Genomics **55,** 78–87 (1999)
Article ID geno.1998.5631, available online at http://www.idealibrary.com on IDEAL®

Integrated Mapping Package—A Physical Mapping Software Tool Kit

Peisen Zhang,[1] Xiaolu Ye, Li Liao, James J. Russo, and Stuart G. Fischer

Columbia Genome Center, Columbia University, New York, New York 10032

Received August 12, 1998; accepted October 13, 1998

We have developed an integrated physical mapping computer software package (IMP), originally designed to support the physical mapping of human chromosome 13 and expanded to support several gene-identification projects based on the positional candidate approach. IMP displays map data in a form that provides useful guidelines to the end users. An integrated map with high resolution and confidence is constructed from different types of mapping data, including hybridization experiments, STS-based PCR assays, genetic linkage mapping, cDNA localization, and FISH data. The map is also designed to provide suggestions for specific experiments that are required to obtain maps with even higher resolution and confidence. To this end, the optimization employs multiple constraints that take into account already established STS "scaffold" maps. This software thus serves as an important general tool kit for physical mapping, sequencing, and gene-hunting projects. © 1999 Academic Press

INTRODUCTION

In recent years, the "positional candidate" strategy has dominated the search for important disease genes (Collins, 1995). Typically, there are multiple stages to such an endeavor, including genetic linkage, physical mapping, cDNA recovery strategies, screening of regional ESTs, and eventually gene characterization and mutation analysis. Multiple approaches exist for each of these processes, and in many gene-hunting projects, several or all of these approaches are used.

For example, physical mapping may consist of restriction fingerprinting (Coulson *et al.*, 1986; Olson *et al.*, 1986; Kohara *et al.*, 1987; Carrano *et al.*, 1989; Stallings *et al.*, 1990; Trask *et al.*, 1992), PCR-based STS content mapping (Green and Olson, 1990; Foote *et al.*, 1992; Chumakov *et al.*, 1992), and clone-to-clone hybridization (Fischer *et al.*, 1994) and often incorporates maps and mapping reagents from external sources. Similarly, identification of candidate genes can include such methods as cDNA selection, exon trapping, use of public databases of regionally mapped genes or ESTs, homology searches (Altschul *et al.*, 1990) based on sample sequencing, and the use of exon-prediction programs such as GRAIL (Xu *et al.*, 1994), Genefinder (Solovyev *et al.*, 1994), and GeneScan (Burge and Karlin, 1997).

We present here an expanded version of our mapping programs (Zhang *et al.*, 1994) now assembled into an integrated mapping package (IMP), which has the capability of incorporating data from several mapping studies using different approaches. A unique feature of the package is the option for the user to adjust scoring parameters depending on which data are considered the most likely to be correct. For instance, genetic linkage or radiation hybrid data may form the boundary conditions within which more error-prone approaches such as matrix hybridization of end-labeled clones must be contained. Thus, initial runs of the mapping algorithms may precede further optimization runs. The data can be imported into a relational database management system such as SYBASE or ORACLE. Finally, given the dynamic nature of the genomics field, the IMP has been designed to be flexible enough to incorporate newer techniques and data (functional genomics databases, structural and motifs-based databases, etc.) as the need arises in a given project. We provide three example projects, undertaken at Columbia, in which the IMP has been instrumental: (i) the generation of a high resolution YAC-cosmid map over most of the length of human chromosome 13 (Cayanis *et al.*, 1998), (ii) the development of two highly annotated maps in the region surrounding BRCA2 (13q12.2; Fischer *et al.*, 1996), and (iii) the CLL deletion locus (13q14.3; Kalachikov *et al.*, 1997). This software developed on the UNIX platform is available by contacting the first author. A web page to use the software can be found at the following URL: http://genome1.ccc.columbia.edu/~genome/IMAP/imap.html.

DATA, MAPS, AND METHODS

We developed an efficient methodology to assemble ordered cosmid contigs aligned to YACs, which integrates

[1] To whom correspondence should be addressed at the Columbia Genome Center, College of Physicians & Surgeons, 630 West 168th Street, New York, NY 10032. Telephone: (212) 304-7555. Fax: (212) 304-5515.

0888-7543/99 $30.00

```
C1 C2 end
C2 C1 C3 C4 end
C3 C2 C4 end
C4 C2 C3 end
C5 C6 end
C6 C5 C7 end
C7 C6 end
C8 C9 end
C9 C8 end
C10 C11 C12 end
C11 C12 C13 C14 end
C12 C13 C14 end
C15 C16 end
C16 C15 end
END
```

```
     cc
     11
     56
c15  *-
c16  -*
```

```
     cc
     89
c8   *-
c9   -*
```

```
     ccc
     567
c5   *-
c6   -*-
c7    -*
```

```
     cccc
     1234
c1   *-
c2   -*--
c3   - *-
c4    --*
```

```
     cccc
     1111
     1432
c11  *--
c14  -* -
c13  - *-
c12   --*
```

FIG. 1. Input data format (left) and output matrix maps (right).

experimental methods and computer technology, and have used this general approach to construct high-resolution physical maps of human chromosome 13 (Fischer *et al.*, 1996). In support of this work, we developed the IMP software. The details of the biological experimental method and contig assembly program have previously been published (Zhang *et al.*, 1994). To assemble physical maps, the inter-*Alu* PCR probes of YACs are hybridized to arrayed cosmids from the Los Alamos chromosome 13 cosmid library gridded on filters at high density. Contigs are assembled from the subsets of cosmids hybridized by the YAC probes. Riboprobes from each cosmid are hybridized to filters containing this subset of cosmids (Fisher *et al.*, 1994). The assembly software, Cmap, previously described (Zhang *et al.*, 1994), is used to generate cosmid contigs aligned to overlapping YACs. The maps generated by IMP consist of two panels. The cosmid hybridization matrix forms the base; the map of cDNAs, markers, and YACs overlying these cosmids form the upper panel.

