 and P_5–P_7 (see Fig. 2), where P_5 had the intrapool mates. The protocol we followed was to first determine which primers in P_5 were mates and then create a new pool without those two primers. We can then combine this pool with the other pools and interpret the results using multiplex sequencing, as explained for case 2a above. In our experiments, we created the new pool $P_N = P_5 - (p_{25}, p_{29})$ and performed multiplex sequencing with P_N and P_3 and again with P_N and P_7.

Case 3c. Both products are the result of interpool mates. This can be determined by observing that neither product appears in all reactions involving a single pool. Several interpretation strategies are available for this case; the most direct is to run at most K additional reactions, each of which eliminates one primer. Some primers might be eliminated beforehand in the interpretation of other reactions. In our experiments, pools P_2 and P_3 shared two interpool mates. The interpretation reactions between pools P_2 and P_3 are shown in Fig. 3B. These two pools contained primers 7–18, but from these 12 primers, 5 were eliminated by multiplex sequencing of PCR products obtained in Fig. 2:

P_1–P_2 eliminated primer p_{10} from P_2; P_2–P_4 eliminated primer p_9 from P_2; P_2–P_6 eliminated primer p_7 from P_2; P_1–P_3 eliminated primer p_{15} from P_3; and P_3–P_6 eliminated primer p_{17} from P_3. The product from

P_3–P_4 could have been eliminated, but no satisfactory sequencing result was obtained. The product in P_2–P_5 was due to intra-P_5 primers (see above). P_3–P_5 could have been sequenced using P_3 and P_5–(p_{25}, p_{29}), which would have eliminated one more primer from P_3 (no products were obtained for the pairs of pools P_2–P_7, P_2–P_8, P_3–P_7, and P_3–P_8). Therefore only seven additional reactions were run (illustrating one advantage of the redundancy in POMP).

Figure 3B shows the results of removing each of the seven remaining primers from the pools P_2 and P_3. As indicated in the figure, the smaller PCR product (toward the bottom of the figure) was generated by primer p_{13} from P_3 and p_{11} from P_2 (this product is missing from those lanes in the gel). The larger product resulted from primers p_{16} and p_8. These four primers were then used separately for sequencing.

Note that for large values of K, it might be costly to run K additional reactions for every reaction that produced two PCR products. One can substantially reduce the number of reactions by interpreting the results in rounds. For example, if $K = 40$ (the pool size is 20), rather than running 40 reactions, one could create 8 subpools of 5 primers each. If we label the pools A and B and the subpools A_1, A_2, A_3, A_4, B_1, B_2, B_3, B_4, then we could run an initial round of 8 reactions: the 4 A_i subpools with B and the 4 B_i subpools with A, for $1 \leq i \leq 4$. The expected result is that these 8 reactions would produce zero or one reaction each, and the PCR product could then be sequenced directly. If multiple products still appeared in a subpool, one could run a second round of five reactions, one for each primer in the subpool, to resolve those products. Obviously other subpool sizes could be used as well; the optimal choice depends on how many PCR products appeared in the original reaction.

Case 4

More than two products are observed. None of the PCR experiments for *S. pneumoniae* yielded more than two products, but this situation could occur in some rare cases. The interpretation of the results is performed with the same methods as described for case 3, starting by identifying the intrapool mates followed by the interpool mates. There will be more steps required to identify each primer pair. Also, it will be harder to obtain pools of primers with only one primer involved in the generation of a product, thus reducing the advantage of multiplex sequencing.

We devised a straightforward alternative strategy for the interpretation of multiple PCR products. This strategy should be effective when any number of PCR products appear in a single reaction. The approach is to create multiple sequencing reactions using each primer from both pools individually and to sequence the mix of PCR products. For example, with pools of 6 and 12 primers per multiplex PCR, we would create 12 sequencing reactions, each using one primer. This strategy might seem somewhat wasteful of sequencing reactions and gels; however, if multiple products are observed, then eventually all these products will have to be sequenced.

Results of the POMP Experiment

For the *S. pneumoniae* project, the actual number of physical gaps at the time we applied the POMP protocol was 38; therefore there were 76 primers in total. We initiated the POMP experiments when the first 48 primers were available. From these experiments, 14 gaps were closed, accounting for 28 primers, and 20 primers were left unpaired. The 28 additional primers were subsequently synthesized, 6 of which were adjacent to large repeat regions and were not used until the final long-range PCR experiments described below. The 22 new primers, together with the 20 remaining primers from the initial round of experiments, were split into 7 pools of 6 primers, and POMP experiments were run again. These 21 PCR produced 14 new PCRs products and allowed us to close another 14 gaps. In total, over the 70 primers tested, 28 gaps were closed, accounting for 56 of the primers. To attempt to close the remaining gaps, the 14 primers that did not generate products, together with the 6 primers adjacent to repeats, were tested in a long-range POMP experiment.

