(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2005-527904
(P2005-527904A)
(43)公表日　平成17年9月15日(2005.9.15)

(51)Int.Cl.[7]	F I			テーマコード（参考）
G06F 19/00	G06F	19/00	600	4B063
C12N 15/00	G01N	33/53	D	5B075
G01N 33/53	G01N	33/561		
G01N 33/561	G01N	37/00	102	
G01N 37/00	G06F	17/30	170F	

審査請求 未請求　予備審査請求 有　（全 126 頁）　最終頁に続く

(21)出願番号　特願2004-507945 (P2004-507945)
(86)(22)出願日　平成15年5月20日 (2003.5.20)
(85)翻訳文提出日　平成17年1月17日 (2005.1.17)
(86)国際出願番号　PCT/US2003/015768
(87)国際公開番号　WO2003/100557
(87)国際公開日　平成15年12月4日 (2003.12.4)
(31)優先権主張番号　60/382,036
(32)優先日　平成14年5月20日 (2002.5.20)
(33)優先権主張国　米国(US)
(31)優先権主張番号　60/460,304
(32)優先日　平成15年4月2日 (2003.4.2)
(33)優先権主張国　米国(US)

(71)出願人　504291720
ロゼッタ　インファーマティクス　エルエルシー
アメリカ合衆国　９８０３４　ワシントン州，カークランド，エヌ．イー．，１１５　ティーエイチ　アベニュー　１２０４０
(74)代理人　100091096
弁理士　平木　祐輔
(74)代理人　100096183
弁理士　石井　貞次
(74)代理人　100118773
弁理士　藤田　節

最終頁に続く

(54)【発明の名称】複雑性疾患を構成疾患に細分するコンピュータ・システムおよび方法

(57)【要約】
集団内の複数の生物が示す複合形質に対する量的形質遺伝子座を特定する方法。この集団を、分類体系を用いて複数の亜集団に分割する。この集団について知られていることに応じて、教師付き分類または教師なし分類を使用する。この分類体系は、集団内の各生物から得られる複数の細胞構成成分測定値から誘導される。複数の亜集団中の各亜集団に対して、複合形質に対する1個または複数の量的形質遺伝子座を特定するために、亜集団についての量的遺伝解析を実施する。

【特許請求の範囲】
【請求項１】
集団内の複数の生物が示す複合形質の量的形質遺伝子座を特定する方法であって、
(a)前記集団内の各生物を前記亜集団の少なくとも1個に分類する分類体系を用いて、前記集団を複数の亜集団に分割するステップであって、前記分類体系が、前記各生物から得られる複数の各細胞構成成分の各々に対する複数の細胞構成成分測定値から誘導されるステップと、
(b)前記複数の亜集団中の少なくとも1個の亜集団に対して、前記複合形質に対する前記量的形質遺伝子座を特定するために、前記亜集団について量的遺伝解析を実施するステップとを含む、方法。
【請求項２】
前記各生物から得られる前記細胞構成成分測定値が転写状態測定値または翻訳状態測定値である、請求項1に記載の方法。
【請求項３】
前記翻訳状態測定値を抗体アレイまたは二次元ゲル電気泳動を用いて実施する、請求項2に記載の方法。
【請求項４】
前記細胞構成成分が複数の代謝産物を含み、前記複数の細胞構成成分測定値が細胞の表現型技術によって得られる、請求項1に記載の方法。
【請求項５】
前記細胞の表現型技術がメタボロミクス技術を含み、前記各生物における複数の代謝産物レベルが測定される、請求項4に記載の方法。
【請求項６】
前記代謝産物がアミノ酸、金属、可溶性糖または複合糖質を含む、請求項5に記載の方法。
【請求項７】
前記複数の代謝産物レベルを、熱分解質量分析法、フーリエ変換赤外分光法、ラマン分光法、ガスクロマトグラフィー-質量分析法、キャピラリー電気泳動法、高圧液体クロマトグラフィー/質量分析法(HPLC/MS)、液体クロマトグラフィー(LC)-エレクトロスプレー質量分析法、またはcap-LC-タンデム・エレクトロスプレー質量分析法を使用して測定する、請求項5に記載の方法。
【請求項８】
前記複数の細胞構成成分測定値が、遺伝子発現レベル、mRNA存在量、タンパク質発現レベルまたは代謝産物レベルを含む、請求項1に記載の方法。
【請求項９】
前記複合形質が、前記集団中の不完全浸透率を示す対立遺伝子を特徴とする、請求項1に記載の方法。
【請求項１０】
前記複合形質が、前記集団内の生物が罹る疾患であり、前記生物が前記疾患の素因となる対立遺伝子を受け継いでいない、請求項1に記載の方法。
【請求項１１】
前記複合形質が、前記複数の生物のゲノム中の複数の異なる遺伝子のいずれかが突然変異するときに生じる、請求項1に記載の方法。
【請求項１２】
前記複合形質が、前記複数の生物のゲノム中の複数の遺伝子の突然変異が同時に存在することを必要とする、請求項1に記載の方法。
【請求項１３】
前記複合形質が、前記集団における高頻度の疾患原因対立遺伝子に関連する、請求項1に記載の方法。
【請求項１４】

前記複合形質が、単一の遺伝子座に起因し得るメンデルの劣性遺伝または優性遺伝を示さない表現型である、請求項1に記載の方法。

【請求項１５】

前記複合形質が、喘息、血管拡張性失調症、双極性障害、癌、一般的な遅発性アルツハイマー病、糖尿病、心疾患、遺伝性早期発症型アルツハイマー病、遺伝性非腺腫性大腸癌、高血圧、感染症、若年成人発症型糖尿病、真性糖尿病、片頭痛、非アルコール性脂肪肝、非アルコール性脂肪性肝炎、インシュリン非依存性糖尿病、肥満、多発性嚢胞腎、乾せん、精神分裂病または色素性乾皮症である、請求項1に記載の方法。

【請求項１６】

前記各生物から得られる前記複数の細胞構成成分測定値が、前記各生物中の10個以上の細胞構成成分の細胞構成成分レベルの測定値を含む、請求項1に記載の方法。

【請求項１７】

前記各生物から得られる前記複数の細胞構成成分測定値が、前記各生物中の1000個以上の細胞構成成分レベルの細胞構成成分レベルの測定値を含む、請求項1に記載の方法。

【請求項１８】

前記分割ステップが、クラス予測変数が利用可能であるかどうかを判定するステップを含み、

クラス予測変数が利用可能なときに、教師付き分類体系を使用して前記集団内の各生物を前記複数の亜集団内の1個の亜集団に分類し、

クラス予測変数が利用不可能なときに、教師なし分類体系を使用して前記集団内の各生物を前記複数の亜集団内の1個の亜集団に分類する、請求項1に記載の方法。

【請求項１９】

前記分類体系が教師付き分類体系である、請求項1に記載の方法。

【請求項２０】

前記分類体系が教師なし分類体系である、請求項1に記載の方法。

【請求項２１】

前記教師付き分類体系が線形判別分析または線形回帰法を使用する、請求項18または19に記載の方法。

【請求項２２】

前記線形回帰法が多重線形回帰、部分最小二乗回帰または主成分回帰である、請求項21に記載の方法。

【請求項２３】

前記教師なし分類体系が、階層型クラスター分析、非階層型クラスター分析、人工ニューラル・ネットワークおよび自己組織化地図からなる群から選択される、請求項18または20に記載の方法。

【請求項２４】

前記教師なし分類体系が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムを使用して、(i)前記複数の生物中の1つの生物から得られる複数の細胞構成成分測定値と、(ii)前記複数の生物中の別の生物から得られる複数の細胞構成成分測定値との類似度を求める階層型クラスター分析である、請求項23に記載の方法。

【請求項２５】

前記教師なし分類体系が、凝縮型クラスタリング、多形質分割型クラスタリングおよび単形質分割型クラスタリングからなる群から選択される階層型クラスター分析である、請求項23に記載の方法。

【請求項２６】

前記階層型クラスター分析が、ピアソン相関係数、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量または二乗ピアソン相関係数を用いて、(i)前記複数の生物中の1つの生物から得られる複数の細胞構成成分測定値と、(ii)前記複数の生物中の別の生物から得られる複数の細胞構成成分測定値との類似度を求める凝縮型

クラスタリングである、請求項25に記載の方法。
【請求項２７】
前記教師なし分類体系が、K平均クラスタリング、ファジーk平均クラスタリングおよびJarvis-Patrickクラスタリングからなる群から選択される非階層型クラスター分析である、請求項23に記載の方法。
【請求項２８】
前記教師なし分類体系が、Kohonen人工ニューラル・ネットワークまたは自己連想ニューラル・ネットワークである人工ニューラル・ネットワークである、請求項23に記載の方法。
【請求項２９】
前記分割ステップが、さらに、前記複数の亜集団への前記集団の分割を検定するステップを含む、請求項1に記載の方法。
【請求項３０】
前記量的遺伝解析を、連鎖解析、前記複数の細胞構成成分測定値を表現型形質として使用する量的形質遺伝子座(QTL)解析方法、および関連解析からなる群から選択される方法を使用して実施する、請求項1に記載の方法。
【請求項３１】
前記量的遺伝解析を前記QTL解析によって実施し、前記QTL解析方法が、
(a)複数のQTL解析から得られるQTLデータをクラスター化してQTL相互作用地図を作成するステップであって、
前記複数のQTL解析の各QTL解析を、前記QTLデータを作成するために、前記複数の生物のゲノム中の複数の遺伝子中の遺伝子Gに対して、遺伝マーカー地図および量的形質を用いて実施し、各QTL解析では、前記量的形質が、前記複数の生物中の各生物の前記QTL解析を実施する前記遺伝子Gの発現統計量を含み、
前記遺伝マーカー地図を、前記複数の生物に関連する遺伝マーカー・セットから構築するステップと、
(b)前記QTL相互作用地図を解析して、前記量的形質に関連する前記QTLを特定するステップとを含む、請求項30に記載の方法。
【請求項３２】
前記クラスター化ステップの前に、さらに、前記複数の生物に関連する前記遺伝マーカー・セットから前記遺伝マーカー地図を作成するステップを含む、請求項31に記載の方法。
【請求項３３】
前記クラスター化ステップの前に、さらに、前記複数のQTL解析の前記各QTL解析を実施するステップを含む、請求項31に記載の方法。
【請求項３４】
前記遺伝子Gの前記発現統計量を、前記複数の生物中の各生物から得られる前記遺伝子Gの発現レベルの測定値を変換するステップを含む方法によって計算する、請求項31に記載の方法。
【請求項３５】
前記遺伝子Gの発現レベルの測定値を変換する前記ステップが、前記発現統計量を作成するために、前記遺伝子Gの前記発現レベルの測定値を正規化するステップを含む、請求項34に記載の方法。
【請求項３６】
前記発現統計量を作成するために前記遺伝子Gの前記発現レベルの測定値を正規化するステップを、強度のZ-スコア、強度中央値、強度中央値の対数、強度のZ-スコア標準偏差対数、対数強度のZ-スコア平均絶対偏差、較正DNA遺伝子セット、ユーザー正規化遺伝子セット、強度中央値の比率補正、および強度バックグラウンド補正からなる群から選択される正規化技術によって実施する、請求項35に記載の方法。
【請求項３７】

前記各QTL解析が、
(i)前記複数の生物のゲノム中の位置とQTL解析に使用する量的形質との関連を検定するステップと、
(ii)前記ゲノム中の前記位置をある量だけ進めるステップと、
(iii)前記ゲノムのすべてまたは一部を検定するまでステップ(i)と(ii)を繰り返すステップとを含む、請求項31に記載の方法。
【請求項３８】
前記量が100センチモルガン未満である、請求項37に記載の方法。
【請求項３９】
前記量が10センチモルガン未満である、請求項37に記載の方法。
【請求項４０】
前記量が5センチモルガン未満である、請求項37に記載の方法。
【請求項４１】
前記量が2.5センチモルガン未満である、請求項37に記載の方法。
【請求項４２】
各QTL解析から生成する前記QTLデータが、前記各位置において計算される統計スコアを含む、請求項37に記載の方法。
【請求項４３】
前記染色体において検定される各量的形質のQTLベクトルを作成するステップであって、前記QTLベクトルが、前記量的形質に対応する前記QTL解析によって検定される各位置の統計スコアを含むステップをさらに含む、請求項37に記載の方法。
【請求項４４】
QTLデータをクラスター化する前記ステップが、前記各QTLベクトルをクラスター化するステップを含む、請求項43に記載の方法。
【請求項４５】
前記クラスター化のための基礎として使用される類似度計測量が、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量、ピアソン相関係数または二乗ピアソン相関係数であって、QTLベクトル対間で計算される、請求項43に記載の方法。
【請求項４６】
QTLデータをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップ、k平均法を適用するステップ、ファジーk平均法を適用するステップ、Jarvis-Patrickクラスタリングを適用するステップ、自己組織化地図技術を適用するステップ、またはニューラル・ネットワーク技術を適用するステップを含む、請求項31または44に記載の方法。
【請求項４７】
QTLデータをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が凝縮型クラスタリング手順である、請求項46に記載の方法。
【請求項４８】
前記凝縮型クラスタリング手順が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである、請求項47に記載の方法。
【請求項４９】
前記階層型クラスタリング法が分割型クラスタリング手順である、請求項46に記載の方法。
【請求項５０】
前記変換ステップによって作成される各発現統計量から遺伝子発現クラスター地図を構築するステップをさらに含む、請求項35に記載の方法。
【請求項５１】