Data Formats and Maps

The IMP has functions to generate different maps according to the data sets provided and the users' needs. Cmap generates an ordered cosmid contig map using cosmid-to-cosmid riboprobe matrix hybridization data (from the hybridization of riboprobes from the ends of cosmid inserts to high-density cosmid colony filter membranes). Ymap generates a YAC-cosmid map based on hybridization of YAC inter-*Alu* probes to cosmids. Imap produces an integrated map based on various sets of mapping data, such as cosmid-to-cosmid riboprobe matrix hybridization data, YAC cosmid inter-*Alu* hybridization data, preordered and floating sequence-tagged-site (STS) data, cDNA data, genetic marker data, and radiation hybrid data. For these integrated maps, the cosmid-to-cosmid data and the YAC-to-cosmid data are essential. Other data sets are optional. With a very simple model, we explain below input data formats and output maps. To distinguish different input data types, we adopted some naming conventions for the input files. While we use YACs and cosmids in the descriptions here, the user can substitute PAC, BAC, P1, phage, and even plasmid clones. The important caveat is that what is substituted for the YAC should on average be substantially longer than what is substituted for the cosmid.

1. Cosmid matrix hybridization data and cosmid contigs. An input file of cosmid matrix hybridization data is created with a text editor (such as emacs) using the following format:

input_file:=[input_sentence. . .] END

where ":=" means definition, ". . ." means the last unit can be repeated as often as needed, "[ ]" means that choosing one or more of the enclosed items is optional, and END is necessary to terminate the input_file;

input_sentence:= cosmid [cosmids. . .] end

where the word "end" is necessary to complete the sentence, cosmid is represented by an alphanumeric string, and the first cosmid in the string is by definition the probe.

For example, in Fig. 1, Cmap, using a very simple model data file, yields five ordered contigs and a singleton C10. Note that C10 is not attached to the contig containing C11 and C12 because of the absence of two-way hybridization results.

2. YAC cosmid inter-Alu hybridization data and cosmid-to-YAC maps. Create an input data file with the following format:

input_file:=[input_sentence. . .] END

input_sentence:= YAC [cosmids. . .] end

where "YAC" is represented by an alphanumeric string, and other notations are defined as before. Our model YAC-cosmid inter-*Alu* input data are shown in Fig. 2.

By running Ymap, the output map (Fig. 3) is obtained.

```
Y1 C1 C2 C3 C4 C5 C6 end
Y2 C2 C3 C4 C5 C6 C7 C8 end
Y3 C4 C5 C6 C7 C8 C9 end
Y4 C11 C12 C13 C14 end
Y5 C11 C13 C14 end
Y6 C9 C12 C15 C16 end
END
```

FIGURE 2

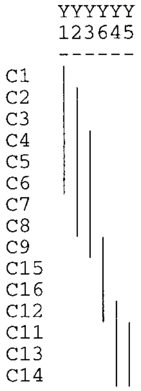


FIGURE 3

3. Basic integrated map. By combining cosmid-matrix hybridization data with YAC-cosmid inter-*Alu* hybridization data, a supercontig can be assembled and a basic integrated map (as in Fig. 4) can be generated, without extra mapping information, as those presented in a comprehensive integrated map (as in Fig. 3). A supercontig refers to a set of cosmids, YACs, markers, cDNAs, and other segments of DNA (such as STSs) that are connected by any length of physical overlap. For the model data sets shown in the previous sections, Imap will assemble a supercontig and generate a basic integrated map as shown in Fig. 4.

4. cDNA, genetic marker, and other supplemental data sets. These data sets can be treated in a way similar to YACs and have a format similar to the YAC-to-cosmid inter-*Alu* data. In Fig. 5, the cDNA and marker data for a model are shown in this format.

5. Annotated data sets and comprehensive integrated maps. The following mapping data sets have been merged into our integrated map as annotations: preordered and floating STS data, genetic mapping data, radiation hybrid data, and EST data. The STS data have the following format:

```
Y5  |---
Y4  |----
Y6  |   ----
Y3  |      ------
Y2  |       -------
Y1  |         ------

     CCCCCCCCCCCCCCC
     111111987654321
     143265
     ---------------
C11 |*--
C14 |-* -
C13 |- *-
C12 | --*
C16 |    *-
C15 |    -*
C9  |      *-
C8  |      -*
C7  |        *-
C6  |        -*-
C5  |         -*
C4  |           *--
C3  |           -*-
C2  |           --*-
C1  |             -*
```

FIGURE 4

```
cD5 c8 c9 end
D13 c12 c15 c16 end
END
```

FIGURE 5

input_file:=[input_sentence. . .] END

input_sentence:= STS [YAC. . .]; [YAC. . .] . [order] end where STS, as a cosmid or pseudo-cosmid, is represented by an alphanumeric string. The YACs in the first sequence before the semicolon are PCR positive. The YACs in the second sequence after the semicolon are PCR negative. The “order” after the period indicates a relative order of STSs. “Order” should be a nonzero integer. For example, we have the following STS data file shown in Fig. 6 for the model.