Long-Range PCR

Because most of the gaps in a shotgun sequencing project with good coverage (6× or more) are expected to be less than 1 kb in length (Fraser and Fleischmann, 1997), we suggest attempting long-range PCR (LR-PCR) only with primers left over after an initial POMP experiment. In addition, LR-PCR proves more efficient with primers 10 nucleotides longer than the 18- to 25mers used routinely in shotgun sequencing projects. Note that a "multiplex long accurate PCR" method was described by Sorokin *et al.* (1996) for the *B. subtilis* genome project. They obtained PCR products of about 18 kb routinely using seven primers per reaction, but the numbers of reactions and pipettings were not optimized.

We applied our POMP algorithm to the 20 remaining primers of the *S. pneumoniae* project. Four additional primers were added to the LR-PCR to check for correctness of existing contigs that contained long internal repeats. With 24 primers, the POMP strategy calls for setting $K = 10$ and using pools of 5 primers, generating $\binom{5}{2} = 10$ reactions run under LR-PCR conditions (see Materials and Methods).

We obtained four different products from the LR-PCR experiment, three of which were intrapool products, as shown in Fig. 4. Closing these four gaps leaves six gaps remaining after the POMP experiments, which had to be closed with other methods. Note that the LR-PCR did not generate products longer than 3 kb, although we ran a control and successfully generated an 18.7-kb test product.

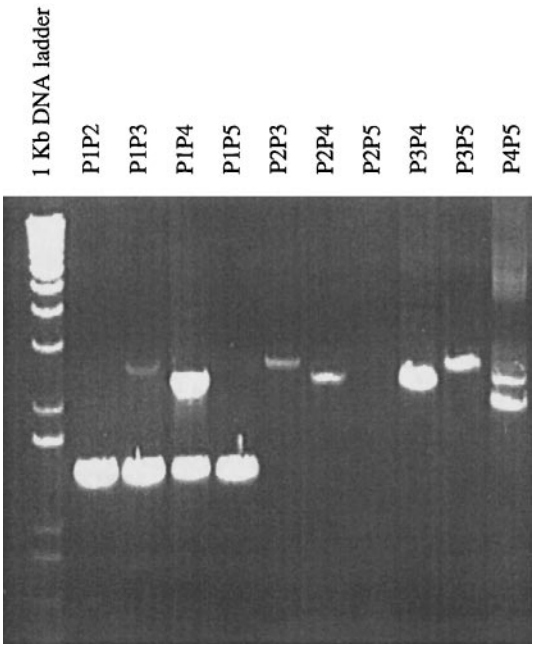

FIG. 4. PCR gel results using 24 primers in a long-range PCR experiment. Five pools were created, with only 4 primers in the last pool. Pools P_1, P_3, and P_4 each had an intrapool product. P_4 and P_5 generated an interpool product.

OPTIMIZING THE NUMBER OF PIPETTINGS

The protocol we used in the lab was based on $K = 12$, primarily because we were not confident that larger values of K would produce clean PCR data. In subsequent tests of multiplex PCR, we have found that we can adjust the conditions to obtain good reactions with as many as 30 primers in a single reaction tube (see Fig. 5), and there are indications that this value can be even greater.

Assuming no restriction on the value of K, the POMP protocol chooses K based on N, the number of primers. (Note that N is always twice the number of gaps. Alternatively, N is twice the number of contigs remaining to be joined, since one primer must be synthesized from both ends of each contig.) To optimize pipetting, choose

$$K = 2 \times \lceil \sqrt{N} \rceil$$

and make pools of size $K/2$. For simplicity, assume N is a perfect square, so that $\lceil \sqrt{N} \rceil = \sqrt{N}$. Then the number of reactions required using POMP will be

$$R = \binom{\sqrt{N}}{2} = \frac{N - \sqrt{N}}{2},$$

and the number of pipettings will be

$$2R + N = 2N - \sqrt{N}.$$

Succinctly stated, the number of reactions used for N primers (N/2 gaps) is less than N/2, and the number of pipettings is less than 2N. Note that for R reactions, at least $2R$ pipettings are required simply to load the tubes, and obviously at least N pipettings are required to test any set of N primers. A proof that POMP is optimal remains an open problem; clearly the number of pipettings is within a factor of 2 of the minimum required. Note that this minimal value refers to the number of pipettings required for the initial round of PCR experiments; we have not included the steps for interpreting POMP results in our calculations.