遺伝子発現クラスター地図を作成する前記ステップが、
複数の遺伝子発現ベクトルを作成するステップであって、前記複数の遺伝子発現ベクトル中の各遺伝子発現ベクトルが、複数の生物の各々における前記複数の遺伝子中の1個の遺伝子の発現レベルの測定値であるステップと、
複数の相関係数を計算するステップであって、前記複数の相関係数の各相関係数を前記複数の遺伝子発現ベクトル中の遺伝子発現ベクトル対間で計算するステップと、
前記遺伝子発現クラスター地図を作成するために、前記複数の相関係数に基づいて前記複数の遺伝子発現ベクトルをクラスター化するステップとを含む、請求項50に記載の方法。

【請求項５２】
前記QTL相互作用地図を解析する前記ステップが、候補経路群を得るためにQTL相互作用地図を選別するステップを含み、前記選別ステップが、前記遺伝子発現クラスター地図中の前記候補経路群においてQTLを特定するステップを含む、請求項51に記載の方法。

【請求項５３】
前記複数の相関係数の各相関係数がピアソン相関係数である、請求項51に記載の方法。

【請求項５４】
遺伝子発現クラスター地図を作成する前記ステップが、
複数の遺伝子発現ベクトルを作成するステップであって、前記複数の遺伝子発現ベクトル中の各遺伝子発現ベクトルが前記複数の遺伝子中の1個の遺伝子を表すステップと、
複数の計測量を計算するステップであって、前記複数の計測量の各計測量を前記複数の遺伝子発現ベクトル中の遺伝子発現ベクトル対間で計算するステップと、
前記遺伝子発現クラスター地図を作成するために、前記複数の計測量に基づいて前記複数の遺伝子発現ベクトルをクラスター化するステップとを含む、請求項50に記載の方法。

【請求項５５】
前記各計測量が、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量、ピアソン相関係数および二乗ピアソン相関係数からなる群から選択される、請求項54に記載の方法。

【請求項５６】
前記複数の遺伝子発現ベクトルをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップ、k平均法を適用するステップ、ファジーk平均法を適用するステップ、Jarvis-Patrickクラスタリングを適用するステップ、自己組織化地図技術を適用するステップ、またはニューラル・ネットワーク技術を適用するステップを含む、請求項51または54に記載の方法。

【請求項５７】
前記複数の遺伝子発現ベクトルをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が凝縮型クラスタリング手順である、請求項56に記載の方法。

【請求項５８】
前記凝縮型クラスタリング手順が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである、請求項57に記載の方法。

【請求項５９】
前記複数の遺伝子発現ベクトルをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が分割型クラスタリング手順である、請求項56に記載の方法。

【請求項６０】
前記QTL相互作用地図を解析する前記ステップが、候補経路群を得るために前記QTL相互作用地図を選別するステップを含む、請求項31に記載の方法。

【請求項６１】
前記候補経路群を得るための前記選別ステップが、前記QTL相互作用地図中の別のQTLと

最も強く相互作用する前記候補経路群のQTLを選択するステップを含む、請求項60に記載の方法。

【請求項６２】

前記QTL相互作用地図中の別のQTLと最も強く相互作用する前記QTLが、前記量的形質遺伝子座相互作用地図中のQTL間で計算される全相関係数の75%よりも高い相関係数を、前記量的形質遺伝子座相互作用地図中の別のQTLと共有する前記QTL相互作用地図中のQTLである、請求項61に記載の方法。

【請求項６３】

前記候補経路群を構成する各QTLが前記候補経路群に属する程度を検定するために、多変量統計モデルを前記候補経路群に適合させるステップをさらに含む、請求項61に記載の方法。

【請求項６４】

前記多変量統計モデルが複数の量的形質を同時に考慮する、請求項63に記載の方法。

【請求項６５】

前記多変量統計モデルが、前記候補経路群中のQTL間のエピスタシス相互作用を探索する、請求項63に記載の方法。

【請求項６６】

前記遺伝マーカー・セットが、一塩基多型(SNP)、ミクロサテライト・マーカー、制限断片長多型、短鎖縦列反復、DNAメチル化マーカーまたは配列長多型を含む、請求項31に記載の方法。

【請求項６７】

家系データを請求項1のステップ(b)に使用し、前記家系データが、前記複数の生物中の各生物間の1つまたは複数の関係を示す、請求項31に記載の方法。

【請求項６８】

家系データを請求項1のステップ(b)に使用し、前記統計スコアがロッド・スコアである、請求項42または43に記載の方法。

【請求項６９】

前記複数の生物がヒトである、請求項1に記載の方法。

【請求項７０】

前記複数の生物がF_2集団を含み、前記複数の生物中の各生物間の前記1つまたは複数の関係が、前記複数の生物中のどの生物が前記F_2集団のメンバーであるかを示す、請求項67に記載の方法。

【請求項７１】

QTLデータをクラスター化する前記ステップが、前記複数の生物に関連する前記遺伝マーカー・セットから前記遺伝マーカー地図を作成するステップと、前記複数のQTL解析の各QTL解析を実施するステップとを含む、請求項34に記載の方法。

【請求項７２】

前記分割ステップ(a)が、

(i)前記複数の生物のすべてまたは一部の表現型データを用いて、前記集団を複数の表現型群に区分化するステップと、

(ii)極端な表現型を示す前記複数の表現型群中の1組の極端生物を特定するステップと、

(iii)前記複数の細胞構成成分中の細胞構成成分を特定するステップであって、特定された各細胞構成成分が、前記極端生物セットから得られた前記各細胞構成成分の細胞構成成分測定値が前記複数の表現型群のすべてまたは一部を識別する特性を有するステップと、

(iv)前記特定された細胞構成成分のすべてまたは一部から導出される確率分布を用いて分類子を構築するステップとを含む、請求項1に記載の方法。

【請求項７３】

前記表現型データがバイナリ・イベントを含む、請求項72に記載の方法。

【請求項７４】
前記表現型データが、前記集団内の各生物の1個を超える表現型測定値を含む、請求項72に記載の方法。
【請求項７５】
前記表現型データが前記複数の生物中の各生物が形質を示すかどうかに関する判定を含み、前記区分化ステップ(i)が、前記生物が前記形質を示すときに第1の表現型群中の前記複数の生物中の1つの生物を配置するステップと、前記生物が前記形質を示さないときに第2の表現型群中の前記複数の生物中の1つの生物を配置するステップとを含む、請求項72に記載の方法。
【請求項７６】
前記表現型データが、前記複数の生物のすべてまたは一部に対して作成される複数の表現型測定値を含み、前記区分化ステップ(i)が、
(A)複数の表現型ベクトルを作成するステップであって、前記複数の表現型ベクトル中の各表現型ベクトルが前記複数の生物中の1つの生物に対応し、前記複数の表現型ベクトル中の各表現型ベクトルが前記各表現型ベクトルに対応する前記生物から得られる複数の表現型測定値を含むステップと、
(B)前記複数の表現型ベクトルを複数のクラスターにクラスター化するステップであって、前記複数のクラスター中の各クラスターが、前記複数の表現型群中の1個の表現型群を表すステップとを含む、請求項72に記載の方法。
【請求項７７】
前記クラスター化ステップが、階層型クラスタリング法、k平均法、ファジーk平均法、Jarvis-Patrickクラスタリング、自己組織化地図技術またはニューラル・ネットワーク技術を含む、請求項76に記載の方法。
【請求項７８】
前記クラスター化ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が凝縮型クラスタリング手順である、請求項76に記載の方法。
【請求項７９】
前記凝縮型クラスタリング手順が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである、請求項78に記載の方法。
【請求項８０】
前記階層型クラスタリング法が分割型クラスタリング手順である、請求項78に記載の方法。
【請求項８１】
生物が、前記集団が示す表現型に関して前記集団の上位30パーセンタイルまたは下位30パーセンタイルであるときに、前記極端な表現型である、請求項72に記載の方法。
【請求項８２】
生物が、前記集団が示す表現型に関して前記集団の上位10パーセンタイルまたは下位10パーセンタイルであるときに、前記極端な表現型である、請求項72に記載の方法。
【請求項８３】
前記極端生物セットが5個体を超える生物である、請求項72に記載の方法。
【請求項８４】
前記極端生物セットが2～100個体の生物である、請求項72に記載の方法。
【請求項８５】
前記極端生物セットが1000個未満の生物である、請求項72に記載の方法。
【請求項８６】
前記特定するステップ(iii)が、所定の細胞構成成分の複数の細胞構成成分測定値をt検定にかけるステップを含み、前記複数の細胞構成成分測定値が前記極端生物セットから得られる、請求項72に記載の方法。
【請求項８７】

前記特定するステップ(iii)が、前記複数の細胞構成成分内の特定された細胞構成成分群を多変量解析にかけるステップを含む、請求項72に記載の方法。

【請求項８８】

ステップ(iii)で特定された前記細胞構成成分を、前記構築ステップ(iv)の前に削減する、請求項72に記載の方法。

【請求項８９】

ステップ(iii)で特定された前記細胞構成成分を、段階的回帰、総当り回帰、主成分分析または重判別分析によって削減する、請求項88に記載の方法。

【請求項９０】

ステップ(iii)で特定された前記細胞構成成分を、確率論的検索方法によって削減する、請求項88に記載の方法。

【請求項９１】

前記確率論的検索方法がシミュレーテッド・アニーリングまたは遺伝アルゴリズムである、請求項90に記載の方法。

【請求項９２】

ステップ(iii)で特定された前記細胞構成成分をクラスタリングによって削減するステップであって、前記特定された細胞構成成分ではなく前記クラスタリングによって生成したクラスターを前記構築ステップ(iv)に使用する、請求項88に記載の方法。

【請求項９３】

前記構築ステップ(iv)が、前記確率分布を用いてニューラル・ネットワークを訓練するステップを含む、請求項72に記載の方法。

【請求項９４】

前記構築ステップ(iv)が、前記確率分布が演繹的情報として役立つベイズの決定理論を使用するステップを含む、請求項72に記載の方法。

【請求項９５】

前記構築ステップ(iv)が、線形判別分析、線形プログラミング・アルゴリズムまたはサポート・ベクトル・マシンを使用するステップを含む、請求項72に記載の方法。

【請求項９６】

前記分類体系が、前記分類子を用いて前記集団のすべてまたは一部を分類するステップを含む、請求項72に記載の方法。

【請求項９７】

量的遺伝解析に使用する複数の亜集団を得るために同じ種の複数の生物Sを細分する方法であって、前記複数の生物S中の1つまたは複数の生物が複合形質を示し、

(a)前記複合形質に関してそれぞれが独立した極端である、前記複数の生物S内の2群以上の生物を特定するステップと、

(b)前記複数の生物S内の前記2群以上の生物を識別することができる1組の細胞構成成分Cを決定するステップと、

(c)前記細胞構成成分Cセット中の各細胞構成成分iに対して、前記複合形質と関連する第1のQTLと相互作用または重複するQTLを有する1個または複数の細胞構成成分を特定するために、前記複数の生物Sの少なくとも一部の各生物からそれぞれ測定される前記細胞構成成分iの量を量的形質として用いて、前記細胞構成成分iについてQTL解析を実施するステップと、

(d)ステップ(c)で特定された各細胞構成成分の前記測定量に基づいて前記複数の生物Sをクラスター化し、それによって前記複数の亜集団を得るステップとを含む、方法。

【請求項９８】

前記複合形質に関連する前記第1のQTLを連鎖解析または関連解析によって特定する、請求項97に記載の方法。

【請求項９９】

前記複合形質のQTLを特定するために、前記複数の亜集団中の1個の亜集団について一連のQTL解析を実施するステップであって、前記一連のQTL解析の前記各QTL解析が前記細胞

構成成分Cセット中の1個の細胞構成成分の測定量を量的形質として使用し、前記細胞構成成分の前記測定量を前記亜集団の少なくとも一部の各生物からそれぞれ測定するステップをさらに含む、請求項97に記載の方法。

【請求項１００】

量的形質として前記複合形質を用いた前記亜集団の量的遺伝解析によって、前記量的形質として前記複合形質を用いて前記複数の生物Sの量的遺伝解析によって得られる前記複合形質に関する前記第1のQTLの連鎖スコアよりも高い前記複合形質に関する前記第1のQTLの連鎖スコアが得られる、請求項99に記載の方法。

