



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12N 15/00, 15/09, 15/10, C12Q 1/68	A1	(11) International Publication Number: WO 95/05458 (43) International Publication Date: 23 February 1995 (23.02.95)
(21) International Application Number: PCT/US94/09037 (22) International Filing Date: 5 August 1994 (05.08.94) (30) Priority Data: 08/105,753 12 August 1993 (12.08.93) US (71)(72) Applicant and Inventor: PERLIN, Mark, W. [US/US]; 5904 Beacon Street, Pittsburgh, PA 15217 (US). (74) Agent: SCHWARTZ, Ansel, M.; Suite 301, 425 N. Craig Street, Pittsburgh, PA 15213 (US).		(81) Designated States: CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published <i>With international search report.</i> <i>Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i>
(54) Title: A SYSTEM AND METHOD FOR PRODUCING MAPS AND CLONING GENES THEREFROM		
(57) Abstract		
The present invention pertains to a process for efficiently and accurately constructing indirectly high-resolution maps of genomes and other entities. Specifically, the process comprises the steps of generating three sets of reagents, namely, (1) sample points, (2) covering regions, and (3) probes; performing parallel experiments to efficiently compare these reagents and acquire two data tables, one of (a) probes with covering regions, and another of (b) covering regions with sample points; combining these two independent tables by means of an inner product operation to produce a third table that indirectly compares probes with sample points; analyzing this computed table to construct a high-resolution map of the probes, the sample points, and the covering regions. The resulting integrated map is then used to efficiently search for probes based on their proximity to selected sample points. With genome maps, this search can enable the rapid discovery of genes, their products, and their useful applications.		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

A SYSTEM AND METHOD FOR PRODUCING
MAPS AND CLONING GENES THEREFROM

FIELD OF THE INVENTION

The present invention is related to a process for
5 efficiently and accurately constructing high-resolution maps
of genomes and other entities. More specifically, the
present invention is related to an indirect mapping process
that takes two efficiently obtainable data tables, and
computationally combines them into a third table, from which
10 an indirect high-resolution map can be constructed.

BACKGROUND OF THE INVENTION

For the positional cloning of genes and other novel
types of genetic experiments, in humans and other organisms,
there is a crucial need for techniques that can rapidly
15 construct genome-wide high-resolution ordered clone maps.
Current best methods, such as sequence-tagged site (STS)
content mapping, entail a large number of experiments, and,
in practice, require large low-resolution yeast artificial
chromosome (YAC) clones and very many STSs. Here, Inner
20 Product Mapping (IPM) is introduced that overcomes these
limitations. IPM uses radiation hybrids (RHs) to provide
localizing signatures for YACs. Two independent data tables
are obtained which compare YACs against RHs, and RHs against
STSs; these tables are combined to produce a computed map of
25 the YACs against ordered STSs. IPM maps each YAC
independently, requires relatively few RH comparisons to map
a YAC, and can work with small (or large) YACs and few (or
many) STSs.

30 There is currently considerable effort directed
toward constructing genetic and physical genome maps for

-2-

humans and model organisms. By localizing a phenotype on the genetic map, the corresponding physical and ordered clone maps can be used to retrieve chromosomal segments containing a gene of interest. DNA sequencing of these segments can
5 determine the gene's DNA sequence, thus enabling functional studies of the gene's product and its mutations. A genetic map currently exists at 5 centiMorgan (cM) resolution (NIH/CEPH Collaborative Mapping Group (1992). A Comprehensive Genetic Linkage Map of the Human Genome. *Science*, **258**: 67-86;
10 Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., and Lathrop, M. (1992). A second generation linkage map of the human genome. *Nature*, **359**: 794-801), and a greater resolution map is currently being developed. Physical and ordered clone mapping of
15 chromosomes, however, is still in its infancy (Maier, E., Hoheisel, J.D., McCarthy, L., Mott, R., Grigoriev, A.V., Monaco, A., Larin, Z., and Lehrach, H. (1992). Complete coverage of the *Schizosaccharomyces pombe* genome in yeast artificial chromosomes. *Nature Genetics*, **1**: 273-277).
20 Current technologies are enabling the construction of a 1 megabase (mb) resolution map (Bellanne-Chantelot, C., Lacroix, B., Ougen, P., Billault, A., Beaufigli, S., Bertrand, S., Georges, S., Glibert, F., Gros, I., Lucotte, G., Susini, L., Codani, J.-J., Gesnouin, P., Pook, S., Vaysseix, G., Lu-
25 Kuo, J., Ried, T., Ward, D., Chumakov, I., Le Paslier, D., Barillot, E., and Cohen, D. (1992). Mapping the Whole Genome by Fingerprinting Yeast Artificial Chromosome. *Cell*, **70**: 1059-1068), but require tremendous experimental effort (Green, E.D., and Green, P. (1991). Sequence-tagged site
30 (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods and Applications*, **1**: 77-90). For routine use, a higher resolution 250 kilobase (kb) map would be far more useful for

-3-

locating genes. There is, therefore, a great and immediate need for ordered clone mapping strategies that are both more efficient and achieve higher resolution.

5 Crucial to the success of large scale mapping experiments is the availability of numerous sequence tagged sites (Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989). A common language for physical mapping of the human genome. *Science*, **245**: 1434-35) (STSs), which are unique,
10 very short 100-500 base pair (bp) genomic sequences that can be amplified from sample DNA by the polymerase chain reaction (PCR). Polymorphic STS markers have proved highly successful in genetic linkage map construction (Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P.,
15 Vaysseix, G., and Lathrop, M. (1992). A second generation linkage map of the human genome. *Nature*, **359**: 794-801). Further, when also used as physical mapping reagents, these STSs can correlate the genetic and physical maps.

20 High resolution (< 250 kb) physical maps of STSs can be efficiently constructed via radiation hybrid (RH) mapping (Cox, D.R., Burmeister, M., Price, E.R., Kim, S., and Myers, R.M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of
25 mammalian chromosomes. *Science*, **250**: 245-250). A RH clone contains several very large (5 to 50 mb), nonoverlapping human chromosome fragments from a specific chromosome in a rodent cell line. These chromosome fragments are randomly formed, and cover 25-50% of the chromosome. To physically
30 map STSs, a data table is constructed which compares a set of RH clones against a set of unordered STSs. Each entry in the table records whether or not an STS is present in some fragment of a RH. By permuting the order of the STSs (i.e.,

-4-

the table's columns) (Chakravarti, A., and Reefer, J.E. (1992). A Theory for Radiation Hybrid (Goss-Harris) Mapping: Application to Proximal 21q Markers. In Genetic Analysis Workshop 7: Issues in Gene Mapping and the Detection of Major Genes. Cytogenet Cell Genet, 99-101. vol. 59, (MacCluer, J. W., Chakravarti, A., Cox, D., Bishop, D. T., Bale, S. J., and Skolnick, M. H., eds.; Cox, D.R., Burmeister, M., Price, E.R., Kim, S., and Myers, R.M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science*, 250: 245-250), an ordering can be found either by minimizing the number of obligate breaks (Boehnke, M. (1992). Radiation hybrid mapping by minimization of the number of obligate chromosome breaks. In Genetic Analysis Workshop 7: Issues in Gene Mapping and the Detection of Major Genes. Cytogenet Cell Genet, 96-98. vol. 59, (MacCluer, J. W., Chakravarti, A., Cox, D., Bishop, D. T., Bale, S. J., and Skolnick, M. H., eds.; Weeks, D.E., Lehner, T., and Ott, J. (1992). Preliminary ranking procedures for multilocus ordering based on radiation hybrid data. In Genetic Analysis Workshop 7: Issues in Gene Mapping and the Detection of Major Genes. Cytogenet Cell Genet, 125-127. vol. 59, (MacCluer, J. W., Chakravarti, A., Cox, D., Bishop, D. T., Bale, S. J., and Skolnick, M. H., eds.)) present in all RH rows, or by using maximum likelihood methods (Chakravarti et al., 1992; Cox, D.R., Burmeister, M., Price, E.R., Kim, S., and Myers, R.M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science*, 250: 245-250). Inter-STS distances can also be estimated (Boehnke, M., Lange, K., and Cox, D.R. (1991). Statistical Methods for Multipoint Radiation Hybrid Mapping. *Am. J. Hum. Genet.*, 49: 1174-1188; Chakravarti et al., 1992; Cox, D.R., Burmeister, M., Price, E.R., Kim, S.,

-5-

and Myers, R.M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science*, **250**: 245-250) by counting the RH breaks occurring between STSs.

5 Yeast artificial chromosome (YAC) clones (Bellanne-Chantelot et al., 1992; Burke, D.T., Carle, G.F., and Olson, M.V. (1987). Cloning of large exogenous DNA into yeast by means of artificial chromosomes. *Science*, **236**: 806-
10 812) can perpetuate a large (100kb to 1.2mb) linear sequence of human DNA. Since YAC DNA inserts are very large, an ordered YAC clone map is an ideal reagent for retrieving genes via positional cloning. Such YAC orderings can be constructed by labor and time intensive techniques such as
15 hybridization fingerprinting, which compares restriction enzyme digests of two YACs at a time for commonalities in their fingerprint patterns, and extends contigs of overlapping YACs. The current best approach to ordering YACs is STS-content mapping (Green, E.D., and Green, P. (1991).
20 Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods and Applications*, **1**: 77-90), which employs YAC vs. STS comparisons. Two YACs overlap when they share a common STS; YACs containing two or more STSs can be
25 used to extend contigs. When the STSs are genetically mapped, there is an immediate correspondence with the physical YAC clone map. The chief drawback to this method is that large YACs and very many STSs must be used to achieve sufficient intersections of YACs with STSs. Another likely
30 candidate approach for mapping YACs at high resolution is RH mapping. Unfortunately, YAC inserts are too large to satisfy the necessary break minimization assumptions.

-6-

This invention utilizes Inner Product Mapping (IPM), a new technique that enables high resolution mapping of YACs using RHs. IPM employs a YAC vs. RH comparison data table, but supplements it with localizing information for
5 each RH, specifying where its fragments reside on the chromosome: e.g., a RH vs. ordered STS table. When these two independent data tables are mathematically combined, a computed YAC vs. ordered STS table is obtained. The positive entries in this computed table localize the YAC along the
10 chromosome or genome.

Lehrach's group has demonstrated the feasibility of performing highly parallelized experiments which can compare, for example, one RH against tens of thousands of gridded YACs in a single experiment (Monaco, A.P., Lam, V.M.S., Zehetner,
15 G., Lennon, G.G., Douglas, C., Nizetic, D., Goodfellow, P.N., and Lehrach, H. (1991). Mapping irradiation hybrids to cosmid and yeast artificial chromosome libraries by direct hybridization of Alu-PCR products. *Nucleic Acids Res*, 19(12): 3315-3318). This group has noted the utility of inner
20 products, but has only used them for consistency checks, and not for building maps (Mott, R., Grigoriev, A., Maier, E., Hoheisel, J., and Lehrach, H. (1993). Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schizosaccharomyces pombe*. *Nucleic*
25 *Acids Research*, 21(8): 1965-1974). Studies of the distribution of Alu and L-1 interspersed repetitive sequences (IRS) in YACs (Arveiler, B., and Porteous, D.J. (1992). Distribution of Alu and L1 repeats in human YAC recombinants. *Mammalian Genome*, 3: 661-668) and chromosomes have
30 demonstrated the utility of using IRS-PCR as a mechanism for extracting human DNA content from various clonal reagents. Statistics for superimposing recombinant chromosomes in

-7-

pedigrees have been used recently in genetic mapping (Nelson, S.F., McCusker, J.H., Sander, M.A., Kee, Y., Modrich, P., and Brown, P.O. (1993). Genomic mismatch scanning: a new approach to genetic linkage mapping. *Nature Genetics*, 4(May): 11-18; 5 Ward, P.J. (1993). Some Developments on the Affected-Pedigree-Member Method of Linkage Analysis. *Am. J. Hum. Genet.*, 52: 1200-1215), but, unlike IPM, make no use of neighborhood information and have not been applied to constructing physical maps.

10

SUMMARY OF THE INVENTION

The present invention pertains to a method for producing a gene. The method comprises the steps of isolating a gene from a map constructed by using radiation hybrids versus probes. Then there is the step of producing 15 the gene isolated from the map.

The present invention also pertains to a system for cloning a gene of a genome. The system comprises means for isolating a sufficiently large set of first DNA sequences having a size between 100b-40kb to achieve a desired 20 resolution. Each first DNA sequence is part of the genome. The system is also comprised of means for isolating a sufficiently large set of second DNA sequences having a size between 40kb-2mb to achieve the desired resolution. Each second DNA sequence is part of the genome. Additionally, the 25 system comprises means for isolating a sufficiently large set of collection of DNA sequence fragments wherein each fragment's average size is greater than or equal to the size of said second DNA sequence fragment to achieve the desired resolution. Each collection of DNA sequence fragments is

-8-

part of the genome. The system comprises a computer having a memory and means for inputting information into the computer about the set of second DNA sequences, the set of first DNA sequences and the set of collections of DNA sequence fragments. There is means for forming a Table A in the memory. Table A determines the containment or overlap of the set of second DNA sequences against the set of collections of DNA sequence fragments. The system comprises means for forming a Table B in the memory. Table B determines the containment or overlap of the set of first DNA sequences against the set of collections of DNA sequence fragments. Additionally, the system comprises means for forming a Table L from Table A and Table B. Table L provides localization information about the set of second DNA sequences relative to the set of first DNA sequences. There is means for producing a map of the gene from Table L. The map determines where the set of first DNA sequences are located within the genome and where the set of second DNA sequences are located within the genome. The system also comprises means for producing a gene in the genome of the map. Preferably, there can also be means for purifying derivatives of the gene produced in the gene producing means, and the first DNA sequences are STSs, the set of second DNA sequences are YACs, and the DNA sequence fragments are radiation hybrids.