One nice feature of this strategy is that, because the number of reaction tubes is slightly less than $N/2$, the expected number of products per tube is approximately 1. This correlates well with our experimental evidence. Tubes with one product, as explained above, are the simplest to interpret. The most common result, in 31 of the 49 PCRs, was a single product.

To obtain a more accurate picture of how many PCR products we could expect in a POMP experiment, we created a simulation to compute the expected number of reactions in each tube. Figure 6 shows results from this simulation using three different sets of values for N and K: 48 primers with 12 primers per reaction, 400 primers with 40 per reaction, and 4356 with 132 per reaction (see Future Perspectives for a discussion of the latter two choices). As Fig. 6 shows, with 48 primers and 12 primers per tube, the most likely result is exactly one PCR product. With the larger values of N and K, the probabilities of one product and two products are approximately the same. In all cases the majority of reactions will contain either one or two PCR products, with only 10–15% of reactions generating no product. Thus the reactions are used quite efficiently in

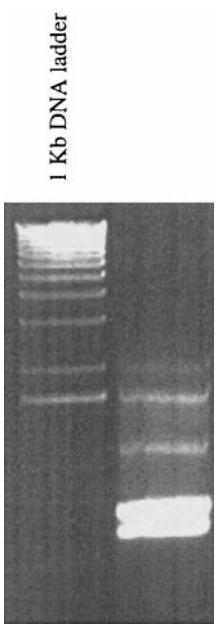

FIG. 5. PCR gel results using a 30-primer mixture, with two pools of 15 primers that have 5 interpool mates. The largest PCR product, about 2 kb in length, is faint but still visible.

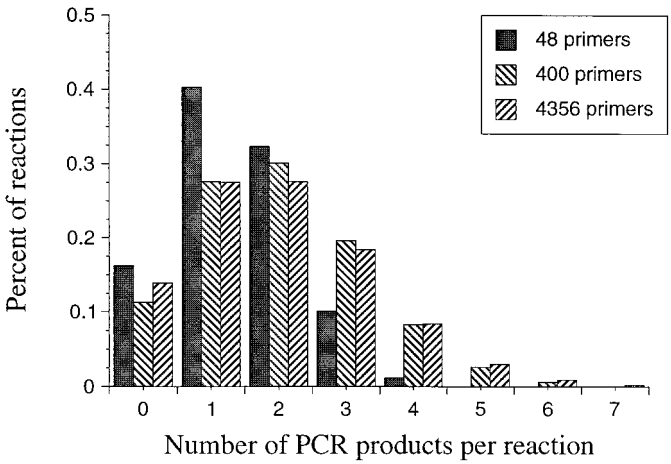

FIG. 6. Simulation results for 48 primers with 12 primers per reaction, 400 primers with 40 per reaction, and 4356 primers with 132 per reaction. The vertical axis shows the percentage of reactions in which a given number of PCR products occur. Results shown are averaged over 1000 simulations.

that few are wasted. Also note that this simulation assumes that all primer mates will yield a product, which is quite unlikely in reality. For example, some gaps may be too large to PCR across (but might be resolved using LR-PCR). Therefore the distributions of actual results in a POMP experiment will probably be shifted to the left of those in the figure, with slightly fewer products per reaction on average.

Using the formulas described above, we see that we could have reduced the number of pipettings in our experiments with *S. pneumoniae*. Instead of pools of size 6, we would choose $K = 2 \times \lceil \sqrt{48} \rceil = 14$, with a pool size of 7. This would then result in $\binom{7}{2} = 21$ reactions, and only $48 + 2 \times 21 = 90$ pipettings rather than the 104 required for pools of size 6. At the time of the experiments, we had not yet developed a protocol that allowed more than 12 primers per reaction, hence our choice of pool size.

OPTIMIZING THE NUMBER OF REACTIONS

A more complicated strategy can test all primer pairs using fewer reactions than POMP. This ROMP strategy might be preferred when the pipetting is performed robotically and is therefore less likely to introduce experimental errors or when the reaction cost is very high due to the large number of gaps. The advantages to using POMP are several: (1) it is simple, (2) it uses fewer pipetting reactions, and (3) the pooling approach allows the experimenter to start running reactions as primers are synthesized. In the ROMP protocol described below, more of the primers need to be available before the experiments begin. The strategy described below uses an optimal number of reactions, but the amount of pipetting required is greater. The solution we have designed draws upon a well-studied mathematical method known as the affine planes technique (Hall, 1996), which has not previously been applied to problems in genome sequencing.

