US 20020119451A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2002/0119451 A1

Usuka et al. (43) Pub. Date: Aug. 29, 2002

(54) SYSTEM AND METHOD FOR PREDICTING CHROMOSOMAL REGIONS THAT CONTROL PHENOTYPIC TRAITS

(76) Inventors: **Jonathan A. Usuka**, Palo Alto, CA (US); **Andrew Grupe**, Redwood City, CA (US); **Gary Allen Peltz**, Redwood City, CA (US)

Correspondence Address:
Pennie & Edmonds, LLP
3300 Hillview Avenue
Palo Alto, CA 94304 (US)

(21) Appl. No.: **09/737,918**

(22) Filed: **Dec. 15, 2000**

Publication Classification

(51) **Int. Cl.**[7] **G06F 19/00**; G01N 33/48; G01N 33/50

(52) **U.S. Cl.** **435/6**; 435/7.1; 702/19

(57) **ABSTRACT**

A method of associating a phenotype with one or more candidate chromosomal regions in a genome of an organism includes the step of deriving a phenotypic data structure that represents differences in phenotypes between different strains of the organism. Further, a genotypic data structure is established. The genotypic data structure corresponds to a locus selected from a plurality of loci in the genome of the organism. The genotypic data structure represents variations of at least one component of the locus between different strains of the organism. The phenotypic data structure is compared to the genotypic data structure to form a correlation value. The process of establishing a genotypic data structure and comparing it to the phenotypic data structure is repeated for each locus in the plurality of loci, thereby identifying one or more genotypic data structures that form a high correlation value relative to all other compared genotypic data structures. The loci that correspond to the one or more genotypic data structures having a high correlation value represent the one or more candidate chromosomal regions.

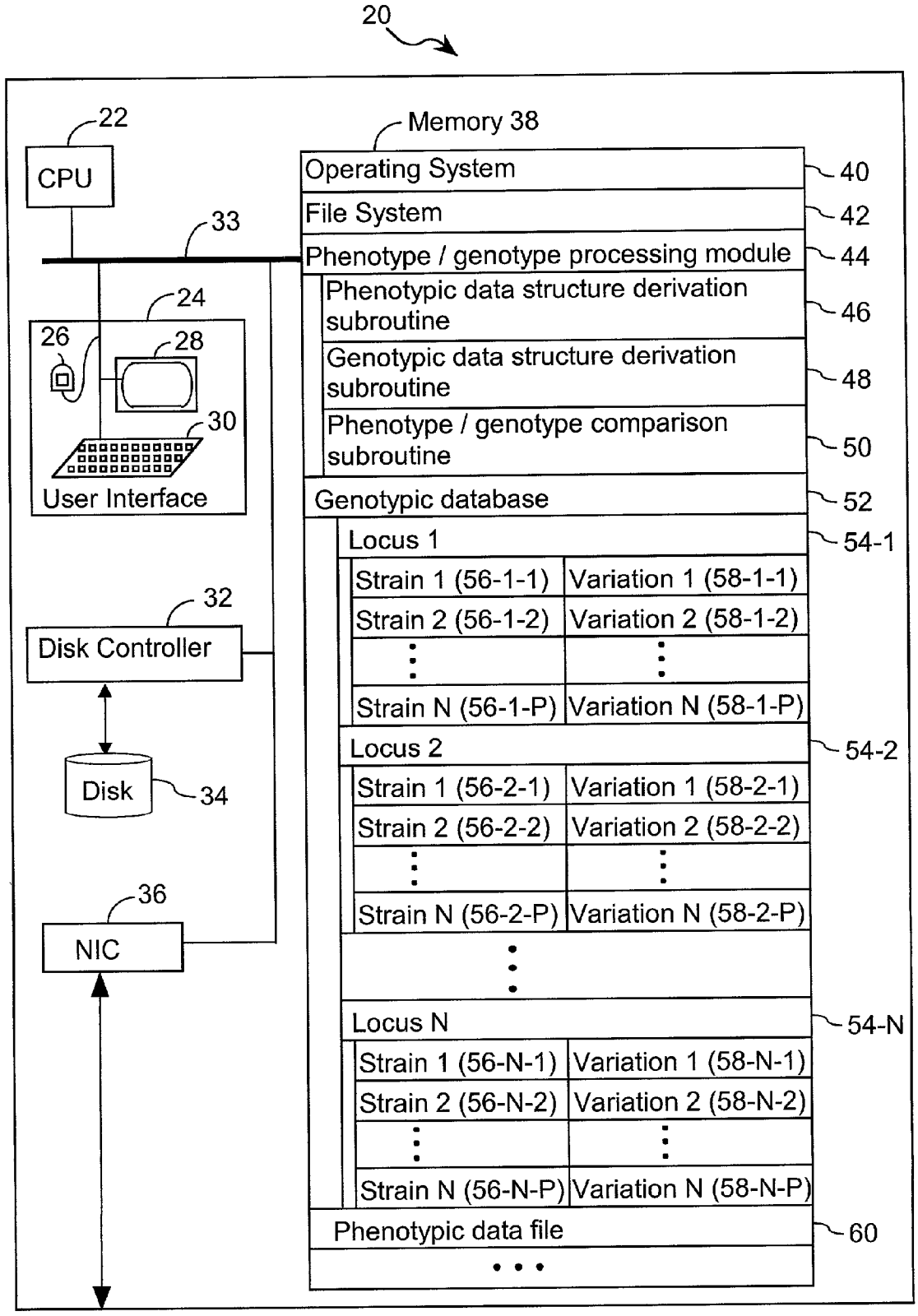
20
22
CPU
33
24
26
28
30
User Interface
32
Disk Controller
Disk
34
36
NIC
Memory 38
Operating System
40
File System
42
Phenotype / genotype processing module
44
Phenotypic data structure derivation subroutine
46
Genotypic data structure derivation subroutine
48
Phenotype / genotype comparison subroutine
50
Genotypic database
52
Locus 1
54-1
Strain 1 (56-1-1)
Variation 1 (58-1-1)
Strain 2 (56-1-2)
Variation 2 (58-1-2)
Strain N (56-1-P)
Variation N (58-1-P)
Locus 2
54-2
Strain 1 (56-2-1)
Variation 1 (58-2-1)
Strain 2 (56-2-2)
Variation 2 (58-2-2)
Strain N (56-2-P)
Variation N (58-2-P)
Locus N
54-N
Strain 1 (56-N-1)
Variation 1 (58-N-1)
Strain 2 (56-N-2)
Variation 2 (58-N-2)
Strain N (56-N-P)
Variation N (58-N-P)
Phenotypic data file
60

FIG. 1

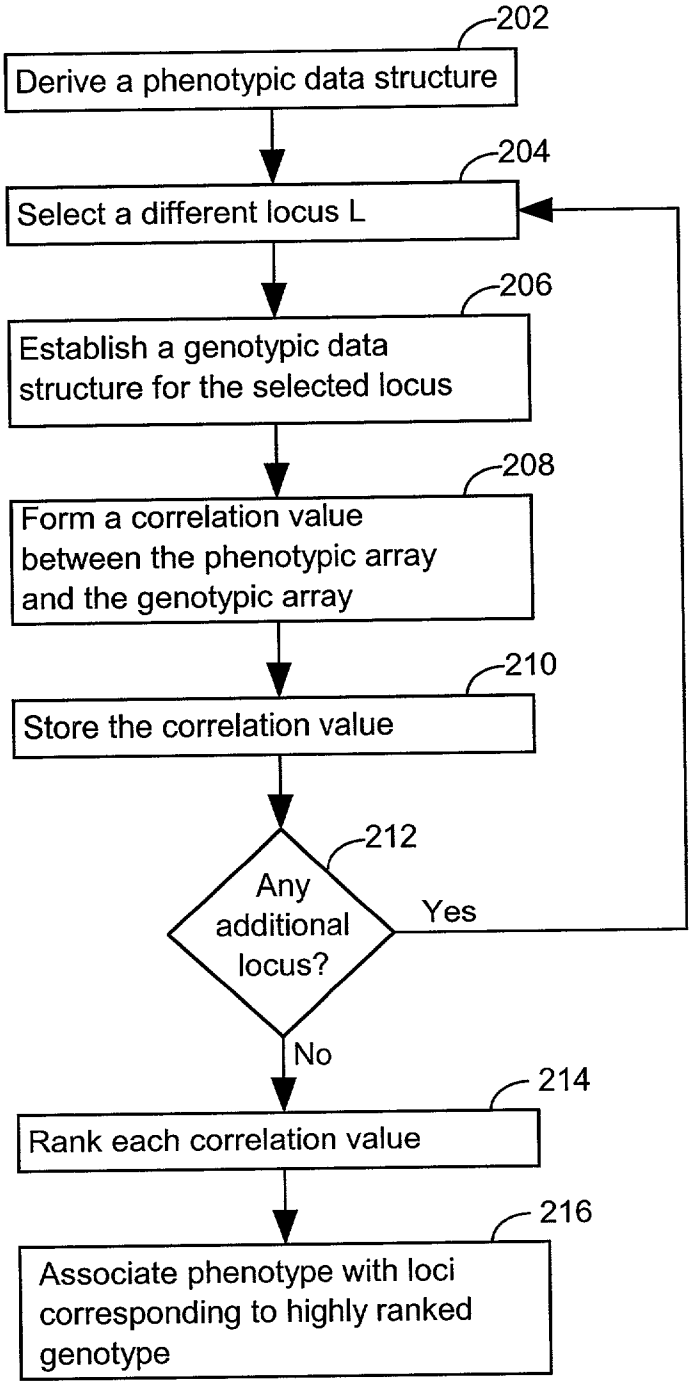
202
Derive a phenotypic data structure
204
Select a different locus L
206
Establish a genotypic data structure for the selected locus
208
Form a correlation value between the phenotypic array and the genotypic array
210
Store the correlation value
212
Any additional locus?
Yes
No
214
Rank each correlation value
216
Associate phenotype with loci corresponding to highly ranked genotype

