A simple and efficient method for constructing high resolution physical maps

Kaoru Yoshida, Michael P. Strathmann, Carol A. Mayeda, Christopher H. Martin and Michael J. Palazzolo*

Human Genome Center, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720, USA

Received March 2, 1993; Revised and Accepted June 21, 1993

ABSTRACT

This paper describes a simple and efficient walking method for constructing high resolution physical maps and discusses its applications to genome analysis. The method is an integration of three strategies: (1) use of a highly redundant library of 3Kb-long subclones; (2) construction of a multidimensional pool from the library; (3) direct application of a PCR (polymerase chain reaction)-based screening technique to the pooled library, with two PCR primers, one from the end of the subcloning vector and the other from the leading edge of the walk. This technique allows not only detection of each overlapping subclone but simultaneous determination of its orientation and the size of its overlap. The end of the subclone with the smallest overlap is sequenced and a primer is designed for the next step in the walk. Iteration of the screening procedure with minimum overlapping subclones results in completion of the high resolution map. Using this method, a 3Kb-resolution map was constructed from an 80Kb region of the bithorax complex of *Drosophila melanogaster*. The method is general enough to be applicable to DNA from other species, and simple enough to be automated.

INTRODUCTION

Since chromosome walking was first introduced and applied to the bithorax complex (BX-C) of *Drosophila* [1], many techniques have been developed for physical mapping. The task of physical mapping is to characterize a set of cloned genomic fragments to determine how they overlap. A set of overlapping clones is called a contig. Both fingerprinting and hybridization-based detection methods have been widely used for building contigs. The recent development of the polymerase chain reaction (PCR) [2] has impacted many physical mapping methods, from traditional chromosome walking to new strategies based on amplification of sequence-tagged sites (STSs) [3,4,5]. Each individual contig is assigned to a corresponding genomic position by either *in situ* hybridization [6,7] or by genetic means [8]. Regardless of the strategy, the end result is an ordered

representation of overlapping clones that cover the genome. These ordered sets of mapping clones can facilitate traditional molecular and genetic experiments in various ways: as probes for extracting expressed sequences, as landmarks to identify sequences corresponding to mutations, and as templates or as intermediates in template generation in large-scale genomic DNA sequencing projects. The numerous methods mentioned above have been developed to construct physical maps of rather low resolution from whole genomes. However, there is no organized, efficient and cost-effective strategy for breaking down a large mapping clone into an ordered set of small subclones, and thereby building a high resolution map. Since many experimental manipulations are easily performed using clones with relatively small inserts, there is a need for high resolution maps. We present a simple and efficient method for constructing such high resolution maps and discuss their applications to genome analysis.

MATERIALS AND METHOD

Construction of a library with high redundancy

The target P1 clone was purified by using the Qiagen plasmid prep kit followed by centrifugation through a cesium chloride gradient. 15µg of the DNA was randomly sheared by passage through a French press (see [9] for experimental details) and fractionated on a 1% low-melting point agarose gel. Fragments around 3Kb in size were excised from the gel and purified using the Qiaex kit (Qiagen). The fragments were treated with the Klenow fragment, phenol-extracted and precipitated with ethanol. These blunt-ended fragments were ligated into the *PvuII* site of the plasmid vector, pSP72 (Promega). Prior to ligation, the vector was treated with phosphatase. Electroporation was used to facilitate transformation of the strain, DH5α, with the ligation products. 960 transformants were transferred separately to individual microtiter wells and grown to saturation in 200µl LB + 100µg/ml ampicillin.

DNA preparation from the pooled library

The library was pooled in 3 dimensions (see below). To make each pool, 20µl of the bacterial culture were collected from the appropriate microtiter wells. Plasmid DNA was extracted from

* To whom correspondence should be addressed

1.5ml of the pooled cultures using the alkaline lysis preparation [10]. The DNA was resuspended in 90 μ l of TE buffer (pH 7.5) containing 10 μ g/ml RNaseA. After incubation for 30 min at room temperature, 50 μ l of this DNA solution was diluted to 1ml with TE buffer + RNaseA. The diluted solution was used for PCR analysis. All solutions were stored at -20°C and mixed thoroughly before use.

PCR-based screening condition

For each step in a walk, two sets of PCRs were carried out on the pools: set A with primers, τ and α , and set B with primers, τ and β . α and β are vector-specific primers with the following sequences: α (on the SP6 site): 5'-ATTAGGTGACACTATA-GAACTCGCGC-3', and β (on the T7 site): 5'-GATGAATT-CGAGCTCGGTACC-3'. τ is a region-specific primer whose theoretical melting temperature is near 60°C (see below).

The PCR reactions contained 10mM Tris-HCl (pH 8.3), 1.5mM MgCl₂, 50mM KCl, 0.01% gelatin, 200 μ M dNTP's, 1.0 μ M primer τ , 1.0 μ M primer α (or β), 0.01mM tetramethylammonium chloride (TMAC), 0.5 units of AmpliTaq DNA polymerase (Perkin-Elmer Cetus). The reaction mixture was combined on ice with 2 μ l of pooled DNA to a final volume of 20 μ l, and transferred to a thermal cycler that was preheated to 95°C. Samples were held at 95°C (initiation) for 5 min and subjected to 30 cycles of PCR with the following thermal profile: 94°C (denaturation) for 15 sec, 55°C (annealing) for 15 sec, 72°C (elongation) for 2 min. Completed reactions were fractionated by gel electrophoresis on 1.5% agarose gels in TBE buffer containing 0.1 μ g/ml ethidium bromide at 10 volts/cm, and then directly visualized by UV fluorescence. The ends of the minimum-overlapping subclones were sequenced using the Dye Primer Cycle Sequencing kit (Applied Biosystems). Region-specific primers for further steps in the walk were designed by using the Primer program [11].