The present invention also pertains to a method for cloning a gene of a genome. The method comprises the steps of identifying a portion of a genome to be mapped. Then there is the step of setting a desired resolution of the map. Next, there is the step of isolating a sufficiently large set of first DNA sequences having a size between 100b-40kb to achieve the desired resolution. Each first DNA sequence is

-9-

part of the genome. Then there is the step of isolating a sufficiently large set of second DNA sequences having a size between 40kb-2mb to achieve the desired resolution. Each second DNA sequence is part of the genome. Next there is the

5 step of isolating a sufficiently large set of collections of DNA sequence fragments wherein each fragment's average size is greater than or equal to the size of each of said second DNA sequences to achieve the desired resolution. Each collection of DNA sequence fragments is part of the genome.

10 Then there is the step of forming a Table A in a memory of a computer. The Table A determines the containment or overlap of a set of second DNA sequences against the collection of DNA sequence fragments. Next there is the step of forming a Table B in the memory wherein Table B determines the

15 containment or overlap of a set of first DNA sequences against the collection of DNA sequence fragments. Next there is the step of operating a processor of the computer on Table A and Table B to produce a Table L which provides localization information about the set of second DNA

20 sequences relative to the set of first DNA sequences. Next there is the step of operating the processor on Table L to produce a map of the genome which determines where the set of first DNA sequences are located within the genome. Then there is the step of producing a gene in the genome map.

25 Preferably, after the producing step, there is the step of purifying derivatives of the gene produced. Also, the first DNA sequences are preferably STSS, the second DNA sequences are YACs, and the DNA sequence fragments are radiation hybrids.

30 The present invention pertains to a process for efficiently and accurately constructing indirectly high-resolution maps of genomes and other entities. Specifically,

-10-

the process comprises the steps of generating three sets of reagents, namely, (1) sample points, (2) covering regions, and (3) probes; performing parallel experiments to efficiently compare these reagents and acquire two data
5 tables, one of (a) probes with covering regions, and another of (b) covering regions with sample points; combining these two independent tables by means of an inner product operation to produce a third table that indirectly compares probes with sample points; analyzing this computed table to construct a
10 high-resolution map of the probes, the sample points, and the covering regions. The resulting integrated map is then used to efficiently search for probes based on their proximity to selected sample points. With genome maps, this search can enable the rapid discovery of genes, their products, and
15 their useful applications.

BRIEF DESCRIPTION OF THE DRAWINGS

In the accompanying drawings, the preferred embodiment of the invention and preferred methods of practicing the invention are illustrated in which:

20 Figure 1 is a Table A of YAC vs. RH comparison data.

 Figure 2 is a Table B of RH vs. STS comparison data.

 Figure 3 is an RH fragment/gap pattern. Each row
25 represents one RH. The tick marks above and below the diagram shows the true locations of the STS sample points.

-11-

Figure 4 is a Table L of the YAC vs. STS comparison matrix computed by IPM using an arithmetic inner product that sums all the matches.

Figure 5 is a Table AxB of the result of the inner product multiplication of Tables A and B, after thresholding Table L row by row. Both the YAC rows and the STS columns are ordered for better visualization.

Figure 6 is a for every row, the predicted YAC center (the row's vertical tick mark) is shown overlain upon the true YAC location (the row's horizontal line).

Figure 7 is a schematic representation of a system of the present invention.

Figure 8 is a flow chart of a method for cloning a gene of the present invention.

15 DESCRIPTION OF THE PREFERRED EMBODIMENT

Referring now to the drawings wherein like reference numerals refer to similar or identical parts throughout the several views, and more specifically to figure 7 thereof, there is shown a schematic representation of a system 10 for cloning a gene of a genome. The system 10 comprises means 12 for isolating a sufficiently large set of first DNA sequences having a size between 100b-40kb to achieve a desired resolution. Each first DNA sequence is part of the genome. For instance, for STSS of well-characterized species (man, mouse, etc.), the first DNA sequence isolating means preferably entails looking up (in journals or databases) the PCR primers (or sequences) of

-12-

known STSs, and then constructing the primers using a DNA synthesizer. Alternatively, the original laboratories can be contacted to provide DNA sequence primers for the STSs. In species that have not yet been well characterized, 5 genetically useful highly polymorphic STSs can be constructed by building CA (or other tandem nucleotide) repeat markers. These are found by using complementary DNA (e.g., GT polymers) as PCR primers, and expressing and isolating matching CA-repeat sequences from digested DNA fragments. 10 Characterization, leading to eventual sequencing, can identify an entire library of polymorphic markers.

The system 10 also comprises means 14 for isolating a sufficiently large set of second DNA sequences having a size between 40kb-2mb to achieve the desired resolution. 15 Each second DNA sequence is part of the genome. For instance, for a cosmid library, the second DNA sequence isolating means entails purchasing the requisite n-hit library from one of many companies which advertise in molecular biology journals; the current cost is about \$5,000. 20 For a human YAC library, this entails contacting one of many Human Genome Centers, and requesting the clones or panels.

Alternatively, one could construct such a library by

- dilution-limited restriction enzyme digestion of a genome (or portion thereof),
- 25 - size selection by gel migration, and
- cloning of appropriate size inserts.

-13-

The system 10 additionally comprises means 16 for isolating a sufficiently large set of collection of DNA sequence fragments wherein each fragment's average size is greater than or equal to the size of said second DNA sequence to achieve the desired resolution. Each collection of DNA sequence fragment is part of the genome. For example, the means for isolating a set of collections of DNA fragments can be to build a set of radiation hybrid clones. Further, with IPM, these need not be chromosome specific. For example, to create human radiation hybrid clones, this is done by lethally irradiating (whole genome or chromosome specific) hamster-human hybrid HPRT+ donor cells, and rescuing the irradiated hybrid cells by cell fusion with HPRT-recipient cells. The surviving cells have (1) a HPRT gene from the donor cells, and (2) possibly, a collection of human chromosome fragments. Subsequent screening with human Alu or Line repeat sequences assure that (2) is the case. Another approach is to pool YACs or other large insert clones from library. Such insert clones are created by the mechanisms described above.

The system 10 is comprised of a computer 18 having a memory 20 and means 22 for inputting information into the computer 18 about the set of second DNA sequences, the set of first DNA sequences and the set of collection of DNA sequence fragments.

The system 10 is also comprised of means 24 for forming a Table A in the memory 20. Table A determines the containment or overlap of the set of second DNA sequences against the set of collections of DNA sequence fragments. The system 10 is comprised of means 26 for forming a Table B in the memory 20. Table B determines the containment or

-14-

overlap of the set of first DNA sequences against the set of collections of DNA sequence fragments. There is also means 28 for forming a Table L from Table A and Table B. Table L provides localization information about the set of second DNA sequences relative to the set of first DNA sequences. The system 10 is also comprised of means for producing a map of the gene from Table L. The map producing means 30 is disposed in the memory 20. The map determines where the set of first DNA sequences are located within the genome and where the set of second DNA sequences are located within the genome. The means for producing Table A, the means for producing Table B, the means for producing Table L and the means for producing the map can preferably be the technique described below. The system 10 additionally is comprised of means 32 for producing a gene of the genome from the map. Preferably, the system 10 also includes means 34 for purifying derivatives of the gene produced in the gene producing means 32. Additionally, the first set of DNA sequences are preferably STSs, the second set of DNA sequences are YACs and the DNA sequence fragments are radiation hybrids.

The present invention also pertains to a method for producing a gene. The method comprises the steps of isolating a gene from a map constructed by using radiation hybrids. Then there is the step of producing a gene isolated from the map. Preferably, the method of cloning a gene of a genome comprises the steps of identifying a portion of a genome to be mapped, as shown in figure 8. Then there is the step of setting a desired resolution of the map. Then there is the step of isolating a sufficiently large set of first DNA sequences having a size between 100b-40kb to achieve this desired resolution. Each first DNA sequence is part of the

-15-

genome. Then there is the step of isolating a sufficiently large set of second DNA sequences having a size between 40kb-2mb to achieve the desired resolution. Each second DNA sequence is part of the genome. Next there is the step of
5 isolating a sufficiently large set of collections of DNA sequence fragments wherein each fragment's average size is greater than or equal to the size of each of said second DNA sequences to achieve the desired resolution. Each collection of DNA sequence fragment is part of the genome. Next there
10 is the step of forming a Table A in the memory of a computer 18. Table A determines the containment or overlap of the set of second DNA sequences against the set of collections of DNA sequence fragments. Next there is the step of forming a Table B in the memory 20. Table B determines the containment
15 or overlap of the set of first DNA sequences against the set of collections of DNA sequence fragments. Next there is the step of operating a processor of the computer 18 on Table A and Table B to produce a Table L which provides localization information about the set of second DNA sequences relative to
20 the set of first DNA sequences. Then there is the step of operating the processor on Table L to produce a map of the gene which determines where the set of first DNA sequences are located within the gene and where the set of second DNA sequences are located within the gene. Then there is the
25 step of producing a gene from the map. Preferably, after the producing step, there is the step of purifying derivatives of the gene produced. Moreover, the first DNA sequences are preferably STSSs, the second DNA sequences are preferably YACs and the DNA sequence fragments are preferably radiation
30 hybrids.

More specifically, the task is to find the location of every probe (belonging to a set of probes) in some (e.g.,

-16-

spatio-temporal, multivariate) coordinate system. A probe may be thought of as some ball of any small (even zero) to medium size volume in the space. Initially,

- 5 - it is known that each probe resides at some fixed location;
- it is not known what this location is;
- the coordinate system is described by a set of sample points;
- 10 - it is not necessary to know the location of these coordinate sample points.

Under these conditions, the objective is to build a map of where the probes (and, possibly, the sample points) reside in the coordinate system.

15 There are two possible strategies to determine the locations of the probes (and, possibly, the sample points). The first is to *directly* compare each probe against each sample point, and determine whether or not this probe/point pair overlaps, recording the +/- response. When there is no guarantee that each probe overlaps at least one sample point
20 (e.g., too few sample points, or very small volume probes), this direct method cannot work.

25 The second, novel, strategy presented here is to perform *indirect* comparison experiments via an intermediate set of independent *covering regions*, and then computationally combine these comparison results to build a map of the probes (and, when necessary, the sample points). A set of covering

-17-

regions is selected, which are neither probes or sample points. One covering region is comprised of a set of separate, individually connected and fairly large, subregions. The total volume of all these subregions (for one covering region) covers a large volume (say, half) of the coordinate space.

- Each probe is compared against each covering region entity for overlap, constructing a data table A.
- 10 - Each covering region is compared against each sample point for overlap, constructing a data table B.
- Tables A and B are then multiplied as matrices under a suitable generalized inner product operation that performs a matching operation.
- 15
- The result is Table L, which indirectly compares each probe with each sample point, thereby localizing the probe in the coordinate space.
- If the locations of the sample points are not initially known, and there is a sufficient number of comparison experiments, the method can additionally determine the position of each sample point in the coordinate space. I.e., it can map out the coordinate space via its sample points.
- 20
- 25 This indirect mapping technique is called *Inner Product Mapping*, or *IPM*. It works as follows.

-18-

Suppose that the location of the sample points are entirely known (eventually this assumption will be relaxed). To fix ideas, suppose that the coordinate space is finite, two-dimensional, and covered by a uniform grid of sample
5 points. Further suppose that the probes are too small relative to the sample points to be reliably detected. I.e., direct mapping cannot work here, and the indirect IPM method is required.

To use IPM, a suitable set of covering regions are first
10 identified. They are selected to be properties of the coordinate space that roughly cover 50% of the area. For concreteness, suppose that the two-dimensional coordinate space is actually some small geographical area. And, that the probes can perform measurements in this area to ascertain
15 whether or not they reside in a covering region. In this example, likely covering regions might be:

- Elevation - Is the altitude above 100 feet?
- Vegetation - Does some (fairly ubiquitous) plant grow here?
- 20 - Pollution - Is the region polluted by some chemical?

Further assume that these covering region properties are entirely independent of one another.

Now, one tries to locate a given probe.

- 25 (1) The probe tests the elevation, and returns a "+"; it is indeed over 100 feet. This partial

-19-

localization reduces the search area to $1/2$ of the total space.

- 5 (2A) The probe then tests the vegetation, and records a
"-": the plant does not grow at the probe's
location. This is another partial localization,
to the complement of the vegetation covering
region, that reduces the possible area by $1/2$.
- 10 (2B) By combining the independent results of (1) and
(2), and intersecting the elevation covering
region, together with the complement of the
vegetation covering region, the probe is localized
to $1/4$ ($= 1/2 \times 1/2$) the total area.
- 15 (3) Continuing in this way, the probe would then test
the third covering region. Suppose the result was
"+", indicating pollution. Intersecting this
covering region with the previous results would
localize the probe $1/8$ ($= 1/2 \times 1/2 \times 1/2$) the
area.

20 Thus, the +/- signature vector of the probe against each
of the covering regions provides a mechanism for rapid,
precise locationalization. By intersecting regions or their
complements, according to the "+" or "-" result, IPM effects
localization. And, since each new region incrementally
reduces the possible probe area by (roughly) $1/2$, the IPM
25 method has exponential power in the number of covering
regions.

Mathematically, this result can be described as the
inner product of the probe's signature vector **a** against a
matrix **B** of covering regions. Each row of **B** characterizes

-20-

one covering region in terms of the space's sample points. Multiplication of a "+" vector component with a row selects the (sampled) covering region, while multiplication by "-" selects the complement of the covering region. This can be
5 thought of as an equivalence operation: + with +, or - with -, yields +; + with -, or - with +, yields -.