The genetic mapping location data file has the following format:

input_file:=[input_sentence. . .]

input_sentence:= cosmid location

where location is a string with a unit of centimorgans (cM). For example, in the given model, we will have the following genetic mapping location file: C7 20. This data file has only one input_sentence. It means that C7 has a genetic location of 20 cM. If there is only an approximate location, a tilde is used. For example, C7 ~20.

From the cosmid-to-cosmid data, YAC-to-cosmid data, STS data, and genetic location data, Imap will generate the map shown in Fig. 7. Users have the option to show or hide the STS order numbers.

On the first line of the integrated map, beginning with the abbreviation cM, a number represents the genetic distance in centimorgans from the centromere of the chromosome. There is only one cosmid, C7, that has a known genetic position, 20 cM, in this model. The second line, beginning with cR (centiray), gives the radiation hybrid location. The data are empty in this sample.

Then in the left margin above the cosmid hybridization matrix, this map contains a list of YAC names (e.g., Y3). The bottom block of the map contains a matrix of cosmids (e.g., C12) reading in the same order from left to right and from top to bottom. A positive hybridization is indicated as a dash at the intersection between a YAC and a cosmid or between cosmids. A capital “S” indicates that part of the cosmid sequence is

```
C8 Y2 Y3 ; . 10 end
C3 Y1 Y2 ; Y3 . 50 end
C5 Y1 Y2 Y3  . 10 end
C15 Y6 ; Y1 . 5 end
C4  Y2 Y1 Y3 ; . 20 end
C2  Y1 Y2 ; . 30 end
END
```

FIGURE 6

```
cM          20
cR

          S S  SSSS

D13 |   ---
cD5 |      --
Y5  |---
Y4  |----
Y6  |   --+-
Y3  |      -+---+ n
Y2  |       +--++++
Y1  |     n   -++++-
     CCCCCCCCCCCCCCC
     111111987654231
     143265
     ---------------
C11 |*--
C14 |-* -
C13 |- *-
C12 | --*
C16 |    *-
C15 |    -*
C9  |      *-
C8  |      -*
C7  |        *-
C6  |        -*-
C5  |         -*
C4  |           *--
C2  |           -*--
C3  |           --*
C1  |            - *
```

FIGURE 7

an STS. Testing by PCR back to the YACs is indicated by a small “s”, “+”, “o”, “-”, “n”, or empty space.

s, STS is positive, but cosmid did not hybridize to YAC inter-*Alu* PCR probe; n, STS is negative and cosmid did not hybridize to YAC inter-*Alu* PCR probe (sometimes, an empty space is used instead of “n”); +, STS is positive and cosmid did hybridize to YAC inter-*Alu* PCR probe; o, STS is negative although cosmid did hybridize to YAC inter-*Alu* PCR probe; -, STS was not tested; cosmid did hybridize to YAC inter-*Alu* PCR probe. An empty space means that the STS was not tested and the inter-*Alu* PCR test was negative.

Asterisks (*) refer to cosmid self-hybridization and form a diagonal in the matrix. The cosmid hybridization matrix consists of five cosmid contigs in this model.

Algorithms Underlying the IMP

1. Double search contig assembly. This algorithm (Zhang *et al.*, 1994), the foundation of this package, is outlined for the reader’s convenience. It is based on a double breadth-first search. Given a set of matrix cross-hybridization data, starting at any probe as a root to build a search tree, any clone (probe or nonprobe) that overlaps the root is put in the first layer of the tree, and this clone and the root are connected by adding an edge between them. Then a search is done for every clone in the first layer, any clone (not yet on the tree) that overlaps a clone of the first layer being put in the second layer, and it and its parent in the first layer are connected by adding an edge between them. This process is repeated for the clones in the second layer, the third layer, and so on until all clones have been searched. All the clones on the tree will form a contig. The boundary clones of the contig will be put on the tree last. By initiating a second breadth-first search from the boundary clones, the boundary clones on the other side will be obtained. Then alternative spanning paths will be generated, which will become the backbone for the ordering of all the clones in the contig. This algorithm is very fast and effective. A matrix has been chosen to represent the ordered contig, but the form of the output can be customized. For example, the column coordinate can represent the probes and the row coordinate the nonprobes. As an alternative, a square matrix with identical coordinates for all clones can be used without distinguishing between probe and nonprobe. The ordered contig matrix representation allows one to: (1) visualize the location of all the clones in the contig, (2) identify “weak points” in the contig where there are a relatively small number of connections, (3) identify outliers, which may indicate erroneously assigned hybridizing pairs, and (4) visualize the false positives and/or repeat units. A couple of techniques have been used to reduce the noise level (Zhang *et al.*, 1994). All the cosmid contigs in the human chromosome 13 mapping project have been generated by this algorithm. A mathematical graph, which can be abstracted from the hybridization data (Zhang *et al.*, 1994), has been studied (McMorris *et al.*, 1999).

2. Contig assembly with cDNAs, genetic markers, and YACs. To incorporate cDNA, genetic marker, and YAC data into cosmid contigs, the primary contig assembly algorithm has been enhanced in the following two aspects. First, cDNAs and genetic markers are allowed to participate in the process of cosmid contig assembly. In the double search cosmid contig assembly, cDNAs and genetic markers are considered the same as cosmids except in the output format. For example, given a cosmid and genetic marker layout as shown in Fig. 8, the marker m1 is equivalent to a cosmid that overlaps with cosmids a2 and a4. The data of cosmid hybridization, genetic marker, and cosmid contig were obtained as shown in Fig. 9.