Note that the above PCR conditions are optimized for the case of walking in a single direction from one end of the target genomic region to the other, in which only minimum-overlapping subclones need to be mapped. To more reliably map other overlapping subclones, the protocol can be modified as follows: (1) The PCR elongation time is extended from 2 min to 8 min; (2) After completing the PCR, adjust the buffer to 1mM ZnSO₄ by the addition of 10mM ZnSO₄, add 1 unit mung bean nuclease (New England Biolabs) and incubate at room temperature for 30 min before gel electrophoresis.

PRINCIPLE OF THE METHOD

The important considerations in the development of a clone-based mapping method are: (1) how to detect and position overlapping subclones in an efficient and reliable manner, (2) how to efficiently assemble them into a single continuous contig, and (3) how to systematize the overall process. Contig-assembly strategies have traditionally employed either fingerprinting or probe-based binary detection methods.

Fingerprinting involves the analysis of individual clones, usually by digestion with restriction enzymes, to reveal a pattern of DNA fragments [12,13,14,15]. Similarities between the patterns from different clones indicates an overlap. Data handling and contig assembly, however, are not trivial. In addition, the exact size of overlaps is not necessarily determined.

Probe-based binary detection methods utilize a 'positive or negative' assay to detect overlapping clones. Traditionally these

methods have employed hybridization as the binary assay [1,16]. More recently, the polymerase chain reaction has achieved wide use in binary detection schemes because of its increased sensitivity and reliability; PCR is less prone to false-positives and false-negatives [17,18,19,20,4]. Probe-based binary detection methods frequently include a pooling strategy in which clones (and/or probes) are grouped in order to minimize the number of assays [21,22,23]. Information is gained when some pools are positive and some pools are negative to a given probe. If any probe will detect a large fraction of the clones in a library, then the library has to be divided into a large number of pools each containing a few clones. Otherwise, all the pools will give positive signals and no information is gained. In this case, the efficiency of any pooling strategy is very limited.

To circumvent these limitations, we have developed a probe-based *non*-binary detection method that integrates the following strategies: (1) use of a library with high redundancy, (2) construction of a multidimensional pool from the library, and (3) direct application of a PCR-based screening technique to the pooled library, which allows quantitative assessment of all overlapping subclones in one step. Below we describe these strategies, and explain why their integration results in an efficient high-resolution physical mapping method.

Construction of a library with high redundancy

It has been frequently observed that as much effort is spent to close a few gaps as is spent to assemble the majority of a map. A gap simply means the absence of a subclone in a required position of the map. The absence of this subclone may be due to the inability of the mapping method to detect it or to the incompleteness of the library from which the map is constructed. The only statistical solution to reduce the possibility of a gap, independent of the mapping method, is the use of a highly redundant library.

The redundancy of a library required for map completion can be calculated from theoretical considerations [24]. However, in practice, the distribution of clone length and bias introduced by bacterial propagation of clones may necessitate a higher redundancy than the estimated value. Consequently it is often most cost-effective to supplement the original mapping approach with different map completion strategies. If an efficient pooling/detection method is introduced, such that the handling cost is not proportional to the redundancy, then it becomes feasible to efficiently screen a library with high redundancy. For example, if the characterization of a 30-fold redundant library requires only 30% more assays than a 10-fold redundant library, then one need not resort to a more costly completion strategy.

Construction of multidimensional pools

In order to reduce the mapping cost of a large library, an efficient pooling strategy is required. We employ a multidimensional pooling scheme [25] such that the library is projected to K dimensions and a member of the library is uniquely determined by the K pools (one per dimension) in which it resides.

Given N subclones, standard K -dimensional (K -D) pools are composed of $K\sqrt{N}$ pools per dimension and $K \cdot K\sqrt{N}$ pools as a whole. 'Standard' means that the number of pools is equal in each dimension. For example, given a library of 1000 subclones, the 3-D pooling is composed of 10 pools per dimension and 30 pools as a whole. Thus the handling cost is reduced from 1000 individual subclones to 30 pools.

Figure 1. Pooling a library in 3 dimensions. 960 subclones are arrayed in 10 microtiter plates. The collection of subclones comprising plate pool 2, column pool 4 and row pool B are shown. When these three pools raise positive signals, then clone 2_4B is identified as the positive subclone.

Figure 2. Determining the sizes and orientations of overlaps. The walk extends from subclone W. Overlapping subclones fall into two classes: those overlapping clone W on their α side (clone X), and the others overlapping on their β side (clone Y). Thus all possible overlaps with the right edge of clone W can be PCR-amplified by applying two sets of PCR, set A with primers α and τ , and set B with primers β and τ , where τ is a region-specific PCR primer with the indicated orientation and α and β are vector-specific PCR primers surrounding the cloning site. T_x and T_y represent the sizes of the PCR products.

Consider a library of 960 subclones arrayed in 10 96-well microtiterplates. 3-D pooling is convenient and easily adapted to this arrangement as shown in Figure 1. 10 plate pools, 12 column pools, and 8 row pools constitute a 3-dimensional space. Each plate pool contains 96 subclones, each column pool contains 80 ($= 8 \times 10$) subclones and each row pool contains 120 ($= 12 \times 10$) subclones. For example, plate pool P_2 is a collection of all subclones in the 2nd plate; column pool C_4 is

a collection of all subclones in the 4th column of each plate; row pool R_B is a collection of all subclones in the B-th row of each plate. By adopting this modified 3-D pooling scheme, the handling cost is reduced from 960 subclones to 30 ($= 10 + 12 + 8$) pools.

In some cases, a pooling scheme of higher dimension may be appropriate. Introduction of an additional dimension, *group*, allows extension of the pooling strategy from 3 to 4 dimensions. For example, a library of 9600 subclones can be pooled in 4 dimensions by a simple extension of the 3-D pooling scheme outlined above. Let the arrangement of subclones depicted in Figure 1 represent a group. Now, 9600 subclones can be arrayed in 10 groups. Group pool #1 is the collection of all subclones in group #1, plate pool #4 is the collection of all subclones on the 4th plate from each group, and so on. Likewise, 5-D pools can be constructed by introducing another dimension, such as *section*.