The inner product combines these component multiplication results by *intersecting* them. With *conjunctive* intersection, for example, a sample point is
10 inferred to be close to a probe only when its covering region signature is identical to (i.e., equivalent for every covering region) the probe's covering region signature. There are other, less restrictive, ways of combining the partial results. For example, those points which have the
15 greatest number of equivalences could be inferred as closest to the probe.

The common property to all such inner product definitions is the notion of signature *match*. Every probe and sample point has its own signature, determined by
20 independent experiments against the covering regions. In essence, the inner product operation looks for the closest match between the probe's signature vector and one (or more) sample point's signature(s). The point(s) having the best matching signature is inferred to be in closest proximity to
25 the probe. In practice, a local neighborhood, rather than just one point, may be matched.

Notationally, it is written as:

$$I = A \times B,$$

-21-

where

- $\mathbf{l}$ is the probe's localization vector containing the Boolean or numerical localization values. The peak of this curve localizes the probe with respect to the sample points. The shape of the (numerical match) curve is Gaussian-like.
- $\mathbf{a}$ is the probe's signature vector (i.e., covering region comparisons).
- $\mathbf{x}$ is the particular inner product operation.
- $\mathbf{B}$ is the matrix of covering region vs. sample point comparisons.

Each row of $\mathbf{B}$ describes one covering region in terms of sample points. Conversely, each column of $\mathbf{B}$ provides one sample point's comparison signature against the covering regions.

It is often the case that multiple probes must be localized. If each probe can perform its tests on the covering regions independently of the other probes, then much experiment parallelization can be achieved. (The same holds true experimentally for testing multiple sample points against the covering regions.) Instead of just one probe's signature row vector $\mathbf{a}$, a signature matrix $\mathbf{A}$ is written, each row of which provides the signature of a distinct probe.

Mathematically, then, a matrix $\mathbf{A}$ of probe signature row vectors is obtained. Performing all the probe localization computations in one process, it is therefore written:

-22-

$$L = A \times B,$$

where each row of matrix L localizes one probe.

The key mathematical fact is that

THEOREM: Each row of L localizes its probe in the space.

- 5 That is, by obtaining probe vs. covering space comparisons (Table A), and covering space vs. sample point comparisons (Table B), the inner product of A and B does indeed compute an effective comparison of the probes vs. the sample points, i.e., the location of the probe.
- 10 The theorem has several useful corollaries, such as:
- Corollary.** When the probes are uniformly distributed, and each probe resides at a single location, then, up to row permutation, L resembles a (broadened) identity matrix.
- 15 Suppose that the positions of neither the probes nor the sample points are initially known. Further suppose that the covering regions are not explicitly known, and are, in fact, only measurable by performing comparisons against probes or points. That is, much data can be generated, but there is
- 20 absolutely no localization information about probes, covering regions, or sample points.

A key feature of IPM is that from such comparison experiments, full localization information can be accurately inferred.

-23-

For clarity, consider the one dimensional case (e.g., chromosome mapping) where the space is a finite line segment (e.g., a chromosome). When the proper column permutation p is known, the (suitably modified) IPM localization theorem

5 $L = A \times B \times p$

will hold. In that case, every row of L will have the shape of a stretched Gaussian-like curve. However, with the wrong column permutation $p' \neq p$, the product

$$L = A \times B \times p'$$

10 will scatter the row peaks all over the rows.

It is straightforward to construct a heuristic $h(\text{row})$ that is minimized precisely when the row assumes a Gaussian-like, single-peaked, shape. Elements of this heuristic that may cluster maxima include:

- 15 - minimizing the curve's line length;
- forcing the maxima to lie close to one other;
- maximizing separation between putative peaks and surrounding regions.

If a good predicted curve shape is available,

- 20 - an L2 norm can be used (i.e., a function space inner product) to minimize the functional distance between the observed and predicted functions. Specifically, the predicted template function is

-24-

placed at every spatial coordinate, and the L2 difference between it and the data is computed; the smallest such L2 difference is $h(\text{row})$.

Regardless, a function can be readily specified that
5 determines the deviation of a row from the desired peaked shape. A global matrix error function $g(\text{matrix})$ is then computed by summing $h(\text{row})$ over all the rows.

Using (1) the IPM Theorem, together with (2) such a matrix error function, a global search technique that
10 minimizes the total deviation (such as *simulated annealing*) can rapidly and accurately converge to a correct permutation p . (There may be more than one permutation that satisfy the data.) In the one dimensional case, this amounts to properly ordering the sample points. The power of this method
15 increases with the number of probes: the greater the quantity of unlocalized data, the greater the localization capability.

By modeling or simulation, the precise shape of the row function can be determined. This is used to compute the
precise spatial location of the sample points (i.e., not just
20 their relative ordering). This is done by specifying a functional that computes the deviation of each probe's observed row function from its predicted shape. The spatial location of each sample point is slightly adjusted to minimize the overall deviation (for all the probes), say by
25 local relaxation methods. Thus, the underlying space is locally stretched into its proper shape, thereby determining the spatial location of every sample point.

-25-

In summary, then, one starts with a large of set of probes, and a large set of sample points, both with unknown locations in some space. An (also unlocalized) set of intermediate covering regions are selected. Comparison
5 experiments of the probes vs. the regions are performed, to obtain table A, and of the regions vs. the points to obtain table B. Inner product mapping then computationally combines tables A and B to accurately determine:

- 10 - the location of every probe (in terms of the sample points),
- the location of every sample point (as a coordinate in the space), and
- the spatial extent of every covering region.

One concrete application of IPM is in efficiently
15 constructing integrated genetic/physical maps of chromosomes (or entire genomes). That is, the space here is a one dimensional line of finite length. Suppose one has a dense (e.g., 1 megabase (MB) resolution) set of *sample points* along the chromosome comprised of genetic/physical markers, such as
20 the PCR-based sequence-tagged sites (STSS). Further suppose that one also had physical *probes* of DNA sequences, such as yeast artificial chromosomes (YACs) or cosmids. Now, assume that both the sample points and the probes were accurately mapped along the one dimensional chromosome. Then, useful
25 genes could be rapidly found by positional cloning:

- Perform genetic family studies of a disease to localize the causative gene(s) along the genetic point map.

-26-

- Look up the gene point's position in map coordinates.
- Retrieve the corresponding physical probes at that map location.
- 5 - Determine the DNA sequences of the probes.
- Identify the few candidate genes on the probes.
- Select (and verify) the causative gene.

With the causative gene(s), one can (1) immediately devise diagnostic tests based on DNA markers, (2) better
10 understand the disease by translating the DNA into its protein product and performing further experimentation, and (3) ultimately develop a cure, based on either the gene's DNA, its protein product, or resulting improved understanding of the disease mechanism.

15 To map a genome using IPM, then, one needs three reagents:

- *Points*. These can be STSSs, or fluorescence-based *in situ* hybridization (FISH) studies. Any measurement that can sample the chromosome space will suffice.
- 20 - *Lines*. These DNA probes can be large, such as YACs, or smaller sized, such as cosmids or plasmids.
- *Covering Regions*. Radiation hybrids (RHs) are ideally suited to the task. Each RH is a set of

-27-

large (randomly generated) chromosomal subregions that is preserved in a hamster cell culture clone. Other candidate reagents, such as pooled YACs, are discussed below.

- 5 Two sets of experiments are performed to obtain two data tables. These can be done on a single chromosome, or on an entire genome.
- 10 - Table A. The line probes (rows) are compared against the covering regions (columns). The extent of reaction, on a graded scale of -1 to +1 is recorded for each comparison. For parallelization, this can be done by obtaining each row or column in one experiment. For example, to obtain one column at a time, one can
- 15 compare all the rows (probes) with one column (radiation hybrid) at once. This is done by first gridding out the (YAC or cosmid) probe library's DNA onto filters, and then hybridizing with one
- 20 - Table B. The covering regions (rows) are compared against the sample points (columns). The extent of reaction, on a graded scale of -1 to +1 is recorded for each comparison. For parallelization, this can be done by obtaining
- 25 each row or column in one experiment. For example, to obtain one column at a time, one can compare all the rows (covering regions) with one column (STS) at once. This is done by first gridding out the RH library's DNA onto filters,
- 30 and then conducting a PCR experiment for one STS.

-28-

Suppose that radiation mapping was first performed on the STSs, and the point locations are therefore known. To determine the probe positions, the matrices A and B would be multiplied under a suitable inner product (say, the standard
5 additive sum over multiplicative product operation) and obtain the product matrix L. In each row of L, the location of the probe would be suggested by a (Gaussian-like) unimodal curve. Taking the average of the maxima of the curve would give an estimate of the peak's spatial position, i.e., the
10 probe's chromosome location. A more sophisticated approach would fit the observed curve neighborhood to a predicted curve shape, and use the fitting parameters to compute the peak location.

Suppose now that RH mapping were not done first, and
15 that there was absolutely no localization information about the sample points. The inner product of data Tables A and B would still be used to compute L. However, since the columns of L (i.e., the STS point ordering) would be (almost certainly) incorrectly permuted, L would not provide the
20 localization. Rather, an error function would be computed from L (as described above) that estimated the deviation from the expected unimodal curves in every row of L. Coupled with a global search procedure (such as simulated annealing), alternative permutations would be explored to minimize the
25 error function. When minimized, an approximately correct column permutation would be found, and both L and the STS ordering would be found. The distances between the STSs would be computed by local relaxation techniques (outlined above) that minimized an error function of predicted vs.
30 observed curve shapes, over all the probes. Then, with the point locations known, one would proceed as in the preceding paragraph to determine the probe locations.

-29-

The spatial resolution is set by the number of "bins" constructed by the intersections of covering regions. To fix ideas, consider the one dimensional case of radiation hybrids. Suppose that a RH, on average, contains five fragments and five gaps; this generates ten bins per RH. Combining k RHs will then generate, approximately, $10 \cdot k$ bins. Each such multi-RH bin will have a relatively unique signature vector of k "+/-" values, out the 2^k possible signature vectors. An increased number of fragments and gaps per RH will lead to more bins, hence greater resolution, per RH.

The sample points, e.g., the STSs used in chromosome mapping, serve to populate the bins with measurable reagents. The number of points sets the final resolution of the system. As described above, as additional RHs are incrementally deployed, the resolution of the system improves exponentially. Once every point is in its own unique bin, the maximum resolution is attained. Beyond this, additional RHs serve primarily to overcome noise in the system.

How does localization vary with probe volume? With genome mapping, probe volume corresponds to the length of the DNA insert of the YAC, cosmid, or other vector. The best resolution is achieved with point size probes, i.e., at zero volume. Larger probes act to locally smooth over a wider neighborhood of spatial bins. This might appear to worsen resolution as probe size increased. In fact, however, by fitting the expected broadened curve to the observed data, the true probe center can be localized with little or no loss of resolution.

-30-

This example of IPM for chromosome mapping assumes that the order (and distances) of the STS probes are known. If they are not known, then, as detailed above, one can use the IPM product, together with a global search mechanism, to
5 search for the correct ordering. (Specifically, with the correct ordering, the product matrix must have each matrix row take the form of a broad, single peak function. A global search such as simulated annealing can permute the matrix product columns to find the best STS ordering. Subsequent
10 curve fitting with local relaxation methods will then determine inter-point distances.)

This specific example was constructed using our IPM simulation, visualization, and analysis software. Suppose that the chromosome length is 100 megabases (MB). Fifty
15 random STSs are used, which sets an average inter-STS distance of 2.0 MB, i.e., the resolution of the map. Thirty-two radiation hybrids are used, with a fragment retention probability of 0.30, and an average inter-break distance of 5.0MB; this provides sufficient bins for our 50 sample
20 points. Twenty-five YAC insert probes are examined, with an average insert size of 0.100MB, uniformly distributed over a range of +/- 0.020MB.

To construct Table A, the YAC probes are compared for overlap against the RH covering regions. This may be done in
25 parallel one column at a time by first preparing gridded filters, each of which contains a dense array of YAC DNA. This DNA is obtained, for example, by amplification with interspersed repetitive sequence (IRS) (e.g., Alu or L-1) PCR. Then, each RH's DNA is similarly amplified by IRS-PCR,
30 and the labelled product is hybridized against the YAC filters. The hybridizations are scored as +/- comparisons,

-31-

and recorded. The result is the Table A of YACs vs. RHs, shown in Figure 1. Each row of the table provides a YAC with a unique RH signature vector.

To construct Table B, every RH is compared against every
5 sample point to determine whether or not the point is contained in (or overlaps with) some RH fragment. This can be done in parallel either one row or one column at a time. For example, one can test each point against all the RHs at once. With STS sample points, this is done by first
10 preparing gridded filters of RH DNA products; each filter contains the extracted and amplified DNA of all the RH clones, each RH located at its unique coordinate. PCR is then performed using labelled primers with one STS against all the RHs on the filter. The +/- comparisons are then
15 scored and recorded. The resulting Table B of the RHs vs. the STS is shown in Figure 2. Each column of Table B provides an RH signature vector for some STS sample point.

Note how each row of the table provides a sampling of one RH's fragment/gap pattern (contiguous +'s or -'s) along
20 the chromosome. Compare this with the true RH locations shown in Figure 3 below.

The Tables A and B are treated as matrices, and multiplied together under a suitable inner product. Here, we
25 use the sum over the product. The result, shown in Figure 4, is a matrix of nonnegative integers, comparing YAC rows against the STS columns.

Each row is scanned for its peak. For example, with the low noise of this example, it suffices to locate the maxima
30 of each row. This is recorded as a +, and presented in

-32-

matrix form, as shown in the table of Figure 5. In each row, the +'s indicate local matches of the YAC's signature against the STS column signatures. Note that the comparison of ordered YACs against ordered STSs shown does indeed result in
5 an approximate identity matrix, visually suggesting the success of the IPM method.