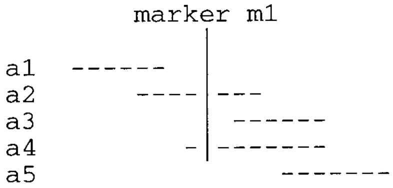
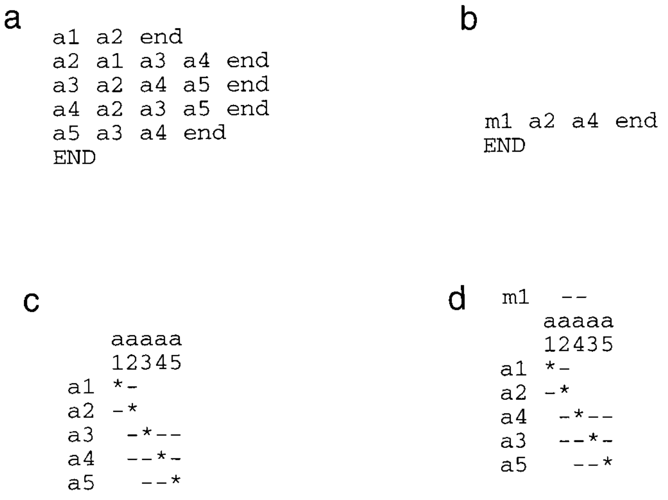


FIGURE 8



a

```
a1 a2 end
a2 a1 a3 a4 end
a3 a2 a4 a5 end
a4 a2 a3 a5 end
a5 a3 a4 end
END
```

b

```
m1 a2 a4 end
END
```

c

```
   aaaaa
   12345
a1 *-
a2 -*
a3  -*--
a4  --*-
a5   --*
```

d

```
m1  --
   aaaaa
   12435
a1 *-
a2 -*
a4  -*--
a3  --*-
a5   --*
```

FIG. 9. (a) Cosmid hybridization data. (b) Genetic marker data. (c) Cosmid contig assembled without the participation of genetic marker m1. (d) Cosmid contig assembled with the participation of genetic marker.

From the double breadth-first searches, several possible spanning paths will be generated. According to the order of the cosmids on the spanning path, the order of all the cosmids in the contig can be decided (Zhang *et al.*, 1994). In general, there are several possible orders. From those orders, one is chosen that best fits the data using a minimum target function that will be a weighted sum of the number of “holes” in the map of cosmids, YACs, cDNAs, and genetic markers. We call a discontinuous point between consecutive overlap symbols a hole. In the sample map, cosmid 133C1 has a hole at the intersection with cosmid 38A12, and cDNA125 has three holes at the intersections with cosmids 61F10, 118D2, and 106F7. The algorithm allows flexibility in assigning weights to different clone types. In our human chromosome 13 maps, the weight assigned to cDNAs and genetic markers is 10 times stronger than the weight assigned to YACs.

3. Contig assembly with ordered STSs. To incorporate ordered STS data into contigs, an alternative contig assembly algorithm has been developed. At Columbia, some cosmids have been used to generate STSs. The ends of the cosmid inserts are sequenced to design PCR primers. PCR assays are then performed on YACs, to validate the cosmid/YAC correspondence and also to generate the relative order of the STSs (cosmids). In addition, we have incorporated ordered STSs on chromosome 13 from outside databases (e.g., Whitehead Institute STS content map) into our integrated chromosome 13 maps. PCR assays are performed on the cosmid library to determine the overlaps among the STSs and cosmids. Thus, we deal with the STSs as pseudo-cosmids. We use the ordered STSs as a scaffold on which to order other cosmids in a manner similar to ordering cosmids by using the spanning path as a framework. An equally weighted order scheme has been adopted, as follows.

3.1. Every ordered STS (cosmid and pseudo-cosmid) is given a weight corresponding to its order. The first has weight 1, the second has weight 2, and so on.

3.2. Each remaining cosmid is inspected separately. If it overlaps with one or more cosmids and pseudo-cosmids that have been weighted, we count its weight as the average of the weights of those cosmids and pseudo-cosmids that it overlaps. If a cosmid does not overlap any weighted cosmids or pseudo-cosmids, assignment of a weight will be delayed until some of the cosmids and pseudo-cosmids that it does overlap have been assigned weights.

3.3. According to the weights, the cosmids and pseudo-cosmids are reordered.

3.4. To refine the order, the weights for each cosmid and pseudo-cosmid are assigned based on their overall order. A new order is obtained from the weights. If the new order is the same as the former one, a stable order has been attained. Otherwise the weights are recounted and reordered until a stable order has been obtained.

4. Statistical distance supercontig assembly. This algorithm can order and orient multiple adjacent cosmid contigs into supercontigs on multiple YACs. From the maximum-likelihood statistical distances (identical to those that Mott *et al.* (1993) have used for physical mapping in a different context), the average maximum-likelihood statistical distances (according to the YAC-to-cosmid inter-*Alu*-PCR hybridization data) between the contig halves (each contig is divided equally into left and right halves for the purpose of calculation of average maximum-likelihood distances) have been introduced to form the base for supercontig assembly. The algorithm consists of the following steps.

4.1. One randomly chosen contig is placed into an initial supercontig.

4.2. The average maximum-likelihood statistic distances (Ads) between the boundary halves of the supercontig and all the halves of the cosmid contigs that are not yet put in the supercontig are calculated (the formulae for calculations are defined below).

4.3. The contig with the shortest distance is placed into the supercontig, the order (left side or right side) and the orientation of the contig depend on the combination of the halves. (If the shortest distance is from the pair at the left boundary half of the supercontig and the left half of some tested cosmid contig, the latter contig is put on the left side of the supercontig with reverse orientation.)