Spectrum PCR: one-step quantitative assessment of overlaps

In order to efficiently screen a highly redundant library that is pooled in K dimensions, it is necessary not only to detect overlapping subclones but also to individualize these subclones. For this purpose, binary detection methods are not adequate. A binary detection scheme does not provide by itself any information that is unique to the overlapping subclone. Consequently, if several overlapping subclones are detected in the pooled library, they cannot be precisely identified without further analysis. Consider the library depicted in Figure 1. If three overlapping subclones are present, then three plate pools, three column pools and three row pools can give positive signals. 27 ($= 3 \times 3 \times 3$) subclones must be further analyzed to find the three correct subclones. Now consider the case in which 30 overlapping subclones are present in the library. The average number of overlapping subclones per pool is about 3. Thus every plate, column and row pool will be positive, providing no information at all.

One way to individualize overlapping subclones is to determine the sizes and orientations of their overlaps since this property, in most cases, is unique to each subclone. We employ a PCR-based screening technique which allows not only detection of overlapping subclones but also simultaneous determination of the size and orientation of each overlap [26,27,28]. As illustrated in Figure 2, suppose we want to find the subclones that overlap the right end of target subclone W. We sequence this end and design a region-specific primer, τ . Now, using the vector-specific primer, α , in addition to τ , overlapping subclones (e.g., subclone X) can be specifically amplified and positioned relative to τ . Subclones with inserts in the opposite orientation relative to the vector (e.g., subclone Y) can be amplified with τ and the other vector-specific primer, β . We apply these two sets of PCR assays on pools of subclones. Each PCR product from a pool is a collection of PCR fragments of different lengths. Using gel electrophoresis, the length of each PCR fragment (T_x and T_y), and hence the size of each overlap, is determined.

For a K -dimensional pooling scheme, when K bands of the same size appear, one for each dimension, then the bands denote a subclone that has an overlap of the observed size in the region of interest. That is, the overlapping subclone can be identified as a K -tuple, *subclone* ($i, j, k, \dots$), from the indices of the pools that show bands of the same size. Figure 3 shows the case in which a 3-D pooled library is used in the walk from the β -end of subclone (3,3,E). Three bands are present at the 400bp position in gel B, one in plate pool #2, one in column pool #4 and the

Figure 3. 3-dimensional Spectrum PCR. Two sets of PCR are applied to the library, which is pooled in 3 dimensions, and their PCR products are fractionated on agarose gels. Three bands of equal size serve to identify a subclone that overlaps the target subclone, W. The size of the bands and the set in which they appear determine the length and orientation of the overlap, as depicted in Figure 2.

other in row pool # B. This triplet indicates that the β -end of subclone (2,4,B) overlaps the target region by about 400bp. Another triplet around 1600bp in gel A identifies a 1600bp overlap at the α -end of subclone (3,11,A). Subclones (3,3,E), (3,11,A) and (2,4,B) correspond to subclones W, X and Y in Figure 2, respectively.

Regarding a K -tuple of PCR bands as the spectral line of an overlapping subclone and the pooled library as a light source, this method functions like a prism. Thus we designate it *Spectrum PCR*. The essence of Spectrum PCR is that two sets of PCR amplification, set A with primers α and τ and set B with primers β and τ , are performed directly on the pooled library. Spectrum PCR not only detects an overlapping subclone but also provides unique information about the size and orientation of its overlap thereby allowing a highly redundant library to be screened efficiently.

Walking with large strides through the library

Since Spectrum PCR allows both detection and positioning of overlapping subclones, it is ideally suited to constructing a map by walking through the highly redundant library. For each

Figure 4. Contig connection. Walks initiated from subclones U1 and W1 can meet in two ways. (A) Two different walks, one from subclone U2 and the other from subclone W2 identify two common overlapping subclones, Q and R, in a consistent manner: one side (α) of the subclone in one walk and the other side (β) in the other walk. (B) When one walk is attempted from U1 and another from W1, no common subclone appears in the two walks. In the next step, the walk from subclone U2 and the walk from W2 identify each other as an overlapping subclone as well as others which already appeared in the previous step. In this case, one step of walking, either from U1 to U2 or from W1 to W2, results in waste. Arrows indicate the region-specific primers that are used to identify overlapping clones.

overlapping subclone, insert length minus overlap length equals its *stride*. For an efficient walk, the subclone with the largest stride should be chosen for the next step in the walk. Since stride is the main concern, a more direct method is to amplify strides by applying Spectrum PCR with region-specific primer σ (Figure 2). However, smaller fragments are often amplified more easily than larger fragments. Indeed, for some applications, the stride length may well exceed the limits of PCR. If the subclones in the library are almost uniform in size, it is sufficient to pay attention merely to the size of the overlap in determining the next subclone. Namely, the minimum-overlapping subclone is almost always one of the largest striding subclones. In Figure 3, among the seven overlapping subclones, subclone Y shows the minimum overlap (400bp) and would be taken for the next step in the walk.

Walking can be performed in a stepwise manner from one edge of the source DNA to the other edge (e.g., one edge of a P1 clone insert to the other). Alternatively, given a single parent clone, in order to complete its high resolution map more quickly, the entire walking process can be split into parallel walking subprocesses by picking more than one starting point. In the case of walking from both edges of the source DNA, it means the initiation of two parallel walking subprocesses. More starting points can be generated at random, or if the source DNA is characterized, starting points can be derived from previously mapped regions such as STSs or cDNAs.