FIG. 2

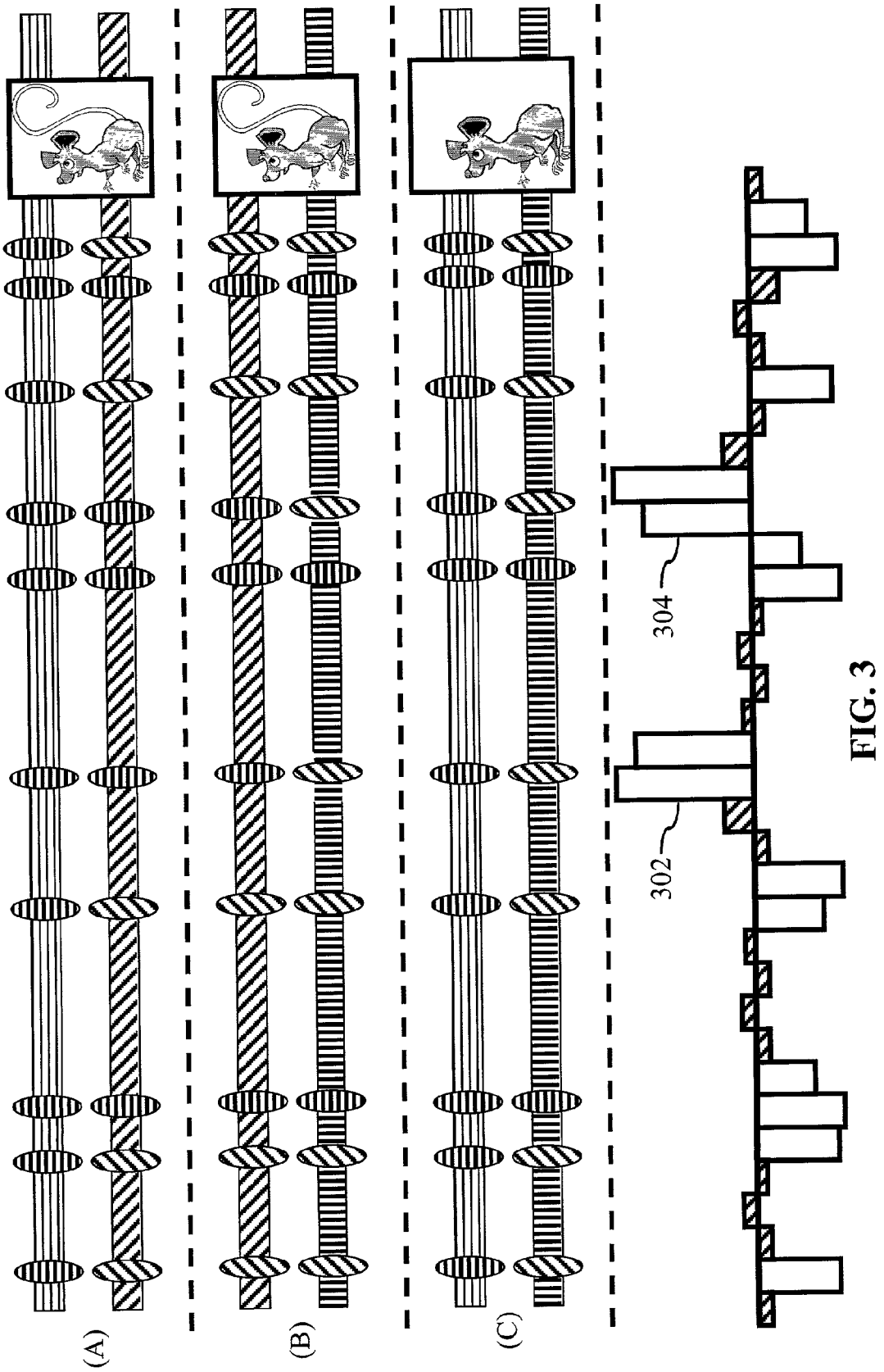

FIG. 3

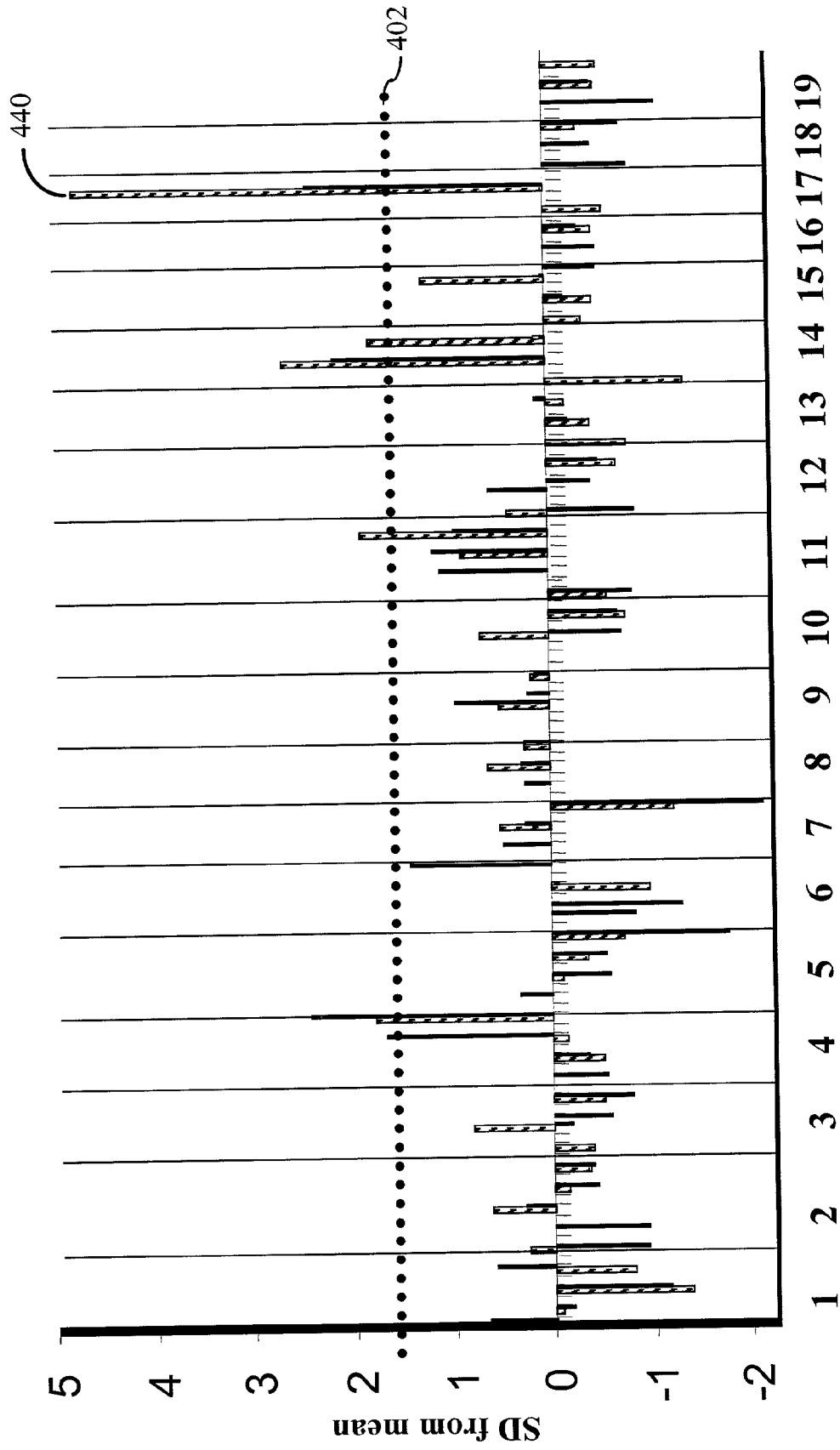
440
402
SD from mean
5
4
3
2
1
0
-1
-2
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

FIG. 4A

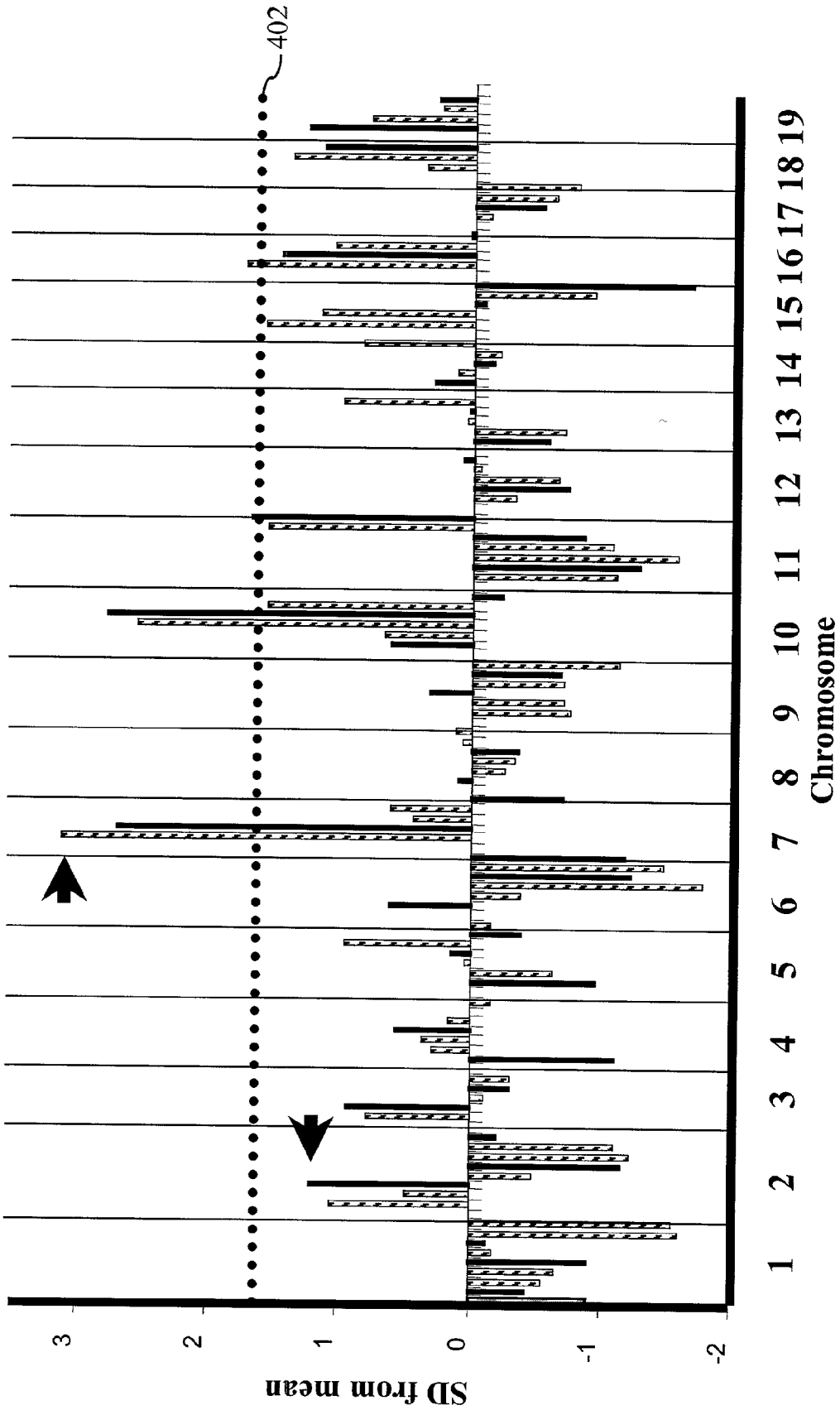
402
SD from mean
3
2
1
0
-1
-2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Chromosome

FIG. 4B

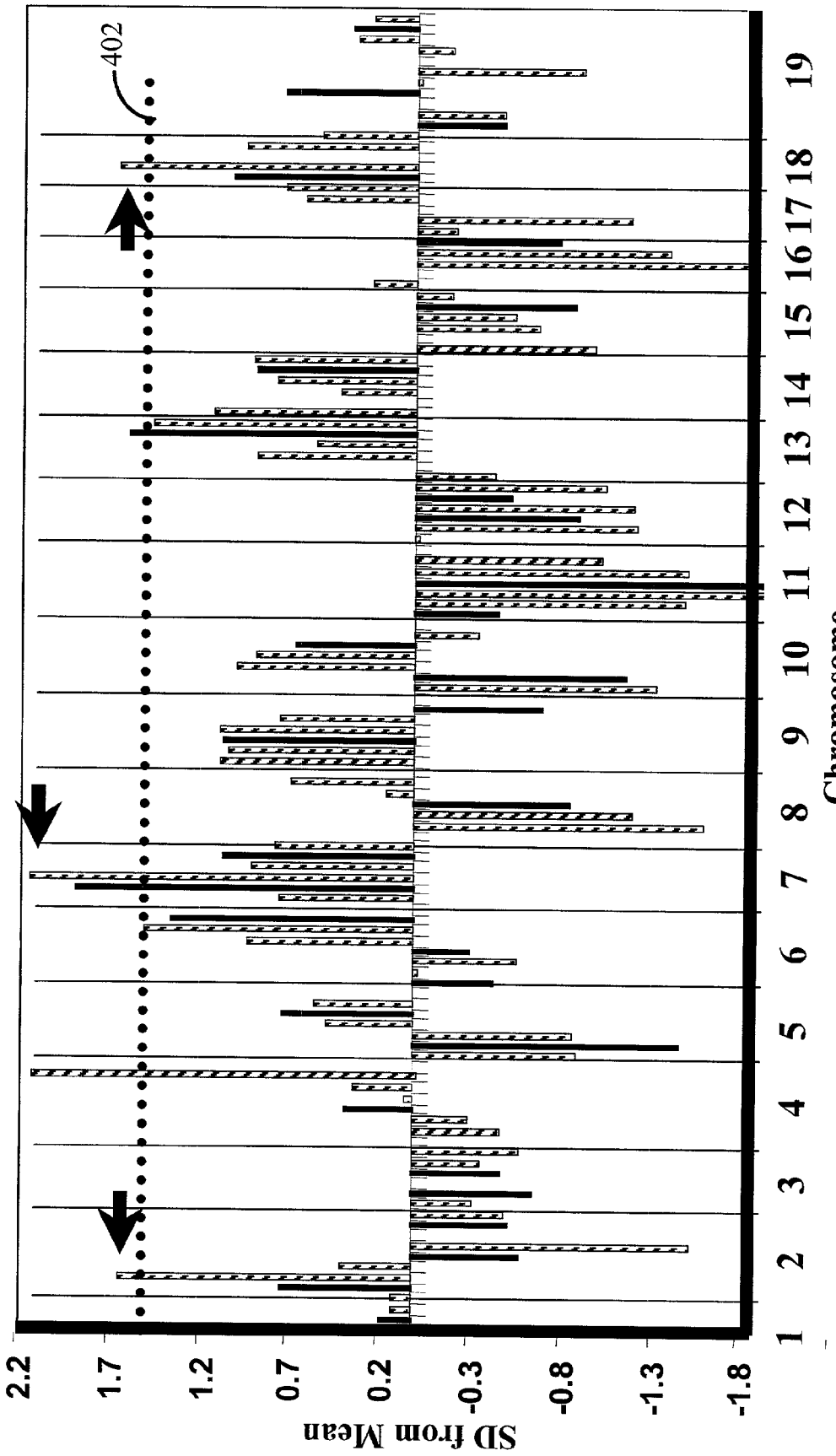
402
SD from Mean
2.2
1.7
1.2
0.7
0.2
-0.3
-0.8
-1.3
-1.8
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