【請求項１０１】

前記各細胞構成成分の前記各測定量を転写状態測定または翻訳状態測定によって求める、請求項97に記載の方法。

【請求項１０２】

前記細胞構成成分Cセット中の細胞構成成分が代謝産物であり、前記複数の生物の少なくとも一部の各生物から測定される前記代謝産物の前記量を求めるのに使用する技術が細胞表現型技術である、請求項97に記載の方法。

【請求項１０３】

前記細胞の表現型技術がメタボロミクス技術を含み、前記各生物における複数の代謝産物レベルが測定される、請求項102に記載の方法。

【請求項１０４】

前記複数の代謝産物を、熱分解質量分析法、フーリエ変換赤外分光法、ラマン分光法、ガスクロマトグラフィー-質量分析法、キャピラリー電気泳動法、高圧液体クロマトグラフィー/質量分析法(HPLC/MS)、液体クロマトグラフィー(LC)-エレクトロスプレー質量分析法またはcap-LC-タンデム・エレクトロスプレー質量分析法によって測定する、請求項102に記載の方法。

【請求項１０５】

前記代謝産物が、アミノ酸、金属、可溶性糖または複合糖質である、請求項102に記載の方法。

【請求項１０６】

複数の生物の少なくとも一部の各生物からそれぞれ測定される前記細胞構成成分iの量が、遺伝子発現レベル、mRNA存在量、タンパク質発現レベルまたは代謝産物レベルである、請求項97に記載の方法。

【請求項１０７】

前記複合形質が、前記複数の生物Sにおいて不完全浸透率を示す対立遺伝子によって特徴付けられる、請求項97に記載の方法。

【請求項１０８】

前記複合形質が、前記複数の生物S中の1つの生物が罹る疾患であり、前記生物が前記疾患の素因となる対立遺伝子を受け継いでいない、請求項97に記載の方法。

【請求項１０９】

前記複合形質が、前記種のゲノム中の複数の異なる遺伝子のいずれかが突然変異するときに生じる、請求項97に記載の方法。

【請求項１１０】

前記複合形質が、前記種のゲノム中の複数の遺伝子の突然変異が同時に存在することを必要とする、請求項97に記載の方法。

【請求項１１１】

前記複合形質が、前記集団における高頻度の疾患原因対立遺伝子に関連する、請求項97に記載の方法。

【請求項１１２】

前記複合形質が、単一の遺伝子座に起因し得るメンデルの劣性遺伝または優性遺伝を示さない表現型である、請求項97に記載の方法。

【請求項１１３】

前記複合形質が、心疾患、高血圧、糖尿病、癌、感染症、多発性嚢胞腎、早期発症型アルツハイマー病、若年成人発症型糖尿病、遺伝性非腺腫性大腸癌、血管拡張性失調症、肥満または色素性乾皮症に対する罹患性である、請求項97に記載の方法。

【請求項１１４】

前記クラスター化ステップが、階層型クラスター分析、非階層型クラスター分析、人工ニューラル・ネットワークおよび自己組織化地図からなる群から選択される技術を使用する、請求項97に記載の方法。

【請求項１１５】

前記クラスター化ステップが、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムを使用して、(i)前記複数の生物S中の1つの生物から得られる前記細胞構成成分Cセットの細胞構成成分測定量と、(ii)前記複数の生物S中の別の生物から得られる前記細胞構成成分Cセットの細胞構成成分測定量との類似度を求める階層型クラスター分析を使用する、請求項114に記載の方法。

【請求項１１６】

前記クラスター化ステップが、凝縮型クラスタリング、多形質分割型クラスタリングおよび単形質分割型クラスタリングからなる群から選択される階層型クラスター分析を使用する、請求項114に記載の方法。

【請求項１１７】

前記階層型クラスター分析が、ピアソン相関係数、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量または二乗ピアソン相関係数を使用して、(i)前記複数の生物S中の1つの生物から得られる前記細胞構成成分Cセットの細胞構成成分測定量と、(ii)前記複数の生物S中の別の生物から得られる前記細胞構成成分Cセットの細胞構成成分測定量との類似度を求める凝縮型クラスタリングである、請求項116に記載の方法。

【請求項１１８】

前記クラスター化ステップが、K平均クラスタリング、ファジーk平均クラスタリングおよびJarvis-Patrickクラスタリングからなる群から選択される非階層型クラスター分析を使用する、請求項114に記載の方法。

【請求項１１９】

前記クラスター化ステップが、Kohonen人工ニューラル・ネットワークまたは自己連想ニューラル・ネットワークである人工ニューラル・ネットワークを使用する、請求項114に記載の方法。

【請求項１２０】

前記実施ステップにおける各QTL解析が、

(i)前記複数の生物のゲノムの染色体中の位置と前記QTL解析に使用される前記量的形質との関連を検定するステップと、

(ii)前記染色体中の前記位置をある量だけ進めるステップと、

(iii)前記染色体の端部に到達するまでステップ(i)と(ii)を繰り返すステップとを含む、請求項97に記載の方法。

【請求項１２１】

前記量が100センチモルガン未満である、請求項120に記載の方法。

【請求項１２２】

前記量が2.5センチモルガン未満である、請求項120に記載の方法。

【請求項１２３】

前記複数の生物Sが分離集団である、請求項97に記載の方法。

【請求項１２４】

前記分離集団が、F_2植物、2つの近交系に由来するマウス、およびヒト系統からなる群から選択される、請求項123に記載の方法。

【請求項１２５】

コンピュータ・システムとともに使用されるコンピュータ・プログラム製品であって、

前記コンピュータ・プログラム製品がコンピュータ読み取り可能な記憶媒体およびその中に埋め込まれたコンピュータ・プログラム機構を含み、前記コンピュータ・プログラム機構が、

集団内の複数の生物を複数の亜集団に、前記集団内の各生物を前記亜集団の少なくとも1個に分類する分類体系を用いて分割する分類モジュールであって、前記分類体系が前記集団内の前記各生物から得られる複数の各細胞構成成分についての各々の複数の細胞構成成分測定値から誘導される分類モジュールと、

前記複数の亜集団中の少なくとも1個の亜集団に対して、前記複数の生物中の1個または複数の生物が示す複合形質に対する量的形質遺伝子座を特定するために、前記亜集団について量的遺伝解析を実施する量的遺伝解析モジュールとを備える、コンピュータ・プログラム製品。

【請求項１２６】

前記各生物から得られる前記細胞構成成分測定値が転写状態測定値または翻訳状態測定値である、請求項125に記載のコンピュータ・プログラム製品。

【請求項１２７】

前記翻訳状態測定を抗体アレイまたは二次元ゲル電気泳動を用いて実施する、請求項126に記載のコンピュータ・プログラム製品。

【請求項１２８】

前記細胞構成成分が複数の代謝産物を含み、前記複数の細胞構成成分測定値が細胞の表現型技術によって得られる、請求項125に記載のコンピュータ・プログラム製品。

【請求項１２９】

前記細胞の表現型技術がメタボロミクス技術を含み、前記各生物における複数の代謝産物レベルが測定される、請求項128に記載のコンピュータ・プログラム製品。

【請求項１３０】

前記代謝産物が、アミノ酸、金属、可溶性糖または複合糖質を含む、請求項129に記載のコンピュータ・プログラム製品。

【請求項１３１】

前記複数の代謝産物レベルを、熱分解質量分析法、フーリエ変換赤外分光法、ラマン分光法、ガスクロマトグラフィー-質量分析法、キャピラリー電気泳動法、高圧液体クロマトグラフィー/質量分析法(HPLC/MS)、液体クロマトグラフィー(LC)-エレクトロスプレー質量分析法またはcap-LC-タンデム・エレクトロスプレー質量分析法を使用して測定する、請求項129に記載のコンピュータ・プログラム製品。

【請求項１３２】

前記複数の細胞構成成分測定値が、遺伝子発現レベル、mRNA存在量、タンパク質発現レベルまたは代謝産物レベルを含む、請求項125に記載のコンピュータ・プログラム製品。

【請求項１３３】

前記複合形質が、前記集団において不完全浸透率を示す対立遺伝子によって特徴付けられる、請求項125に記載のコンピュータ・プログラム製品。

【請求項１３４】

前記複合形質が、前記集団内の生物が罹る疾患であり、前記生物が前記疾患の素因となる対立遺伝子を受け継いでいない、請求項125に記載のコンピュータ・プログラム製品。

【請求項１３５】

前記複合形質が、前記複数の生物のゲノム中の複数の異なる遺伝子のいずれかが突然変異するときに生じる、請求項125に記載のコンピュータ・プログラム製品。

【請求項１３６】

前記複合形質が、前記複数の生物のゲノム中の複数の遺伝子の突然変異が同時に存在することを必要とする、請求項125に記載のコンピュータ・プログラム製品。

【請求項１３７】

前記複合形質が、前記集団における高頻度の疾患原因対立遺伝子に関連する、請求項125に記載のコンピュータ・プログラム製品。

【請求項１３８】
前記複合形質が、単一の遺伝子座に起因し得るメンデルの劣性遺伝または優性遺伝を示さない表現型である、請求項125に記載のコンピュータ・プログラム製品。
【請求項１３９】
前記複合形質が、心疾患、高血圧、糖尿病、癌、感染症、多発性嚢胞腎、早期発症型アルツハイマー病、若年成人発症型糖尿病、遺伝性非腺腫性大腸癌、血管拡張性失調症、肥満または色素性乾皮症に対する罹患性である、請求項125に記載のコンピュータ・プログラム製品。
【請求項１４０】
前記各生物から得られる前記複数の細胞構成成分測定値が、前記各生物中の10個以上の細胞構成成分の細胞構成成分レベルの測定値を含む、請求項125に記載のコンピュータ・プログラム製品。
【請求項１４１】
前記各生物から得られる前記複数の細胞構成成分測定値が、前記各生物中の1000個以上の細胞構成成分レベルの細胞構成成分レベルの測定値を含む、請求項125に記載のコンピュータ・プログラム製品。
【請求項１４２】
前記分類モジュールによって、クラス予測変数が利用可能かどうかを判定し、
クラス予測変数が利用可能なときに、教師付き分類体系を使用して前記集団内の各生物を前記複数の亜集団内の1個の亜集団に分類し、
クラス予測変数が利用不可能なときに、教師なし分類体系を使用して前記集団内の各生物を前記複数の亜集団内の1個の亜集団に分類する、請求項125に記載のコンピュータ・プログラム製品。
【請求項１４３】
前記分類体系が教師付き分類体系である、請求項125に記載のコンピュータ・プログラム製品。
【請求項１４４】
前記分類体系が教師なし分類体系である、請求項125に記載のコンピュータ・プログラム製品。
【請求項１４５】
前記教師付き分類体系が線形判別分析または線形回帰法を使用する、請求項142または143に記載のコンピュータ・プログラム製品。
【請求項１４６】
前記線形回帰法が、多重線形回帰、部分最小二乗回帰または主成分回帰である、請求項145に記載のコンピュータ・プログラム製品。
【請求項１４７】
前記教師なし分類体系が、階層型クラスター分析、非階層型クラスター分析、人工ニューラル・ネットワークおよび自己組織化地図からなる群から選択される、請求項142または144に記載のコンピュータ・プログラム製品。
【請求項１４８】
前記教師なし分類体系が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムを使用して、(i)前記複数の生物中の1つの生物から得られる複数の細胞構成成分測定値と、(ii)前記複数の生物中の別の生物から得られる複数の細胞構成成分測定値との類似度を求める階層型クラスター分析である、請求項147に記載のコンピュータ・プログラム製品。
【請求項１４９】
前記教師なし分類体系が、凝縮型クラスタリング、多形質分割型クラスタリングおよび単形質分割型クラスタリングからなる群から選択される階層型クラスター分析である、請求項147に記載のコンピュータ・プログラム製品。
【請求項１５０】

前記階層型クラスター分析が、ピアソン相関係数、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量または二乗ピアソン相関係数を用いて、(i)前記複数の生物中の1つの生物から得られる複数の細胞構成成分測定値と、(ii)前記複数の生物中の別の生物から得られる複数の細胞構成成分測定値との類似度を求める凝縮型クラスタリングである、請求項149に記載のコンピュータ・プログラム製品。

【請求項１５１】

前記教師なし分類体系が、K平均クラスタリング、ファジーk平均クラスタリングおよびJarvis-Patrickクラスタリングからなる群から選択される非階層型クラスター分析である、請求項147に記載のコンピュータ・プログラム製品。

【請求項１５２】

前記教師なし分類体系が、Kohonen人工ニューラル・ネットワークまたは自己連想ニューラル・ネットワークである人工ニューラル・ネットワークである、請求項147に記載のコンピュータ・プログラム製品。

【請求項１５３】

前記分類モジュールが、前記複数の亜集団への前記集団の分割をさらに検証する、請求項125に記載のコンピュータ・プログラム製品。

【請求項１５４】

前記量的遺伝解析モジュールが、連鎖解析、前記複数の細胞構成成分測定値を表現型形質として使用する量的形質遺伝子座(QTL)解析方法、および関連解析からなる群から選択される方法を使用する、請求項125に記載のコンピュータ・プログラム製品。