4.4. Steps 4.2 and 4.3 are repeated until no further contigs can be placed within the first supercontig or the shortest distance becomes greater than some threshold value.

4.5. For the remainder, another supercontig is initiated.

4.6. The process is repeated until no contigs remain.

Definition 1, SD (the statistical distance between two single cosmids): Given cosmids a and b,

$$\mathrm{SD}_{ab} = 1 - \frac{\text{No. YACs hybridizing both a and b}}{\text{No. YACs hybridizing either a or b}}.$$

Definition 2, AD (average statistic distance between two cosmid groups): Given cosmid groups A and B, A consists of cosmids $a_1, a_2, a_3, \ldots, a_n$ and B consists of $b_1, b_2, b_3, \ldots, b_m$, such that

$$\mathrm{AD}_{AB} = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{SD}_{a_i b_j}\right)/(n \times m).$$

In our IMP, there are two user-defined threshold values for supercontig assembly. One is for a singleton, the single cosmid contig; the other is for nonsingleton contigs. To put a contig into a supercontig, the required distance (threshold value) is shorter for singletons than for nonsingletons.

5. Supercontig assembly with ordered STSs as scaffold. If there are at least two ordered STSs in a cosmid contig (including pseudo-cosmids), the orientation of this contig has been fixed by the order of the STSs. If there are at least two contigs with fixed order and orientation in a supercontig, they are used as a scaffold to generate an ordered supercontig. An alternative "supercontig assembly with scaffold" algorithm has been developed. This algorithm consists of two iterative steps.

5.1. Suppose there are n ordered and oriented contigs as scaffold. In placing one extra contig in the scaffold, there will be $n + 1$ possible locations: before the first, between the first and the second, between the second and the third, and so on until after the last. In addition, the orientation should be determined. There are therefore a total of $2(n + 1)$ configurations. The local AD is calculated to find out which one is the most favorable. For example, for the configuration with location between the first and the second on a reverse orientation, the local AD is half of the sum of the AD between the right half of the first and the right half of the extra and the AD between the left half of the second and the left half of the extra. For each boundary position, only one AD between the halves needs to be calculated. For example, for the left boundary with normal orientation, the local AD is the AD between the right half of the extra and the left half of the first. Given each extra contig, the most favorable configuration can be chosen by calculating all local ADs. From all extra contigs, one can be chosen with the closest, most favorable configuration. This is referred to as the "most favorable contig."

5.2. The most favorable contig is put into its most favorable configuration to form a new scaffold. Then step 5.1 is repeated until all contigs have been located. In the iteration of step 5.1, some previously calculated local ADs can be saved. The penalty for reversal of ordered STSs can be incorporated into the local ADs.

6. Algorithm used in Ymap. Given a set of YAC inter-*Alu* cosmid hybridization data, an algorithm that will generate the orders of all YACs and cosmids is implemented as follows. The user defines the threshold value that will determine if the two YACs overlap. If the data are relatively pure, two YACs that share one cosmid can be considered overlapping. For noisy data, the user can set the threshold value greater than 1, for example, 2. Thus the overlaps among the YACs will be obtained. By a double search algorithm, spanning paths are obtained. From these spanning paths, the order of the YACs can be determined by using weight-order iteration as presented in section 3. To obtain the order of the cosmids and to refine the order of the YACs, an alternative iteration scheme is performed as follows.

6.1. Each cosmid is weighted according to the average orders of YACs that it overlaps. All cosmids are ordered from their weights.

6.2. Each YAC is weighted according to the average orders of cosmids that it overlaps. All YACs are ordered from their weights.

6.3. Steps 6.1 and 6.2 are repeated until a stable order has been achieved. In our definition, if a new order is identical to the previous one, a stable order has been reached.

This algorithm can be used to determine the orders of STSs from a set of STS content mapping data. As in the previous case, the user should define the threshold value that will determine if the two clones overlap. If the data are relatively pure, two clones that share one STS can be considered overlapping. For noisy data, the user can set the threshold value greater than 1. Thus the overlaps among the clones will be obtained. With the above algorithm, the STS set can be treated as a cosmid set and the clone set as a YAC set. The STS order and the clone order can be determined.

RESULTS

IMP has been used to generate integrated YAC maps for human chromosome 13 (Fischer *et al.*, 1994; Cayanis *et al.*, 1998), a high-resolution comprehensive physical map of the human chromosome 13q12–q13 region containing the breast cancer susceptibility locus BRCA2 (Fischer *et al.*, 1996), and an integrated physical map of the human chromosome 13q14 region containing the deletion in chronic lymphocytic leukemia (Kalachikov *et al.*, 1997). These maps can be found on our WWW page (http://genome1.ccc.columbia.edu/~genome/).

YAC Maps for Human Chromosome 13

Since the supercontigs are often too extensive to visualize, YAC maps have been adopted as the basic unit for presentation of data. These YAC maps are integrated components of the Columbia human chromosome 13 database. With a Web browser, certain features of the map may be selected to allow for information searches extending from the original supercontig beyond its boundaries to overlapping YAC maps. Where a related YAC seems to run off the map, clicking on the "|" following that YAC's name calls its own map and often allows visualization of its extension; clicking on the YAC's name yields all the YAC inter-*Alu* PCR probe hybridization data, STS-to-YAC hybridization data, and marker (including cDNA)-to-YAC hybridization data available in the human chromosome 13 database. With this approach, one can walk along the chromosome until the boundary of the supercontig—a gap between YACs not spanned by any other clone—is reached. Clicking on the cosmid's name will lead to three sets of data, cosmid-to-cosmid, cosmid-to-YAC, and cosmid-to-marker (including cDNA), all from the same chromosome 13 database.