For more precise localization, these STS positions can be averaged together into a estimate of the YAC's center location. This is shown in Figure 6, which draws (1) the
10 true STS locations as vertical tick marks at the top and bottom, (2) the true YAC locations as a horizontal line in each row, and (3) the predicted YAC center position as a vertical tick mark shown on the YAC's row. As shown in every row, the predicted center positions (i.e., item 3) lie near
15 or on the true YAC locations (i.e., item 2).

How good is the result? Analysis of the simulation shows that, in reality, only one YAC was even hit by an STS. I.e., direct comparison methods such as STS-content mapping would have completely failed. However, IPM's match-based
20 indirect comparison approach that combines two independent data tables worked quite well. Specifically, the distance between the predicted and true YAC centers ranged from 0.03MB to 2.80MB, with an average distance of 0.93MB. This is the best that one can expect from a 2.0MB inter-STS sampling
25 distance.

The key components of a useful software system for IPM are:

- A *simulation* module that can generate test examples according to different experimental reagents and conditions.

-33-

- A *mapping* module that can solve problems using IPM and construct maps.
- An *analysis* module that can assess the utility of the solutions.
- 5 - A *visualization* module that can present results and analyses in a visually perspicuous way.

These have been developed and implemented. The current version runs in Common LISP on the Macintosh computer.

Several refinements and extensions to the basic IPM
10 method are now described. These increase the utility, applicability, and power of IPM.

IPM is based on intersections of regions. Positive and negative comparisons of probes against covering regions are equally informative. Therefore, a covering region is most
15 informative when the volume of its covering subregions is about 50% of the total volume. In this case, the volume of the complement of the covering subregions is also about 50%.

The more informative the covering regions, the fewer covering regions required. For example, with only 30%
20 coverage per covering region, the minimal multiplicative reduction by a region or its complement is $\text{MAX}\{p, q\}$, where $p=.30$, and $q=.70$. Here, that figure is .70. Thus, two covering regions, having a combined localization power of .49 ($= .70 \times .70$) would be needed to achieve about the same
25 localization power of a single 50% covering region. Fortunately, the localization is still exponential: only

-34-

twice as many (i.e., a constant factor) covering regions are needed.

This combination of covering regions suggests one mechanism for reducing experiments: pool covering regions
5 into fewer (combined) covering regions. With genome mapping, this can be done by pooling the DNA of RHs to effect 50% coverage.

Another pooling strategy enables the elimination of the third set of reagents (i.e., RHs) entirely. Rather than pool
10 covering regions, one can just pool the probes, thereby creating synthetic covering regions with optimal properties. In the case of genome mapping, this would allow the RHs to be entirely dispensed with. Instead, the DNA insert probes (e.g., YACs, cosmids, plasmids, BACs, etc.) would be pooled
15 to form synthetic RHs that cover 50% of the genome or chromosome. This has several advantages:

- (1) The need for a third reagent is eliminated. That is, only DNA insert probes and STSs are needed.
- (2) The number of experiments can be reduced by using
20 ideal (i.e., 50% coverage) covering regions.
- (3) The resolution of each covering region is increased, since the number of bins (fragments or gaps) is increased with synthetic covering regions.

25 By simulation, mathematical, statistical, neural network, or any other modelling technique, the expected shape of the localization distribution that appears in each row of

-35-

the product matrix can be determined. This has considerable utility:

- 5 (1) The probe center position can be estimated by fitting the observed data to the expected distribution, and then recording the spatial coordinates at the distribution's peak.
- 10 (2) The probe size can be estimated by similarly fitting the data to a family of expected distributions that are indexed by probe size. The best fit estimates the probe size.
- 15 (3) Suppose that the sample point locations are unknown. Then, a global search mechanism (e.g., simulated annealing) can use a minimization function over the probe vs. point IPM product to determine the point locations. The *a priori* expected distribution function of the probes can be used (e.g., as an L2 minimization heuristic) to better fit the point locations to the observed data.
- 20 (4) In the case of one dimensional spaces such as genomes or chromosomes, item 3 implies both the ordering of STSs, as well as determining the inter-STs distances.

Both false negative and false positives are possible, in
25 both Table A and Table B. Using more RHs will provide more unique signature vectors for the probes and points, and thus overcome much of the experimental noise. When coupled with

-36-

curve fitting, the background noise is essentially eliminated.

Another approach, discussed next, is to modify the inner product operation.

5 A variety of inner products may be used with IPM. The Boolean inner product (and over equivalence) only works with perfect data. Numerical inner products such as sum over product, where +/- comparisons are encoded as +1/-1, allow for more informative post-processing of the rows in the
10 product matrix.

When additional information is available, the inner products can be adapted to use it. For example, with some noisy experiments, the average false negative and false positive rates can be determined experimentally. Then
15 combinations of various observable results can be predicted, and written down as constants. Specifically,

- $a = \text{Prob}\{ \text{true} + \text{in A} \} \times \text{Prob}\{ \text{true} + \text{in B} \}$
- $b = \text{Prob}\{ \text{true} - \text{in A} \} \times \text{Prob}\{ \text{true} - \text{in B} \}$
- $c = \text{Prob}\{ \text{true} + \text{in A} \} \times \text{Prob}\{ \text{true} - \text{in B} \}$
- 20 - $d = \text{Prob}\{ \text{true} - \text{in A} \} \times \text{Prob}\{ \text{true} + \text{in B} \}$

The standard numerical inner product can be rewritten as the equivalent sum of four terms:

$$\text{SUM}(+,+) + \text{SUM}(-,-) - \text{SUM}(+,-) - \text{SUM}(-,+),$$

-37-

where the (r,s) indicates one of the four possible pairwise comparisons between a +/- value r in matrix A and a +/- value s in matrix B. Accounting for combinations of false negative and false positive rates in tables A and B amounts to
5 weighting this inner product by the constants a, b, c, and d:

$$a\text{SUM}(+,+) + b\text{SUM}(-,-) - c\text{SUM}(+,-) - d\text{SUM}(-,+)$$

The resulting sum is then a better predictor of true match.

Another inner product operation can be constructed to account for the a priori coverage at the covering regions.
10 Specifically, a probability or likelihood can be computed from the Table A or Table B data that adjusts each component inner product multiplicand by the odds at the observed result. For example, if p is the fraction of +'s for the Kth covering region and ++ match is observed, then the Kth multiplicand in
15 the inner product could be adjusted by a factor such as 1/p.

The experimental results are not always discrete all-or-none comparisons. Uncertainty and intermediate results are observed. Therefore, inner products that accommodate continuous values (say, between -1 and +1) are
20 useful. One such inner product is the sum of the products. This can be further adapted, say for false negative or false positive distributions, as shown above.

When the sample point locations are unknown, a global search mechanism (e.g., simulated annealing) can use a
25 minimization function over the probe vs. point IPM product to determine the point locations. The key idea is that, in each row, the shape of the product forms a broad Gaussian-like peak, and not a delta function. This broadening in the probe

-38-

vs. point product can be exploited to determine sample point positions.

When all the points are correctly localized, then the IPM product will localize every probe to its proper location.
5 That is, every probe will be characterized by a broad Gaussian-like distribution, with the center of the distribution located at the true center of the probe. I.e., the probe occupies some localized density in the space.

When the point locations are *not* correctly localized,
10 then, for each probe, this density will be scattered across the space. An error function (between observed and predicted) can compute the extent of this scattering. The error function can be heuristic (looking merely for densities), and/or highly specific (e.b., the L2 deviation
15 from a probe's predicted distribution function).

A global search mechanism (such as simulated annealing) can shuffle the sample points around the space. When the points move toward their correct position, the probe densities are less scattered, and the error function is
20 minimized. Iterating this process to minimize the error will unscatter an entire set of probe densities, and thus correctly localize (within some error) all the sample points.

Importantly, the more probes tested, the better the error function. That is, as the amount of data increases,
25 little extra computational effort is required, but the algorithm better converges to the true locations.

-39-

The current best method for Radiation Hybrid mapping is break minimization and its derivatives. The equivalent of Table B (RHs vs. STSs) is experimentally constructed, and then the STS columns are permuted to minimize the number of obligate breaks. That is, a global search (e.g., using simulated annealing) is done, where the error function is the number of transitions between + and - in Table B.

IPM provides a more principled mechanism that can reduce the number of required experiments. Specifically, obtain Table B. Then set Table A equal to the transpose of Table B. When the correct STS ordering (and distances) are obtained, each row of the product matrix should resemble a unimodal Gaussian-like function. Thus, using the method of previous subsection (on determining sample point position), RH mapping can be effected.

Advantages include:

(1) IPM knows the shape of the predicted row product distributions, where as the break minimization method has no such distribution information available to it. This can improve the accuracy of the STS ordering and inter-STS distances.

(2) The use of IPM may (exponentially) reduce the number of required RHs and RH experiments, thereby reducing expense.

Or, it can be for constructing ordered probe maps, wherein only Table A (the comparison of DNA sequence probes against radiation hybrids) is obtained. Table B is then set equal to the transpose of Table A. The product matrix L is

-40-

then a symmetric matrix which compares each probe with every other probe. Each entry in L is the inner product of the radiation hybrid signatures of two probes. The resulting map then localizes the probes, without the need for STS reagents
5 nor with comparisons of the radiation hybrids against STSs.

The objective for three-dimensional visualization of chromosomes is to effectively visualize many genomic probes against the entire genome in a single fluorescent microscopy experiment. The technology for visualizing DNA probes on
10 chromosomes is still in its infancy. The current best techniques use two dimensional (2D) light microscopy to measure multiple fluorescent DNA probes at precise locations along a chromosome. One key limitation is the resource intensive search for (rare) neatly arrayed metaphase
15 chromosomes. Restricting precision is the microscope's 2D projection of three dimensional (3D) chromosomes. Further, assigning one fluor to one chromosome (for twenty three chromosomes) is beyond the resolving capabilities of fluorescence detection. These problems can be solved by
20 using 3D microscopy and multiplexed pooling of fluorescent DNA probes, coupled with Inner Product Mapping (IPM), to visualize (common) interphase nuclei.

One chromosome can be readily visualized using a single fluor attached to multiple DNA markers along its length.
25 Ideally, then, all 23 human chromosomes could be visualized by using 23 distinct flours. With IPM's pooled multiplexing arrangement, though, only five (i.e., $\log_2(23)$) are needed; further, this smaller number (i.e., 5) is experimentally feasible. The idea is to use each fluor to label multiple
30 chromosomes in a single interphase nucleus, obtain five separate images, and then use IPM to combine images and their

-41-

intersections to mathematically visualize each individual chromosome.

The result is a 3D map of the interphase chromosomes. Using additional fluors, labeled DNA probes can then be
5 precisely mapped to chromosomal locations. For example, inter-Alu PCR of genomic RHs can be used to localize the RH fragments on the chromosomes. Similarly, DNA disease markers can be efficiently localized. An alternative allocation would be using just one fluor for one chromosome, with
10 multiple fluors to map an exponential number of IPM encoded DNA probes on that chromosome.

A procedure of many possible procedures is now given for constructing maps, identifying disease genes and using such disease genes.

15 Isolate and purify the DNA from a chromosome, whole genome, or portion thereof.

Form the DNA reagents from said genome for constructing map:

- 20 (1) DNA point reagents. These include STSs (e.g., microsatellite repeats used for genetic mapping), and other small DNA inserts.
- (2) DNA probe reagents. These are comprised of genomic inserts contained within cosmid, P1, or YAC vectors.

-42-

- (3) DNA covering region reagents. These include radiation hybrids, and pools of genomic clones, e.g., YACs.

Form Table A by comparing the DNA probe reagents against the DNA covering region reagents. This can be done by the steps of:

- Extracting the DNA from the probe reagents. For species-specific DNA amplification, interspersed repetitive sequence (IRS) PCR can be performed.
- 10 - Preparing gridded filters of the probe DNA reagents.
- Extracting labelled DNA from the covering region reagents. This can be done using labelled PCR primers. For species-specific DNA amplification, use IRS-PCR.
- 15
- Hybridizing the labelled DNA product of each covering region reagent against one or more gridded filters that contain the entire probe library. Cot1 DNA can be used to block repetitive sequences. This detects the intersection or overlap of the probes against the covering regions.
- 20
- Scoring the hybridization results, and recording them in a computer.

25 Alternatively, form Table A as above, but transpose the DNA reagents by instead putting the covering region DNA onto

-43-

gridded filters, labelling the probe DNA, and then conducting the hybridizations.

Form Table B by comparing the DNA covering region reagents against the DNA point reagents. This can be done by
5 the steps of:

- Preparing PCR primers corresponding to the STS point reagents.
- Extracting the DNA from the covering region reagents. IRS-PCR need not be used here.
- 10 - Preparing gridded filters of the covering region DNA reagents.
- Performing labelled-PCR with each STS point reagent's PCR primers against one or more gridded filters that contain the entire covering region
15 library. This detects the presence or absence of a point within a covering region.
- Scoring the PCR amplification results, and recording them in a computer.

Alternatively, form Table B as above, but transpose the
20 DNA reagents by instead testing each covering regions's DNA against an array of STSs using labelled STS primers in a multiplexed PCR reaction.

Forming Table L by combining Tables A and B using Inner Product Mapping.

-44-

Constructing a map of the DNA reagents. This is done by using Table L, and, possibly, iteratively using a global search mechanism (e.g., simulated annealing), together with search heuristics for permuting the rows and columns of Table L into the known shape of the desired solution. The final permutations determine the locations (i.e., both order and position, up to the desired map resolution) on the map of the DNA reagents.