【請求項１５５】

前記量的遺伝解析を前記QTL解析によって実施し、前記QTL解析方法が、

(a)複数のQTL解析から得られるQTLデータをクラスター化してQTL相互作用地図を作成するステップであって、

前記複数のQTL解析の各QTL解析を、前記QTLデータを作成するために、前記複数の生物のゲノム中の複数の遺伝子中の遺伝子Gに対して、遺伝マーカー地図および量的形質を用いて実施し、各QTL解析では、前記量的形質が、前記複数の生物中の各生物の前記QTL解析を実施する前記遺伝子Gの発現統計量を含み、

前記遺伝マーカー地図を、前記複数の生物に関連する遺伝マーカー・セットから構築するステップと、

(b)前記QTL相互作用地図を解析して、前記量的形質に関連する前記QTLを特定するステップとを含む、請求項125に記載のコンピュータ・プログラム製品。

【請求項１５６】

前記遺伝子Gの前記発現統計量を、前記複数の生物中の各生物から得られる前記遺伝子Gの前記発現レベルの測定値を変換するステップを含む方法によって計算する、請求項155に記載のコンピュータ・プログラム製品。

【請求項１５７】

前記遺伝子Gの前記発現レベルの測定値を変換する前記ステップが、前記発現統計量を作成するために、前記遺伝子Gの前記発現レベルの測定値を正規化するステップを含む、請求項156に記載のコンピュータ・プログラム製品。

【請求項１５８】

前記発現統計量を作成するための前記遺伝子Gの前記発現レベルの測定値を正規化するステップを、強度のZ-スコア、強度中央値、強度中央値の対数、強度のZ-スコア標準偏差対数、対数強度のZ-スコア平均絶対偏差、較正DNA遺伝子セット、ユーザー正規化遺伝子セット、強度中央値の比率補正、および強度バックグラウンド補正からなる群から選択される正規化技術によって実施する、請求項157に記載のコンピュータ・プログラム製品。

【請求項１５９】

前記各QTL解析が、

(i)前記複数の生物のゲノムの染色体中の位置とQTL解析に使用される前記量的形質との関連を検定するステップと、

(ii)前記染色体中の前記位置をある量だけ進めるステップと、
(iii)前記染色体の端部に到達するまでステップ(i)と(ii)を繰り返すステップとを含む、請求項155に記載のコンピュータ・プログラム製品。
【請求項１６０】
前記量が100センチモルガン未満である、請求項159に記載のコンピュータ・プログラム製品。
【請求項１６１】
前記量が10センチモルガン未満である、請求項159に記載のコンピュータ・プログラム製品。
【請求項１６２】
前記量が5センチモルガン未満である、請求項159に記載のコンピュータ・プログラム製品。
【請求項１６３】
前記量が2.5センチモルガン未満である、請求項159に記載のコンピュータ・プログラム製品。
【請求項１６４】
各QTL解析から生成する前記QTLデータが、前記各位置において計算される統計スコアを含む、請求項159に記載のコンピュータ・プログラム製品。
【請求項１６５】
QTLベクトルが、前記染色体において検定される各量的形質に対して作成され、前記QTLベクトルが、前記量的形質に対応する前記QTL解析によって検定される各位置の統計スコアを含む、請求項159に記載のコンピュータ・プログラム製品。
【請求項１６６】
QTLデータをクラスター化する前記ステップが、前記各QTLベクトルをクラスター化するステップを含む、請求項165に記載のコンピュータ・プログラム製品。
【請求項１６７】
前記クラスター化ステップのための基礎として使用される類似度計測量が、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量、ピアソン相関係数または二乗ピアソン相関係数であって、QTLベクトル対間で計算される、請求項165に記載のコンピュータ・プログラム製品。
【請求項１６８】
QTLデータの前記クラスタリングが、階層型クラスタリング法を適用するステップ、k平均法を適用するステップ、ファジーk平均法を適用するステップ、Jarvis-Patrickクラスタリングを適用するステップ、自己組織化地図技術を適用するステップ、またはニューラル・ネットワーク技術を適用するステップを含む、請求項155または166に記載のコンピュータ・プログラム製品。
【請求項１６９】
QTLデータをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が凝縮型クラスタリング手順である、請求項168に記載のコンピュータ・プログラム製品。
【請求項１７０】
前記凝縮型クラスタリング手順が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである、請求項169に記載のコンピュータ・プログラム製品。
【請求項１７１】
前記階層型クラスタリング法が分割型クラスタリング手順である、請求項168に記載のコンピュータ・プログラム製品。
【請求項１７２】
前記変換ステップによって作成される各発現統計量から遺伝子発現クラスター地図を構築するステップをさらに含む、請求項156に記載のコンピュータ・プログラム製品。

【請求項１７３】
遺伝子発現クラスター地図を構築する前記ステップが、
複数の遺伝子発現ベクトルを作成するステップであって、前記複数の遺伝子発現ベクトル中の各遺伝子発現ベクトルが、複数の生物の各々における前記複数の遺伝子中の1個の遺伝子の発現レベルの測定値であるステップと、
複数の相関係数を計算するステップであって、前記複数の相関係数の各相関係数を前記複数の遺伝子発現ベクトル中の遺伝子発現ベクトル対間で計算するステップと、
前記遺伝子発現クラスター地図を作成するために、前記複数の相関係数に基づいて前記複数の遺伝子発現ベクトルをクラスター化するステップとを含む、請求項172に記載のコンピュータ・プログラム製品。
【請求項１７４】
前記QTL相互作用地図を解析する前記ステップが、候補経路群を得るためにQTL相互作用地図を選別するステップを含み、前記選別ステップが、前記遺伝子発現クラスター地図中の前記候補経路群においてQTLを特定するステップを含む、請求項173に記載のコンピュータ・プログラム製品。
【請求項１７５】
前記複数の相関係数の各相関係数がピアソン相関係数である、請求項173に記載のコンピュータ・プログラム製品。
【請求項１７６】
遺伝子発現クラスター地図を構築する前記ステップが、
複数の遺伝子発現ベクトルを作成するステップであって、前記複数の遺伝子発現ベクトル中の各遺伝子発現ベクトルが、前記複数の遺伝子中の1個の遺伝子であるステップと、
複数の計測量を計算するステップであって、前記複数の計測量の各計測量を前記複数の遺伝子発現ベクトル中の遺伝子発現ベクトル対間で計算するステップと、
前記遺伝子発現クラスター地図を作成するために、前記複数の計測量に基づいて前記複数の遺伝子発現ベクトルをクラスター化するステップとを含む、請求項172に記載のコンピュータ・プログラム製品。
【請求項１７７】
前記各計測量が、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量、ピアソン相関係数および二乗ピアソン相関係数からなる群から選択される、請求項176に記載のコンピュータ・プログラム製品。
【請求項１７８】
複数の遺伝子発現ベクトルをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップ、k平均法を適用するステップ、ファジーk平均法を適用するステップ、Jarvis-Patrickクラスタリングを適用するステップ、自己組織化地図技術を適用するステップ、またはニューラル・ネットワーク技術を適用するステップを含む、請求項173または176に記載のコンピュータ・プログラム製品。
【請求項１７９】
前記複数の遺伝子発現ベクトルをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が凝縮型クラスタリング手順である、請求項178に記載のコンピュータ・プログラム製品。
【請求項１８０】
前記凝縮型クラスタリング手順が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである、請求項179に記載のコンピュータ・プログラム製品。
【請求項１８１】
前記複数の遺伝子発現ベクトルをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が分割型クラスタリング手順である、請求項178に記載のコンピュータ・プログラム製品。
【請求項１８２】

前記QTL相互作用地図を解析する前記ステップが、候補経路群を得るために、前記QTL相互作用地図を選別するステップを含む、請求項155に記載のコンピュータ・プログラム製品。

【請求項１８３】

前記候補経路群を得るための前記選別ステップが、前記QTL相互作用地図中の別のQTLと最も強く相互作用する前記候補経路群のQTLを選択するステップを含む、請求項182に記載のコンピュータ・プログラム製品。

【請求項１８４】

前記QTL相互作用地図中の別のQTLと最も強く相互作用する前記QTLが、前記量的形質遺伝子座相互作用地図中のQTL間で計算される全相関係数の75%よりも高い相関係数を、前記量的形質遺伝子座相互作用地図中の別のQTLと共有する前記QTL相互作用地図中のQTLである、請求項183に記載のコンピュータ・プログラム製品。

【請求項１８５】

前記候補経路群を構成する各QTLが前記候補経路群に属する程度を検定するために、多変量統計モデルを前記候補経路群に適合させる、請求項183に記載のコンピュータ・プログラム製品。

【請求項１８６】

前記多変量統計モデルが複数の量的形質を同時に考慮する、請求項185に記載のコンピュータ・プログラム製品。

【請求項１８７】

前記多変量統計モデルが、前記候補経路群中のQTL間のエピスタシス相互作用を探索する、請求項185に記載のコンピュータ・プログラム製品。

【請求項１８８】

前記遺伝マーカー・セットが、一塩基多型(SNP)、ミクロサテライト・マーカー、制限断片長多型、短鎖縦列反復、DNAメチル化マーカーまたは配列長多型を含む、請求項155に記載のコンピュータ・プログラム製品。

【請求項１８９】

前記分類モジュールが、

(i)前記複数の生物のすべてまたは一部の表現型データに基づいて前記集団を複数の表現型群に区分化する命令と、

(ii)極端な表現型を示す前記複数の表現型群中の1組の極端生物を特定する命令と、

(iii)前記複数の細胞構成成分中の細胞構成成分を特定する命令であって、特定された各細胞構成成分が、前記極端生物セットから得られた前記各細胞構成成分の細胞構成成分測定値が前記複数の表現型群のすべてまたは一部を識別する特性を有する命令と、

(iv)前記特定された細胞構成成分のすべてまたは一部から導出される確率分布を用いて分類子を構築する命令とを含む、請求項125に記載のコンピュータ・プログラム製品。

【請求項１９０】

前記表現型データがバイナリ・イベントを含む、請求項189に記載のコンピュータ・プログラム製品。

【請求項１９１】

前記表現型データが、前記集団内の各生物の1個を超える表現型測定値を含む、請求項189に記載のコンピュータ・プログラム製品。

【請求項１９２】

前記表現型データが前記複数の生物中の各生物が形質を示すかどうかに関する判定を含み、前記区分化命令が、前記生物が前記形質を示すときに第1の表現型群中の前記複数の生物中の1つの生物を配置するステップと、前記生物が前記形質を示さないときに第2の表現型群中の前記複数の生物中の1つの生物を配置するステップとを含む、請求項189に記載のコンピュータ・プログラム製品。

【請求項１９３】

前記表現型データが、前記複数の生物のすべてまたは一部に対して作成される複数の表

現型測定値を含み、前記区分化命令が、
(A)複数の表現型ベクトルを作成する命令であって、前記複数の表現型ベクトル中の各表現型ベクトルが前記複数の生物中の1つの生物に対応し、前記複数の表現型ベクトル中の各表現型ベクトルが前記各表現型ベクトルに対応する前記生物から得られる複数の表現型測定値を含む命令と、
(B)前記複数の表現型ベクトルを複数のクラスターにクラスター化する命令であって、前記複数のクラスター中の各クラスターが、前記複数の表現型群中の1個の表現型群を表す命令とを含む、請求項189に記載のコンピュータ・プログラム製品。
【請求項１９４】
前記クラスター化ステップが、階層型クラスタリング法、k平均法、ファジーk平均法、Jarvis-Patrickクラスタリング、自己組織化地図技術またはニューラル・ネットワーク技術を含む、請求項193に記載のコンピュータ・プログラム製品。
【請求項１９５】
前記クラスター化ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が凝縮型クラスタリング手順である、請求項193に記載のコンピュータ・プログラム製品。
【請求項１９６】
前記凝縮型クラスタリング手順が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである、請求項195に記載のコンピュータ・プログラム製品。
【請求項１９７】
前記階層型クラスタリング法が分割型クラスタリング手順である、請求項195に記載のコンピュータ・プログラム製品。
【請求項１９８】
生物が、前記集団が示す表現型に関して前記集団の上位30パーセンタイルまたは下位30パーセンタイルであるときに、前記極端な表現型である、請求項189に記載のコンピュータ・プログラム製品。
【請求項１９９】
生物が、前記集団が示す表現型に関して前記集団の上位10パーセンタイルまたは下位10パーセンタイルであるときに、前記極端な表現型である、請求項189に記載のコンピュータ・プログラム製品。
【請求項２００】
前記極端生物セットが5個体を超える生物である、請求項189に記載のコンピュータ・プログラム製品。
【請求項２０１】
前記極端生物セットが2～100個体の生物である、請求項189に記載のコンピュータ・プログラム製品。
【請求項２０２】
前記極端生物セットが1000個体未満の生物である、請求項189に記載のコンピュータ・プログラム製品。
【請求項２０３】
前記特定命令(iii)が、所定の細胞構成成分の複数の細胞構成成分測定値をt検定にかけるステップを含み、前記複数の細胞構成成分測定値が前記極端生物セットから得られる、請求項189に記載のコンピュータ・プログラム製品。
【請求項２０４】
前記特定命令が、前記特定された細胞構成成分を多変量解析にかけるステップを含む、請求項189に記載のコンピュータ・プログラム製品
【請求項２０５】
前記特定命令(iii)によって特定された前記細胞構成成分を、前記構築命令(iv)を実行する前に削減する、請求項189に記載のコンピュータ・プログラム製品。