FIG. 4C

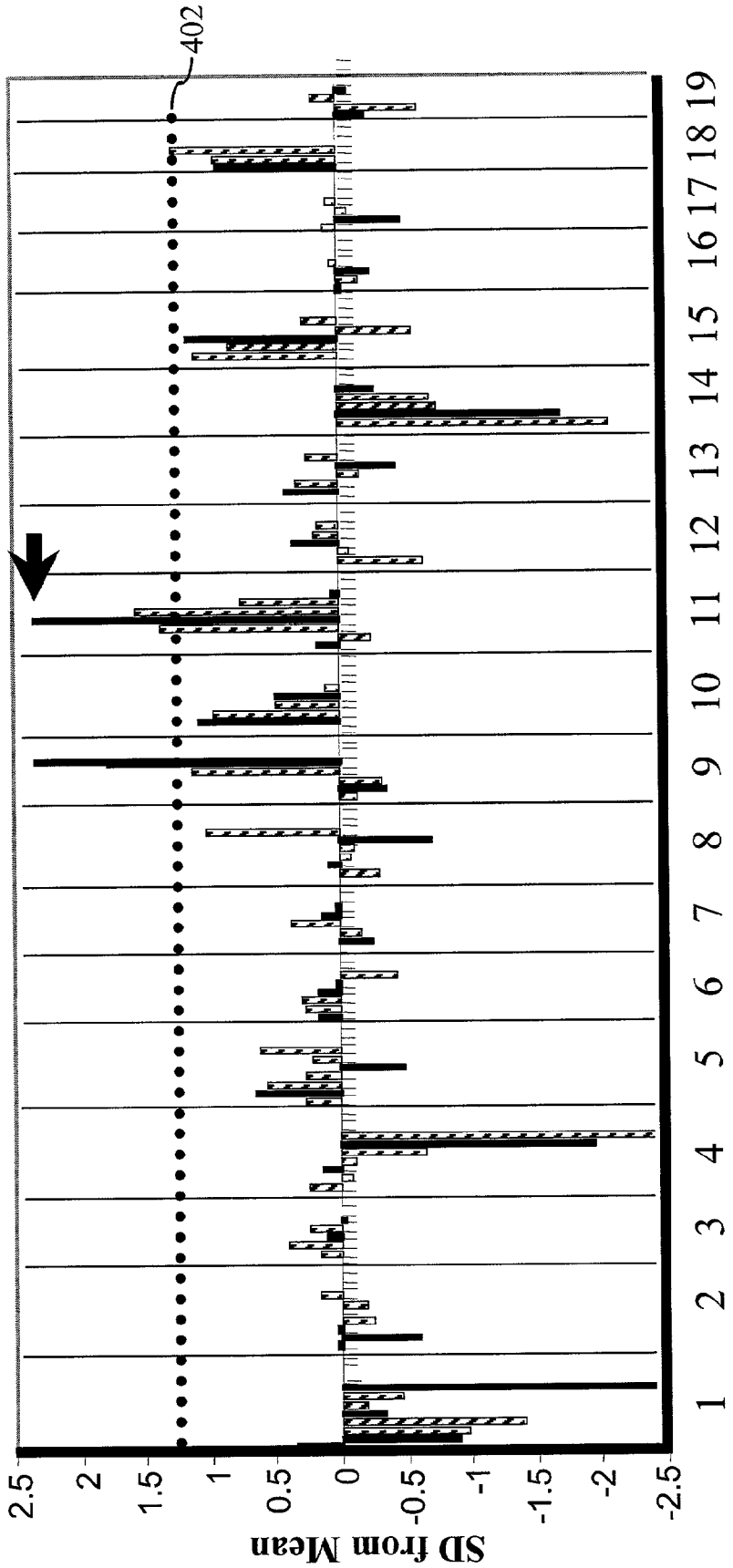
402
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
SD from Mean
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Chromosome

FIG. 4D

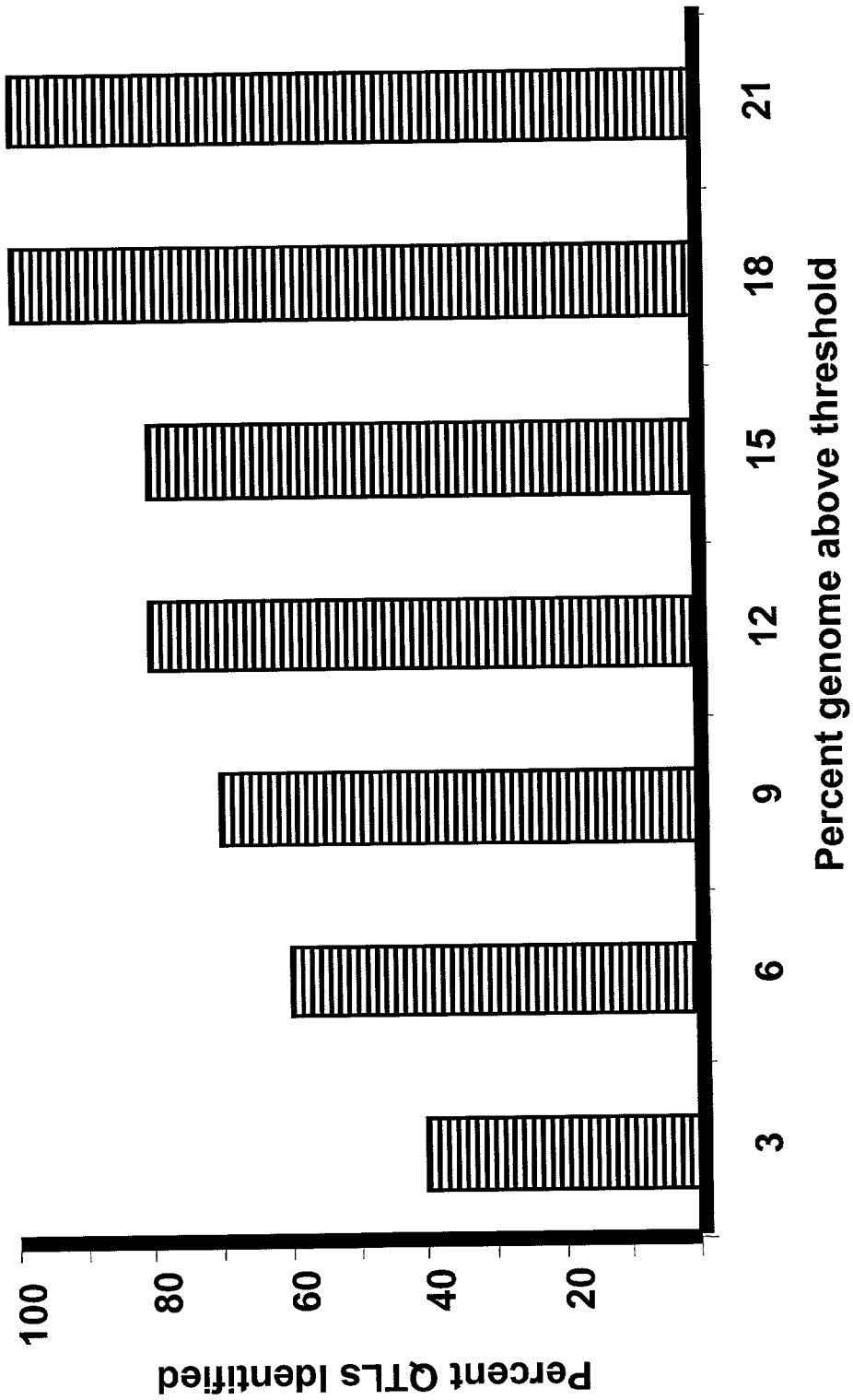
Percent QTLs identified
100
80
60
40
20
Percent genome above threshold
3
6
9
12
15
18
21

FIG. 5

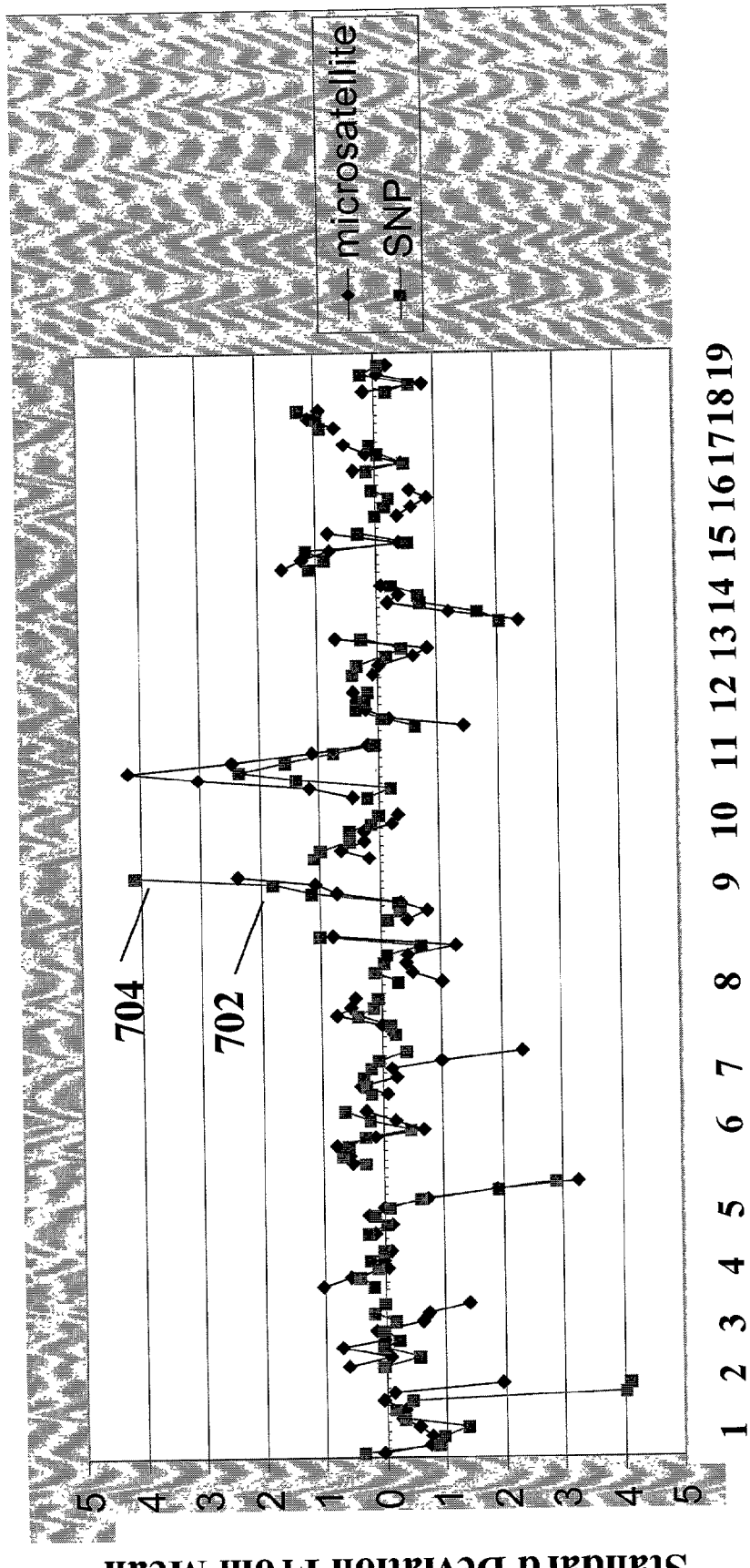
microsatellite
SNP
704
702
5 4 3 2 1 0 1 2 3 4 5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Standard Deviation From Mean
Chromosomal Locus

FIG. 6

SYSTEM AND METHOD FOR PREDICTING CHROMOSOMAL REGIONS THAT CONTROL PHENOTYPIC TRAITS

BACKGROUND OF THE INVENTION

[0001] Identification of genetic loci that regulate susceptibility to disease has promised insight into pathophysiologic mechanisms and the development of novel therapies for common human diseases. Family studies clearly demonstrate a heritable predisposition to many common human diseases such as asthma, autism, schizophrenia, multiple sclerosis, systemic lupus erythematosus, and type I and type II diabetes mellitus. For a review, see Risch, Nature 405, 847-856, 2000. Over the last 20 years, causative genetic mutations for a number of highly penetrant, single gene (Mendelian) disorders such as cystic fibrosis, Huntington's disease and Duchene muscular dystrophy have been identified by linkage analysis and positional cloning in human populations. These successes have occurred in relatively rare disorders in which there is a strong association between the genetic composition of a genome of a species (genotype) and one or more physical characteristics exhibited by the species (phenotype).