Note that the ROMP method is not completely general in the sense that it needs to be customized for each value of N and K (see Materials and Methods). It is not clear that any completely general method will generate optimal solutions for all combinations of N and K, although we have had no trouble generating optimal strategies for all values that we have considered.

The optimal number of reactions for $N = 48$ and $K = 12$ is 20, not the 28 used by POMP. This can be proven as follows. To make sure all primer pairs are tested, each primer must be paired with all 47 other primers. Since a reaction tube can only pair a primer with 11 others, that means each primer will have to be placed into at least $\lceil 47/11 \rceil = 5$ tubes. Clearly the best we can do is to place all 48 primers into 5 tubes each with no redundancy (no two primers go into the same tube more than once). Thus $48 \times 5 = 240$ "virtual tubes" are required, and with 12 primers per tube, the optimal number of tubes is $48 \times 5/12 = 20$ tubes.

The ROMP procedure relies on a sophisticated framework in combinatorial optimization referred to as block design. In particular, we utilize an algebraic technique known as the affine planes technique (Hall, 1996) to allocate the primers to reaction tubes. Similar pooling strategies using related mathematical methods have been used previously to develop library screening techniques (Barillot *et al.*, 1991; Bruno *et al.*, 1995; Knill *et al.*, 1996). These strategies had to account for much higher false-positive and false-negative rates than the procedure here, but the overall strategic approach is similar.

We first group the 48 primers into 16 pools, numbered 0–15, each containing 3 primers. This reduces our problem to matching the 16 pools with one another using tubes containing 4 pools each. The allocation of pools to tubes is as follows (note that tubes are numbered in base 4, as explained under Materials and Methods):

$T(00) = 0,4,8,12$	$T(13) = 3,4,9,14$	$T(32) = 2,5,8,15$
$T(01) = 1,5,9,13$	$T(20) = 0,6,8,14$	$T(33) = 3,6,9,12$
$T(02) = 2,6,10,14$	$T(21) = 1,7,9,15$	$T(1) = 0,1,2,3$
$T(03) = 3,7,11,15$	$T(22) = 2,4,10,12$	$T(2) = 4,5,6,7$
$T(10) = 0,5,10,15$	$T(23) = 3,5,11,13$	$T(3) = 8,9,10,11$
$T(11) = 1,6,11,12$	$T(30) = 0,7,10,13$	$T(4) = 12,13,14,15$
$T(12) = 2,7,8,13$	$T(31) = 1,4,11,14$	

Each pool in this design appears in a tube with each of the 15 other pools. Consequently, each primer will appear at least once in a tube with each of the other 47 primers.

Clearly, this design uses the optimal number of reactions (20). However, the number of pipetting operations is greater than with the larger pools that we actually used. Forty-eight pipettings are required to create the 16 pools, and four pipettings are required for each of the 20 tubes. This gives $48 + (4 \times 20) = 128$ total pipettings, versus 104 for the other method.

FUTURE PERSPECTIVES

The POMP method can be extended to larger genomes or to earlier stages in a sequencing project when

the number of gaps is larger than the 76/2 = 38 that were present in this study. A typical bacterial genome in a whole-genome shotgun sequencing project will have approximately 50–100 physical gaps at the end of the random sequencing phase, after careful editing and reassembly, and assuming that the genome does not have an inordinate number of repeat sequences. For example, the *Haemophilus influenzae* (Fleischmann *et al.*, 1995) and *Borrelia burgdorferi* (Fraser *et al.*, 1997) projects had 42 and 85 physical gaps, respectively, at the end of the random sequencing phase. Both larger genome size and nonrandomness in the clone library will result in more physical gaps. The *S. pneumoniae* sequencing project, whose large insert library suffered from nonrandomness, had many more gaps than the 38 described here, but the POMP protocol was applied only after extensive, time-consuming genomic walking experiments were used to close more than 100 gaps.

The current plans for sequencing the human, mouse, *Arabidopsis,* and rice genomes use a BAC-based approach. Although our methods are not unnecessary for closing a single BAC, they could be usefully applied to close multiple BACs that were sequenced together, allowing BAC-based sequencing projects to proceed in larger increments than the 100- to 150-kb size of an average BAC.

The following illustrates the POMP protocol for a project with 200 physical gaps. First, 400 primers would be synthesized, one from each end of every contig. We then choose $K = 2 \times \sqrt{400} = 40$ and create 20 pools of size 20 (note that we have yet to verify that $K =$ 40 primers per reactions will yield good results; our experiments with up to 30 primers were successful, but we have not attempted to use more). Creating reactions for all pairs of pools would result in 190 reactions, and 780 pipettings would be required to prepare all the reactions. If reaction cost is critical, then a ROMP design would use only 110 reactions, which is optimal. The ROMP strategy involves the creation of 100 pools of size 4, where each tube contains 10 pools, and 1500 pipettings would be needed to create the 110 reactions.