Using said map, together with the well-known technique of positional cloning (also termed "reverse genetics"), to clone disease genes. This can be done in several ways. Each approach meets the three requirements for the cloning of a disease gene:

- (a) genetic inheritance in affected family,
- (b) functional expression in affected tissues, and
- (c) localization on a genetic/physical map.

Four illustrative (out of many possible) approaches are presented. In these (and virtually all possible) approaches, the use of said map is crucial to enablement.

Approach 1. Proceed directly from the constructed map.

- (1) Establish phenotypic criteria for affected individuals.
- (2) Identify and characterize families with affected individuals.

-45-

- 5 (3) Use said constructed map to perform genotyping and linkage analyses of said families to map the disease to a map location. Selection of the genetic markers may be random (their selection would be based on the optimal distribution along the genetic map) or based upon proximity to suspected DNA breakpoints (based upon cytogenetic cases), or related to known candidate genes (directly using the information of the physical map to select informative markers near specific genes).
- 10
- (4) Use said constructed map to look up the corresponding physical DNA sequences corresponding to the map location.
- 15 (5) Use standard laboratory-based techniques on said physical DNA sequences to:
- (5a) Find gene regions via promoter sequences. This can be done using rare restriction cutters to identify CpG-rich sequences.
- 20 (5b) Find gene regions via species conservation.
- (5c) Clone candidate genes from these selected sequences.
- (6) As an alternative to Step (5), sequence all said physical DNA sequences, and use standard computer-based techniques to:
- 25

-46-

- (6a) Find gene regions by searching the sequences containing promoter and/or coding regions.
- (6b) Find gene regions via species conservation. This can be done by comparing the sequences against sequence databases.
- (6c) Clone candidate genes, say by direct DNA synthesis of these selected sequences.
- (7) Express the candidate gene as messenger RNA (mRNA). Compare the expression of the mRNA (or its protein products) in appropriate tissues of affected and unaffected individuals to determine which candidate gene correlates with disease activity. The one which correlates is the disease gene.
- 15 Approach 2. Proceed directly from the constructed map, but also use candidate genes derived from affected tissues.
- (8) Proceed as in Steps 1-6.
- (9) Obtain candidate genes derived from appropriate tissues of affected and unaffected individuals. Said candidate genes can be constructed by building cDNAs from the expressed mRNA. Compare the expressed candidate genes with the map-derived candidate genes, and find the disease gene by confirming identity.
- 25 Approach 3. Proceed directly from candidate genes derived from affected tissues, but also use the constructed map.

-47-

- 5 (10) Construct candidate genes derived from appropriate tissues of affected and unaffected individuals. Said candidate genes can be constructed by building cDNAs from expressed mRNA in said tissues.
- 10 (11) Compare said expressed candidate genes with the physical DNA reagents (or their derivatives) that comprise the physical sequence component of said constructed genetic/physical map. This will determine the map location of said candidate gene on the map.
- 15 (12) Select genetic markers directly from the map. Alternatively, new markers can be constructed from the physical sequences or from the candidate gene DNA. A set of highly polymorphic genetic markers works best; since this is most likely not present on the candidate gene, use of the map is required to select or construct said polymorphic markers.
- 20 (13) Establish criteria for phenotypic expression in affected individuals.
- (14) Identify families having affected individuals who express the disease.
- 25 (15) Use the said constructed genetic/physical map to verify that the phenotypic disease, and the candidate gene markers (derived using the map), both map to the same location. This confirms that the candidate gene is, in fact, the disease gene.

-48-

Approach 4. A combined positional and candidate gene cloning strategy using the constructed map.

(16) Proceed as in Steps 1-3.

5 (17) Using the map, localize the disease-related gene
between two flanking markers, or demonstrate
linkage with one or more candidate genes. (Direct
testing of a candidate gene, independently of
linkage, would not require the use of the physical
map. This situation would be the exception rather
10 than the rule, since it is unusual for a specific
candidate gene to be clearly established for a
given disease.)

(18) If a specific candidate gene is identified by
linkage analysis then one would proceed to step
15 19. If a specific candidate gene has not been
strongly linked, or one or more of the initial
candidate genes have been excluded by the criteria
stated below, then one must identify additional
candidate genes within the flanking genetic
20 markers.

(18a) Potential candidate genes include those that
have
been previously localized to the region on the said
physical map.

25 (18b) Additional candidate genes, which have not been
previously identified and mapped would be
identified
by using the genomic contigs on the physical map

-49-

5 corresponding to the region within the flanking
markers to identify other potential genes either by
sequence analysis (i.e., algorithms designed to
identify regions of genomic DNA that are likely to
10 encode genes or transcribed regions), or by
experimental criteria including but not limited to
CpG islands, exon trapping, cross-species
conservation, and hybridization to specific clones
from a suitable cDNA library. The experimental
15 screening of the genomic region for candidate genes
would generally include the use of expressed
sequences from tissues that are thought to be
affected by the disease.

15 (19) Specific candidate genes would be tested by
mutation analysis, including but not limited to,
alteration of genomic structure by insertion,
deletion, or rearrangement, alteration of
transcript size or abundance, loss of protein
product or enzymatic function, point mutation
20 detection by indirect (SSCP, DGGE, chemical
cleavage, etc.) or direct methods (DNA sequencing,
dideoxy-fingerprinting, ASO, etc.).

25 (20) Confirmation of the candidate gene by linkage
analysis of the suspected mutation, and
demonstration that the mutation is not present in
the normal population. Additional functional
studies may be employed to demonstrate that the
particular mutation alters the normal function or
structure of the gene.

-50-

- 5 (21) Perpetuation. The identified disease gene can be perpetuated by inserting the DNA sequence of the disease gene into a vector clone. Suitable vector clones are readily available either by purchase from any of a number of vendors, or by direct construction using well-known recombinant DNA techniques. With *E. coli*, for example, the insert DNA is maintained as a plasmid.
- 10 (22) Amplification. Following Step 21, the disease gene can be amplified by permitting the vector (e.g., *E. coli*) to grow to the desired quantity.
- 15 (23) Purification. There are dozens of established techniques for this step. With *E. coli* vectors, one approach would be to first disrupt the cell integrity, and the precipitate out the DNA component with salts. The plasmid DNA is then isolated by ultracentrifugation with a cesium chloride gradient, and then retrieving the plasmids from the appropriate band in the layered
- 20 centrifuged product. Further plasmid-specific isolation may also be performed.
- 25 (25) Protein products. Expression systems are used to produce recombinant proteins. The most popular expression systems are the bacteria *E. coli*, and *Bacillus subtilis*, yeast, and cultured insect and mammalian cells. These exploit the transcriptional (DNA to mRNA) and translational (mRNA to protein) machinery of the intact cell to manufacture protein products. These protein
- 30 products are then isolated and purified from this

-51-

amplification procedure using well-known techniques, including ionic, column, electrophoretic, and centrifuge separations.

5 (24) DNA-based diagnostic tests. DNA-based diagnostic tests can be derived directly from the sequence of the isolated disease gene.

(24a) ASOs are developed from mutations in the exon region of the gene, which can then be detected using
10 hybridization, SSCP, LCR, RFLP, or any other appropriate mutation assay.

(24b) Linkage markers are developed from polymorphic tandem repeat regions in the intron region of the gene, which can then be detected using differential
15 sizing, hybridization, SSCP, RFLP, or any other appropriate mutation assay.

(24c) Direct sequencing of the DNA will identify allelic variations.

(24c) To use the diagnostic tests described in Steps
20 24a, 24b, and 24c, the disease gene region is amplified from an individual's blood using PCR (with
labelled primers). The DNA-based diagnostic test
is
25 then applied to this DNA product to determine the individual's alleles. With linkage analysis, additional testing of family members may be required.

-52-

(25) Protein-based diagnostic tests. The protein product produced from the gene's DNA can also form the basis of a diagnostic test. This is done by sampling an individual's blood, and assaying for protein mutations. There are very many such specific assays, including immunodetection, specific binding assays, functional assays, and electrophoretic migration assays.

More generally, the inventive procedure involves:

- 10 - Identifying a physical space which is to be mapped.
- Setting a desired resolution of the map.
- Isolating a sufficiently large set of sample
15 points to achieve this desired resolution, each
 component sample point representative of a point
 which physically resides within the space and has
 a corresponding mass localized to a first volume
 within a predetermined range of the space.
- 20 - Isolating a set of probes, probe representative of
 a probe which physically resides within the space
 and has a corresponding mass localized within a
 second volume of a predetermined range, said
 predetermined range of said second volume is
25 greater than or equal to said predetermined range
 of said first volume.
- Isolating a sufficiently large set of covering
 regions to achieve the desired resolution, each

-53-

5 component covering region representative of a
region which is comprised of a set of subregions
each of which has an associated mass; the mass of
each subregion extends over a connected volume
that physically resides within the space and has a
corresponding mass localized within a third volume
of a predetermined range, said predetermined range
of said third volume is greater than or equal to
said predetermined range of said second volume.
10 The total volume of said covering region [which
does not exceed the total volume of its
subregions] has a corresponding mass localized
within a fourth volume of a predetermined range;
said predetermined range of said fourth volume is
15 sufficiently large to be within several orders of
magnitude of the volume of the entire space.

In constructing the map, the three reagents are
described by their relative sizes. Specifically, the size
range (i.e., lengths) of the DNA sequences of the points,
20 lines, and subregions of the covering regions must generally
maintain the following relationship:

$$\text{range}(\text{point sequence}) < \text{range}(\text{line sequence}) < \text{range}(\text{subregion sequence}).$$

For example, with a map constructed from STS points, YAC
25 lines, and RH covering regions, one satisfies this inequality
since:

$$\begin{aligned} \text{range}(\text{STS sequence}) &= [100\text{B}, 500\text{B}], \\ \text{range}(\text{YAC sequence}) &= [100\text{KB}, 2000\text{KB}], \\ \text{range}(\text{RH fragment sequence}) &= [5\text{MB}, 50\text{MB}], \end{aligned}$$

-54-

where B denotes DNA base units, KB denotes kilobases, and MB denotes megabases. The IPM will also tolerate some deviation from this strict inequality relation.

The following measurements are physically realizable
5 within this method:

- 10 - Direct comparison measurement of containment or overlap of a probe against a covering region. A positive comparison indicates containment or overlap, a negative comparison indicates no containment or overlap, and an intermediate value indicates partial containment or overlap. The method accommodates uncertain comparison results, as well as false negative or false positive results.
- 15 - Direct comparison measurement of containment or overlap of a covering region against a sample point. A positive comparison indicates containment or overlap, a negative comparison indicates no containment or overlap, and an
20 intermediate value indicates partial containment or overlap. The method accommodates uncertain comparison results, as well as false negative or false positive results.

Using the physically realizable measurement procedures,
25 two separate data tables are then acquired. For efficiency, this acquisition is best done in parallel manner, obtaining one row or column (or portions thereof) at a time. Specifically:

-55-

- Table A which determines the containment or overlap of the probes against the covering regions.
- Table B is obtained, which determines the containment or overlap of the sample points against the covering regions.

A combination to obtain Table L from Tables A and B, which provides localizing information about the probes relative to the sample points. One mechanism for determining Table L is by combining Tables A and B using any form of inner product combination. Typical inner product operations might include Boolean, arithmetic, or thresholding procedures.

A localization of the volumes of the sample points within the space. When the location or ordering of the sample points is not known, computation of Table L, together with a global search procedure such as simulated annealing, is used to localize the sample points. This is done by repeatedly permuting the columns of Table L (equivalently, the columns of Table B) to best cluster the densities in every row of Table L. Permutations may additionally be performed upon the rows of Table L (equivalently, the rows of Table A) to cluster the densities along the diagonal of Table L. Clustering measures on Table L may include operations for finding, grouping, and separating relatively larger or smaller values, preserving continuity, and fitting to predetermined functional forms.

A postprocessing of Table L to construct a map. This postprocessing step may be effected by any manner of curve

-56-

fitting mechanisms, or the use of curve statistics derived therefrom. (Typical statistics include the location of relatively larger or smaller values, and estimates of variance.) Specifically, within said desired resolution, the
5 map construction constructs a map which determines:

- the location of the sample point volumes within the space,
- the location of the probe volumes within the space, relative to said sample point volumes,
- 10 - and, if so desired, the location of the covering region volumes within the space, relative to said sample point volumes.
- All probes that are physically obtained using, at least in part, said map or localization process.
- 15 - The purified derivatives of said physically obtained probes.
- All useful applications of said physically obtained probes or their purified derivatives.

The probes can be clones containing nucleic acid inserts
20 of any size, the sample points are small nucleic acid sequences such as STSSs, and the covering regions are radiation hybrids or pooled nucleic acid fragments. Further, all these reagents are derived from the same chromosome or genome.

-57-

The probes can be clones containing nucleic acid inserts of any size, the covering regions are radiation hybrids or pooled nucleic acid fragments, and the rows of Table B are obtained by fluorescent *in situ* hybridization of the
5 radiation hybrids' nucleic acid content against the genome. That is, the sample points are locations along the fluorescently derived radiation hybrid characterizations. Further, all these reagents are derived from the same chromosome or genome.

10 Alternatively, only Table B (the comparison of radiation hybrids against STSs) is obtained. Table A is then set equal to the transpose of Table B, i.e., the probes are themselves the STSs. The resulting map then localizes both the STSs and the radiation hybrid fragments.

15 Or, it can be for constructing ordered probe maps, wherein only Table A (the comparison of DNA sequence probes against radiation hybrids) is obtained. Table B is then set equal to the transpose of Table A. The product matrix L is then a symmetric matrix which compares each probe with every
20 other probe. Each entry in L is the inner product of the radiation hybrid signatures of two probes. The resulting map then localizes the probes, without the need for STS reagents nor with comparisons of the radiation hybrids against STSs.