[0002] It was hoped that the same methods could be used to identify genetic variants associated with susceptibility to common diseases in the general population. For a review, see Lander and Schork, Science 265, 2037-2048, 1994. Genetic variants associated with susceptibility to subsets of some common diseases such as breast cancer (BRCA-1 and -2), colon cancer (FAP and HNPCC), Alzheimer's disease (APP) and type II diabetes (MODY-1, -2, -3) have been identified by these methods, which has raised expectations. However, these genetic variants have a very strong effect in only a very limited subset of individuals suffering from these diseases (Risch, Nature, 405, 847-856, 2000).

[0003] Despite considerable effort, genetic variants accounting for susceptibility to common, non-Mendelian disorders in the general population have not been identified. Since multiple genetic loci are involved, and each individual locus makes a small contribution to overall disease susceptibility, it will be quite difficult to identify common disease susceptibility loci by applying conventional linkage and positional cloning methods to human populations. Mapping of disease susceptibility genes in human populations has also been hampered by variability in phenotype, genetic heterogeneity across populations, and uncontrolled environmental influences. The variable reports of linkage between the chromosome 1q42 region and systemic lupus erythematosus illustrate the difficulties encountered in human genetic studies. One group reported strong linkage between the 1q42 region (Tsao, J.Clin.Invest, 99, 725-731, 1997) and to microsatellite alleles of a gene (PARP) within that region (Tsao, J.Clin.Invest. 103, 1135-1140, 1999). In contrast, no evidence for association with the PARP microsatellite marker was noted (Criswell et al., J.Clin.Invest, Jun;105, 1501-1502, 2000; Delrieu et al., Arthritis & Rheumatism 42, 2194-2197, 1999); and minimal (Mucenski, et al., Molecular & Cellular Biology 6, 4236-4243, 1986) or no linkage (Lindqvist, et al., Journal of Autoimmunity, Mar;14, 169-178, 2000) to the 1q42 region was found in several other SLE populations analyzed. It is likely that additional tools and approaches will be needed to identify genetic factors underlying common human diseases.

[0004] Analysis of experimental murine genetic models of human disease biology should greatly facilitate identification of genetic susceptibility loci for common human diseases. Experimental murine models have the following advantages for genetic analysis: inbred (homozygous) parental strains are available, controlled breeding, common environment, controlled experimental intervention, and ready access to tissue. A large number of murine models of human disease biology have been described, and many have been available for a decade or more. Despite this, relatively limited progress has been made in identifying genetic susceptibility loci for complex disease using murine models. Genetic analysis of murine models requires generation, phenotypic screening and genotyping of a large number of intercross progeny. Using currently available tools, this is a laborious, expensive and time-consuming process that has greatly limited the rate at which genetic loci can be identified in mice, prior to confirmation in humans. For a review, see Nadeau and Frankel, Nature Genetics Aug;25, 381-384, 2000.

[0005] The difficulties encountered in associating phenotypic variations, such as susceptibility to common diseases, with genetic variations gives rise to a need in the art for additional tools for identifying chromosomal regions that are most likely to contribute to quantitative traits or phenotypes. In view of this situation, it would be highly desirable to provide a technique for associating a phenotype with one or more candidate chromosomal regions in the genome of an organism without reliance on time consuming techniques such as cross breeding experiments or laborious post-PCR manipulation.

SUMMARY OF THE INVENTION

[0006] The present invention provides a system and method for associating a phenotype with one or more candidate chromosomal regions in the genome of an organism. In the method, phenotypic differences between a plurality of strains of the organism are correlated with variations and/or similarities in the respective genomes of the plurality of strains of the organism. The invention relies on the use of a genotypic database that includes variations and similarities of representative strains of the organism of interest. Representative genotypic databases include, but are not limited to, single nucleotide polymorphism databases, microsatellite marker databases, restriction fragment length polymorphism databases, short tandem repeat databases, sequence length polymorphism databases, expression profile databases, and DNA methylation databases.

[0007] One embodiment of the present invention provides a method for associating a phenotype with one or more candidate chromosomal regions in a genome of an organism. In this method, a phenotypic data structure that represents a difference in one or more phenotypes between different strains of the organism is derived. In its simplest form, the phenotypic data structure comprises a definition of one or more phenotypes exhibited by the organism together with a measure of each of these phenotypes. For example, a hypothetical phenotypic data structure for rabbits could include the phenotypes "tail length" and "hair color" and the respective measure for each of these phenotypes could be "7 centimeters" and "brown."

[0008] A genotypic data structure is established in accordance with one embodiment of the present invention. The

genotypic data structure is identified by a particular locus selected from a plurality of loci present in the genome of the organism. The genotypic data structure includes one or more positions within the locus. For each of these positions, the genotypic data structure provides information on the extent of a variation between different strains of the organism. A hypothetical example of a genotypic data structure in accordance with the present invention is an data structure for a locus that includes genes A and B. In such an example, the genotypic data structure includes the positions of genes A and B within the locus as well as some measurement related to genes A and B, such as the mRNA expression level that has been measured for each of these genes. In this example, the mRNA expression-level defines the extent of variation between different strains of the organism.

[0009] The phenotypic and genotypic data structures are then compared to form a correlation value. The process continues with the establishment of another genotypic data structure that corresponds to a different loci and the concomitant comparison of this genotypic data structure to the phenotypic structure until several of the loci in the genome of the organism have been tested in this manner. In this way, one or more genotypic data structures are identified that form a high correlation value relative to all other genotypic data structures that have been compared to the phenotypic data structure. Further, the loci in the genome of the organism that correspond to the highly correlated genotypic data structures represent one or more candidate chromosomal regions that may be associated with the phenotype of interest.

[0010] In some embodiments of the present invention, each element in a phenotypic data structure represents a variation in the phenotype between a different first and second strain of the organism of interest. Such variations may be determined by measurement of an attribute corresponding to the phenotype in the respective strains of the organism. Representative phenotypic variations include, for example, eye color, hair color, and susceptibility to a particular disease. In other embodiments, each element in a phenotypic data structure represents a variation in the phenotype between a different first and second cluster of strains of the organism of interest.

[0011] In additional embodiments of the present invention, the genotypic data structure represents a variation of at least one component of a locus between two strains of the organism of interest. In other embodiments, each element in the genotypic data structure represents a variation of at least one component of the locus between a different first cluster of strains of the organism and a different second cluster of strains of the organism. In some embodiments, the phenotypic and genotypic data structures represent a subset of all strains of the organism of interest.

[0012] The present invention contemplates a considerable number of different methods for comparing the phenotypic and genotypic data structures. In one embodiment the correlation value between the phenotypic data structure and a particular genotypic data structure is formed in accordance with the expression:

$$c(P, G^L) = \frac{\sum^i (p(i) - < P >)(g(i) - < G^L >)}{\{[\sum^i (p(i) - < P >)^2][\sum^i (g(i) - < G^L >)^2]\}^{1/2}}$$

[0013] where,

[0014] $c(P, G^L)$ is the correlation value;

[0015] p(i) is a value of the i^{th} element of the phenotypic data structure;

[0016] g(i) is a value of the i^{th} element of the genotypic data structure;

[0017] <P> is a mean value of all elements in the phenotypic data structure;

[0018] $<G^L>$ is a mean value of all elements in the genotypic data structure;

[0019] and

$$\sum^i = \sum_{i=1}^{N}.$$

[0020] Other methods for forming a correlation value between the phenotypic data structure and a particular genotypic data structure include but are not limited to regression analysis, regression analysis with data transformations, a Pearson correlation, a Spearman rank correlation, a regression tree and concomitant data reduction, partial least squares, and canonical analysis.

[0021] In some embodiments of the present invention, statistical methods are used to identify which of the genotypic data structures that have been compared to a phenotypic data structure are highly correlated. In one such embodiment, a mean correlation value that represents a mean of correlation values is computed between the phenotypic data structure and a particular genotypic data structure. Further, a standard deviation of the mean correlation is computed. Genotypic data structures having a correlation value that is a number of standard deviations above the mean correlation value are considered to be the data structures that correspond to loci that are associated with the genotypic trait. The number of standard deviations that is chosen for the cutoff is dynamically chosen so that a specific percentage of the genome, such as ten percent, is identified as positive.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] **FIG. 1** illustrates a computer system for associating a phenotype with one or more candidate chromosomal regions in a genome of an organism in accordance with one embodiment of the present invention.

[0023] **FIG. 2** illustrates the processing steps for associating a phenotype with one or more candidate chromosomal regions in a genome of an organism in accordance with one embodiment of the present invention.

[0024] **FIG. 3** illustrates a hypothetical representation of the method for computational prediction of QTL intervals in accordance with one embodiment of the present invention.

[0025] **FIGS. 4A -4**D illustrate the computational prediction of chromosomal regions containing genes that determine MHC haplotype

[0026] **(FIG. 4A)**, lymphoma susceptibility

[0027] **(FIG. 4B)**, airway hyperresponsiveness

[0028] **(FIG. 4C)** and retinal ganglion number

[0029] **(FIG. 4D)** in accordance with one embodiment of the present invention.

[0030] **FIG. 5** illustrates an analysis of the sensitivity of the computational genome scanning method for prediction using ten experimentally verified QTL intervals. A graph of the percentage of correct predictions as a function of the amount of genomic sequence (percent) contained within the predicted regions is plotted.

[0031] **FIG. 6** illustrates the comparison of a genotypic database **52** that includes SNP data versus a genotypic database that includes microsatellite data in identifying the murine chromosomal location for the phenotypic trait of retinal ganglion cell formation, in accordance with one embodiment of the present invention.

[0032] Like reference numerals refer to corresponding parts throughout the several views of the drawings.

DETAILED DESCRIPTION OF THE INVENTION

[0033] A key aspect of research in genetics is associating sequence variations with heritable phenotypes. The most common variations are single nucleotide polymorphisms (SNPs), which occur approximately once every 100 to 300 bases in a genome. Because SNPs are expected to facilitate large-scale association genetics studies, there has recently been great interest in SNP discovery and detection. The present invention contemplates the use of genotypic databases such as SNP databases in order to correlate genetic variances in an organism with one or more phenotypic variances. As an example, a searchable database of mouse SNPs that contains alleles for 15 common inbred mouse strains and information for performing high throughput, inexpensive genotyping assays for each SNP was built. Using pooled DNA samples and SNP genotyping assays in the database, a genome scan on phenotypically extreme progeny from an experimental intercross was completed. SNP-based genotyping of pooled samples requires at least twenty-fold fewer assays than genotyping individual samples with microsatellite markers, and identified the same linkage regions.