Once parallel walks are initiated, they will result in simultaneous construction of more than one contig. In order to see whether different contigs are connected, it is necessary to keep record of not only the minimum-overlapping subclone but

Figure 5. 3-dimensional Spectrum PCR applied to the SP6-side edge of subclone F2_4B. Subclone F2_4B was found to have a minimum-overlap with subclone F3_3E on its T7 side and chosen as the target for the next step in the walk. To walk from the SP6-side edge of subclone F2_4B, two sets of PCR were applied to the library of 960 subclones, which were pooled in 3 dimensions: (A) set A with region-specific primer τ (F3_3E:sp6.r) and vector-specific primer α ; (B) set B with τ (F3_3E:sp6.r) and another vector-specific primer, β . The PCR products were produced with the modified PCR protocol (see Materials and Methods) and subjected to gel electrophoresis in 1.2% agarose at 10 volts/cm. Two different markers were used: the 100bp DNA ladder (lane m) and the 1Kb DNA ladder (lane M), both from GIBCO/BRL.

also other overlapping subclones. If one or more overlapping subclones appear in different walks, then these subclones bridge the two contigs. Figure 4 illustrates the two ways in which walks extending from subclones U2 and W2 can meet. In case 1 (see Figure 4-A), U2 and W2 do not overlap each other, and the distance between their edges is less than the average subclone size. Subclones P, Q and R are found to overlap subclone U2, while subclones Q, R and S are found to overlap subclone W2. Thus, two subclones Q and R bridge the contigs. In case 2 (see Figure 4-B), subclones M, N, O and U2 were found to overlap subclone U1, while subclones P, Q, R, S and W2 overlapped subclone W1. Since there are not any subclones shared by the two contigs, further steps were taken from subclones U2 and W2. Subclones P, Q, R and W2 are found to overlap subclone U2, while subclones O, N, P and U2 overlap subclone W2. The cross reference between subclones U2 and W2 implies that they already overlap. In addition, the occurrences over more than one step of subclones N, O, P, Q and R support this connection. In this case, the step from subclone U2 (or W2) supplies redundant information. Consequently, it is evident that parallel walking can result in extra steps when contigs are bridged. Thus parallel

walking should be attempted when the number of parent subclones to map is too few for the available task force so that the throughput can be raised.

Overlaps should be confirmed when information concerning them is ambiguous or inconsistent. Individual subclones can be amplified with the appropriate primers to verify the results of Spectrum PCR. In addition, if an overlap is less than 700–800bp, then it can be confirmed by sequencing the overlapping site of the candidate subclone (e.g., the β -end of clone Y in Figure 2). If the overlap is too large for sequence confirmation, then either stride amplification or conventional STS screening can be applied to the candidate subclone. In either case, a second primer, σ (Figure 2), must be synthesized. The stride of clone X (Figure 2) is amplified with primers σ and β , whereas STS amplification is performed with primers σ and τ .

EXPERIMENTAL RESULTS

The proposed mapping method was used to construct a high resolution map from the abdominal region [29] of the *Drosophila* bithorax complex (BX-C) [1] that had been cloned into a P1 bacteriophage vector [30]. The P1 clone carried 80Kb of *Drosophila* DNA in a 17Kb vector. Physical fragmentation methods, such as sonication [31,32] and pressure shearing [9], provide a library whose subclones are more uniform in size and more random in distribution than enzymatic methods such as partial digestion by restriction endonucleases [10]. DNA from the P1 clone was sheared using the French press and approximately 3Kb fragments were subcloned into a plasmid vector. 960 subclones were individually picked and placed into 10 96-well microtiter plates. This implies a 30-fold redundant library. The subclones were pooled in 3 dimensions according to the strategy illustrated in Figure 1.

Figure 5 shows an example of Spectrum PCR performed on the pooled library. Each step of the walk detected on average 15 triplets or multiplets, and about 10 of them could be clearly identified. Based on the number of subclones in the library, we should have detected on average 30 overlapping subclones per step. This suggests that 50% of the library did not contain inserts. PCR amplification of 10% of the library (using primer α and β in Figure 2) showed that 40% of the subclones did not contain inserts.

A complete map of the P1 insert is depicted in Figure 6-A. The map covers 80Kb and consists of 30 overlapping subclones. Walks were initiated at 10 different points. 2 points were at the edges of the P1 vector, and 8 more starting points were located at the edges of 4 starter subclones that contained known separate regions of the bithorax complex. These 4 subclones were isolated by hybridization. In Figure 6-B, a portion of the map (indicated with a dotted box in Figure 6-A) is expanded to include other overlapping subclones. In this region, two walks originating from F3_3E and F8_4A cross at the intersection of subclones F2_4B and F5_1F. In this case, one redundant walking step was made from subclone F2_4B (or subclone F5_1F). All minimum overlaps were confirmed by sequencing except for three large overlaps. One of these overlaps was confirmed by conventional STS amplification, whereas the other two were confirmed by neighborhood information (shared overlapping subclones). Later, this entire region of the bithorax complex was sequenced, providing additional confirmation of the map [33].

By the completion of the map, we had synthesized 40 region-specific primers and performed 2400 PCRs. A total of 40

Figure 6. Plasmid subclone map of a 80Kb bacteriophage P1 clone from the bithorax complex (BX-C) of *Drosophila melanogaster*. (A) The map extends from +140Kb to +220Kb in the infra abdominal (iab) region [31] of the *Drosophila* bithorax complex. The subclones comprising the map are shown as boxes. Walks were initiated from 10 points, 2 points from the edges of the P1 vector (retR and retL) and 8 more points from the edges of 4 starter subclones (T1_6A, T1_9B, T1_1A and T1_5B). A walk from clone A to clone B is indicated by an arrow from the edge of clone A to the overlapping shaded region of clone B. The white region of a clone indicates its stride. The sizes of overlaps and strides are given in base pairs. Symbols S and T at the ends of each clone indicate the orientation of the insert in the subcloning vector. (B) The subclones that could be identified during the mapping process are shown for the portion of the map outlined in (A).

subclones were identified by Spectrum PCR and sequenced at both edges in order to confirm the overlaps and design primers. These numbers are higher than expected because the map was being constructed while we were optimizing the Spectrum PCR protocol. Upon sequencing, twelve subclones were found not to overlap the region of interest. Each of these 'false positives' showed strong homology to the 3'-end of the region- or vector-specific primer (12–16bp identity). These mis-priming events were practically eliminated by two improvements in the PCR protocol. We adopted a *pseudo-hot start* technique in which reaction components were combined on ice and then transferred to a *hot* thermal cycler. 8 of the 12 false positives disappeared when reexamined using this technique. The remaining 4 false positives produced weak signals and could have been avoided simply by choosing intense triplets. By adding tetramethylammonium chloride (TMAC) to the PCR reactions, which is known to increase primer specificity [34], we found that weak signals either disappeared completely or were reduced to even lower intensities relative to true positives.