In another embodiment, the probes can be phenotypically
25 observable traits or genes, the sample points are a dense (possibly ordered) array of genetic markers (such as STSs), and the covering regions are comprised of each individual's recombinant chromosomes, with the subregions thereof given by the chromosomal portions inherited from the founder(s)
30 possessing the disease gene. Then, using identity-by-descent

-58-

or identity-by-state allele comparisons, the IPM-based superposition of the genetic marker alleles, with backprojection onto the alleles of the founder(s) possessing the disease chromosomes, will localize (i.e., map) the traits
5 or genes.

In yet another embodiment, the present invention can be for constructing two or three dimensional maps of chromosomes, nucleic acid reagents (e.g., radiation hybrids) visualized on chromosomes, or other cellular organelles using
10 fluorescent light microscopy. Each covering region corresponds to a fluorescent probe. Table A is the combinatorial encoding of each chromosome in terms of the fluorescent probes that represent it, and each row of Table B is a two or three dimensional microscopic image of the cell
15 taken through a filter that selectively highlights one fluorescent probe. The inner product L of Tables A and B, taken across the common covering region dimension, extracts the spatial position of each chromosome.

Preferably, this method can be for constructing
20 radiation hybrid maps, wherein only Table B (the comparison of hybrids against STSs) is obtained. Table A is then set equal to the transpose of Table B, i.e., the probes are themselves the STSs. The resulting map then localizes both the STSs and the radiation hybrid fragments. Or, it can be
25 for constructing ordered probe maps, wherein only Table A (the comparison of DNA sequence probes against radiation hybrids) is obtained. Table B is then set equal to the transpose of Table A. The product matrix L is then a symmetric matrix which compares each probe with every other
30 probe. Each entry in L is the inner product of the radiation hybrid signatures of two probes. The resulting map then

-59-

localizes the probes, without the need for STS reagents nor with comparisons of the radiation hybrids against STSs.

An apparatus for constructing maps comprises:

A computer, a memory, and data transfer devices.

- 5 A space to be mapped needs a desired resolution to be set out and useful physical reagents. Specifically:
- A physical space which is to be mapped is identified.
 - A predetermined resolution of the resulting map.
 - 10 - A sufficiently large set of sample points to achieve this desired resolution, wherein each component sample point physically resides within the space and has a corresponding mass localized to a first volume within a predetermined range.
 - 15 - A set of probes, wherein each component probe physically resides within the space and has a corresponding mass localized within a second volume of a predetermined range. Specifically,
20 said predetermined range of said second volume is greater than or equal to said predetermined range of said first volume.
 - A sufficiently large set of covering regions to achieve the desired resolution, wherein each component covering region is comprised of a set of
25 subregions. The mass of each subregion extends

-60-

5 over a connected volume that physically resides
within the space and has a corresponding mass
localized within a third volume of a predetermined
range. Specifically, said predetermined range of
said third volume is greater than or equal to said
predetermined range of said second volume. The
total volume of said covering region (which does
not exceed the total volume of its subregions) has
a corresponding mass localized within a fourth
10 volume of a predetermined range; said
predetermined range of said fourth volume is
sufficiently large to be within several orders of
magnitude of the volume of the entire space.

Physically realizable measurements, specifically:

- 15 - Direct comparison measurement of containment or
overlap of a probe against a covering region. A
positive comparison indicates containment or
overlap, a negative comparison indicates no
containment or overlap, and an intermediate value
20 indicates partial containment or overlap. The
method accommodates uncertain comparison results,
as well as false negative or false positive
results.
- 25 - Direct comparison measurement of containment or
overlap of a covering region against a sample
point. A positive comparison indicates
containment or overlap, a negative comparison
indicates no containment or overlap, and an
intermediate value indicates partial containment
30 or overlap. The method accommodates uncertain

-61-

comparison results, as well as false negative or false positive results.

Data buffers which may contain three data tables, A, B, and L. Specifically:

- 5 - Table A, which contains the containment or overlap of the probes against the covering regions,
- Table B, which contains the containment or overlap of the sample points against the covering regions, and
- 10 - Table L, which is a combination of Tables A and B.

A first program resident in the memory component, which can compute a combination of Tables A and B to produce a localization Table L. The particular combination operation may be comprised of an inner product combination, which can
15 entail Boolean, arithmetic, and/or thresholding procedures. Further processing of the rows of Table L may include curve fitting mechanisms for determining curve statistics, such as the location of maxima or minima.

A second program resident in the memory component, which
20 can compute a localization of the volumes of the sample points within the space. Specifically, when the location or ordering of the sample points is not known, said second program can localize the sample points by using said program 1, together with a global search procedure such as simulated
25 annealing, and thereby localize the sample points. This is done by repeatedly permuting the columns of Table L

-62-

(equivalently, the columns of Table B) to best cluster the densities in every row of Table L. Permutations may additionally be performed upon the rows of Table L (equivalently, the rows of Table A) to cluster the densities
5 along the diagonal of Table L. Clustering measures on Table L may include operations for finding, grouping, and separating relatively larger or smaller values, preserving continuity, and fitting to predetermined functional forms.

A postprocessing computation of Table L to construct a
10 map. This postprocessing step may be effected by any manner of curve fitting mechanisms, or the use of curve statistics derived therefrom. (Typical statistics include the location of relatively larger or smaller values, and estimates of variance.) Specifically, within said desired resolution, the
15 map construction determines:

- the location of the sample point volumes within the space,
- the location of the probe volumes within the space, relative to said sample point volumes,
- 20 - and, if so desired, the location of the covering region volumes within the space, relative to said sample point volumes.

An optional output data buffer for containing the map formed by said postprocessing computation of Table L.

25 From the map, probes can be obtained.

-63-

The means of effecting the said combination of Tables A and B can be by means of an inner product operation. Said inner product operation may include matrix multiplication, graph processing, set manipulation, list processing, vector
5 processing, or any other formulation or data representation that is mathematically equivalent to performing any step of inner product operation.

Preferably, any of the matrices A, B, or L can be determined, provided that the other two are known. This
10 holds, for example, when the inner product operation is the usual matrix multiplication "sum over product", and arithmetic values (say, -1 to +1) are used in the matrices. The determination of matrices A or B is effected by standard or least squares matrix inversion. This may be useful, for
15 example, in ascertaining probe or sample point signatures.

Such an apparatus can also include:

- A *simulation* module that can generate IPM test examples according to different experimental reagents and conditions.
- 20 - A *mapping* module that can solve problems using IPM and construct maps.
- An *analysis* module that can assess the utility of IPM's solutions.
- 25 - A *visualization* module that can present IPM's results and analyses in a visually perspicuous, and interactive, way.

-64-

- Pattern recognition or pattern matching.
- Machine learning application.
- Clustering operations in one or more dimensions.
- 5 - Neural or connectionist network learning procedure.
- Neural or connectionist network operation.
- Fuzzy logic application.

A system for constructing maps would utilize the following:

10

- A physical space which is to be mapped.
- A predetermined resolution of the resulting map.
- A sufficiently large set of sample points to achieve this desired resolution, wherein each component sample point physically resides within the space and has a corresponding mass localized to a first volume within a predetermined range.
- 15
- A set of probes, wherein each component probe physically resides within the space and has a corresponding mass localized within a second volume of a predetermined range. Specifically, said predetermined range of said second volume is greater than or equal to said predetermined range of said first volume.
- 20

-65-

- A sufficiently large set of covering regions to achieve the desired resolution, wherein each component covering region is comprised of a set of subregions. The mass of each subregion extends over a connected volume that physically resides within the space and has a corresponding mass localized within a third volume of a predetermined range. Specifically, said predetermined range of said third volume is greater than or equal to said predetermined range of said second volume. The total volume of said covering region (which does not exceed the total volume of its subregions) has a corresponding mass localized within a fourth volume of a predetermined range; said predetermined range of said fourth volume is sufficiently large to be within several orders of magnitude of the volume of the entire space.

Physically realizable measurements, specifically:

- Direct comparison measurement of containment or overlap of a probe against a covering region. A positive comparison indicates containment or overlap, a negative comparison indicates no containment or overlap, and an intermediate value indicates partial containment or overlap. The method accommodates uncertain comparison results, as well as false negative or false positive results.
- Direct comparison measurement of containment or overlap of a covering region against a sample point. A positive comparison indicates

-66-

containment or overlap, a negative comparison indicates no containment or overlap, and an intermediate value indicates partial containment or overlap. The method accommodates uncertain
5 comparison results, as well as false negative or false positive results.

A parallel experimentation approach for efficiently acquiring two data tables. Specifically:

- 10 - Table A, which contains the containment or overlap of the probes against the covering regions, and
- Table B, which contains the containment or overlap of the sample points against the covering regions.

An apparatus comprised of a computer, a memory, and data transfer devices that can combine Tables A and B to construct
15 a map. This apparatus has input data buffers for Tables A and B, as well as an intermediate buffer for Table L. It contains one program resident in memory which forms the inner product Table L by combining Tables A and B. It may have a second program resident in memory which can use Table
20 L, together with a global search procedure such as simulated annealing, to correctly localize the volumes of the sample points. Additional post-processing programs may be present in memory to effect the final probe localization, which constructs the map. The output of this apparatus is a map
25 that integrates the locations of the volumes of the sample points, probes, and, possibly, the covering regions.

An integrated map of any resolution, comprised of the volumes of the sample points, the probes, and, possibly, the

-67-

covering regions, which can be used to localize probes and sample points on the same map.

The physical probes obtained by retrieving a small subset of probes based on their location in the original or
5 derived map. This map location may be based on proximity to one or more sample point locations.

The further purification or characterization of the physical construction of this small subset of physical probes.

10 All useful products and applications that can be derived from said physically characterized probes or their purified derivatives.

The probes can be clones containing nucleic acid inserts of any size, the sample points are small nucleic acid
15 sequences such as STSs, and the covering regions are radiation hybrids or pooled nucleic acid fragments. Further, all these reagents are derived from the same chromosome or genome. Specifically, the system comprises:

A parallel experimentation approach for efficiently
20 acquiring two data tables of the requisite genomic reagents.

An apparatus comprised of a computer, a memory, and data transfer devices that can perform the IPM operation for genome mapping.

The resulting integrated genome map of any resolution,
25 comprised of the nucleic acid sample points, nucleic acid

-68-

probes, and nucleic acid covering regions, which can be used to localize probes and sample points on the same map.

The physical probes (e.g., YACs, cosmids, phagemids, BACs, or STSs) obtained by retrieving a small subset of probes based on their location in the original or derived map. This map location may be based on proximity to one or more sample point (e.g., STS) locations, especially STS markers that are selected because of their association with a particular genetic trait or disease. The genetic association may be determined by genetic mapping studies in families that integrate the STS's genetic recombination distance with their physical distances on the integrated genome map.

The further characterization of the physical construction of this small subset of physical probes, such as the determination of their nucleic acid sequences.

All genes that are found to reside within these nucleic acid sequences. This specifically includes the gene (or genes) that causes the particular trait or disease.

All useful purified gene products, and applications of those purified gene products, that can be derived from said genes. Such products and applications include:

- Nucleic acid markers obtained from within (or near) the gene, used, for example, as diagnostic tests for the particular trait or disease.

-69-

- The protein (and its derivatives) that the gene codes for, used, for example, to better understand the mechanism of the particular trait or disease.
- 5 - Diagnostic and therapeutic mechanisms derived from either the gene (e.g., gene therapy), its protein (e.g., immunological diagnostics, or protein-derived pharmaceutical therapeutics), or the improved understanding of the mechanism of the particular trait or disease (e.g., indirect
10 pharmaceutical agents).

Any form of such map construction can be used for:

- (a) ordering probe reagents via ordered sample point reagents,
- 15 (b) ordering sample point reagents via ordered probe reagents,
- (c) simultaneously ordering both sample point and probe reagents,
- (d) determining distances along said constructed maps,
- 20 (e) constructing a map of sample point or probe reagents.

Covering regions can be characterized for the purpose of using Table B to construct maps by any of these reagents:

-70-

- (a) Ordered sample points, obtained by any sample point ordering process based on covering regions.
- (b) Fluorescent visualization of the covering regions.
- (c) Unordered sample points, which are ordered by said
5 map construction system in any manner.

A manufacturing process for constructing maps would utilize the following:

- A physical space which is to be mapped.
- A predetermined resolution of the resulting map.
- 10 - A sufficiently large set of sample points to achieve this desired resolution, wherein each component sample point physically resides within the space and has a corresponding mass localized to a first volume within a predetermined range.
- 15 - A set of probes, wherein each component probe physically resides within the space and has a corresponding mass localized within a second volume of a predetermined range. Specifically,
20 said predetermined range of said second volume is greater than or equal to said predetermined range of said first volume.
- A sufficiently large set of covering regions to achieve the desired resolution, wherein each component covering region is comprised of a set of
25 subregions. The mass of each subregion extends

-71-

5 over a connected volume that physically resides within the space and has a corresponding mass localized within a third volume of a predetermined range. Specifically, said predetermined range of said third volume is greater than or equal to said predetermined range of said second volume. The total volume of said covering region (which does not exceed the total volume of its subregions) has a corresponding mass localized within a fourth volume of a predetermined range; said predetermined range of said fourth volume is sufficiently large to be within several orders of magnitude of the volume of the entire space.

Physically realizable measurements, specifically:

- 15 - Direct comparison measurement of containment or overlap of a probe against a covering region. A positive comparison indicates containment or overlap, a negative comparison indicates no containment or overlap, and an intermediate value indicates partial containment or overlap. The method accommodates uncertain comparison results, as well as false negative or false positive results.
- 20
- 25 - Direct comparison measurement of containment or overlap of a covering region against a sample point. A positive comparison indicates containment or overlap, a negative comparison indicates no containment or overlap, and an intermediate value indicates partial containment or overlap. The method accommodates uncertain
- 30

-72-

comparison results, as well as false negative or false positive results.