Integrated YAC maps, of a suitable size for viewing, have been chosen as windows to see the whole map of chromosome 13. Although the integrated YAC maps are not real geometric maps due to the lack of length information, nor real topological maps given the lack of absolute continuity for every DNA fragment, the integrated organization of different mapping data into one map provides a very useful picture to visualize the complicated relationships among the different mapping data sets. The maps can thus be viewed as quasi-topological maps. According to the nature of the data, the orders of clones and markers shown in those maps, although still subject to being modified by later experimental data, can serve as a good starting point for further analysis at the gene or sequence level.

At the time of writing, we have generated more than 700 integrated YAC maps and an equal number of YAC-to-cosmid maps. We present here one integrated YAC map, that of Y841a5 (Fig. 10).The remaining YAC maps can be found through our Web page. Fifty-five cosmids are hit by YAC 841a5, there are nine other YACs hitting those 55 cosmids, and 27 cDNAs and markers are related to those cosmids. Among the 55 cosmids, 13 have been used to generate STSs. On the first line of this YAC map, labeled cM, a number represents the genetic distance from the centromere. In YAC map Y841a5, cosmid 106F7 has a genetic distance of 23 cM relative to the centromere on the q arm of chromosome 13. Similarly, on the second line labeled cR, a number represents the radiation hybrid distance from the centromere. In map Y841a5, cosmid 15F8 has a radiation hybrid distance of 70 cR and 106F7 has a radiation hybrid distance of 74 cR. On the third line, an S tag indicates that the cosmid in the same column has been used to generate an STS. On the Sequence line, a "-" indicates that an end, usually T7, of the cosmid in the same column has been sequenced. On the line beginning with the tag q12.2, a "-" indicates that the cosmid has been mapped *in situ* to 13q12.2. On the upper half of the upper block of the map, in lines beginning with marker or cDNA names, a "-" is used to denote hybridization between the marker or cDNA and the corresponding cosmid. In map Y841a5, marker D13S120 hybridizes to cosmids 10E12 and 69G12, while cDNA125 hybridizes to cosmids 61F10 and 106F7. On the lower half of the upper block of the map, in lines beginning with YAC names, a "-" signifies hybridization between the YAC and the corresponding cosmid. Since it is a map of YAC 841a5, this YAC overlaps all 55 cosmids. YAC 930e2 overlaps 28 cosmids. Nine of them have been used to generate STSs. Four of those 9 cosmids, 110A10, 104G4, 116F7, and 124A7, were not hit by YAC 930e2 *Alu*-PCR hybridization, although their STS assays were positive. Such a discrepancy could be due to insufficient representation of inter-*Alu* PCR products in this portion of the YAC, false-positive PCR results, or deletions within the YAC. This YAC map consists of 6 cosmid contigs, their order generated by IMP.

A Physical Map of the Human Chromosome 13q12–q13 Region

By assembling various types of physical mapping data, a high-resolution annotated physical map of a 4-Mb region of human chromosome 13 that includes the BRCA2 locus has been generated. This map consists of a YAC contig with 42 members spanning the 13q12–q13 region and aligned contigs of 399 cosmids established by cross-hybridization between the cosmids, which were selected from a chromosome 13-specific cosmid library using inter-*Alu* PCR probes from the YACs. The end sequences of 60 cosmids spaced nearly evenly across the map were used to generate STSs, which were mapped to the YACs by PCR. A contig framework was generated by STS content mapping, and the map was assembled on this scaffold. Additional annotation was provided by 72 expressed sequences and 10 genetic markers that were positioned on the map by hybridization to cosmids. For additional details, see the BRCA2 map (December 1995) at our Human Chromosome 13 Web page (http://genome1.ccc.columbia.edu/∼genome/) (Fischer *et al.*, 1996).

A Physical Map of the Human Chromosome 13q14 Region

An integrated high-resolution annotated map of YAC, PAC, and cosmid contigs has been constructed. This map covers 600 kb of the 13q14 genomic region where an unknown tumor suppressor gene for B-cell chronic lymphocytic leukemia (CLL) is expected. In addition to densely positioned genetic markers and STSs, this map was further annotated by localization

```
cM                                                           23
cR                       70                                  74

            S      S          S S          SSSS  SS     SS      S

SEQUENCE  | -  -   ---       --            ----  ---   ---      -
q12.2     |                      -   -            --      --     -
BL834     |------
D13S120   | --
D13S904   |         -   ---
fB-16E2   |                - --
pT-1D7c   |                  ---- --
br2       |                         ----
pB-20F2   |                           - --
pFB-1G2c  |                               -  - --
47C4Alu   |                               -----
pB-19H3   |                                        ---
pB-19B9   |                                        ---
pB-19A10  |                                        ---
pB-20F5   |                                        ---
fH/P-17E4 |                                           ----
pH/P-17G5 |                                           ---
pH/P-17F6 |                                           ---
pB-19E5   |                                           ---
pFB-1D7c  |                                           ---
pT1-A10c  |                                           - --
br5       |                                           ----
pT-1B12c  |                                            --
D13S6     |                                              --
cDNA125   |                                                -   ---
fH/P-14B11|                                                - -- -
pT-1A7y   |                                                 - -
D13S290   |                                                 - -
pB-20F6   |                                                    --
930e2     |-+------+--------  s-+   -   -  ss+s- +
940g4     |-+------+----------+-+-------- -++++--++-----++------+--
841a5     |-+------+----------+-+---------++++--++-----++------+--
964b2     |        s        --+-+  --- -   ++++--++-----++--- --s
910h2     |                    -   --   -  - --- - -- ----    ---
765d6     |                                ++s+--++-- --++  ----+ -
753d5     |                                      s+-- - ++      s
755c1     |                                       --------   -----
365b5B    |                                           ---
931f4     |                                                - - -
           11664435118336512471529164425929111121619911492116112212
           90997783307649653031977108921106013212C15441837331101B32
           DEGHBBAG37CHBAAFBAF0BBC6BEBAHFDC4684E174CD07BEB61F86H29A
           31114312CA156388989A837G77758111GFBA1D C63DE562GD1DF5 B8
            220  2 112        1   1     001471711 1  71   54027  1
                    0         0                 0 1              2
```