Using the improved PCR protocol, overlaps up to 1.5Kb in length were consistently detected. Within this range, at least one overlapping subclone was found in all cases. By increasing the elongation time from 2 min to 8 min, overlaps over 4Kb in length could be visualized; even two overlaps per pool that differed in size by more than 2Kb were amplified successfully (see Figure 5). In addition, the treatment of PCR products with mung bean nuclease before gel electrophoresis enabled bands of similar sizes to be resolved.

PROPERTIES OF THE METHOD

Performance

The cost for contig building by Spectrum PCR can be measured in terms of the numbers of PCR assays, sequencing reactions and oligonucleotides (or walking steps) required to construct a complete map. Note that the cost for library construction and pool construction is quite small compared to that for contig building.

Consider a source DNA of total length L_P consisting of an insert of length L_{PI} and a vector length L_{PV} , that is $L_P = L_{PI} + L_{PV}$. We want to construct a map of the insert from a c -fold redundant library consisting of subclones of average insert size L_{SI} , derived from the source DNA. The number of subclones contained in the library is

$$N = c \cdot \frac{L_P}{L_{SI}}$$

The average size of the minimum overlap, T , and the average size of the maximum stride, S , are expressed as:

$$T = \frac{L_{SI}}{c}$$

$$S = L_{SI} - T = L_{SI} \cdot \left(1 - \frac{1}{c}\right)$$

The average number of walking steps (or oligonucleotide syntheses) per map is

$$W = \frac{L_{PI}}{S} = \frac{L_{PI}}{L_{SI} \cdot \left(1 - \frac{1}{c}\right)}$$

Note each walking step requires two sequencing reactions, one for PCR primer design and the other for verification of the overlap, so the number of sequencing reactions is double the number of walking steps.

If the library is grouped according to the standard K -dimensional pooling strategy, then the number of pools per dimension is $K\sqrt[N]{N}$ and the total number of pools is

$$V = K \cdot K\sqrt[N]{N} = K \cdot K \sqrt[c]{\frac{L_P}{L_{SI}}}$$

The average number of PCR assays required to complete the map is

$$E_{PCR} = 2V \cdot W = 2 \left(K \cdot K \sqrt[c]{\frac{L_P}{L_{SI}}} \right) \left(\frac{L_{PI}}{L_{SI} \cdot \left(1 - \frac{1}{c}\right)} \right)$$

Let T_{min} be the minimum overlap length required for overlap detection, then the number of gaps is estimated as [24]:

$$G \approx \rho \cdot N \cdot e^{-c \cdot \left(1 - \frac{T_{min}}{L_{SI}}\right)}$$

where $\rho = L_{PI}/L_P$. In case of PCR-based detection, T_{min} is about 30bp, so the above equation reduces to:

$$G_{PCR} \approx \rho \cdot N \cdot e^{-c}$$

Note that this estimation is based on the assumption that the size of subclones, L_{SI} , and the redundancy throughout the library, c , are constant. In practice, biases introduced during the construction of the library and its propagation in a bacterial host affect both the size of subclones and the local densities in the library.

Consider the case in which a 3Kb subclone map is constructed from a parent clone carried in a P1 vector. Then $L_P = 100,000$; $L_{PI} = 80,000$ ($\rho = 0.8$); $L_{PV} = 20,000$; $L_{SI} = 3,000$. Figure 7 shows the above properties when Spectrum PCR is applied to a c -fold redundant subclone library that is pooled in 3 dimensions.

Notice that as c increases, W approaches a constant, L_{PI}/L_{SI} , and E_{PCR} approaches a value proportional to $K\sqrt[c]{c}$. Since the cost for sequencing the ends of subclones is proportional to W , the total mapping cost will be proportional to $K\sqrt[c]{c}$ for large c . In the example above, only 30% more work (E_{PCR}) is required to construct a map from a 30-fold redundant ($c=30$) library than from a 10-fold redundant ($c=10$) library. Although the estimated number of gaps in a 10-fold redundant library is 2×10^{-2} , for reasons mentioned above, the actual number of gaps can be far greater than this value. Construction of a complete map is more likely when a 30-fold redundant library is used, so the 30% additional work for this purpose is well justified. For example, if a region of the P1 insert is underrepresented in the library by

Figure 7. Performance. Consider a source DNA of total length L_P consisting of an insert of length L_{PI} and a vector length L_{PV} (i.e., $L_P = L_{PI} + L_{PV}$). A high resolution map of the insert is constructed from a c -fold redundant library consisting of subclones of average insert size L_{SI} , derived from the source DNA. If the library is grouped according to a 3-dimensional pooling strategy, then the performance of 3-dimensional Spectrum PCR can be measured with the following properties: (1) Number of subclones in library, $N = c \cdot L_P/L_{SI}$; (2) Total number of pools, $V = 3 \times 3\sqrt[N]{N}$; (3) Number of walking steps (or oligonucleotide syntheses) per map, $W = \rho \cdot L_P/L_{SI} (1 - 1/c)$, where $\rho = L_{PI}/L_P$; (4) Total number of PCRs, $E_{PCR} = 2V \cdot W$; (5) Number of gaps in the map, $G = \rho \cdot N \cdot e^{-c}$. These properties are graphed for the case in which a 3Kb subclone map is constructed from a 100Kb P1 clone containing an 80Kb insert: $L_P = 100,000$, $L_{PI} = 80,000$ ($\rho = 0.8$) and $L_{SI} = 3,000$.

a factor of 5, then to cover this region with 6-fold redundancy, we have to work with a 30-fold redundant library.