A highly parallel experimentation approach for very efficiently acquiring two data tables. Specifically:

- 5 - Table A contains the containment or overlap of the probes against the covering regions,
- Table B contains the containment or overlap of the sample points against the covering regions.

Entire rows or columns (or portions thereof) of Table A
10 can be determined in a single experiment. Entire rows or columns (or portions thereof) of Table B can be determined in a single experiment. Further, under suitable conditions, the number of covering regions is proportional to just the logarithm of the number of sample points. That is, the
15 manufacturing process is designed to greatly reduce the acquisition cost of the data required for constructing a useful map.

An apparatus comprised of a computer, a memory, and data transfer devices that can combine Tables A and B to construct
20 a map. This apparatus has input data buffers for Tables A and B, as well as an intermediate buffer for Table L. It contains one program resident in memory which forms the inner product Table L by combining Tables A and B. It may have a second program resident in memory which can use Table L,
25 together with a global search procedure such as simulated annealing, to correctly localize the volumes of the sample points. Additional post-processing programs may be present in memory to effect the final probe localization, which

-73-

constructs the map. The output of this apparatus is a map that integrates the locations of the volumes of the sample points, probes, and, possibly, the covering regions.

Applying the apparatus to the data tables constructs an
5 integrated map of any resolution, comprised of the volumes of the sample points, probes, and covering regions, which can be used to localize probes and sample points on the same map.

The physical probes obtained by retrieving a small
subset of probes based on their location in the original or
10 derived map. This map location may be based on proximity to one or more sample point locations.

The further purification or characterization of the physical construction of this small subset of physical probes.

15 All useful products and applications that can be derived from said physically characterized probes or their purified derivatives.

The probes can be clones containing nucleic acid inserts of any size, the sample points are small nucleic acid
20 sequences such as STSs, and the covering regions are radiation hybrids or pooled nucleic acid fragments. Further, all these reagents are derived from the same chromosome or genome. Genetic mapping studies in families can integrate the STS's genetic recombination distance with their physical
25 distances on the integrated genome map. Specifically, the output of the manufacturing process is an integrated genetic/physical map comprised of genetically useful markers

-74-

(e.g., STSs) and of reagents (e.g., cosmid) or indices to (e.g., YACs) sequence-ready nucleic acid insert clones.

The manufacturing process can be for identifying probes of interest that are located near sample points of interest
5 comprising:

An integrated map that localizes both probes and sample points.

Using this integrated map to spatially identify localized probes by their neighboring sample points. The
10 physical probes are obtained by retrieving a small subset of probes based on their location in the original or derived map. This map location may be based on proximity to one or more sample point locations.

Further characterizing the physical construction of each
15 of these physical probes.

Developing useful products and applications that are derived from these characterized probes and their derivatives.

Preferably, the manufacturing process can be for
20 identifying nucleic acid probes of interest, the genes residing on these probes, and their useful applications, that are located near nucleic acid sample points of interest comprising:

The integrated genetic/physical map that localizes both
25 probes and sample points.

-75-

Physical probes (e.g., YACs, cosmids, phagemids, BACs, or STSs) are obtained by retrieving one or more probes based on their map location. This map location may be based on proximity to one or more sample point (e.g., STS) locations,
5 selected as genetic markers that are associated with a particular genetic trait or disease.

The further characterization of these physical probes, effected, in part, by determining their nucleic acid sequences.

10 Genes that are found within these nucleic acid sequences by scanning for open coding regions. Such genes specifically includes the gene (or genes) that causes the particular trait or disease.

Purified gene products, and useful applications of those
15 purified gene products, can then be derived from said genes. Such products and applications that are obtainable from this manufacturing process include:

- Nucleic acid markers obtained from within (or near) the gene, used, for example, as diagnostic
20 tests for the particular trait or disease.
- The protein (and its derivatives) that the gene codes for, used, for example, to better understand the mechanism of the particular trait or disease.
- Diagnostic and therapeutic mechanisms derived from
25 either the gene (e.g., gene therapy), its protein (e.g., immunological diagnostics, or protein-derived pharmaceutical therapeutics), or the

-76-

improved understanding of the mechanism of the particular trait or disease (e.g., indirect pharmaceutical agents).

Although the invention has been described in detail in
5 the foregoing embodiments for the purpose of illustration, it
is to be understood that such detail is solely for that
purpose and that variations can be made therein by those
skilled in the art without departing from the spirit and
scope of the invention except as it may be described by the
10 following claims.

-77-

WHAT IS CLAIMED IS:

1. A method for cloning a gene of a genome comprising the steps of:

identifying a portion of a genome to be mapped;

setting a desired resolution of the map;

isolating a sufficiently large set of first DNA sequences having a size between 100b-40kb to achieve this desired resolution, each first DNA sequence part of the genome;

isolating a sufficiently large set of second DNA sequences having a size between 40kb-2mb to achieve the desired resolution, each second DNA sequence part of the genome;

isolating a sufficiently large set of collection of DNA sequence fragments wherein each fragment's average size is greater than or equal to the size of each of said second DNA sequences to achieve the desired resolution, each collection of DNA sequence fragments part of the genome;

forming a Table A in a memory of a computer, Table A determining the containment or overlap of the set of second DNA sequences against the collection of DNA sequence fragments;

forming a Table B in the memory, Table B determining the containment or overlap of the set of first DNA sequences against the collection of DNA sequence fragments;

-78-

operating a processor of the computer on Table A and Table B to produce a Table L which provides localization information about the set of second DNA sequences relative to the set of first DNA sequences;

operating the processor on Table L to produce a map of the gene which determines where the set of first DNA sequences are located within the gene and where the set of second DNA sequences are located within the gene; and

producing the gene from the map.

2. A method as described in Claim 1 including after the producing step, there is the step of purifying derivatives of the gene produced.

3. A method as described in Claim 2 wherein the first DNA sequences are STSs, the second DNA sequences are YACs, and the DNA sequence fragments are radiation hybrids.

4. A system for cloning a gene of a genome comprising:

means for isolating a sufficiently large set of first DNA sequences having a size between 100b-40kb to achieve a desired resolution, each first DNA sequence part of the genome;

means for isolating a sufficiently large set of second DNA sequences having a size between 40kb-2mb to achieve the desired resolution, each second DNA sequence part of the genome;

-79-

means for isolating a sufficiently large set of collection of DNA sequence fragments wherein each fragment's average size is greater than or equal to the size of said second DNA sequence to achieve the desired resolution, each collection of DNA sequence part of the genome;

a computer having a memory and means for inputting information into the computer about the set of second DNA sequences, the set of first DNA sequences and the set of collection of DNA sequence fragments;

means for forming a Table A in the memory, Table A determining the containment or overlap of the set of second DNA sequences against the collection of DNA sequence fragments;

means for forming a Table B in the memory, Table B determining the containment or overlap of the set of first DNA sequences against the collection of DNA sequence fragments;

means for forming a Table L in the memory from Table A and Table B, Table L providing localization information about the set of second DNA sequences relative to the set of first DNA sequences;

means for producing a map of the gene from Table L, said map producing means disposed in said memory said map determines where the set of first DNA sequences are located within the genome and where the set of second DNA sequences are located within the genome; and

means for producing a gene of the genome from the map.

-80-

5. A system as described in Claim 4 including means for purifying derivatives of the gene produced from the gene producing means.

6. A system as described in Claim 5 wherein the first DNA sequences are STSs, the set of second DNA sequences are YACs, and the DNA sequence fragments are radiation hybrids.

7. A method of producing a gene comprising the steps of:

isolating a gene from a map constructed by using radiation hybrids versus probes; and

producing the gene isolated from the map.

8. A system for producing a map of a genome comprising:

means for isolating a sufficiently large set of first DNA sequences having a size between 100b-40kb to achieve a desired resolution, each first DNA sequence part of the genome;

means for isolating a sufficiently large set of second DNA sequences having a size between 40kb-2mb to achieve the desired resolution, each second DNA sequence part of the genome;

means for isolating a sufficiently large set of collection of DNA sequence fragments wherein each fragment's average size is greater than or equal to the size of said second DNA sequence to achieve the desired resolution, each collection of DNA sequence part of the genome;

-81-

a computer having a memory and means for inputting information into the computer about the set of second DNA sequences, the set of first DNA sequences and the set of collection of DNA sequence fragments;

means for forming a Table A in the memory, Table A determining the containment or overlap of the set of second DNA sequences against the collection of DNA sequence fragments;

means for forming a Table B in the memory, Table B determining the containment or overlap of the set of first DNA sequences against the collection of DNA sequence fragments;

means for forming a Table L in the memory from Table A and Table B, Table L providing localization information about the set of second DNA sequences relative to the set of first DNA sequences; and

means for producing a map of the gene from Table L, said map producing means disposed in said memory said map determines where the set of first DNA sequences are located within the genome and where the set of second DNA sequences are located within the genome.

9. A system for producing a map of an ordered clone map of a genome comprising:

means for isolating a sufficiently large set of first DNA sequences having a size between 10kb-2mb to achieve the desired resolution, each second DNA sequence part of the genome;

-82-

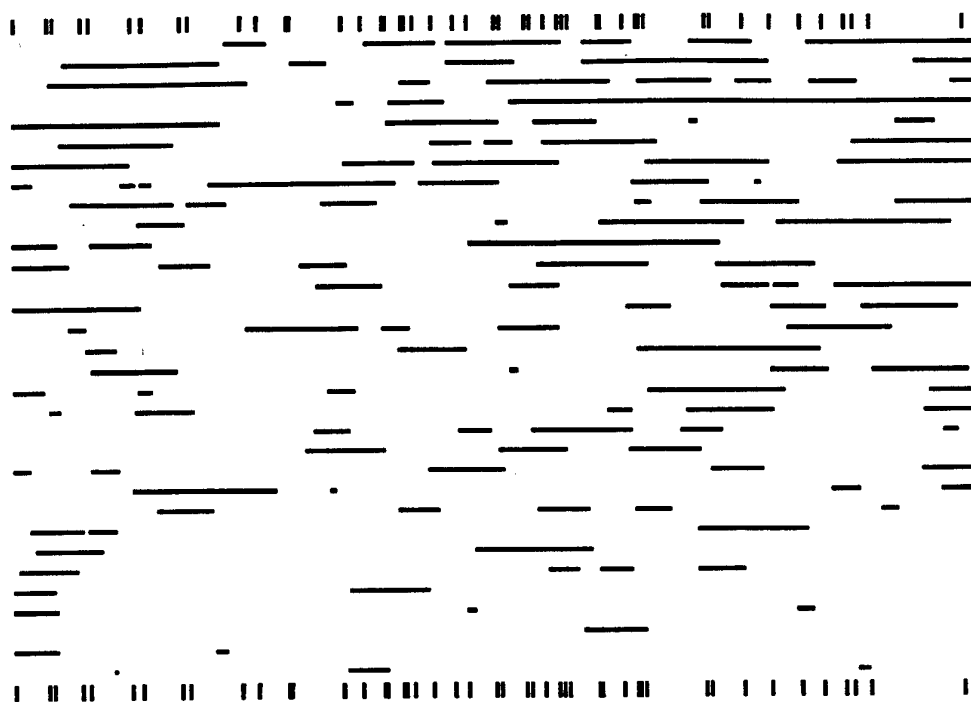
means for isolating a sufficiently large set of collection of DNA sequence fragments wherein each fragment's average size is greater than or equal to the size of said first DNA sequence to achieve the desired resolution, each collection of DNA sequence part of the genome;

a computer having a memory and means for inputting information into the computer about the set of first DNA sequences and the set of collection of DNA sequence fragments;

means for forming a Table A in the memory, Table A determining the containment or overlap of the set of first DNA sequences against the collection of DNA sequence fragments;

means for forming a Table L in the memory from Table A and a transpose of Table A, Table L providing localization information about the set of second DNA sequences relative to the set of first DNA sequences; and

means for producing a map of the set of first DNA sequences from Table L, said map producing means disposed in said memory said map determines where the set of first DNA sequences are located within the genome.