【請求項２０６】
前記特定命令(iii)によって特定された前記細胞構成成分を、段階的回帰、総当り回帰、主成分分析法または重判別分析によって削減する、請求項205に記載のコンピュータ・プログラム製品。
【請求項２０７】
前記特定命令(iii)によって特定された前記細胞構成成分を確率論的検索方法によって削減する、請求項205に記載のコンピュータ・プログラム製品。
【請求項２０８】
前記確率論的検索方法がシミュレーテッド・アニーリングまたは遺伝アルゴリズムである、請求項207に記載のコンピュータ・プログラム製品。
【請求項２０９】
前記特定命令(iii)によって特定された前記細胞構成成分をクラスタリングによって削減し、前記特定された細胞構成成分ではなく、前記クラスタリングによって生成したクラスターを前記構築命令(iv)に使用する、請求項205に記載のコンピュータ・プログラム製品。
【請求項２１０】
前記構築命令(iv)が、前記確率分布を用いてニューラル・ネットワークを訓練するステップを含む、請求項189に記載のコンピュータ・プログラム製品。
【請求項２１１】
前記構築命令(iv)が、前記確率分布が演繹的情報として役立つベイズの決定理論を使用するステップを含む、請求項189に記載のコンピュータ・プログラム製品。
【請求項２１２】
前記構築命令(iv)が、線形判別分析、線形プログラミング・アルゴリズムまたはサポート・ベクトル・マシンを使用するステップを含む、請求項189に記載のコンピュータ・プログラム製品。
【請求項２１３】
前記分類体系が、前記分類子を用いて前記集団のすべてまたは一部を分類するステップを含む、請求項189に記載のコンピュータ・プログラム製品。
【請求項２１４】
コンピュータ・システムとともに使用されるコンピュータ・プログラム製品であって、前記コンピュータ・プログラム製品がコンピュータ読み取り可能な記憶媒体およびその中に埋め込まれたコンピュータ・プログラム機構を含み、前記コンピュータ・プログラム機構が、
量的遺伝解析に使用する複数の亜集団を得るために同じ種の複数の生物Sを細分する分類モジュールであって、前記複数の生物S中の1つまたは複数の生物が複合形質を示す分類モジュールを含み、前記分類モジュールが、
前記複合形質に関してそれぞれが独立した極端である前記複数の生物S内の2群以上の生物を特定する命令と、
前記複数の生物S内の前記2群以上の生物を識別することができる1組の細胞構成成分Cを決定する命令と、
前記細胞構成成分Cセット中の各細胞構成成分iに対して、前記複合形質と関連する第1のQTLと相互作用または重複するQTLを有する1個または複数の細胞構成成分を特定するために、前記複数の生物Sの少なくとも一部の各生物からそれぞれ測定される前記細胞構成成分iの量を量的形質として用いて、前記細胞構成成分iについてのQTL解析を実施する命令と、
前記実施命令によって特定された各細胞構成成分の前記測定量に基づいて前記複数の生物Sをクラスター化し、それによって前記複数の亜集団を得る命令とを含む、コンピュータ・プログラム製品。
【請求項２１５】
前記複合形質に関連する前記第1のQTLを連鎖解析または関連解析によって特定する、請

求項214に記載のコンピュータ・プログラム製品。
【請求項２１６】
前記分類モジュールが、前記複合形質のQTLを特定するために、前記複数の亜集団内の1個の亜集団について一連のQTL解析を実施する命令であって、前記一連のQTL解析の前記各QTL解析が前記細胞構成成分Cセット中の1個の細胞構成成分の測定量を量的形質として使用し、前記細胞構成成分の前記測定量を前記亜集団の少なくとも一部の各生物からそれぞれ測定する命令をさらに含む、請求項215に記載のコンピュータ・プログラム製品。
【請求項２１７】
量的形質として前記複合形質を用いた前記亜集団の量的遺伝解析によって、前記量的形質として前記複合形質を用いて前記複数の生物Sの量的遺伝解析によって得られた前記複合形質に関する前記第1のQTLの連鎖スコアよりも高い前記複合形質に関する前記第1のQTLの連鎖スコアが得られる、請求項216に記載のコンピュータ・プログラム製品。
【請求項２１８】
前記各細胞構成成分の前記各測定量を転写状態測定または翻訳状態測定によって求める、請求項214に記載のコンピュータ・プログラム製品。
【請求項２１９】
前記細胞構成成分Cセット中の細胞構成成分が代謝産物であり、前記複数の生物の少なくとも一部の各生物から測定される前記代謝産物の前記量を求めるのに使用する技術が細胞表現型技術である、請求項214に記載のコンピュータ・プログラム製品。
【請求項２２０】
前記細胞の表現型技術がメタボロミクス技術を含み、前記各生物における複数の代謝産物レベルが測定される、請求項219に記載のコンピュータ・プログラム製品。
【請求項２２１】
前記複数の代謝産物を、熱分解質量分析法、フーリエ変換赤外分光法、ラマン分光法、ガスクロマトグラフィー-質量分析法、キャピラリー電気泳動法、高圧液体クロマトグラフィー/質量分析法(HPLC/MS)、液体クロマトグラフィー(LC)-エレクトロスプレー質量分析法またはcap-LC-タンデム・エレクトロスプレー質量分析法によって測定する、請求項219に記載のコンピュータ・プログラム製品。
【請求項２２２】
前記代謝産物が、アミノ酸、金属、可溶性糖または複合糖質を含む、請求項219に記載のコンピュータ・プログラム製品。
【請求項２２３】
複数の生物の少なくとも一部の各生物からそれぞれ測定される前記細胞構成成分iの量が、遺伝子発現レベル、mRNA存在量、タンパク質発現レベルまたは代謝産物レベルである、請求項214に記載のコンピュータ・プログラム製品。
【請求項２２４】
前記複合形質が、前記複数の生物Sにおいて不完全浸透率を示す対立遺伝子によって特徴付けられる、請求項214に記載のコンピュータ・プログラム製品。
【請求項２２５】
前記複合形質が、前記複数の生物S中の1つの生物が罹る疾患であり、前記生物が前記疾患の素因となる対立遺伝子を受け継いでいない、請求項214に記載のコンピュータ・プログラム製品。
【請求項２２６】
前記複合形質が、前記種のゲノム中の複数の異なる遺伝子のいずれかが突然変異するときに生じる、請求項214に記載のコンピュータ・プログラム製品。
【請求項２２７】
前記複合形質が、前記種のゲノム中の複数の遺伝子の突然変異が同時に存在することを必要とする、請求項214に記載のコンピュータ・プログラム製品。
【請求項２２８】
前記複合形質が、前記集団における高頻度の疾患原因対立遺伝子に関連する、請求項21

4に記載のコンピュータ・プログラム製品。
【請求項２２９】
前記複合形質が、単一の遺伝子座に起因し得るメンデルの劣性遺伝または優性遺伝を示さない表現型である、請求項214に記載のコンピュータ・プログラム製品。
【請求項２３０】
前記複合形質が、心疾患、高血圧、糖尿病、癌、感染症、多発性嚢胞腎、早期発症型アルツハイマー病、若年成人発症型糖尿病、遺伝性非腺腫性大腸癌、血管拡張性失調症、肥満または色素性乾皮症に対する罹患性である、請求項214に記載のコンピュータ・プログラム製品。
【請求項２３１】
前記クラスター化命令が、階層型クラスター分析、非階層型クラスター分析、人工ニューラル・ネットワークおよび自己組織化地図からなる群から選択される技術を使用する、請求項214に記載のコンピュータ・プログラム製品。
【請求項２３２】
前記クラスター化命令が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムを使用して、(i)前記複数の生物S中の1つの生物から得られる1組の細胞構成成分Cの細胞構成成分測定量と、(ii)前記複数の生物S中の別の生物から得られる1組の細胞構成成分Cの細胞構成成分測定量との類似度を求める階層型クラスター分析を使用する、請求項231に記載のコンピュータ・プログラム製品。
【請求項２３３】
前記クラスター化命令が、凝縮型クラスタリング、多形質分割型クラスタリングおよび単形質分割型クラスタリングからなる群から選択される階層型クラスター分析を使用する、請求項231に記載のコンピュータ・プログラム製品。
【請求項２３４】
前記階層型クラスター分析が、ピアソン相関係数、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量または二乗ピアソン相関係数を使用して、(i)前記複数の生物S中の1つの生物から得られる前記細胞構成成分Cセットの細胞構成成分測定量と、(ii)前記複数の生物S中の別の生物から得られる1組の細胞構成成分Cの細胞構成成分測定量との類似度を求める凝縮型クラスタリングである、請求項233に記載のコンピュータ・プログラム製品。
【請求項２３５】
前記クラスター化命令が、K平均クラスタリング、ファジーk平均クラスタリングおよびJarvis-Patrickクラスタリングからなる群から選択される非階層型クラスター分析を使用する、請求項231に記載のコンピュータ・プログラム製品。
【請求項２３６】
前記クラスター化命令が、Kohonen人工ニューラル・ネットワークまたは自己連想ニューラル・ネットワークである人工ニューラル・ネットワークを使用する、請求項231に記載のコンピュータ・プログラム製品。
【請求項２３７】
前記実施命令によって計算される各QTL解析が、
(i)前記複数の生物のゲノムの染色体中の位置と前記QTL解析に使用される前記量的形質との関連を検定するステップと、
(ii)前記染色体中の前記位置をある量だけ進めるステップと、
(iii)前記染色体の端部に到達するまでステップ(i)と(ii)を繰り返すステップとを含む、請求項214に記載のコンピュータ・プログラム製品。
【請求項２３８】
前記量が100センチモルガン未満である、請求項237に記載のコンピュータ・プログラム製品。
【請求項２３９】

前記量が10センチモルガン未満である、請求項237に記載のコンピュータ・プログラム製品。

【請求項２４０】

前記量が5センチモルガン未満である、請求項237に記載のコンピュータ・プログラム製品。

【請求項２４１】

前記量が2.5センチモルガン未満である、請求項237に記載のコンピュータ・プログラム製品。

【請求項２４２】

前記複数の生物が分離集団である、請求項214に記載のコンピュータ・プログラム製品。

【請求項２４３】

前記分離集団が、F_2植物、2つの近交系に由来するマウス、およびヒト系統からなる群から選択される、請求項242に記載のコンピュータ・プログラム製品。

【請求項２４４】

集団内の複数の生物が示す複合形質に対する量的形質遺伝子座を特定するコンピュータ・システムであって、

中央処理装置と、

前記中央処理装置に接続され、分類モジュールおよび量的遺伝解析モジュールを保存するメモリとを含み、

前記分類モジュールは、集団内の複数の生物を複数の亜集団に、前記集団内の各生物を前記亜集団の少なくとも1個に分類する分類体系を用いて分割する命令を含み、前記分類体系が前記集団内の前記各生物から得られる複数の各細胞構成成分の各々の複数の細胞構成成分測定値から誘導され、

前記量的遺伝解析モジュールは、前記複合形質に対する前記量的形質遺伝子座を特定するために、前記複数の亜集団中の少なくとも1個の亜集団に対して、前記亜集団についての量的遺伝解析を実施する命令を含む、コンピュータ・システム。

【請求項２４５】

前記各生物から得られる前記細胞構成成分測定値が転写状態測定値または翻訳状態測定値である、請求項244に記載のコンピュータ・システム。

【請求項２４６】

前記翻訳状態測定を抗体アレイまたは二次元ゲル電気泳動を用いて実施する、請求項245に記載のコンピュータ・システム。

【請求項２４７】

前記細胞構成成分が複数の代謝産物を含み、前記複数の細胞構成成分測定値が細胞の表現型技術によって得られる、請求項244に記載のコンピュータ・システム。

【請求項２４８】

前記細胞の表現型技術がメタボロミクス技術を含み、前記各生物における複数の代謝産物レベルが測定される、請求項247に記載のコンピュータ・システム。

【請求項２４９】

前記複数の代謝産物が、アミノ酸、金属、可溶性糖または複合糖質を含む、請求項248に記載のコンピュータ・システム。

【請求項２５０】

前記代謝産物レベルを、熱分解質量分析法、フーリエ変換赤外分光法、ラマン分光法、ガスクロマトグラフィー-質量分析法、キャピラリー電気泳動法、高圧液体クロマトグラフィー/質量分析法(HPLC/MS)、液体クロマトグラフィー(LC)-エレクトロスプレー質量分析法またはcap-LC-タンデム・エレクトロスプレー質量分析法によって測定する、請求項248に記載のコンピュータ・システム。