Limitations

Size of subclones. For certain applications, it may be useful to construct a map from subclones that are larger than 3Kb. However, the practical limit of PCR amplification is about 3Kb, so subclones that overlap the edge of a walk by more than this limit will go undetected. Nevertheless, it should be feasible to use Spectrum PCR to construct a lower resolution map as long as walking is done from one end of the source DNA to the other and the library is sufficiently redundant to ensure that the minimum overlaps are less than 3Kb. Parallel walking may be inefficient as it requires amplification of other overlapping subclones for contig connection. Obviously, as techniques for longer-distance PCR improve [35], the range in which Spectrum PCR can be applied to lower resolution mapping will expand.

Sensitivity of Spectrum PCR. The sensitivity of Spectrum PCR is limited by the number of overlaps that PCR can reliably amplify

in each pool and by the ability of gel fractionation to resolve the overlaps. The former influences the pooling strategy, whereas the latter affects the redundancy of the library. During the development of this method, we encountered mis-priming events in which PCR-amplification occurred from sequences that were similar to the primer sequence. This problem was reduced significantly by adopting the *pseudo-hot start* technique and by adding tetramethylammonium chloride to the PCR reactions, as described above. In addition, the employment of a longer extension time and the treatment of the PCR products with mung bean nuclease increased the number of identifiable overlaps. Further improvements in the sensitivity of Spectrum PCR can be expected. However, as long as many of the overlapping subclones can be identified, then a reliable map can be constructed efficiently from any number of starting points.

Generality

The method worked successfully when applied to a P1 clone that contained a region of the *Drosophila* bithorax complex. However, when applying the method to mammalian genomes, further considerations must be given to their repetitive DNA sequences [36,37,38,39,40]. If repeats are interspersed uniformly in the genome, then a P1 clone (about 80–100Kb long) derived from human DNA should contain about 13–33 Alu repeats (<300bp in length) [41,40] and 1–3 partial L1 repeats (<1.5kb in length) [42,40]. Since most of these repetitive elements are less than 300 base pairs, the sequence obtained from the end of the subclone should almost always contain some unique region from which a primer can be designed. However, the distributions of these repetitive elements are not completely uniform [43,44,39]; they form tandem repeats and clusters. In these regions, the chances are greater that the sequence from the end of a subclone will contain only repeats. As long as some region-specific sequence exists within the subclone, then a walk can be continued either by sequencing the entire insert or by choosing another overlapping subclone whose edge might be specific to the region. For instance, Edwards *et al* [45] found region-specific sequences within every 500bp in a 57Kb long continuous region, rich in Alu repeats (with 49 occurrences). Occasionally, one may encounter repetitive sequences that exceed the length of the subclone (e.g., the complete 7Kb L1 repeat). These long repeats will be rare and usually will be unique within the region of DNA (e.g., a P1 clone) that is being mapped. We are currently constructing a high-resolution map from human DNA.

APPLICATIONS OF HIGH RESOLUTION MAPS

A target for primer-walking or transposon-facilitated sequencing methods

A large region of DNA can readily be sequenced by first constructing a high resolution map consisting of minimum-overlapping subclones. Then, each subclone can be sequenced by either a primer-walking [46] or transposon-facilitated method [47].

Primer-walking requires the synthesis of many oligonucleotide primers. Each primer serves to extend the sequence about 300bp at which point another primer can be designed. Recently solutions to reduce the cost for primer synthesis and to introduce parallelism to the sequencing process have been studied [46]. Small subclones, such as those that comprise the high resolution map,

can be maintained in high-copy-number plasmids (or M13), and thus are excellent substrates for primer-directed sequencing.

In contrast, one example of transposon-facilitated sequencing involves inserting the transposon $\gamma\delta$ at random into the subclone and locating its position by PCR [47]. The ends of the transposon serve as priming sites for Sanger-sequencing. Since it does not require synthesis of region-specific primers and allows parallel sequencing from each transposon insertion, and moreover the experimental manipulations are simple, this method was used to obtain the complete sequence of the subclones depicted in Figure 6-A.

An alternative or supplementary method for sequencing

The proposed method has the potential to facilitate sequencing by efficiently positioning sequence priming sites throughout a region of DNA. Consider the case in which a library of 1000 3Kb subclones is constructed from a 100Kb P1 clone with an 80Kb insert. If we obtain 500bp of sequence from the ends of all the subclones, then we have performed *shotgun* sequencing [31,32] with 5-fold redundancy on each strand. We can eliminate over 80% of these sequencing reactions by first applying Spectrum PCR to map the subclones in the library, and then sequencing only the ends of subclones that are spaced 500bp apart. Since we are sequencing mapped regions, the data assembly is trivial when compared with the shotgun strategy. Spectrum PCR accurately positions only one end of a subclone. The opposite end can be positioned by determining the exact size of the subclone. This is easily accomplished by amplifying the subclone with the two vector-specific primers, α and β (Figure 2). Thus to sequence the 80Kb P1 insert, 800 PCRs are needed in addition to those required to construct the map.

The scenario outlined above is an oversimplification as we have not discussed gaps nor the fact that sequencing reactions frequently do not extend to 500bp. Consider the 100Kb P1 clone with an 80Kb insert (insert ratio 0.8). For both a 30-fold redundant library of 3Kb fragments and a 20-fold redundant library of 2Kb fragments, the expected number of sequencing gaps, G_{seq} , for a given readable sequencing ladder of length R_{seq} is as follows: $G_{seq}=11$ for $R_{seq}=500bp$, $G_{seq}=30$ for $R_{seq}=400bp$, and $G_{seq}=80$ for $R_{seq}=300bp$. We have mentioned the library of 2Kb fragments because 20 overlaps spread over 2Kb will be easier to accurately position by Spectrum PCR than 30 overlaps covering 3Kb. Of course, this benefit must be weighed against the 19 additional walking steps that are needed to complete the 2Kb resolution map and the increased probability that the map itself will contain a gap.