[0034] Although the examples provided herein utilize a genotypic database that includes fifteen mouse strains, it will be appreciated that the methods of the present invention allow for the use of any number of different types of genetic information. For example, suitable genotypic databases include databases that have various types of gene expression data from platform types such as spotted microarray (microarray), high-density oligonucleotide array (HDA), hybridization filter (filter) and serial analysis of gene expression (SAGE) data. Gene expression changes often reflect genotypic variation. Therefore, databases of gene expression among tissues obtained from different individuals (mouse strains or humans), can also be utilized by this method. The chromosomal position of all human genes is known for human genes, as a result of physical mapping or sequencing of the human genome. For gene expression data for mouse or other species, the chromosomal location is either known (physical mapping or mouse genomic sequencing) or can be estimated by syntenic mapping based upon homology with human genes. Another example of a genetic database that can be used is a DNA methylation database. For details on a representative DNA methylation database, see Grunau et al., "MethDB—a public database for DNA methylation data," Nucleic Acids Research, in press; or the URL:

[0035] http://genome.imb-jena.de/public.html.

[0036] To accelerate the process of analyzing experimental genetic models in order to identify the genetic causes of complex human disease, the present invention provides tools for scanning genotypic databases, such as SNP databases, to predict quantitative trait loci (QTL) after phenotypic information obtained from common strains of the organism is provided. The computational QTL prediction method is capable of correctly predicting the chromosomal regions that have been previously identified by tedious and laborious analysis of experimental intercross populations for the multiple traits that are analyzed. Thus, the present invention bypasses the burdensome requirement for generation and characterization of intercross progeny, enabling QTL regions to be predicted within a millisecond time frame.

[0037] **FIG. 1** shows a system **20** for associating a phenotype with one or more candidate chromosomal regions in a genome of an organism.

[0038] System **20** preferably includes:

[0039] a central processing unit **22**;

[0040] a main non-volatile storage unit **34**, preferably a hard disk drive, for storing software and data, the storage unit **34** controlled by disk controller **32**;

[0041] a system memory **38**, preferably high speed random-access memory (RAM), for storing system control programs, data, and application programs, including programs and data loaded from non-volatile storage unit **34**; system memory **38** may also include read-only memory (ROM);

[0042] a user interface **24**, including one or more input devices **(26, 30)** and a display **28**;

[0043] a network interface card **36** for connecting to any wired or wireless communication network; and

[0044] an internal bus **33** for interconnecting the aforementioned elements of the system.

[0045] Operation of system **20** is controlled primarily by operating system **40**, which is executed by central processing unit **22**. Operating system **40** may be stored in system memory **38**. In a typical implementation, system memory **38** includes:

[0046] operating system **40**;

[0047] file system **42** for controlling access to the various files and data structures used by the present invention;

[0048] phenotype/genotype processing module **44** for associating a phenotype with one or more candidate chromosomal regions in a genome of an organism;

[0049] genotypic database **52** for storing variations in genomic sequences of a plurality of strains of an organism; and

[0050] phenotypic data **60** that includes measured differences in one or phenotypic traits associated with the organism.

[0051] In a preferred embodiment, phenotype/genotype processing module **44** includes:

[0052] a phenotypic data structure derivation subroutine **46** for deriving a phenotypic data structure that represents a variation in a phenotype between different strains of an organism of interest;

[0053] a genotypic data structure derivation subroutine **48** for establishing a genotypic data structure that corresponds to a locus in the genome of the organism of interest; and

[0054] a phenotype/genotype comparison subroutine **50** for comparing the phenotypic array to the genotypic array to form a correlation value.

[0055] The operation of these subroutines is described below in connection with the description for **FIG. 2**.

[0056] Genotypic database **52** is any type of genetic database that tracks variations in the genome of an organism of interest. Information that is typically represented in genotypic database **52** is a collection of loci **54** within the genome of the organism of interest. For each locus **54**, strains **56** for which genetic variation information is available are represented. For each represented strain **56**, variation information **58** is provided. Variation information **58** is any type of genetic variation information. Representative genetic variation information **58** includes, but is not limited to, single nucleotide polymorphisms, restriction fragment length polymorphisms, microsatellite markers, restriction fragment length polymorphisms, and short tandem repeats. Therefore, suitable genotypic databases **52** include, but are not limited to:

Genetic variation type	Uniform resource location
SNP	http://bioinfo.pal.roche.com/usuka_bioinformatics/cgi-bin/msnp/msnp.pl
SNP	http://snp.cshl.org/
SNP	http://www.ibc.wustl.edu/SNP/
SNP	http://www-genome.wi.mit.edu/SNP/mouse/
SNP	http://www.ncbi.nlm.nih.gov/SNP/
Microsatellite markers	http://www.informatics.jax.org/searches/polymorphism_form.shtml
Restriction fragment length polymorphisms	http://www.informatics.jax.org/searches/polymorphism_form.shtml
Short tandem repeats	http://www.cidr.jhmi.edu/mouse/mmset.html
Sequence length polymorphisms	http://mcbio.med.buffalo.edu/mit.html
DNA methylation database	http://genome.imb-jena.de/public.html

[0057] In addition, the genetic variations used by the methods of the present invention may involve differences in the expression levels of genes rather than actual identified variations in the composition of the genome of the organism of interest. Therefore, genotypic databases **52** within the scope of the present invention include a wide array of expression profile databases such as the one found at the URL:

[0058] http://www.ncbi.nlm.nih.gov/geo/

[0059] It will be appreciated that when the variation tracked by genotypic database **52** is a variation in the expression level of a gene rather than a variation in the genome, there is no requirement that genomic database **52** be populated with elements such as locus **54**.

[0060] Referring to **FIG. 2**, the processing steps that are performed in accordance with one embodiment of the present invention are illustrated. In processing step **202**, a phenotypic data structure is derived from phenotypic data **60** (**FIG. 1**) using phenotypic data structure derivation subroutine **46** (**FIG. 1**). The phenotypic data structure tracks measured differences in traits between strains of an organism of interest.

[0061] In one embodiment, the phenotypic data structure used is a phenotypic array. In this embodiment, the phenotypic array is formed in a stepwise fashion by subroutine **46**. First, an N×N phenotypic distance matrix, P, is established where both the ith row and the jth column are associated with a given strain for which quantitative information t_i exists for a given trait.

[0062] This matrix is populated with the differences between strains in regard to the examined trait as follows:

$$p(i,j)=|t_i-t_j|$$

[0063] Therefore, each element in the matrix corresponds to a distance between strains using the quantitative trait as a metric for the space. This matrix has the following properties:

[0064] All of its diagonal elements are zero, because

$$p(i,i)=|t_i-t_i=0\forall i$$

[0065] The matrix is symmetric, because

$$p(i,j)=|t_i-t_j|=|t_j-t_i|=p(j,i)$$

[0066] As an example, consider phenotypic information on the lifespan of five mouse strains:

Strains	Lifespan (days)
A/J	777
AKR/J	282
C3H/HeJ	510
C57BL/6J	895
DBA/2J	568

[0067] A phenotypic distance matrix that tracks the lifespan for these five species members has the form:

P	A/J	AKR/J	C3H/HeJ	C57BL/6J	DBA/2J
A/J	0	495	267	118	209
AKR/J	495	0	228	613	286

-continued

P	A/J	AKR/J	C3H/HeJ	C57BL/6J	DBA/2J
C3H/HeJ	267	228	0	385	58
C57BL/6J	118	613	385	0	327
DBA/2J	209	286	58	327	0

[0068] Each value in this illustrative phenotypic distance matrix represents the difference in life span between the designated members.

[0069] The phenotypic data structure derivation subroutine **46** converts the phenotypic matrix to the phenotypic array by taking the non-redundant, non-diagonal elements of the matrix and arranging them into a vector P:

$$P=p(1,2),\ p(1,3),\ \ldots,\ p(1,N),\ p(2,3),\ p(2,4),\ \ldots,\ p(2,N),\ \ldots\ p(N-1,\ N)$$

[0070] The vector P obtained for the illustrative distance matrix set forth above is P=(495, 267, 118, 209, 228, 613, 286, 385, 58, 327). The linear format of P facilitates the ordered comparison of the phenotype and genotype of respective strains of an organism of interest in subsequent computational steps.

[0071] In some embodiments of the present invention, the phenotypic data used by phenotypic data structure derivation subroutine **46** (**FIG. 1**) in processing step **202** (**FIG. 2**) is entered by hand into system **20** by a computer operator. In other embodiments, the phenotypic data is read from a source such as phenotypic data file **60** (**FIG. 1**). It will be appreciated that there are no limitations on the format of the phenotypic data. The phenotypic data can, for example, represent a series of measurements for a quantifiable phenotypic trait in a collection of strains of a species. Such quantifiable phenotypic traits may include, for example, murine tail length, lifespan, eye color, size and weight. Alternatively, the phenotypic data can be in binary form that tracks the absence or presence of some phenotypic trait. As an example, a "1" may indicate that a particular species of the organism of interest possesses a given phenotypic trait and a "0" may indicate that a particular species of the organism of interest lacks the phenotypic trait. The phenotypic data structure can be populated with any form of biological data that is representative of the phenotype of the organism of interest. Thus, in some embodiments of the present invention, the phenotypic data can be expression data such as mRNA expression data or protein expression level data. In such embodiments, each element in the phenotypic data structure is populated with differences in mRNA or protein expression levels between strains of the organism of interest or of cells cultured from the organism of interest.

[0072] At processing step **204**, a particular locus, or position, is selected within the genome of the organism of interest. Processing step **204** is the first step of a repetitive loop formed by processing steps **204** through **212** that is repeated for several different loci, or positions, within the genome of the organism of interest.

[0073] In processing step **206**, a genotypic data structure is established for the selected locus. In one embodiment, processing step **206** is performed by genotypic data structure derivation subroutine **48** (**FIG. 1**). The genotypic data structure is typically formed in a method similar to the construction of the phenotypic data structure. While the phenotypic data structure's values are typically the differences in quantitative traits exhibited by several strains of an organism of interest, the values in the genotypic data structure correspond to counts of the polymorphic differences between strains for a given locus L that contains M genetic variations, such as SNPs. That is, a given locus L may have several independent genetic variations M, and the goal of the genotypic array that corresponds to this locus is to quantify the number of these independent genetic variations. To accomplish this, an individual variation matrix S^x is established for each variation in every position x within locus L. In each such matrix, S^x, the i^{th} row and the j^{th} column are associated with the allele value $l^x(i)$ for strain i at locus position x. The elements of these S matrices are populated according to the following rule:

$$s_x(i,j)=\tfrac{1}{2} \text{ if } l^x(i)=\emptyset \text{ or } l^x(j)=\emptyset=0 \text{ if } l^x(i)=l^x(j) =1 \text{ if } l^x(i)\neq l^x(j)$$

[0074] where Ø indicates the allelic value for strain i at locus position x is not known at the present time. Therefore, if the alleles for two strains i and j are identical at position x, the entry in the individual variation matrix for x would be:

$$s_x(i,j)=s_x(j,i)=0$$

[0075] and if the two alleles are different, a "1" is entered. A complication arises regarding variations for which not all allelic information is known at the present time (symbolized by Ø). For example, locus position l^x may contain information on the allele for strain i, but not for strain j. In this situation, the assumption is made that strain j has equal probability of containing either allele, and the corresponding entry is set equal to one half.

[0076] At this point, in some embodiments, each individual variation matrix S contains elements that take on one of three values: 0, ½, or 1. It will be appreciated that many other types of schemes may be used when allelic information is not presently known and use of the value "½" in such instances merely illustrates one example of a scheme that may be used in such instances. Similarly, any number of weighting schemes can be used rather than a "0" or "1" and all such weighting schemes are within the scope of the present invention.