【請求項２５１】

前記複数の細胞構成成分測定値が、遺伝子発現レベル、mRNA存在量、タンパク質発現レ

ベルまたは代謝産物レベルを含む、請求項244に記載のコンピュータ・システム。
【請求項２５２】
前記複合形質が、前記集団において不完全浸透率を示す対立遺伝子によって特徴付けられる、請求項244に記載のコンピュータ・システム。
【請求項２５３】
前記複合形質が、前記集団内の生物が罹る疾患であり、前記生物が前記疾患の素因となる対立遺伝子を受け継いでいない、請求項244に記載のコンピュータ・システム。
【請求項２５４】
前記複合形質が、前記複数の生物のゲノム中の複数の異なる遺伝子のいずれかが突然変異するときに生じる、請求項244に記載のコンピュータ・システム。
【請求項２５５】
前記複合形質が、前記複数の生物のゲノム中の複数の遺伝子の突然変異が同時に存在することを必要とする、請求項244に記載のコンピュータ・システム。
【請求項２５６】
前記複合形質が、前記集団における高頻度の疾患原因対立遺伝子に関連する、請求項244に記載のコンピュータ・システム。
【請求項２５７】
前記複合形質が、単一の遺伝子座に起因し得るメンデルの劣性遺伝または優性遺伝を示さない表現型である、請求項244に記載のコンピュータ・システム。
【請求項２５８】
前記複合形質が、心疾患、高血圧、糖尿病、癌、感染症、多発性嚢胞腎、早期発症型アルツハイマー病、若年成人発症型糖尿病、遺伝性非腺腫性大腸癌、血管拡張性失調症、肥満または色素性乾皮症に対する罹患性である、請求項244に記載のコンピュータ・システム。
【請求項２５９】
前記各生物から得られる前記複数の細胞構成成分測定値が、前記各生物中の10個以上の細胞構成成分の細胞構成成分レベルの測定値を含む、請求項244に記載のコンピュータ・システム。
【請求項２６０】
前記各生物から得られる前記複数の細胞構成成分測定値が、前記各生物中の1000個以上の細胞構成成分の細胞構成成分レベルの測定値を含む、請求項244に記載のコンピュータ・システム。
【請求項２６１】
前記分類モジュールが、クラス予測変数が利用可能かどうかをさらに判定し、
クラス予測変数が利用可能なときに、教師付き分類体系を使用して前記集団内の各生物を前記複数の亜集団内の1個の亜集団に分類し、
クラス予測変数が利用不可能なときに、教師なし分類体系を使用して前記集団内の各生物を前記複数の亜集団内の1個の亜集団に分類する、請求項244に記載のコンピュータ・システム。
【請求項２６２】
前記分類体系が教師付き分類体系である、請求項244に記載のコンピュータ・システム。
【請求項２６３】
前記分類体系が教師なし分類体系である、請求項244に記載のコンピュータ・システム。
【請求項２６４】
前記教師付き分類体系が線形判別分析または線形回帰法を使用する、請求項261または262に記載のコンピュータ・システム。
【請求項２６５】
前記線形回帰法が、多重線形回帰、部分最小二乗回帰または主成分回帰である、請求項

264に記載のコンピュータ・システム。
【請求項２６６】
前記教師なし分類体系が、階層型クラスター分析、非階層型クラスター分析、人工ニューラル・ネットワークおよび自己組織化地図からなる群から選択される、請求項261または263に記載のコンピュータ・システム。
【請求項２６７】
前記教師なし分類体系が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムを使用して、(i)前記複数の生物中の1つの生物から得られる複数の細胞構成成分測定値と、(ii)前記複数の生物中の別の生物から得られる複数の細胞構成成分測定値との類似度を求める階層型クラスター分析である、請求項266に記載のコンピュータ・システム。
【請求項２６８】
前記教師なし分類体系が、凝縮型クラスタリング、多形質分割型クラスタリングおよび単形質分割型クラスタリングからなる群から選択される階層型クラスター分析である、請求項266に記載のコンピュータ・システム。
【請求項２６９】
前記階層型クラスター分析が、ピアソン相関係数、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量または二乗ピアソン相関係数を用いて、(i)前記複数の生物中の1つの生物から得られる複数の細胞構成成分測定値と、(ii)前記複数の生物中の別の生物から得られる複数の細胞構成成分測定値との類似度を求める凝縮型クラスタリングである、請求項268に記載のコンピュータ・システム。
【請求項２７０】
前記教師なし分類体系が、K平均クラスタリング、ファジーk平均クラスタリングおよびJarvis-Patrickクラスタリングからなる群から選択される非階層型クラスター分析である、請求項266に記載のコンピュータ・システム。
【請求項２７１】
前記教師なし分類体系が、Kohonen人工ニューラル・ネットワークまたは自己連想ニューラル・ネットワークである人工ニューラル・ネットワークである、請求項266に記載のコンピュータ・システム。
【請求項２７２】
前記分類モジュールが、前記複数の亜集団への前記集団の分割を検証する命令をさらに含む、請求項244に記載のコンピュータ・システム。
【請求項２７３】
前記量的遺伝解析モジュールが、連鎖解析、前記複数の細胞構成成分測定値を表現型形質として使用する量的形質遺伝子座(QTL)解析方法、および関連解析からなる群から選択される方法を使用する、請求項244に記載のコンピュータ・システム。
【請求項２７４】
前記量的遺伝解析を前記QTL解析によって実施し、前記QTL解析方法が、
(a)複数のQTL解析から得られるQTLデータをクラスター化してQTL相互作用地図を作成するステップであって、
前記複数のQTL解析の各QTL解析を、前記QTLデータを作成するために、前記複数の生物のゲノム中の複数の遺伝子中の遺伝子Gに対して、遺伝マーカー地図および量的形質を用いて実施し、各QTL解析では、前記量的形質が、前記複数の生物中の各生物の前記QTL解析を実施する前記遺伝子Gの発現統計量を含み、
前記遺伝マーカー地図を前記複数の生物に関連する遺伝マーカー・セットから構築するステップと、
(b)前記QTL相互作用地図を解析して、前記量的形質に関連する前記QTLを特定するステップとを含む、請求項244に記載のコンピュータ・システム。
【請求項２７５】
前記遺伝子Gの前記発現統計量を、前記複数の生物中の各生物から得られる前記遺伝子G

の前記発現レベルの測定値を変換するステップを含む方法によって計算する、請求項274に記載のコンピュータ・システム。

【請求項２７６】

前記遺伝子Gの前記発現レベルの測定値を変換する前記ステップが、前記発現統計量を作成するために、前記遺伝子Gの前記発現レベルの測定値を正規化するステップを含む、請求項275に記載のコンピュータ・システム。

【請求項２７７】

前記発現統計量を作成するための前記遺伝子Gの前記発現レベルの測定値を正規化するステップを、強度のZ-スコア、強度中央値、強度中央値の対数、強度のZ-スコア標準偏差対数、対数強度のZ-スコア平均絶対偏差、較正DNA遺伝子セット、ユーザー正規化遺伝子セット、強度中央値の比率補正、および強度バックグラウンド補正からなる群から選択される正規化技術によって実施する、請求項276に記載のコンピュータ・システム。

【請求項２７８】

前記各QTL解析が、

(i)前記複数の生物のゲノムの染色体中の位置と前記QTL解析に使用される前記量的形質との関連を検定するステップと、

(ii)前記染色体中の前記位置をある量だけ進めるステップと、

(iii)前記染色体の端部に到達するまでステップ(i)と(ii)を繰り返すステップとを含む、請求項274に記載のコンピュータ・システム。

【請求項２７９】

前記量が100センチモルガン未満である、請求項278に記載のコンピュータ・システム。

【請求項２８０】

前記量が10センチモルガン未満である、請求項278に記載のコンピュータ・システム。

【請求項２８１】

前記量が5センチモルガン未満である、請求項278に記載のコンピュータ・システム。

【請求項２８２】

前記量が2.5センチモルガン未満である、請求項278に記載のコンピュータ・システム。

【請求項２８３】

各QTL解析から生成する前記QTLデータが、前記各位置において計算される統計スコアを含む、請求項278に記載のコンピュータ・システム。

【請求項２８４】

QTLベクトルが、前記染色体において検定される各量的形質に対して作成され、前記QTLベクトルが、前記量的形質に対応する前記QTL解析によって検定される各位置の統計スコアを含む、請求項278に記載のコンピュータ・システム。

【請求項２８５】

QTLデータをクラスター化する前記ステップが、前記各QTLベクトルをクラスター化するステップを含む、請求項284に記載のコンピュータ・システム。

【請求項２８６】

前記クラスター化ステップのための基礎として使用される類似度計測量が、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量、ピアソン相関係数または二乗ピアソン相関係数であって、QTLベクトル対間で計算される、請求項284に記載のコンピュータ・システム。

【請求項２８７】

QTLデータをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップ、k平均法を適用するステップ、ファジーk平均法を適用するステップ、Jarvis-Patrickクラスタリングを適用するステップ、自己組織化地図技術を適用するステップ、またはニューラル・ネットワーク技術を適用するステップを含む、請求項274または285に記載のコンピュータ・システム。

【請求項２８８】

QTLデータをクラスター化する前記ステップが、階層型クラスタリング法を適用するス

テップを含み、前記階層型クラスタリング法が凝縮型クラスタリング手順である、請求項287に記載のコンピュータ・システム。

【請求項２８９】

前記凝縮型クラスタリング手順が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである、請求項288に記載のコンピュータ・システム。

【請求項２９０】

前記階層型クラスタリング法が分割型クラスタリング手順である、請求項287に記載のコンピュータ・システム。

【請求項２９１】

前記変換ステップによって作成される各発現統計量から遺伝子発現クラスター地図を構築するステップをさらに含む、請求項275に記載のコンピュータ・システム。

【請求項２９２】

遺伝子発現クラスター地図を構築する前記ステップが、

複数の遺伝子発現ベクトルを作成するステップであって、前記複数の遺伝子発現ベクトル中の各遺伝子発現ベクトルが、複数の生物の各々における前記複数の遺伝子中の1個の遺伝子の発現レベルの測定値であるステップと、

複数の相関係数を計算するステップであって、前記複数の相関係数の各相関係数を前記複数の遺伝子発現ベクトル中の遺伝子発現ベクトル対間で計算するステップと、

前記遺伝子発現クラスター地図を作成するために、前記複数の相関係数に基づいて前記複数の遺伝子発現ベクトルをクラスター化するステップとを含む、請求項291に記載のコンピュータ・システム。

【請求項２９３】

前記QTL相互作用地図を解析する前記ステップが、候補経路群を得るためにQTL相互作用地図を選別するステップを含み、前記選別ステップが、前記遺伝子発現クラスター地図中の前記候補経路群においてQTLを特定するステップを含む、請求項292に記載のコンピュータ・システム。

【請求項２９４】

前記複数の相関係数の各相関係数がピアソン相関係数である、請求項292に記載のコンピュータ・システム。

【請求項２９５】

遺伝子発現クラスター地図を構築する前記ステップが、

複数の遺伝子発現ベクトルを作成するステップであって、前記複数の遺伝子発現ベクトル中の各遺伝子発現ベクトルが、前記複数の遺伝子中の1個の遺伝子であるステップと、

複数の計測量を計算するステップであって、前記複数の計測量の各計測量を前記複数の遺伝子発現ベクトル中の遺伝子発現ベクトル対間で計算するステップと、

前記遺伝子発現クラスター地図を作成するために、前記複数の計測量に基づいて前記複数の遺伝子発現ベクトルをクラスター化するステップとを含む、請求項292に記載のコンピュータ・システム。

【請求項２９６】

前記各計測量が、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量、ピアソン相関係数および二乗ピアソン相関係数からなる群から選択される、請求項295に記載のコンピュータ・システム。

【請求項２９７】

複数の遺伝子発現ベクトルをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップ、k平均法を適用するステップ、ファジーk平均法を適用するステップ、Jarvis-Patrickクラスタリングを適用するステップ、自己組織化地図技術を適用するステップ、またはニューラル・ネットワーク技術を適用するステップを含む、請求項295に記載のコンピュータ・システム。

【請求項２９８】

前記複数の遺伝子発現ベクトルをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が凝縮型クラスタリング手順である、請求項297に記載のコンピュータ・システム。

【請求項２９９】

前記凝縮型クラスタリング手順が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである、請求項298に記載のコンピュータ・システム。

【請求項３００】

前記複数の遺伝子発現ベクトルをクラスター化する前記ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が分割型クラスタリング手順である、請求項297に記載のコンピュータ・システム。