The formula for gaps, mentioned earlier, can be rewritten as:

$$G \approx \rho \cdot N \cdot e^{\left(-c \frac{S_{max}}{L_{SI}} \right)}$$

where S_{max} is the maximum stride that can be made without leaving a gap. For mapping purposes, a gap is defined as a portion of the source DNA not contained in any library subclones. Hence, for PCR-based detection $S_{max} \approx L_{SI}$. However for sequencing purposes, if a stride exceeds the length of a readable sequencing ladder, R_{seq} (typically 300–500bp), then a gap will exist in the final sequence. For example, if the stride of subclone X (Figure 2) is greater than R_{seq} , then the sequence obtained from subclone X with the β primer will not overlap subclone W. Hence $S_{max} = R_{seq}$. Because we want the sequence from both strands

of DNA, the number of sequencing gaps, G_{seq} , is double the value of G :

$$G_{seq} = 2G \approx 2p \cdot N \cdot e^{\left(-c \cdot \frac{R_{seq}}{L_{sl}}\right)}$$

The above estimation is based on the assumption that all overlapping subclones can be detected by Spectrum PCR. The actual number of sequencing gaps will be greater than the estimated value due to the inability to detect some subclones and to biases in the library (see Performance section). To close these sequencing gaps, it is reasonable to apply both primer-walking and transposon-based strategies. In particular, if a specific region of the P1 insert is underrepresented in the library, then only a few subclones will map there. This region can be quickly sequenced by inserting transposons into these subclones. It is also worth noting that the high resolution map, without any additional effort, can complement the transposon-facilitated sequencing method. Since $\gamma\delta$ does not transpose completely at random, some regions will be *cold spots* for insertions. Other factors can also contribute to the need for more sequence priming sites in a particular region. For instance, in sequencing the 80Kb P1 clone, 60 oligonucleotides were synthesized for primer-walking. The ends of some subclones may map to the cold spots and can help fill other sequencing gaps.

Isolation of expressed sequences

The construction of high resolution maps from regions of genomic DNA will simplify the search for expressed sequences within these regions. Each one of the minimum-overlapping subclones can be screened against random-primed cDNA libraries. Since the probes are low in complexity, the libraries can be screened at very high densities [28]. In this way, exons can be easily identified.

Rapid characterization of mutations

A point mutation can often be genetically mapped to a particular chromosomal region, but more precise localization is usually very difficult. If a high resolution map is available, then primers can be made that permit PCR-amplification of this region in 2–3Kb segments. It may then be possible to employ a sensitive heteroduplex mismatch assay to localize the mutation to one of the segments [48]. If the mutation has occurred in a well-defined chromosomal background, then polymorphisms will not complicate the analysis. This technique could facilitate the characterization of mutant loci in model organisms.

CONCLUSION

The enormous amount of effort for mapping and sequencing the human genome demands the development of new tools. For any procedure to be useful on such a large scale, it must be as simple and efficient as possible so that it can be systematized and/or automated. The proposed method satisfies these criteria. The non-binary detection scheme (based on quantitative assessment of overlaps) permits the use of a compact pooling strategy, which in turn leads to the very efficient characterization of overlapping subclones in a highly redundant library. Each step of the walk is performed by applying Spectrum PCR against the same set of pools in the same PCR mixture except for the region-specific primer. As long as the PCR products can be fractionated on a gel well enough to identify and quantify most bands, then the

walking step is completed in this phase. The iteration of this one-phase single-pattern experiment leads directly to the completion of a high-resolution map. Using this method, we were able to easily construct a 3Kb-resolution map from an 80Kb region of the *Drosophila* bithorax complex. This method is general enough to be applicable to DNA from other species and simple enough to be automated. The construction of a high resolution map from human DNA is in progress. We believe that this method will contribute to a wide range of applications in genome analysis.

ACKNOWLEDGEMENTS

We thank Terry Speed, Gerry Rubin, Jasper Rine and Ed Theil for their careful reading and helpful comments and suggestions. M.J.P. is a Lucille P. Markey Scholar and his part in this work was supported by a Scholar's Grant from the Lucille P. Markey Charitable Trust. M.P.S. is supported by the Alexander Hollaender Distinguished Postdoctoral Fellowship Program sponsored by the US Department of Energy, Office of Health and Environmental Research, and administered by the Oak Ridge Institute for Science and Education. This work was supported by Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, US Department of Energy under Contract DE-AC03-76SF00098 and also by Grant GNM 1 R01 HG00837-01 to M.J.P. from National Center for Human Genome Research.

REFERENCES

- Bender, W., Spierer, P., and Hogness, D.S. (1983) *Journal of Molecular Biology* **168**:17–33.
- Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H. (1988) *Science* **239**:487–491.
- Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989) *Science* **245**:1434–1435.
- Palazzolo, M.J., Sawyer, S.A., Martin, C.H., Smoller, D.A., and Hartl, D.L. (1991) *Proceedings of the National Academy of Sciences U.S.A.* **88**:8034–8038.
- Rosenthal, A. (1992) *Trends in Biotechnology* **10**:44–48.
- Boyle, A.L., Feltquite, D.M., Dracopoli, N.C., Housman, D.E., and Ward, D.C. (1992) *Genomics* **12**:106–115.
- Lichter, P., Cremer, T., Borden, J., Manuelidis, L., and Ward, D.C. (1988) *Science* **247**:64–69.
- Copeland, N.G. and Jenkins, N.A. (1991) *Trends in Genetics* **7**(4):113–118.
- Schrieffer, L.A., Gebauer, B.K., Quie, L.Q.Q., Waterston, R.H., and Wilson, R.K. (1990) *Nucleic Acids Research* **18**(24):7455–7456.
- Maniatis, T., Fritsch, E. F. and Sambrook J. (1982) *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory, New York.
- Lincoln, S.E., Daly, M.J. and Lander, E.S. (1991) available through ftp by anonymous login at genome.wi.edu
- Bellanné-Chantelot, C., Lacroix, B., Ougen, P., Billault, A., Beaufils, S., Bertrand, S., Georges, I., Gilbert, F., Gros, I., Lucotte, G., Susini, L., Codani, J.-J., Gesnouin, P., Pook, S., Vaysseix, G., Lu-Kuo, J., Ried, T., Ward, D., Chumakov, I., Le Paslier, D., Barillot, E., and Cohen, D. (1992) *Cell* **70**:1059–1068.
- Coulson, A., Sulston, J., Brenner, S., and Karn, J. (1992) *Proceedings of the National Academy of Sciences U.S.A.* **83**:7821–7825.
- Kohara, Y., Akiyama, K., and Isono, K. (1987) *Cell* **50**:495–508.
- Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., and Frank, T. (1986) *Proceedings of the National Academy of Sciences U.S.A.* **83**:7826–7830.
- Nizetić, D., Zehentner, G., Monaco, A.P., Gellen, L., Young, B.D., and Lehrach, H. (1991) *Proceedings of the National Academy of Sciences U.S.A.* **88**:3233–3237.
- Green, E.D., and Olson, M.V. (1990) *Proceedings of the National Academy of Sciences U.S.A.* **87**:1213–1217.
- Bentley, D.R., Todd, C., Collins, J., Holland, H., Dunham, I., Hassock, S., Bankier, A., and Giannelli, F. (1992) *Genomics* **12**:534–541.