[0077] A variation matrix S that tracks an individual locus position l^x for five members (M1 through M5) of a species may have the form:

S	M1	M2	M3	M4	M5
M1	0	0.5	0.5	1	0
M2	0.5	0	0.5	0	1
M3	0.5	0.5	0	1	1
M4	1	0	1	0	0.5
M5	0	1	1	0.5	0

[0078] To assemble the overall genotypic matrix for this locus, the M individual variation matrices, representing the M variations in locus L, are summed:

$$g(i,j)=\Sigma_m{}^M s_m(i,j)$$

[0079] Therefore, an illustrative genotypic matrix G that represents a specific locus in five species members (M1 through M5) has the form:

G	M1	M2	M3	M4	M5
M1	0	3.5	2	4	3
M2	3.5	0	3	2.5	1
M3	2	3	0	1	1
M4	4	2.5	1	0	0.5
M5	3	1	1	0.5	0

[0080] In viewing the illustrative genotypic matrix G, it is apparent that there is relatively little genotypic variance between members M5 and M4 (0.5) whereas there is more variance between M1 and M2 (3.5).

[0081] Finally, one embodiment of genotypic data structure derivation subroutine **48** converts the genotypic matrix to a genotypic array by taking the non-redundant, non-diagonal elements of the matrix and arranging them into the vector G:

$$G=g(1,2),\ g(1,3),\ \ldots,\ g(1,N),\ g(2,3),\ g(2,4),\ \ldots,\ g(2,N),\ \ldots\ g(N-1,\ N)$$

[0082] The vector G obtained for the illustrative genotypic matrix set forth above is G=(3.5, 2, 4, 3, 3, 2.5, 1, 1, 1, 0.5). Once a genotypic matrix such as G has been established in processing step **206**, a correlation value is formed between the phenotypic array and the genotypic array (processing step **208**). This correlation value is typically computed by phenotype/genotype comparison subroutine **50 (FIG. 1)**. In one embodiment, this correlation is determined by linear regression and the correlation coefficient is calculated as:

$$c(P,\ G^L)=\frac{\sum^i (p(i)-<P>)(g(i)-<G^L>)}{\{[\sum^i (p(i)-<P>)^2][\sum^i (g(i)-<G^L>)^2]\}^{1/2}}$$

[0083] where,

[0084] $<A>=\Sigma^i a(i)/I$, $1 \leqq i \leqq I$; (The mean of the scalars comprising vector A);

[0085] $c(P, G^L)$ is the correlation value between the phenotypic array and the genotypic array that corresponds to locus L;

[0086] p(i) is a value of the i^{th} element of the phenotypic array;

[0087] g(i) is a value of the i^{th} element of the genotypic array;

[0088] <P> is a mean value of all elements in the phenotypic array;

[0089] $<G^L>$ is a mean value of all elements in the genotypic array; and

$$\sum^i = \sum_{i=1}^{N}.$$

[0090] It will be appreciated that the phenotypic and genotypic arrays can be compared in processing step **208** using any number of algorithms other than linear regression. For example, alternative methods for forming a correlation value in processing step **208** include, but are not limited to, regression analysis, regression analysis with data transformations, Pearson correlations, Spearman rank correlation, a regression tree and concomitant data reduction, partial least squares, and canonical analysis. (See e.g. Lui, "Statistical Genomics," CRC Press LLC, New York, 1998; Stuart & Ord, "Kendall's Advanced Theory of Statistics," Arnold, London, England, 1994).

[0091] While processing steps **202** through **206** have been described with reference to linear phenotypic and genotypic arrays, it will be appreciated that the methods of the present invention are not limited to the comparison of such arrays. Indeed, any form of data structure having elements that preserve the information in the above described matrices and arrays may be used. For example, rather than using the genotypic array described above, the individual variation matrices can be used. Further, rather than using the phenotypic array, a phenotypic distance matrix can be used.

[0092] Once a correlation value between the phenotypic data structure and a genotypic data structure that corresponds to a particular locus L has been formed, the correlation value is stored in processing step **210** so that it can be subsequently ranked with the correlation value of each of the other loci that are analyzed.

[0093] Processing step **212** is provided so that the procedure can be repeated in an iterative fashion for all suitable loci vectors L in genotypic database **52**. Thus, in processing step **212**, a decision is made whether to test an additional locus by asking whether all of the loci present in genotypic database **52 (FIG. 1)** have been tested. In one embodiment, when additional loci **54** are present in genotypic database **52**, processing step **212** returns a "yes" and the process continues by looping back to processing step **204** where an additional, untested locus is selected from genotypic database **52**.

[0094] When there are no additional loci to test (**212**-No), the correlation value for each of the comparisons of genotypic data structures to the phenotypic data structure are ranked with respect to each other in processing step **214**. In one embodiment, processing step **214** comprises the arrangement of the tested loci in a vector K according their correlation scores:

$$K=(L^t,\ L^u,\ L^v,\ \ldots)$$

[0095] where $c(P, G^{Lt}) \geqq c(P, G^{Lu}) \geqq c(P, G^{Lv}) \geqq \ldots$

[0096] In another embodiment of the present invention, processing step **214** includes the computation of (i) a mean correlation value that represents a mean of each correlation value formed during instances of processing step **208**; and (ii) a standard deviation of the mean correlation value based on each of the correlation values formed during instances of processing step **208**.

[0097] In processing step **216**, the genotypic data structures that achieve the highest correlation values are selected. Since each genotypic data structure corresponds to a particular locus in the genome, the selection process in processing step **216** results in the association of the phenotype with particular loci in the organism of interest. In one embodiment, the selection process in processing step **216** is performed by selecting genotypic data structures that form a

correlation value that is a predetermined number of standard deviations above the mean correlation value. Typically, the predetermined number is chosen so that a small percentage of the genome of the organism, such as five percent, will be selected during processing step **216**.

EXAMPLES

[**0098**] Building a Murine SNP Database.

[**0099**] The methods of the present invention are particularly useful in embodiments that make use of genetic information from inbred strains of an organism of interest. Thus, a genotypic database **52** was developed that contains allele information across 15 inbred strains. At Roche Bioscience, 293 SNPs at defined locations were identified in the mouse genome. The SNPs were identified by direct sequencing of PCR amplification products from defined chromosomal locations. This database also incorporates published allele information for 2848 SNPs, 45% of which are characterized in a subset of *M. Musculus* strains, and 55% of the SNPs are polymorphic between *M. castaneus* and one or more *M. musculus* subspecies (Lindblad-Toh, et al., Nature Genetics Apr;24, 381-386, 2000). User queries regarding SNPs found within a specified chromosomal region or between selected inbred strains are executed in real time and provided via a user interface **24**.

Example 1

Hypothetical Example of the Method for Prediction of QTL Regions

[**0100**] To aid in the understanding of the methods of the present invention, **FIG. 3** is provided. **FIG. 3** shows hypothetical comparisons, in accordance with the methods of the present invention, between three mouse strains (A, B, C) using SNP information found in the murine SNP database. Each of the two chromosomes sets for a given mouse strain is represented by a horizontal box along the horizontal axis of **FIG. 3**. Each chromosome set is characterized by the hatching type (horizontal, diagonal, and vertical). Chromosomes with the same hatching style in each of the mouse strains are identical. Cross hatched or diagonally hatched ovals respectively represent alleles at specific chromosomal positions. A dashed horizontal line is used to differentiate each of the mouse strains and the accompanying chart at the bottom of **FIG. 3**.

[**0101**] In the hypothetical example provided in **FIG. 3**, two of the three strains, (A) and (B), exhibit a similar phenotype. That is, strains A and B exhibit a similar phenotype (full size tail), while strain C has a different phenotype (short tail). SNP alleles at particular chromosomal regions are represented as cross hatched or diagonally hatched ovals. A series of pairwise comparisons, in accordance with the algorithm illustrated in **FIG. 2**, are made to establish the correlation value between the phenotype and genotype for each locus. In each of these series of pairwise comparisons, allelic differences in a respective segment of the chromosome of each of the mouse strains is correlated with the phenotypic difference between each mouse strain. Graphic analysis of the correlation data between the respective strains is shown at the bottom of **FIG. 3**. The analysis indicates that while most sites exhibit a negative correlation with respect to murine tail length, two chromosomal regions (**302**) and (**304**) have a strong positive correlation. In fact, **302** and **304** are the chromosomal regions predicted to have genes regulating tail length.

[**0102**] The following four examples, (Examples 2 through 5) are made with reference to **FIG. 4**. **FIG. 4** illustrates the correlation between the genotype and phenotype distributions for all 19 mouse autosomal chromosomes for a given trait. Loci are arranged proximal to distal for each chromosome. Each bar represents a 30 cM interval of the respective chromosome and neighboring bars are offset by 10 cM. Dotted line **402** represents a useful cutoff for analyzing the data, with the highest correlated ten percent of the genome being above this line.

Example 2

Predicting the Chromosomal Location of the MHC Complex

[**0103**] The methods of the present invention were used to predict the chromosomal location of the MHC complex, which has been mapped to murine chromosome 17, using the H2 haplotypes for the MHC K locus for 10 inbred strains (Anonymous, JAX Notes 475, 1998). Phenotypic distances for strains that shared a haplotype were set to zero, and a distance of one was used for strains of different haplotypes. The SNPs within and near the MHC region had a genotypic distribution which were highly correlated with the phenotypic distances; the correlation value for interval **440** (**FIG. 4A**) was 5.35 standard deviations above the average for all loci analyzed. There were no other peaks throughout the mouse genome that exhibited a comparable correlation with the phenotype. The computational analysis, executed in accordance with the methods of the present invention, excluded 96% of the mouse genome from consideration without missing the genomic region known to contain the MHC.

Example 3

Identification of the QTLs that Correspond to Allergic Asthma

[**0104**] The chromosomal positions that regulate susceptibility to experimental allergic asthma have been investigated using prior art techniques. For example, published analyses of intercross progeny between susceptible (A/J) and resistant (C3H/HeJ) mouse strains identified QTL intervals on chromosomes 2 and 7 (Ewart, et al., Am J Respir Cell Mol Biol 23, 537-545, 2000; Karp, et al., Nature Immunology 1, 221-226, 2000). The ability of the methods of the present invention to identify these chromosomal regions was investigated.

[**0105**] The phenotypic distance used to populate the phenotypic matrix was the absolute difference between the measured airway response after allergen-challenge for each strain pair. The experimentally identified QTL intervals on chromosomes 2 and 7 were among the strongest peaks identified by the methods of the present invention (**FIG. 4B**). The computational method excluded 80% of the mouse genome from consideration without missing the experimentally mapped QTL regions using airway responsiveness data from only 5 inbred mouse strains.

Example 4

Lifespan Data

[0106] Lifespan data for five mouse strains, which reflected susceptibility to T cell lymphoma, has been published (Chrisp et al., Veterinary Pathology 33, 735-743, 1996). Using conventional techniques, three susceptibility regions were experimentally identified by analysis of intercross progeny (Wielowieyski et al., Mammalian Genome 10, 623-627, 1999; Gilbert, et al., J.Virol. 67, 2083-2090, 1993; Mucenski et al., Molecular & Cellular Biology 6, 4236-4243, 1986; Mucenski et al, Molecular & Cellular Biology 8, 301-308, 1988); and all three regions were predicted by the computational genome scan (**FIG. 4C**). In this example, over ninety percent of the genome could be excluded from consideration by the computational method without overlooking any experimentally verified QTL interval.

Example 5

Retinal Ganglion Cells

[0107] In another example, the measured density of retinal ganglion cells was used as a phenotype. Using conventional techniques, the QTLs associated with this phenotype have been localized to chromosome 11 in the mouse genome (Williams et al., Journal of Neuroscience 18, 138-146, 1998). The experimentally verified QTL interval on chromosome 11 was contained in the chromosomal regions predicted by the methods of the present invention, while 96% of the mouse genome was excluded (**FIG. 4D**).

Example 6

Additional Phenotypic Traits

[0108] The ability of the computational method of the present invention to identify candidate chromosomal regions that are associated with six additional quantitative traits was performed. The chromosomal positions for these six additional quantitative traits are derived from published studies that provided mapped QTL intervals and phenotypic data across multiple inbred strains for each trait (Table 1). As shown in Table 1, a total of 10 QTLs from 6 published phenotypic studies are identified from the literature. Each QTL resides on a different chromosome. Centimorgan positions were interpreted from published marker locations on physical maps.