【請求項３０１】

前記QTL相互作用地図を解析する前記ステップが、候補経路群を得るために、QTL相互作用地図を選別するステップを含む、請求項274に記載のコンピュータ・システム。

【請求項３０２】

前記候補経路群を得るための前記選別ステップが、前記QTL相互作用地図中の別のQTLと最も強く相互作用する前記候補経路群のQTLを選択するステップを含む、請求項301に記載のコンピュータ・システム。

【請求項３０３】

前記QTL相互作用地図中の別のQTLと最も強く相互作用する前記QTLが、前記QTL相互作用地図中のQTL間で計算される全相関係数の75%よりも高い相関係数を、前記量的形質遺伝子座相互作用地図中の別のQTLと共有する前記QTL相互作用地図中のQTLである、請求項302に記載のコンピュータ・システム。

【請求項３０４】

前記候補経路群を構成する各QTLが前記候補経路群に属する程度を検定するために、多変量統計モデルを前記候補経路群に適合させる、請求項302に記載のコンピュータ・システム。

【請求項３０５】

前記多変量統計モデルが複数の量的形質を同時に考慮する、請求項304に記載のコンピュータ・システム。

【請求項３０６】

前記多変量統計モデルが、前記候補経路群中のQTL間のエピスタシス相互作用を探索する、請求項304に記載のコンピュータ・システム。

【請求項３０７】

前記遺伝マーカー・セットが、一塩基多型(SNP)、ミクロサテライト・マーカー、制限断片長多型、短鎖縦列反復、DNAメチル化マーカーまたは配列長多型を含む、請求項274に記載のコンピュータ・システム。

【請求項３０８】

前記分類モジュールが、

(i)前記複数の生物のすべてまたは一部の表現型データに基づいて前記集団を複数の表現型群に区分化する命令と、

(ii)極端な表現型を示す前記複数の表現型群中の1組の極端生物を特定する命令と、

(iii)前記複数の細胞構成成分中の細胞構成成分を特定する命令であって、特定された各細胞構成成分が、前記極端生物セットから得られた前記各細胞構成成分の細胞構成成分測定値が前記複数の表現型群のすべてまたは一部を識別する特性を有する命令と、

(iv)前記特定された細胞構成成分のすべてまたは一部から導出される確率分布を用いて分類子を構築する命令とを含む、請求項244に記載のコンピュータ・システム。

【請求項３０９】

前記表現型データがバイナリ・イベントを含む、請求項308に記載のコンピュータ・システム。

【請求項３１０】
前記表現型データが、前記集団内の各生物の1個を超える表現型測定値を含む、請求項308に記載のコンピュータ・システム。
【請求項３１１】
前記表現型データが前記複数の生物中の各生物が形質を示すかどうかに関する判定を含み、前記区分化命令が、前記生物が前記形質を示すときに第1の表現型群中の前記複数の生物中の1つの生物を配置するステップと、前記生物が前記形質を示さないときに第2の表現型群中の前記複数の生物中の1つの生物を配置するステップとを含む、請求項308に記載のコンピュータ・システム。
【請求項３１２】
前記表現型データが、前記複数の生物のすべてまたは一部に対して作成される複数の表現型測定値を含み、前記区分化命令が、
(A)複数の表現型ベクトルを作成する命令であって、前記複数の表現型ベクトル中の各表現型ベクトルが前記複数の生物中の1つの生物に対応し、前記複数の表現型ベクトル中の各表現型ベクトルが前記各表現型ベクトルに対応する前記生物から得られる複数の表現型測定値を含む命令と、
(B)前記複数の表現型ベクトルを複数のクラスターにクラスター化する命令であって、前記複数のクラスター中の各クラスターが、前記複数の表現型群中の1個の表現型群を表す命令とを含む、請求項308に記載のコンピュータ・システム。
【請求項３１３】
前記クラスター化ステップが、階層型クラスタリング法、k平均法、ファジーk平均法、Jarvis-Patrickクラスタリング、自己組織化地図技術またはニューラル・ネットワーク技術を含む、請求項312に記載のコンピュータ・システム。
【請求項３１４】
前記クラスター化ステップが、階層型クラスタリング法を適用するステップを含み、前記階層型クラスタリング法が凝縮型クラスタリング手順である、請求項312に記載のコンピュータ・システム。
【請求項３１５】
前記凝縮型クラスタリング手順が、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである、請求項314に記載のコンピュータ・システム。
【請求項３１６】
前記階層型クラスタリング法が分割型クラスタリング手順である、請求項315に記載のコンピュータ・システム。
【請求項３１７】
生物が、前記集団が示す表現型に関して前記集団の上位30パーセンタイルまたは下位30パーセンタイルであるときに、前記極端な表現型である、請求項308に記載のコンピュータ・システム。
【請求項３１８】
生物が、前記集団が示す表現型に関して前記集団の上位10パーセンタイルまたは下位10パーセンタイルであるときに、前記極端な表現型である、請求項308に記載のコンピュータ・システム。
【請求項３１９】
前記極端生物セットが5個体を超える生物である、請求項308に記載のコンピュータ・システム。
【請求項３２０】
前記極端生物セットが2～100個体の生物である、請求項308に記載のコンピュータ・システム。
【請求項３２１】
前記極端生物セットが1000個体未満の生物である、請求項308に記載のコンピュータ・

システム。
【請求項３２２】
前記特定命令(iii)が、所定の細胞構成成分の複数の細胞構成成分測定値をt検定にかけるステップを含み、前記複数の細胞構成成分測定値が前記極端生物セットから得られる、請求項308に記載のコンピュータ・システム。
【請求項３２３】
前記特定命令(iii)が、前記特定された細胞構成成分を多変量解析にかけるステップを含む、請求項308に記載のコンピュータ・システム。
【請求項３２４】
前記特定命令(iii)によって特定された前記細胞構成成分を、前記構築命令(iv)を実行する前に削減する、請求項308に記載のコンピュータ・システム。
【請求項３２５】
前記特定命令(iii)によって特定された前記細胞構成成分を、段階的回帰、総当り回帰、主成分分析または重判別分析によって削減する、請求項324に記載のコンピュータ・プログラム製品。
【請求項３２６】
前記特定命令(iii)によって特定された前記細胞構成成分を、確率論的検索方法によって削減する、請求項324に記載のコンピュータ・プログラム製品。
【請求項３２７】
前記確率論的検索方法がシミュレーテッド・アニーリングまたは遺伝アルゴリズムである、請求項326に記載のコンピュータ・プログラム製品。
【請求項３２８】
前記特定命令(iii)によって特定された前記細胞構成成分をクラスタリングによって削減し、前記特定命令(iii)によって特定された前記細胞構成成分ではなく、前記クラスタリングによって生成したクラスターを前記構築命令(iv)に使用する、請求項324に記載のコンピュータ・プログラム製品。
【請求項３２９】
前記構築命令(iv)が、前記確率分布を用いてニューラル・ネットワークを訓練するステップを含む、請求項308に記載のコンピュータ・プログラム製品。
【請求項３３０】
前記構築命令(iv)が、前記確率分布が演繹的情報として役立つベイズの決定理論を使用するステップを含む、請求項308に記載のコンピュータ・プログラム製品。
【請求項３３１】
前記構築命令(iv)が、線形判別分析、線形プログラミング・アルゴリズムまたはサポート・ベクトル・マシンを使用するステップを含む、請求項308に記載のコンピュータ・プログラム製品。
【請求項３３２】
前記分類体系が、前記分類子を用いて前記集団のすべてまたは一部を分類するステップを含む、請求項308に記載のコンピュータ・プログラム製品。
【請求項３３３】
量的遺伝解析に使用する複数の亜集団を得るために同じ種の複数の生物Sを細分するコンピュータ・システムであって、前記複数の生物S中の1つまたは複数の生物が複合形質を示し、前記コンピュータ・システムが、
中央処理装置と、
前記中央処理装置に接続され、分類モジュールを保存するメモリとを含み、前記分類モジュールが、
(a)前記複合形質に関してそれぞれが独立した極端である前記複数の生物S内の2群以上の生物を特定する命令と、
(b)前記複数の生物S内の前記2群以上の生物を識別することができる1組の細胞構成成分Cを決定する命令と、

(C)前記細胞構成成分Cセット中の各細胞構成成分iに対して、前記複合形質と関連する第1のQTLと相互作用または重複するQTLを有する1個または複数の細胞構成成分を特定するために、前記複数の生物Sの少なくとも一部の各生物からそれぞれ測定される前記細胞構成成分iの量を量的形質として用いて、前記細胞構成成分iについてのQTL解析を実施する命令と、

(d)ステップ(c)で特定された各細胞構成成分の測定量に基づいて前記複数の生物Sをクラスター化し、それによって前記複数の亜集団を得る命令とを含む、コンピュータ・システム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明の分野は、複合形質に関連する遺伝子および生物学的経路を特定するコンピュータ・システムおよび方法に関する。特に、本発明は、複合形質を構成形質に細分するコンピュータ・システムおよび方法に関する。次いで、遺伝子発現データおよび遺伝子データを使用して、構成形質と関連する遺伝子-遺伝子相互作用、遺伝子-表現型相互作用および生物学的経路を特定する。

【背景技術】

【０００２】

本願は、参照によりその全体を本明細書に援用する、2002年5月20日に出願された米国仮出願第60/382,036号の利益を35 U.S.C. §119(e)の下に主張するものである。本願は、参照によりその全体を本明細書に援用する、2003年4月2日に出願された米国仮出願第60/460,304号の利益も35 U.S.C.§119(e)の下に主張するものである。

【０００３】

2.1. 現行の生物学的標的選択手法

医薬品研究に適切な生物学的標的(タンパク質)の探索は、重要な課題である。標準的な医薬手法においては、特定のタンパク質が疾患と関連している。生物学的標的が特定されると、生物学的標的の活性を変える化合物が開発され、そのような化合物が病態を解消するように標的の活性を変えることが期待される。

【０００４】

生物学的標的を特定する一手法は、病態において上方制御または下方制御される生物学的カスケードを理解することである。このような生物学的カスケードが解明されると、カスケード内の個々のタンパク質を生物学的標的として選択することができる。例えば、IL-15は、究極的に炎症性疾患をもたらす現象のカスケードにおいて初期に出現する免疫系シグナル分子であるので、重要な疾患標的である。IL-15は、炎症において中心的な役割を果たすことが判明している別のサイトカインであるTNFアルファの産生、ならびに炎症性T細胞の動員を引き起こす。これらのT細胞は、より多くのIL-15の産生を促進し、このサイクルが段階的に増大する。IL-15がある役割を果たす炎症性疾患としては、関節炎、炎症性ボウル疾患、セリアック病、乾せんなどがある。生物学的標的を特定するこの手法の欠点は、往々にして、疾患に関連する生物学的カスケードについての情報が知られておらず、あるいは十分理解されていないことである。この詳細な知識が得られないと、重篤な、さらには致死的な副作用をもたらす恐れがある。例えば、十分理解されていない生物学的標的に対して開発された阻害剤は、実際に、いくつかの基本的な生物学的プロセスを阻害し、あるいは変化させる。

【０００５】

別の標的選択手法では、細胞表面受容体クラスなどの特定のクラスの遺伝子およびそれらの産物が、薬物標的としてのそのクラスの歴史的適合性に基づいてまず選択される。この選択がなされた後、次のステップは、選択されたタイプのすべての可能なタンパク質候補を特定するために、配列データベースならびに文献を調査することである。重要な新しい遺伝子ファミリーはいつでも発見される可能性があるので、最初の包括的な調査が完了した後でも、配列データベースおよび文献の定期的な調査が必要である。100個以上の候

補タンパク質を含むこともある配列データベースおよび文献調査によって特定される可能なタンパク質候補のリストから、最も有望な標的が特定され提示される。ほとんどの候補についての利用可能な情報は不十分であり、常に追加され、見直されているので、この調査は繰り返されることが多い。この手法の欠点の1つは、これが検討中の特定のクラスの遺伝子によって影響を受ける疾患のみに限られることである。

【０００６】

別の標的選択手法では、特定の疾患、または少なくとも特定の治療カテゴリーの疾患にまず焦点が絞られる。初期の実験は、遺伝子発現パターンを解析すること、様々な疾患段階から得られる組織を正常組織と比較すること、および遺伝子発現に対する現行の有効な薬物の効果がある場合にはそれを調べることに焦点が絞られている。このプロセスによって、数百の遺伝子についての情報がもたらされる。この情報を、適切な標的を決定するために解析する。この手法の欠点の1つは、データ解析にかなりの投資をしても、適切な標的を特定できない恐れがあることである。これは、特に、癌、関節炎などの複雑性疾患に当てはまる。このような疾患の問題は、この疾患を有する集団に異質性があることである。例えば、癌集団においては、個々の患者は異なるタイプの癌に罹っている可能性がある。この異質性がはなはだしい場合、大集団から得られる数百の遺伝子についての情報解析は、異質性を考慮しない限り、適切な薬物標的を特定するには至らない。