19. Butler, R., Ogilvie, D.J., Elvin, P., Riley, J.H., Finnear, R.S., Slyn, G., Morten, J.E.N., Markham, A.F. and Anand, R. (1992) *Genomics* **12**:42–51.
20. Coffey, A.J., Roberts, R.G., Green, E.D., Cole, C.G., Butler, R., Anand, R., Giannelli, F. and Bentley, D.R. (1992) *Genomics* **12**:474–484.
21. Evans, G.A., and Lewis, K.A. (1989) *Proceedings of the National Academy of Sciences U.S.A.* **86**:5030–5034.
22. Kwiatkowski Jr., T.J., Zoghbi, H.Y., Ledbetter, S.A., Ellison, K.A., and Chinault, A.C. (1990) *Nucleic Acids Research* **18**(23):7191–7192.
23. Amamiya, C.T., Alegria-Hartman, M.J., Aslanidis, C., Chen, C., Nikolic, J., Gingrich, J.C., and de Jong, P.J. (1992) *Nucleic Acids Research* **20**(10):2559–2563.
24. Lander, E.S., and Waterman, M.S. (1988) *Genomics* **2**:231–239.
25. Barillot, E., Lacroix, B., and Cohen, D. (1991) *Nucleic Acids Research* **19**(22):6241–6247.
26. Frohman, M.A., Dush, M.K., and Martin, G.R. (1988) *Proceedings of the National Academy of Sciences U.S.A.* **85**:8998–9002.
27. Gibbons, I.R., Asai, D.J., Ching, N.S., Dolecki, G.J., Mocz, G., Phillipson, C.A., Ren, H., Tang, W.Y., and Gobbons, B.H. (1991) *Proceedings of the National Academy of Sciences U.S.A.* **88**:8563–67.
28. Hamilton, B.A., Palazzolo, M.J., and Meyerowitz, E.M. (1991) *Nucleic Acids Research* **19**(8):1951–1952.
29. Karch, F., Weiffenbach, B., Peifer, M., Bender, W., Duncan, I., Celniker, S., Crosby, M., and Lewis, E.B. (1985) *Cell* **43**:81–96.
30. Smoller, D.A., Petrov, D., and Hartl, D.L. (1991) *Chromosoma* **100**:487–494.
31. Deininger, P.L. (1983) *Analytical Biochemistry* **129**:216–223.
32. Bankier, A.T. and Barrell, B.G. (1983) *Nucleic Acid Biochemistry*, **B508**:1–34.
33. Martin, C.H., Celniker, S.E., Davis, C.A., Mayeda, C.A., Strathmann, M.P., Yoshida, K. and Palazzolo, M.J. (1992) GenBank entry (accession number L07835, locus DROABDB).
34. Hung, T., Mak, K. and Fong, K. (1990) *Nucleic Acids Research* **18**(16):453.
35. Ohler, L.D., and Rose, E.A. (1992) *PCR Methods and Applications* **2**:51–59.
36. Milosavljevic, A. and Jurka, J. (1992) PYTHIA: server for identification of human repetitive DNA. (send help message to phythia@anl.gov)
37. Jurka, J., Walkichiewicz, J. and Milosavljevic, A. (1992) *Journal of Molecular Evolution* **35**:286–91.
38. Moyzis, R.K., Torney, D.C., Meyne, J., Buckingham, J.M., Wu, J.-R., Burks, C., Sirotkin, K.M. and Goad, W.B. (1989) *Genomics* **4**:273–289.
39. Chen, T.L. and Manuelidis, L. (1989) *Chromosoma* **98**:309–316.
40. Jurka, J. and Milosavljevic, A. (1991) *Journal of Molecular Evolution* **32**:105–21.
41. Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T. and Schmid, C.W. (1981) *Journal of Molecular Biology* **151**:17–33.
42. Maio, J.J., Brown, F.L., McKenna, W.G. and Musich, P.R. (1981) *Chromosoma* **83**:127–144.
43. Daniels, G.E. and Deininger, P.L. (1985) *Nucleic Acids Research* **13**:8939–8954.
44. Furano, A.V., Somerville, C.C., Tschlis, P.N., and D'Ambrosio, E. (1986) *Nucleic Acids Research* **14**:3717–3727.
45. Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., Zimmermann, J., Erfle, H., Caskey, C.T. and Ansorge, W. (1990) *Genomics* **6**(4):593–608.
46. Kieleczawa, J. and Studier, F.W. (1992) *Science* **258**:1787–1791.
47. Strathmann, M.P., Hamilton, B.A., Mayeda, C.A., Simon, M.I., Meyerowitz, E.M. and Palazzolo, M.J. (1991) *Proceedings of the National Academy of Sciences U.S.A.* **88**:1247–1250.
48. Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K. and Sekiya, T. (1989) *Proceedings of the National Academy of Sciences U.S.A.* **86**(8):2766–2770.