FIGURE 10

of 32 transcribed sequences (ESTs) using a combination of exon trapping, direct cDNA selection, sample sequencing of cosmids and PACs, and homology searches. For further details, see the B-CLL map (July 1997) at our Human Chromosome 13 Web page (http://genome1.ccc.columbia.edu/~genome/) (Kalachikov *et al.*, 1997).

DISCUSSION

The suite of programs in IMP has been developed to incorporate additional types of experimental data collected from external databases and from the Columbia Genome Center. The new features of IMP include: (1) supercontig assembly, (2) gap minimization, (3) STS scaffolding, (4) cDNA localization, and (5) optimization. A statistical distance approach has been used to generate supercontigs from adjacent cosmid contigs based on the YAC and cosmid inter-*Alu* PCR data. To enhance the contig assembly program, an optimization procedure was adopted by assigning weighted penalties for the number of holes in the upper panel. We assign strong weights to ungapped cDNAs and genetic markers and weak weights to ungapped YACs. To incorporate the ordered STS data in the maps, three additional algorithms have been used. Inside a cosmid contig, an iterative optimization protocol has been developed, using the ordered STSs as an initial framework, to generate the best cosmid order. Among the cosmid contigs, a penalty function has been added to the statistical distance. If enough ordered STS data are available to form a scaffold, a best statistic distance fitting will be performed. To eliminate false positive links caused by repeat units or chimerism, a mecha-

```
            --------------------------------------------------------
19D3       |*-----
10E12      |-*------
69G12      |--*-----
69H10      |---*-----
47B4       |----*----
47B3       |-----*---
38A12      | -----*-
53G2       | ------*-----
133C1      |   --- -*-------
107A10     |       --*------
87C12      |       ---*-----
36H5       |       ----*----
34B6       |       -----*---
69A3       |        -----*--
56A8       |        ------*-
15F8       |        -------*
23B9       |                *--- -
40A8       |                -*-- -
73F9       |                --*--- -
110A10     |                ---* ---
59B8       |                  - *- -
27B3       |                -----*--- -
97C7       |                   - -*-
116G1      |                  -----*--
60B7       |                     - -*
48E7       |                       - *-- -
49B7       |                     -   -*---
22A5       |                         --*--
51H8       |                          --*-
91F10      |                         ----*-
20D10      |                             -*
96C11      |                               *-
104G4      |                               -*-
116F7      |                                -*- -
138B1      |                                 -*--
124A7      |                                  -*-
21E11      |                                 ---*-
121D10     |                                    -*-
6C7        |                                     -*
114C11     |                                       *--
95C6       |                                       -*-
94D3       |                                       --*-
140D7      |                                         -*
117E1      |                                           *---
48B5       |                                           -*--
93E6       |                                           --*-
27B2       |                                           ---*-
136G5      |                                              -*
131D4      |                                                *------
61F10      |                                                -*---- -
118D2      |                                                --*-----
106F7      |                                                ---*----
21H5       |                                                ----*---
2B2        |                                                -----*--
139B12     |                                                - ----*-
22A8       |                                                 ------*
```

FIGURE 10—*Continued*

nism is provided to indicate "bad" YACs. There are two customizable factors. One is the minimum length and the other is the maximum number of holes (or relative number of holes). For example, if a YAC hybridizes with only one or two cosmids, it is very possible that the YAC does not belong to the region. Users can adjust the length and holes to fit their needs. IMP has been designed to generate reasonable maps from noisy data sets. The algorithms used by IMP have polynomial-time complexity. Most are linear or quadratic. An integrated map of about 500 cosmids, 100 YACs, and 50 cDNAs will be generated in a couple of minutes on a DEC Alpha workstation.

IMP also provides a few plain file database maintenance functions: merger, with which new sets of data can be merged into the database; filter, with which a specific data set can be selected from the database; and purifier, with which a subset of the data can be eliminated from the database. Furthermore, an option is available for copying this plain file database into a relational database management system, to make the manipulation of the data more flexible and more effective. We are using IMP database maintenance functions and SYBASE, a relational database management system, to maintain chromosome 13 mapping data. Web browsers have been used as user interfaces for the human chromosome 13 database and the databases for gene identification projects.

We have developed an integrated physical mapping computer software package (IMP) designed to support physical mapping of human chromosome 13 as well as several gene identification projects based on the positional candidate approach. In the previous sections, IMP was presented and details concerning data formats, maps, and the underlying algorithms were expounded. We have described two applications in highly annotated maps which were constructed by IMP in disease gene loci. Our methodology is a natural expansion of developments in physical mapping (see Primrose, 1995). The ability to handle noisy data is a key criterion for physical mapping software. The algorithm used to construct the cosmid hybridization matrix has the ability to deal with false negatives and false positives (Zhang *et al.*, 1994). Since the algorithm for the supercontig assembly is based on the maximum likelihood statistical distance, it is a tool for managing noisy data. In our experience, for a low-noise data set, IMP will generate a high-quality map; for a modestly noisy data set, a reasonable map can be expected. We have used IMP to generate local maps for several gene identification projects; in fact, IMP has become a routine tool in our research.