At the high level of coverage currently planned for the *Drosophila melanogaster* complete genome shotgun sequencing project under way at Celera Genomics and Berkeley (Pennisi, 1999), the Poisson model of Lander and Waterman (1988) predicts that only 219 gaps will remain at the end of the random phase. It is quite possible that the actual number of gaps will be as high as 10 times that amount; for example, the theoretical model indicated that the *B. burgdorferi* genome sequencing project would have just 8 gaps rather than 85 (Fraser and Fleischmann, 1997). If we are able to increase the effectiveness of multiplex PCR to allow as many as 132 primers in a single tube, then sequencing projects as large as the entire *Drosophila* genome could be closed using POMP. With 2178 contigs and 4356 primers, we would create pools of size $\sqrt{4356} = 66$ and set $K = 132$. The total number of reactions required to test all pairs of primers would then be 2145, and the total number of pipettings would be 8646. The ROMP strategy would use only 1122 reactions but would require 1089 pools and 41,382 pipettings. Improvements in PCR technology and the use of robotics should make the use of POMP and ROMP feasible for genome projects of this scale.

ACKNOWLEDGMENTS

We thank Noga Alon for suggesting the methodology of block design as a framework for ROMP. S.L.S. was supported in part by NIH Grants K01-HG00022-1 and R01 LM06845-01 and by NSF Grants IRI-9530462 and IIS-9902923. S.K. was supported by NSF Grants IRI-9616254 and KDI-9980088. H.T., D.R., and H.K., were supported in part by Merck Genome Research Institute Proposal 72 and by NIH Grant 1 R01 AI40645-01A1.

REFERENCES

Barillot, E., Lacroix, B., and Cohen, D. (1991). Theoretical analysis of library screening using a N-dimensional pooling strategy. *Nucleic Acids Res.* **19**(22): 6241–6247.

Bruno, W., Knill, E., Balding, D., Bruce, D., Doggett, N., Sawhill, W., Stallings, R., Whittaker, C., and Torney, D. (1995). Efficient pooling designs for library screening. *Genomics* **26:** 21–30.

Burgart, L., Robinson, R., Heller, M., Wilke, W., Iakoubova, O., and Cheville, J. (1992). Multiplex polymerase chain reaction. *Mod. Pathol.* **5:** 320–323.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J.-F., Dougherty, B., Merrick, J., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J., Weidman, J., Phillips, C., Spriggs, T., Hedblom, E., Cotton, M., Utterback, T., Hanna, M., Nguyen, D., Saudek, D., Brandon, R., Fine, L., Fritchman, J., Fuhrmann, J., Geoghagen, N., Gnehm, C., McDonald, L., Small, K., Fraser, C., Smith, H., and Venter, J. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd. Science* **269:** 496–512.

Fraser, C., Casjens, S., Huang, W., Sutton, G., Clayton, R., Lathigra, R., White, O., Ketchum, K., Dodson, R., Hickey, E., Gwinn, M., Dougherty, B., Tomb, J.-F., Fleischmann, R., Richardson, D., Peterson, J., Kerlavage, A., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M., Horst, K., Roberts, K., Hatch, B., Smith, H., and Venter, J. (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi. Nature* **390**(6660): 580–586.

Fraser, C., and Fleischmann, R. (1997). Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* **18:** 1207–1216.

Hall, M. (1996). "Combinatorial Theory," 2nd ed., Wiley–Interscience, New York.

Heiner, C., Hunkapiller, K., Chen, S.-M., Glass, J., and Chen, E. (1998). Sequencing multimegabase-template DNA with BigDye terminator chemistry. *Genome Res.* **8:** 557–561.

Knill, E., Schliep, A., and Torney, D. (1996). Interpretation of pooling experiments using the Markov chain monte carlo method. *J. Comp. Biol.* **3**(3): 396–406.

Lander, E., and Waterman, M. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2:** 231–239.

Pennisi, E. (1999). Fruit fly researchers sign pact with Celera. *Science* **283:** 767.

Sorokin, A., Lapidus, A., Capuano, V., Galleron, N., Pujic, P., and Ehrlich, S. (1996). A new approach using multiplex long range accurate PCR and yeast artificial chromosomes for bacterial chromosome mapping and sequencing. *Genome Res.* **6:** 448–453.