TABLE 1

Published chromosomal positions of QTLs that have been associated with particular phenotypes using conventional techniques

Phenotype	Chromosome (cM)	Notes
AHR	2 (23.5), 7 (1)	Allergen induced airway response (APTI)
Eye weight	5 (0–10)	Mouse eye weight (grams), day 75
Retinal anglion	11 (57.5)	Retinal ganglion cell #
Lymphoma	1 (62–73), 6 (30), 16 (50)	Tumor incidence, lifespan
MHC	17 (10)	H2 K serotyping
PKC	11 (66), 3 (16.4, 45)	PKC-α protein amount, activity

[0109] The ability of the methods of the present invention to correctly predict chromosomal regions containing experimentally verified QTL intervals associated with the six phenotypic traits is presented in Table 2.

TABLE 2

Summary of predictions made in accordance with the methods of the present invention

	Experimentally	Methods of the Present Invention		
Phenotype	Verified	Correct	Predicted	Threshold (%)
AHR	2	2	8	19
Eye weight	1	1	6	17
Ganglion	1	1	2	4
Lymphoma	3	3	4	8
MHC	1	1	1	2
PKC	2	2	6	2,11
Totals	10	10	27	

[0110] As shown in Table 2, the methods of the present invention identified all ten experimentally characterized QTL intervals. In addition, seventeen other chromosomal regions were predicted by this computational method. Whether these predicted regions affect phenotypic traits has not yet been experimentally verified. The threshold required for correct identification of a QTL varied from two percent to nineteen percent of the complete mouse genome.

[0111] The percentage of correct predictions as a function of the percentage of the mouse genome contained within the predicted chromosomal regions was examined. If predicted regions contained eighteen percent of the mouse genome (by selecting eighteen percent of the peaks with the highest correlation), all ten experimentally verified QTL intervals were correctly identified (**FIG. 5**). As the threshold was raised, limiting the number of predicted candidate chromosomal regions, the methods of the present invention missed some experimentally verified QTL intervals for these traits. When only three (or nine) percent of the genome was above the threshold, the method identified four (or seven) of the ten verified QTL intervals for these traits (**FIG. 5**).

[0112] When a genome-wide threshold of ten percent was used, the genomic region to search for candidate genes is computationally reduced by an order of magnitude. Since the average size of a predicted genomic region was 38 cM, the 1500 cM mouse genome could be subdivided into approximately forty regions. The computational method was used for seven different phenotypes, so approximately 280 genomic intervals (38-cM in size) were examined. This method correctly identified seven of ten experimentally validated QTL intervals, while missing three, at the ten percent genome-wide threshold. The algorithm further pre-

dicted 23 genomic intervals were involved in a phenotypic trait where no QTL had been experimentally characterized. Finally, the computational method and experimental analysis agreed on 240 loci that were not QTL intervals for the phenotypes examined. This data can be assembled into a 2×2 matrix to assess the ability of the computational method to predict QTL intervals. A Fisher Exact test yields a highly significant P value (7.0×10^{-6}) for the computationally predicted intervals.

[0113] In summary, the methods of the present invention were able to identify ten QTLs for seven phenotypic traits that had been previously identified by prior art techniques. Each of the experimentally verified QTL intervals was identified by the methods of the present invention. The genotypic array used to identify these chromosomal regions was derived from a murine SNP genotypic database. In each case, the conventionally identified QTL interval exhibited a computational SNP distribution that was highly correlated with the tested phenotype. The correlation was well above the mean value for the entire genome, and nine of ten were greater than a full standard deviation above the mean.

Example 7

Use of Alternative Genotypic Databases **52**

[0114] Although the examples provided herein utilize a genotypic database of 15 inbred mouse strains, other types of genotypic databases may be used. For example, suitable genotypic databases include various databases that have various types of gene expression data from platform types such as spotted microarray (microarray), high-density oligonucleotide array (HDA), hybridization filter (filter) and serial analysis of gene expression (SAGE) data.

[0115] As a proof of concept, 315 microsatellite polymorphisms were downloaded from the Center for Inherited Disease Research URL

[0116] http://www.cidr.jhmi.edu/download/CIDR_mouse.xls

[0117] Genotypic database **52** was populated in manner analogous to the case when SNP data was used to populate database **52**: if the polymorphisms matched between two mouse strains, a "0" was entered, if they differed, a "1" was entered. In this way, the number of differences between mouse strains was counted for a given locus. The remainder of the analysis was performed in accordance with the methods of the present invention. For this trial, the MHC locus was identified on chromosome 17. Although the QTL for the MHC region was not as clearly distinguished when using microsatellite information as it was for SNP data, it should be noted that the microsatellite data used for the trial was sparser than the information currently available in the mouse SNP database.

Example 8

Comparison of the Performance of a Genotypic Database **52** Populated with SNP Data to a Genotypic Database **52** Populated with Microsatellite Data

[0118] The genotypic database **52** populated with microsatellite data as described in Example 7 was compared to the previously described genotypic database **52** that contains allele information across 15 inbred strains for 287 SNPs at defined locations in the mouse genome. In this case, the phenotype is the formation of retinal ganglion cells in infant mice. The experimentally verified QTL that correlates with this phenotype is on chromosome 11. As illustrated in **FIG. 6**, the genotypic database **52** populated with the microsatellite information more strongly identifies the correct QTL peak than the genotypic database **52** populated with SNP data (4.2 standard deviations with microsatellites versus 2.3 standard deviations with SNPs). Furthermore, the results using the microsatellite data are less noisy than the results using the SNP data. See, for example, the reduced positive peak on chromosome 9 using the microsatellite data (702 versus 704).

Discussion

[0119] Computational analysis of genotypic databases **54** using phenotypic data from sources such as inbred parental strains and the methods of the present invention rapidly identifies candidate QTL intervals. This can eliminate many months to years of laboratory work required for generation, characterization and genotyping of intercross progeny. In effect, the methods of the present invention reduce the time required for QTL interval identification from many months to milliseconds.

[0120] There are several factors contributing to the successful QTL predictions by computational scanning of the murine SNP genotypic database using the methods of the present invention. The use of inbred mouse strains limits variability due to environment, and timed experimental intervention and sampling limits error in phenotypic assessment. The inbred strains are homozygous at all loci, which eliminates confounding effects due to heterozygosity found in human populations. However, there is no absolute requirement that inbred strains be used to populate genotypic database **52**.

[0121] The methods of the present invention will greatly accelerate analysis of complex traits and mammalian disease biology. Recently, there has been increased emphasis on using chemical mutagenesis in the mouse as a method for studying complex biology. This has occurred as a result of the difficulties noted by investigators searching for complex trait loci using standard methods for QTL analysis. For a review, see Nadeau and Frankel, Nature Genetics Aug;25, 381-384, 2000. However, analysis of genetic variation among existing inbred mouse strains can be markedly accelerated by application of the methods of the present invention. Of course, understanding the genetic basis of complex disease requires additional steps beyond computational prediction of genomic intervals. Specific gene candidates must be identified and evaluated before the underlying mutations can be identified and effective treatment strategies can be designed, tested in animal models, and developed for use with humans.

Alternative Embodiments

[0122] The foregoing descriptions of specific embodiments of the present invention are presented for purposes of illustration and description. They are not intended to be exhaustive or to limit the invention to the precise forms disclosed, obviously many modifications and variations are possible in view of the above teachings. For example, the

techniques of the invention may be applied using pooled or clustered genetic variation information as a source for the genotypic data structure or genetic variation information from individual samples. Similarly, the phenotypic information provided from sources, such as phenotypic data file **60**, may be in the form of pooled or clustered phenotypic data or phenotypic data from individual species. Furthermore, genotypic database **52** may represent inbred species of the organism of interest or randomized species of the organism of interest that have not been inbred. Because of the overwhelming homology between murine and human genomes, the examples provided herein clearly demonstrate that the methods of the present invention provide an invaluable tool for correlating human phenotypic traits with specific loci in the human genome.

[**0123**] While the examples provided herein describe the comparison of a plurality of genotypic data structures to a phenotypic data structure, one of skill in the art will appreciate that many other types of comparisons may be practiced in accordance with the present invention. For instance, consider the genotypic to phenotypic data structure comparison as a two-dimensional comparison. Higher dimensional comparisons than the two-dimensional comparison are possible. For instance, one embodiment of the present invention provides for a three dimensional comparison of the class: "genotypic data structure" versus "phenotypic data structure one" versus "phenotypic data structure two." Another example of a type of comparison within the scope of the present invention includes a comparison of "SNP genotypic data" to "disease phenotypic data" to "microarray data."

Conclusion

[**0124**] The embodiments were chosen and described in order to best explain the principles of the invention and its practical application, to thereby enable others skilled in the art to best utilize the invention and various embodiments with various modifications as are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the following claims and their equivalents.

We claim:

1. A method of associating a phenotype with one or more candidate chromosomal regions in a genome of an organism using a phenotypic data structure that represents a difference in a phenotype between different strains of said organism, said genome including a plurality of loci, said method comprising:

establishing a genotypic data structure, said genotypic data structure corresponding to a locus selected from said plurality of loci, said genotypic data structure representing a variation of at least one component of said locus between different strains of said organism;

comparing said phenotypic data structure to said genotypic data structure to form a correlation value; and

repeating said establishing and comparing steps for each locus in said plurality of loci, thereby identifying one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures that are compared to said phenotypic structure during said comparing step; wherein the loci that correspond to said one or more genotypic structures that form a high correlation value represent said one or more candidate chromosomal regions.

2. The method of claim 1, each element in said phenotypic structure representing a difference in a phenotype between different strains of said organism; wherein, for each element in said phenotypic structure, said different strains of said organism are selected from a plurality of strains of said organism.

3. The method of claim 2, wherein said difference in said phenotype is determined by a measurement of an attribute corresponding to said phenotype in different strains of said organism.

4. The method of claim 1, each element in said phenotypic structure representing a difference in said phenotype between a first cluster of strains of said organism and a different second cluster of strains of said organism; wherein, for each element in said phenotypic structure, said different first and second cluster of strains of said organism are selected from a plurality of clusters of strains of said organism.

5. The method of claim 1, each element in said genotypic structure representing a variation of at least one component of said locus between different strains of said organism; wherein, for each element in said genotypic structure, said different strains of said organism are selected from a plurality of strains of said organism.

6. The method of claim 1, each element in said genotypic structure representing a variation of at least one component of said locus between a first cluster of strains of said organism and a different second cluster of strains of said organism; wherein, for each element in said genotypic structure, said different first and second clusters of strains of said organism are selected from a plurality of strains of said organism.

7. The method of claim 1, wherein said correlation value is formed in accordance with the expression:

$$c(P, G^L) = \frac{\sum^i (p(i) - < P >)(g(i) - < G^L >)}{\{[\sum^i (p(i) - < P >)^2][\sum^i (g(i) - < G^L >)^2]\}^{1/2}}$$

where,

$c(P, G^L)$ is said correlation value;

p(i) is a value of the ith element of said phenotypic data structure;

g(i) is a value of the ith element of said genotypic data structure;

<P> is a mean value of all elements in said phenotypic data structure; and

$<G^L>$ is a mean value of all elements in said genotypic data structure.

8. The method of claim 1, wherein said correlation value is formed using an algorithm selected from the group consisting of regression analysis, regression analysis with data transformations, a Pearson correlation, a Spearman rank correlation, a regression tree and concomitant data reduction, partial least squares, and canonical analysis.

9. The method of claim 1, wherein said repeating step further comprises:

computing (i) a mean correlation value that represents a mean of each said correlation value formed during instances of said comparing step; and (ii) a standard deviation of said mean correlation value based on each said correlation value formed during instances of said comparing step;

wherein, said one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures compared to said phenotypic data structure during said comparing step are identified by selecting genotypic data structures that form a correlation value that is a predetermined number of standard deviations above said mean correlation value.

10. The method of claim 1, wherein each said variation in said genotypic data structure is obtained from a variation in a single nucleotide polymorphism database, a microsatellite marker database, a restriction fragment length polymorphism database, a short tandem repeat database, a sequence length polymorphism database, or an expression profile database.

11. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

a genotypic database for storing variations in genomic sequences of a plurality of strains of an organism;

a phenotypic data structure that represents a difference in a phenotype between different strains of said organism; and

a program module for associating a phenotype with one or more candidate chromosomal regions in a genome of said organism, said genome including a plurality of loci, said program module comprising:

instructions for establishing a genotypic data structure, said genotypic data structure corresponding to a locus selected from a plurality of loci, said genotypic data structure representing a variation of at least one component of said locus between different strains of said organism stored in said genotypic database;

instructions for comparing said phenotypic data structure to said genotypic data structure to form a correlation value; and

instructions for repeating said instructions for establishing and instructions for comparing for each locus in said plurality of loci, thereby identifying one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures that are compared to said phenotypic data structure by said instructions for comparing; wherein the loci that correspond to said one or more genotypic data structures that form a high correlation value represent said one or more candidate chromosomal regions.

12. The computer program product of claim 11, each element in said phenotypic data structure representing a difference in said phenotype between different strains of said organism; wherein, for each element in said phenotypic data structure, said different strains of said organism are selected from said plurality of strains of said organism represented in said genotypic database.

13. The computer program product of claim 12, wherein said difference in said phenotype is determined by a measurement of an attribute corresponding to said phenotype in said different strains of said organism that are represented in said genotypic database.

14. The computer program product of claim 11, each element in said phenotypic data structure representing a difference in said phenotype between a first cluster of strains of said organism and a different second cluster of strains of said organism; wherein, for each element in said phenotypic data structure, said different first and second cluster of strains of said organism are selected from a plurality of clusters of strains of said organism that are represented in said genotypic database.

15. The computer program product of claim 11, each element in said genotypic data structure representing a variation of at least one component of said locus between different strains of said organism; wherein, for each element in said genotypic data structure, said different strains of said organism are selected from said plurality of strains of said organism represented in said genotypic database.

16. The computer program product of claim 11, each element in said genotypic data structure representing a variation of at least one component of said locus between a first cluster of strains of said organism and a different second cluster of strains of said organism; wherein, for each element in said genotypic data structure, said different first and second clusters of strains of said organisms are selected from said plurality of strains of said organism represented in said genotypic database.

17. The computer program product of claim 11, wherein said instructions for comparing include instructions for forming said correlation value in accordance with the expression:

$$c(P, G^L) = \frac{\sum^i (p(i) - < P >)(g(i) - < G^L >)}{\{[\sum^i (p(i) - < P >)^2][\sum^i (g(i) - < G^L >)^2]\}^{1/2}}$$

where,

$c(P, G^L)$ is said correlation value;

p(i) is a value of the i^{th} element of said phenotypic data structure;

g(i) is a value of the i^{th} element of said genotypic data structure;

<P> is a mean value of all elements in said phenotypic data structure; and

$<G^L>$ is a mean value of all elements in said genotypic data structure.

18. The computer program product of claim 11, wherein said instructions for comparing include instructions for forming said correlation value by an algorithm selected from the group consisting of regression analysis, regression analysis with data transformations, a Pearson correlation, a Spearman rank correlation, a regression tree and concomitant data reduction, partial least squares, and canonical analysis.

19. The computer program product of claim 11, wherein said instructions for repeating further comprise:

instructions for computing (i) a mean correlation value that represents a mean of each said correlation value formed during instances of said instructions for comparing; and (ii) a standard deviation of said mean correlation value based on each said correlation value formed during instances of said instructions for comparing;

wherein, said one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures compared to said phenotypic data structure by said instructions for comparing are identified by selecting genotypic data structures that form a correlation value that is a predetermined number of standard deviations above said mean correlation value.

20. The computer program product of claim 11, wherein said genotypic database is a single nucleotide polymorphism database, a microsatellite marker database, a restriction fragment length polymorphism database, a short tandem repeat database, a sequence length polymorphism database, an expression profile database, or a DNA methylation database; and said variation in said genotypic data structure is obtained from said genotypic database.

21. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

a genotypic database for storing variations in genomic sequences of a plurality of strains of an organism;

a phenotypic data structure, each element in said phenotypic data structure representing a difference in said phenotype between different strains of said organism; and

a program module for associating a phenotype with one or more candidate chromosomal regions in a genome of said organism, said genome including a plurality of loci, said program module comprising:

instructions for identifying a genotypic data structure, said genotypic data structure corresponding to a locus selected from said plurality of loci, each element in said genotypic data structure representing a variation of at least one component of said locus between different strains of said organism;

instructions for comparing said phenotypic data structure to said genotypic data structure to form a correlation value; and

instructions for repeating said instructions for identifying and said instructions for comparing, for each locus in said plurality of loci, thereby identifying one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures that are compared to said phenotypic data structure by said instructions for comparing; wherein the loci that correspond to said one or more genotypic data structures that form a high correlation value represent said one or more candidate chromosomal regions.

22. A computer system for associating a phenotype with one or more candidate chromosomal regions in a genome of an organism, said genome including a plurality of loci, the computer system comprising:

a central processing unit;

a memory, coupled to the central processing unit, the memory storing:

a genotypic database for storing variations in genomic sequences of a plurality of strains of said organism;

a phenotypic data structure that represents a difference in a phenotype between different strains of said organism; and

a program module, said program module comprising:

instructions for establishing a genotypic data structure, said genotypic data structure corresponding to a locus selected from a plurality of loci, said genotypic data structure representing a variation of at least one component of said locus between different strains of said organism stored in said genotypic database;

instructions for comparing said phenotypic data structure to said genotypic data structure to form a correlation value; and

instructions for repeating said instructions for establishing and said instructions for comparing, for each locus in said plurality of loci, thereby identifying one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures that are compared to said phenotypic data structure by said instructions for comparing; wherein the loci that correspond to said one or more genotypic data structures that form a high correlation value represent said one or more candidate chromosomal regions.

23. The computer system of claim 22, each element in said phenotypic data structure representing a variation in said phenotype between different strains of said organism; wherein, for each element in said phenotypic data structure, said different strains of said organism are selected from said plurality of strains of said organism represented in said genotypic database.

24. The computer system of claim 23, wherein said difference in a phenotype is determined by a measurement of an attribute corresponding to said phenotype in said different strains of said organism that are represented in said genotypic database.

25. The computer system of claim 22, each element in said phenotypic data structure representing a variation in said phenotype between a first cluster of strains of said organism and a different second cluster of strains of said organism; wherein, for each element in said phenotypic data structure, said different first and second cluster of strains of said organism are selected from a plurality of clusters of strains of said organism that are represented in said genotypic database.

26. The computer system of claim 22, each element in said genotypic data structure representing a variation of at least one component of said locus between different strains of said organism; wherein, for each element in said genotypic data structure, said different strains of said organism

are selected from said plurality of strains of said organism represented in said genotypic database.

27. The computer system of claim 22, each element in said genotypic data structure representing a variation of at least one component of said locus between a first cluster of strains of said organism and a different second cluster of strains of said organism; wherein, for each element in said genotypic data structure, said different first and second clusters of strains of said organisms are selected from said plurality of strains of said organism represented in said genotypic database.

28. The computer system of claim 22, wherein said instructions for comparing include instructions for forming said correlation value in accordance with the expression:

$$c(P, G^L) = \frac{\sum^i (p(i) - < P >)(g(i) - < G^L >)}{\{[\sum^i (p(i) - < P >)^2][\sum^i (g(i) - < G^L >)^2]\}^{1/2}}$$

where,

$c(P, G^L)$ is said correlation value;

p(i) is a value of the i^{th} element of said phenotypic data structure;

g(i) is a value of the i^{th} element of said genotypic data structure;

<P> is a mean value of all elements in said phenotypic data structure; and

$<G^L>$ is a mean value of all elements in said genotypic data structure.

29. The computer system of claim 22, wherein said instructions for comparing include instructions for forming said correlation value by an algorithm selected from the group consisting of regression analysis, regression analysis with data transformations, a Pearson correlation, a Spearman rank correlation, a regression tree and concomitant data reduction, partial least squares, and canonical analysis.

30. The computer system of claim 22, wherein said instructions for repeating further comprise:

instructions for computing (i) a mean correlation value that represents a mean of each said correlation value formed during instances of said instructions for comparing; and (ii) a standard deviation of said mean correlation value based on each said correlation value formed during instances of said instructions for comparing;

wherein, said one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures compared to said phenotypic data structure by said instructions for comparing are identified by selecting genotypic data structures that form a correlation value that is a predetermined number of standard deviations above said mean correlation value.

31. The computer system of claim 22, wherein said genotypic database is a single nucleotide polymorphism database, a microsatellite marker database, a restriction fragment length polymorphism database, a short tandem repeat database, a sequence length polymorphism database, an expression profile database, or a DNA methylation database; and said variation in said genotypic data structure is obtained from said genotypic database.

32. A method of associating a phenotype with one or more candidate chromosomal regions in a genome of an organism using a phenotypic data structure that represents alterations in phenotypes between different strains in a plurality of strains of said organism,

said phenotypic data structure including a description of each said alteration and individual elements of said phenotypic data structure including an amount of alteration between different strains of said organism selected from said plurality of strains of said organism,

said genome including a plurality of loci, each said loci representing one or more positions within said genome,

said method comprising:

establishing a unique individual variation matrix for each said one or more positions represented by said loci, wherein an element within each said unique individual variation matrix represents an allelic comparison between different strains of said organism that are selected from said plurality of strains of said organism;

summing corresponding elements in each said unique individual matrix to form a genotypic data structure;

comparing said phenotypic data structure to said genotypic data structure to form a correlation value; and

repeating said establishing, summing and comparing steps, for each locus in said plurality of loci, thereby identifying one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures that are compared to said phenotypic data structure during said comparing step; wherein the loci that correspond to said one or more genotypic data structures that form a high correlation value represent said one or more candidate chromosomal regions associated with said phenotype.

33. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

a genotypic database for storing variations in genomic sequences of a plurality of strains of an organism;

a phenotypic data structure that represents alterations in phenotypes between different strains of said organism selected from said plurality of strains of said organism, said phenotypic data structure including a description of each said alteration and individual elements of said phenotypic data structure including an amount of alteration between different strains in said plurality of strains of said organism; and

a program module for associating a phenotype with one or more candidate chromosomal regions in a genome of said organism, said genome including a plurality of loci, each said loci representing one or more positions within said genome, said program module comprising:

instructions for establishing a unique individual variation matrix for each said one or more positions

represented by said loci, wherein an element within each said unique individual variation matrix represents an allelic comparison of values stored in said genotypic database between different strains of said organism that are selected from said plurality of strains of said organism;

instructions for summing corresponding elements in each said unique individual matrix to form a genotypic data structure;

instructions for comparing said phenotypic data structure to said genotypic data structure to form a correlation value; and

instructions for repeating said instructions for establishing, summing and comparing, for each locus in said plurality of loci, thereby identifying one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures that are compared to said phenotypic data structure during said comparing step; wherein the loci that correspond to said one or more genotypic data structures that form a high correlation value represent said one or more candidate chromosomal regions associated with said phenotype.

34. A computer system for associating a phenotype with one or more candidate chromosomal regions in a genome of an organism, said genome including a plurality of loci, each said loci representing one or more positions within said genome, said program module comprising:

a central processing unit;

a memory, coupled to the central processing unit, the memory storing:

a genotypic database for storing variations in genomic sequences of a plurality of strains of said organism;

a phenotypic data structure that represents alterations in phenotypes between different strains in said plurality of strains of said organism, said phenotypic data structure including a description of each said alteration and individual elements of said phenotypic data structure including an amount of alteration between different strains in said plurality of strains of said organism; and

a program module, said program module comprising:

instructions for establishing a unique individual variation matrix for each said one or more positions represented by said loci, wherein an element within each said unique individual variation matrix represents an allelic comparison of values stored in said genotypic database between different strains of said organism that are selected from said plurality of strains of said organism;

instructions for summing corresponding elements in each said unique individual matrix to form a genotypic data structure;

instructions for comparing said phenotypic data structure to said genotypic data structure to form a correlation value; and

instructions for repeating said instructions for establishing, summing and comparing, for each locus in said plurality of loci, thereby identifying one or more genotypic data structures that form a high correlation value relative to all other genotypic data structures that are compared to said phenotypic data structure during said comparing step; wherein the loci that correspond to said one or more genotypic data structures that form a high correlation represent said one or more candidate chromosomal regions associated with said phenotype.

* * * * *