*FIG. 3*

332	26	23	17	17	20	17	17	17	19	21	21	18	20	21	21	18	18	19	19	21	22	21	21	18	16	16	15	16	16	15	16	16	17	16	16	14	14	16	18	18	16	15	14	17	14	13
323	29	32	22	22	19	16	20	18	16	18	13	15	16	16	17	17	18	14	16	17	20	19	19	21	18	15	16	16	15	16	16	17	16	16	14	14	16	18	18	16	15	14	17	14	13	
318	20	21	29	31	23	21	19	21	16	20	19	12	23	21	25	22	25	24	20	19	20	21	20	16	21	22	18	17	20	18	20	19	20	19	20	18	20	19	20	18	15	16	15	13		
319	19	20	26	27	22	20	18	16	18	13	17	16	18	13	21	24	24	21	24	24	24	24	24	17	15	16	17	17	18	19	15	18	20	19	19	21	19	19	19	19	16	17	18	21		
414	14	17	23	23	26	29	27	23	21	21	21	16	18	17	17	18	19	19	19	20	21	20	21	22	16	14	17	18	18	20	20	19	20	18	18	20	18	20	18	22	21	22	19	24		
414	14	17	21	24	27	29	31	19	19	19	16	16	15	15	18	19	19	18	19	20	14	16	19	20	20	21	22	16	16	18	20	18	20	19	20	19	20	17	22	17	22	19	24			
414	14	17	21	24	27	29	31	19	19	19	16	16	15	15	18	19	19	18	19	20	14	16	19	20	20	21	22	16	16	18	20	18	20	19	20	19	20	17	22	17	22	19	24			
414	14	17	21	24	27	29	31	19	19	19	16	16	15	15	18	19	19	18	19	20	14	16	19	20	20	21	22	16	16	18	20	18	20	19	20	19	20	17	22	17	22	19	24			
414	14	17	21	24	27	29	31	19	19	19	16	16	15	15	18	19	19	18	19	20	14	16	19	20	20	21	22	16	16	18	20	18	20	19	20	19	20	17	22	17	22	19	24			
414	14	17	21	24	27	29	31	19	19	19	16	16	15	15	18	19	19	18	19	20	14	16	19	20	20	21	22	16	16	18	20	18	20	19	20	19	20	17	22	17	22	19	24			
414	14	17	21	24	27	29	31	19	19	19	16	16	15	15	18	19	19	18	19	20	14	16	19	20	20	21	22	16	16	18	20	18	20	19	20	19	20	17	22	17	22	19	24			
414	14	17	21	24	27	29	31	19	19	19	16	16	15	15	18	19	19	18	19	20	14	16	19	20	20	21	22	16	16	18	20	18	20	19	20	19	20	17	22	17	22	19	24			
414	14	17	21	24	27	29	31	19	19	19	16	16	15	15	18	19	19	18	19	20	14	16	19	20	20	21	22	16	16	18	20	18	20	19	20	19	20	17	22							

FIG. 4

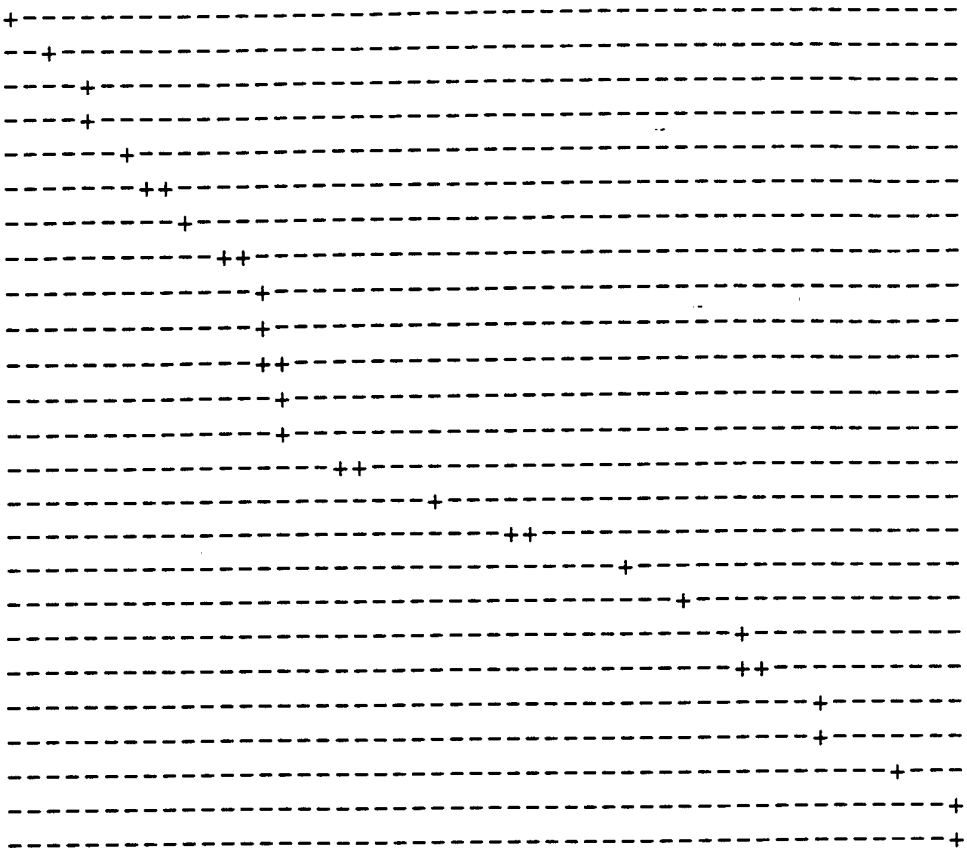
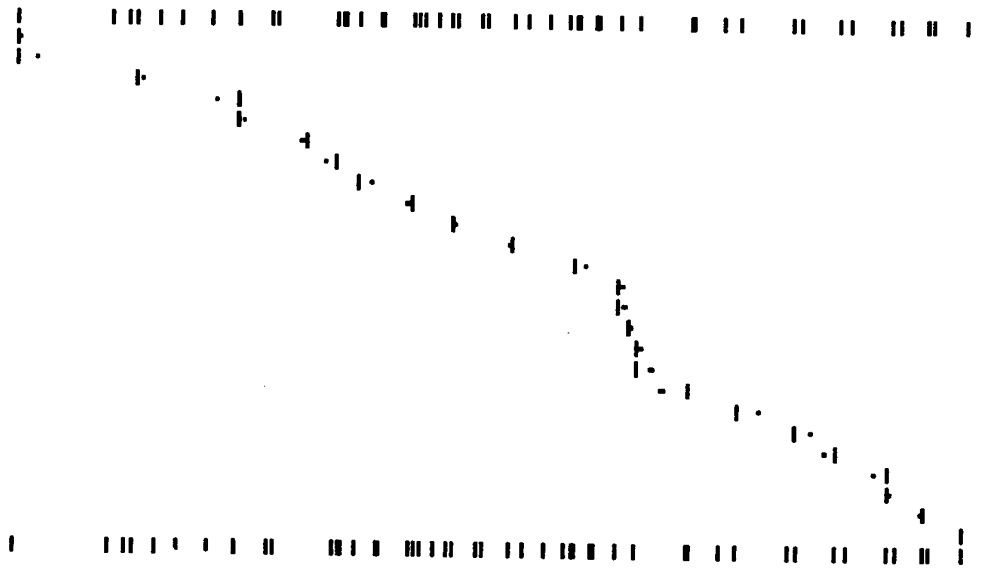


FIG. 5

*FIG. 6*

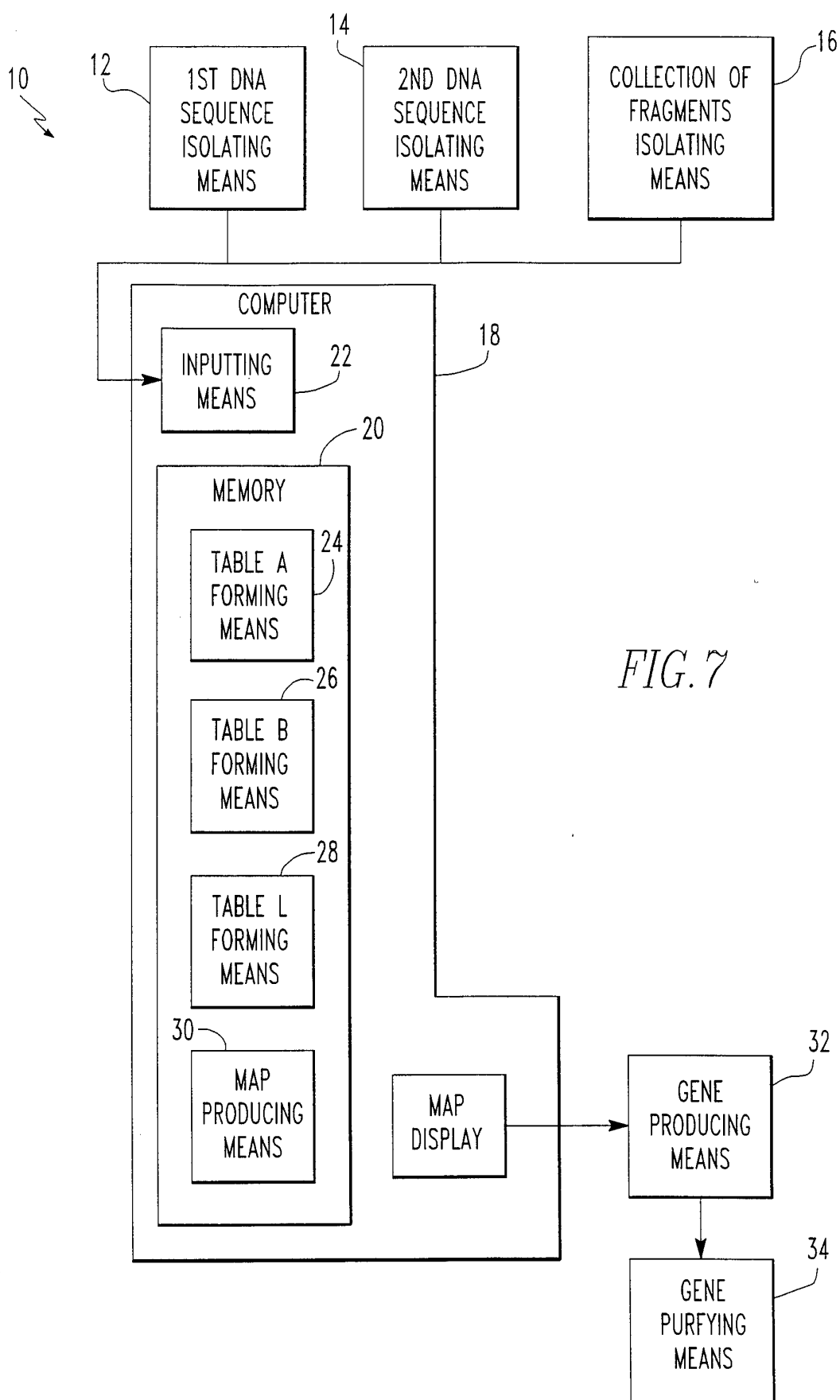


FIG. 7

6 / 8

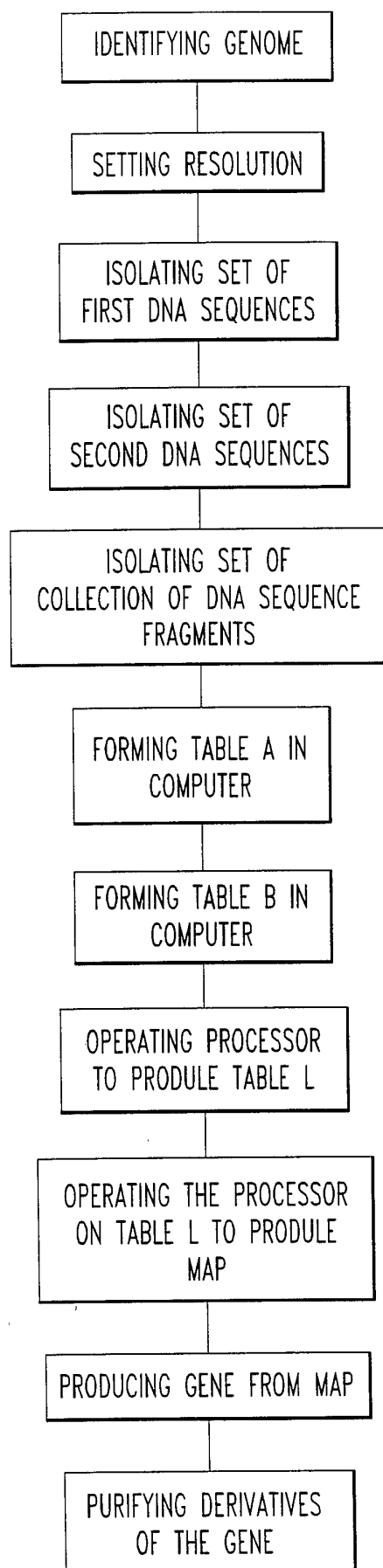


FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US94/09037**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) :C12N 15/00, 15/09, 15/10; C12Q 1/68

US CL :435/6, 172.1, 172.3

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 172.1, 172.3, 320.1; 536/23.1, 24.2, 24.3, 24.31, 24.33

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X ----- Y	Nucleic Acids Research, Volume 21, No. 8, issued 1993, Mott et al, "Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of <i>Schizosaccharomyces pombe</i> ", pages 1965-1974, see pages 1965 and 1972.	1-2, 4-5, 8-9 ----- 1-6, 8-9
X	J. Sambrook et al, "MOLECULAR CLONING, A LABORATORY MANUAL", published 1989 by Cold Spring Harbor Laboratory Press (N.Y.), pages 9.1-9.62, see pages 9.2-9.3 and 9.31.	7
X	Genomics, Volume 13, issued 1992, Altherr et al, "Radiation hybrid map spanning the Huntington Disease gene region of Chromosome 4", pages 1040-1046, see pages 1040-1041.	7

☒ Further documents are listed in the continuation of Box C.
 ☐ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier document published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

08 NOVEMBER 1994

Date of mailing of the international search report

DEC 13 1994

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

TERRY A. MCKELVEY

Telephone No. (703) 308-0196

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US94/09037

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Cytogenetics and Cell Genetics, Volume 59, issued 1992, S.R. Wilson, "A statistical and numerical view of the chromosome 21 physical map data", pages 110-111, see entire document.	1-6, 8-9
Y	Nucleic Acids Research, Volume 19, Number 12, issued 1991, Monaco et al, "Mapping irradiation hybrids to cosmid and yeast artificial chromosome libraries by direct hybridization of Alu-PCR products", pages 3315-3318, see page 3315.	1-6, 8-9
Y	PCR Methods and Applications, Volume 1, Number 2, issued 1991, Green et al, "Sequence-tagged site (STS) content mapping of human chromosomes: Theoretical considerations and early experiences", pages 77-90, see pages 77, 79, and 84.	1-6, 8-9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US94/09037

B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

APS, Dialog, STN

search terms: restriction map, sequencing DNA, sequence tagged site, radiation hybrid map, computer, ordering clone libraries, gene map, indirect high resolution map, STS, ordering clone maps, irradiation hybrids