ACKNOWLEDGMENTS

The authors are grateful to referees for their comments and suggestions.

REFERENCES

Altschul, S. F., Gish, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, in press.

Carrano, A. V., Lamerdin, J., Ashworth, L. K., Watkins, B., Branscomb, E., Slezak, T., Raff, M., DeJong, P. J., Keith, D., McBride, L., Meister, S., and Kronic, M. (1989). A high-resolution fluorescence-based semi-automated method for DNA fingerprinting. *Genomics* **4:** 129–136.

Cayanis, E., Russo J. J., Kalachikov, S., Ye, X., Park, S. H., Sunjevaric, I., Bonaldo, M., Lawton, L., Venkatraj, V. S., Schon, E., Soares, M. B., Rothstein, R., Warburton, D., Edelman, I. S., Zhang, P., Efstratiadis, A., Fischer, S. G. (1998). High resolution YAC–cosmid–STS map of human chromosome 13. In press.

Chumakov, I., *et al.* (1992). Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359:** 380–387.

Collins, S. F. (1995). Positional cloning moves from perditional to traditional. Nat. Genet. 9:347–350.

Coulson, A., Sulston, J., Brenner, S., and Karn, J. (1986). Toward a physical map of the genome of the nematode *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA* **83:** 7821–7825.

Fischer, S. G., Cayanis, E., Russo J., Sunjevaric, I., Boukhgalter, B., Zhang, P., Rothstein, R., Yu, M-T., Warburton, D., Edelman, I. S., and Efstratiadis, A. (1994). Assembly of ordered contigs from YAC-selected cosmids of human chromosome 13. *Genomics* **21:** 525–537.

Fischer S. G., Cayanis, E., Bonaldo, F. M., Bowcock, A. M., Deaven, L. L., Edelman, I. S., Gallardo, T., Kalachikov, S., Lawton, L., Longmire, J. L., Lovett, M., Lawrence, S. O., Rothstein, R., Russo, J., Soares, B., Sunjevaric, I., Venkatraj, V. S., Warburton, D., Zhang, P., and Efstratiadis, A. (1996). A high-resolution annotated physical map of the human chromosome 13q12–13 region containing the breast cancer susceptibility locus BRCA2. *Proc. Natl. Acad. Sci. USA* **93:** 690–694.

Foote, S., Vollrath, D., Hilton, A., and Page, D. C. (1992). The human Y chromosome: Overlapping DNA clones spanning the euchromatic region. *Science* **258:** 60–66.

Green, E. D., and Olson, M. V. (1990). Systematic screening of yeast artificial chromosome libraries by use of the polymerase chain reaction. *Proc. Natl. Acad. Sci. USA* **87:** 1213–1217.

Kalachikov, S., Migliazza, A., Cayanis, E., Fracchiolla, N. S., Bonaldo, M. F., Lawton, L., Jelenc, P., Ye, X., Qu, X., Hauptschein, R., Gaidano, G., Vitolo, U., Saglio, G., Resegotti, L., Brodjansky, V., Yankovsky, N., Zhang, P., Soares, M. B., Russo, J., Edelman, I. S., Efstratiadis, A., Dalla-Favera, R., and Fischer, S. G. (1997). Cloning and gene mapping of the chromosome 13q14 region deleted in chronic lymphocytic leukemia. *Genomics* **42:** 369–377.

Kohara, Y., Akiyama, K., and Isono, K. (1987). The physical map of the whole *E. coli* chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50:** 495–508.

McMorris, F. R., Wang, C., and Zhang, P. (1999). On probe interval graphs. *Discrete Appl. Math.* **89:** 287–295.

Mott, R., Grigoriev, A., Maier, E., Woheisel, J., and Lehrach, H. (1993). Algorithm and software tools for ordering clone libraries, application to the mapping of the genome of *Schizosaccharomyces pombe. Nucleic Acids Res.* **21:** 1965–1974.

Olson, M., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., and Frank, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci. USA* **83:** 7826–7830.

Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989). A common language for physical mapping of human genome. *Science* **245:** 1434–1435.

Primrose, S. B. (1995). “Principles of Genome Analysis,” Blackwell Sci., Oxford.

Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22:** 5156–5163.

Stallings, R. L., Torney, D. C., Hildebrand, C. E., Longmire, J. L., Deaven, L. L., Jett, J. H., Doggett, N. A., and Moyzis, R. K. (1990). Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Natl. Acad. Sci. USA* **87:** 6218–6222.

Trask, B., Christensen, M., Fertitta, A., Bergmann, A., Ashworth, L., Branscomb, E., Carrano, A., and van den Engh, G. (1992). Fluorescence *in situ* hybridization of human chromosome 19: Mapping and verification of cosmid contigs formed by random restriction enzyme fingerprinting. *Genomics* **14:** 162–167.

Xu, Y., Mural, R. J., and Uberbacher, E. C. (1994). Constructing gene models from accurately predicted exons: An application of dynamic programming. *CABIOS* **10:** 613–623.

Zhang, P.**,** Schon, E. A., Fischer, S. G., Cayanis, E., Weiss, J., Kistler, S., and Bourne, P. (1994). An algorithm based on graph theory for the assembly of contigs in physical mapping of DNA. *CABIOS* **10:** 309–317.