【０００７】

遺伝学は、いくつかの標的選択手法において使用される。遺伝学は、形質に関連する遺伝子および経路を特定するのに特に有用である。この知識を使用して、今度は、薬剤開発に適切な標的を特定することができる。1つの遺伝子技術は、遺伝子を表現型全体と結び付ける連鎖解析である。一例では、連鎖解析を用いて、嚢胞性線維症が嚢胞性線維症遺伝子の突然変異に関連付けられた。遺伝学を表現型全体に適用する手法は、嚢胞性線維症など遺伝子の突然変異が疾患を引き起こす特定の状況のみに限られる欠点がある。

【０００８】

生物学的標的を見出す別の手法において、遺伝学は、臨床的に細分された表現型に適用される。すなわち、大きな患者集団において特定の疾患に関連する危険因子を測定し、これらの測定結果を用いて、患者集団のゲノム内の特定の量的形質遺伝子座を遺伝学的手法により危険因子と関連付ける。例えば、心疾患の場合、患者集団におけるトリグリセリド・レベル、HDL/LDLレベル、コレステロール・レベルなどの表現型を測定する。次いで、1個または複数のこれら臨床的に細分された表現型を、ヒト・ゲノム中の量的形質遺伝子座(QTL)に関連付ける。本明細書で使用する量的形質遺伝子座(QTL)は、量的形質に影響を及ぼすゲノム領域である。QTLは、臨床的に細分された表現型と関連する遺伝子を特定するために、QTLマッピングを用いて解析される。QTLマッピング方法は、複合形質に影響を及ぼすゲノム領域を理解し精査するために表現型と遺伝子型の関連を統計解析するものである。

【０００９】

遺伝学分野で重要な進歩は、ヒトなどの種の詳細な遺伝地図を構築するために使用することができる分子/遺伝マーカーの膨大な集合が構築されたことである。これらの地図は、単一マーカー・マッピング(single-marker mapping)、区間マッピング、複合区間マッピング、多形質マッピングなどの量的形質遺伝子座(QTL)マッピング法に使用される。(総説としては、Doerge、2002、Nature Reviews:Genetics 3:43～62を参照されたい)。改善されたマーカー地図が開発されたにもかかわらず、特定の複雑な表現型に関連するすべての領域を特定する目標は、一般に、QTLの絶対数、QTL間の可能なエピスタティス(epistatis)または相互作用、ならびに多数のさらなる変動源のために達成が困難である。

【００１０】

形質に関連する遺伝子および経路を特定するために遺伝学を使用することが標準的な理論的枠組である。まず、ゲノム全般にわたる連鎖解析を、家族を主体とするデータ中の数百の遺伝マーカーを使用して実施して、形質に関連する広範な領域を特定する。この標準的な連鎖解析の結果、形質を制御する領域が特定され、それによって例えばヒト・ゲノム

中の30,000個以上の遺伝子から、形質に関連する特定のゲノム領域中のおそらくはわずか500～1000個の遺伝子に絞り込まれる。しかし、連鎖解析によって特定された領域は、形質に関連する候補遺伝子を特定するには依然としてあまりにも広すぎる。したがって、そのような連鎖研究は、一般に、連鎖領域中のより高密度のマーカーを用いて連鎖領域を詳細にマッピングし、解析における家族数を増やし、別の検定集団を特定することによって追跡調査される。これらの努力によって、形質に関連する特定領域中の約100個の遺伝子のより狭いゲノム領域にさらに焦点が絞られる。このより狭く規定された連鎖領域でさえ、検証すべき遺伝子数は依然として非現実的なほど多い。したがって、この段階での検討は、この領域中の既知遺伝子の推定機能、およびその機能と形質との潜在的な関連性に基づいて、候補遺伝子を特定することに焦点が絞られる。この手法は、遺伝子についての現在の知見に制約されるので問題がある。そのような知見は限られており、解釈によって左右されることが多い。その結果、研究者は道に迷うことが多く、形質に影響を及ぼす遺伝子を特定できない。

【００１１】

ヒトの疾患などの複合形質に関連する遺伝子を特定するのに標準的な遺伝的手法があまり成功していない理由は多数ある。第1に、心疾患、肥満、癌、骨粗しょう症、精神分裂病、その他多くの疾患など一般的なヒトの疾患は、多遺伝子である点で複雑である。すなわち、いくつかの異なる生物学的経路にわたって多数の遺伝子が関与している可能性があり、遺伝的なシグネチャーをあいまいにする複雑な遺伝子-環境相互作用が関与している。第2に、疾患が複雑なために、疾患を起こし得る様々な生物学的経路に異質性が生じる。したがって、あらゆる所与の異質な集団においては、疾患を起こし得るいくつかの異なる経路全体に欠陥が存在し得る。そのため、任意の所与の経路に対して遺伝的シグナルを特定することが困難になる。遺伝的検定に参加している多数の集団は、疾患に関して異質であるので、複数の経路にわたる複数の欠陥が集団内で作用して疾患を引き起こす。第3に、複合形質と分子マーカーの統計的に有意な関連性が特定されたときでも、ゲノム領域は、通常大き過ぎて、これらの領域に焦点を絞るために用いられる後続実験には費用がかかることが多い。この技術の限界によって、一般に、複合形質に影響を及ぼす、または複合形質に関連する少なくとも1個の領域が失われることになる。複合形質変化の原因となる遺伝子を含むゲノム領域が特定された成功事例においては、この方法の開始から終了までの費用および時間は、科学的、経済的、または医学的に重要な問題に広範に適用するにはかかり過ぎることが多い。第4に、形質および病態自体が明確に定義されないことが多い。第5に、患者集団における病態が異質である場合、臨床的に細分された表現型に関連するQTLの決定能力がこの異質性のために失われる。したがって、部分表現型(subphenotype)が、生物学的経路の異なるセットを包含していても、これらの部分表現型は見過ごされることが多い。そのため、関連性を検出する能力が低下する。

【００１２】

2.2. 複合形質

セクション2.1に示したように、適切な生物学的標的を見つけて、複合形質を緩和させる既知の手法は問題が多い。「複合形質」という用語は、単一の遺伝子座に起因し得る古典メンデルの劣性または優性遺伝を示さないあらゆる表現型を意味する。例えば、LanderおよびSchork、1994、Science 265:2037を参照されたい。このような「複合」形質には、心疾患、高血圧、糖尿病、癌、感染などに対する感受性が含まれる。複合形質は、同じ遺伝子型が(偶然、環境または他の遺伝子との相互作用の効果のために)異なる表現型を生じ得ることによって、または異なる遺伝子型が同じ表現型を生じ得ることによって、遺伝子型と表現型の単純な対応が損なわれたときに発生する。

【００１３】

複合形質との完全な同時分離を示す遺伝マーカーを見出すことは困難なことが多い。その理由は、複合形質に関連する基礎的な問題に帰することができる。これらの基礎的問題には、不完全浸透率および表現型模写が含まれる。素因となる対立遺伝子を受け継いだ個体の一部が、疾患を顕在化させないことがある(不完全浸透率)のに対し、素因となる対立

遺伝子を受け継いでいない別の個体が、それにもかかわらず環境または無秩序な原因の結果、疾患に罹る(表現型模写)ことがある。したがって、所与の遺伝子座における遺伝子型は、疾患確率に影響を及ぼし得るが、その結果を特定することは完全にはできない。各遺伝子型Gの疾患確率を規定する浸透度関数f(G)は、年齢、性別、環境、他の遺伝子などの非遺伝的要因にも左右され得る。例えば、BCRA1遺伝子座に突然変異を有する女性における40歳、55歳および80歳での乳癌のリスクは、37%、66%および85%であるのに対し、非変異女性においては0.4%、3%および8%である(Easton等、1993、Cancer Surv. 18:1995;Ford等、1994、Lancet 343:692)。このような症例においては、素因となる対立遺伝子が、罹患していない個体でも存在することがあり、あるいは罹患している個体でも存在しないことがあるので、遺伝マッピングが妨げられる。

【００１４】

複合形質に伴う別の問題は、いくつかの遺伝子のうちいずれか1個の様々な突然変異が同一の表現型を生じ得ることである(遺伝的異質性)。したがって、遺伝的異質性がある場合、2つの患者が、異なる遺伝的理由によって同じ疾患に罹るかどうかは、遺伝子をマッピングするまで判定が困難な場合がある。ヒトにおける遺伝的異質性のために生じる複雑性疾患の例は、多発性嚢胞腎(Reeders等、1987、Human Genetics 76:348)、早期発症型アルツハイマー病(George-Hyslop等、1990、Nature 347:194)、若年発症成人型糖尿病(Barbosa等、1976、Diabete Metab. 2:160)、遺伝性非腺腫性大腸癌(Fishel等、1993、Cell 75:1027)毛細血管拡張性運動失調症(JaspersおよびBootsma、1982、Proc. Natl. Acad. Sci. U.S.A. 79:2641)、肥満、非アルコール性脂肪性肝炎(NASH)(James & Day、1998、J. Hepatol. 29:495～501)、および色素性乾皮症(De Weerd-Kastelein、Nat. New Biol. 238:80)である。遺伝的異質性は、一部の家族においては染色体領域を疾患と同時分離することができるが、別の家族においては同時分離することができないので遺伝マッピングが妨げられる。

【００１５】

いくつかの複合形質に関連するさらに別の問題は、ポリジーン遺伝現象である。ポリジーン遺伝は、複合形質が、複数の遺伝子における突然変異の同時発生を必要とするときに起こる。ヒトにおけるポリジーン遺伝の例は、網膜色素変性症の一形式であり、これには、ペルフェリン(perpherin)/RDSおよびROM1遺伝子において異型接合的な突然変異が存在する必要がある(Kajiwara等、1994、Science 264:1604)。RDSおよびROM1によってコードされるタンパク質は、光受容体外部顔料膜性円板(photoreceptor outer pigment disc membrane)において相互作用すると考えられる。ポリジーン遺伝では、単一遺伝子座が、明確に区別された形質または高い値の量的形質を生じることを厳密に必要としないので、遺伝マッピングが複雑になる。

【００１６】

疾患原因対立遺伝子が集団中に高頻度で存在する場合、疾患原因対立遺伝子の頻度が高いと、簡単な形質でさえマッピングするのが困難になる。これは、複数の独立したDのコピーが系統中で分離し、一部の個体がDに対して同型接合的であり、2個の相同染色体のうちのどちらかが、影響を受けた子孫に受け継がれるので、近傍の遺伝マーカーにおいてDと特異的対立遺伝子の連鎖が見られないという問題によって、予想される疾患のメンデル性遺伝パターンが混乱するからである。遅発性アルツハイマー病は、疾患原因対立遺伝子が高頻度であることによって生じる障害の一例である。初期の連鎖解析によって、19q染色体と連鎖しているわずかな証拠が見出されたが、ロッド・スコア(連鎖の対数尤度比)が比較的低いままであり、どんな精度でも連鎖の位置を正確に示すことが困難であったので、多数の観測者によって棄却された(Pericak-Vance等、1991、Am J. Hum. Genet. 48:1034)。この混乱は、アポリポタンパク質E4型対立遺伝子が、染色体19上の主要な原因遺伝子であると考えられることが発見されて最終的に解決した。対立遺伝子が高頻度(ほとんどの集団においては約16%)であると、従来の連鎖解析は妨害された(Corder等、1993、Science 261:921)。疾患原因対立遺伝子が高頻度であることは、遺伝的異質性が存在する場合にはさらに大きな問題になる。

【発明の開示】
【発明が解決しようとする課題】
【００１７】
したがって、上記背景から、当分野で求められているのは、複雑性疾患などの複合形質に影響を及ぼす遺伝子および生物学的経路を特定する改良方法である。このような遺伝子および生物学的経路を特定する改良方法によって、生物学的標的選択が改善される。
【００１８】
本明細書における参考文献の考察または引用は、このような参考文献が本発明の従来技術であることを認めるものと解釈すべきではない。
【課題を解決するための手段】
【００１９】
本発明は、複合形質に影響を及ぼす遺伝子および生物学的経路を特定する新規なコンピュータ・システムおよび方法を提供する。患者集団は、遺伝子発現測定値、タンパク質発現測定値などの細胞構成成分測定値に基づいて細分される。次いで、細分された患者集団を、各部分群中の標的を特定するように設計された遺伝的方法にかける。本発明に取り入れられたこの手法は、不完全浸透率、表現型模写、遺伝的異質性、ポリジーン遺伝、および高頻度の疾患原因対立遺伝子を含めて、複合形質によって生じる諸問題を回避するのに役立つので有利である。量的遺伝解析にかける前に疾患集団を選別するとさらに有利になる。患者集団の選別によって、量的遺伝解析にかけられる疾患集団に存在する異質性が減少する。これによって、量的遺伝解析を通して得られる結果の正確度および信頼性が改善される。これらの改善によって、創薬標的として役立ち得る複合形質に関連する生物学的経路中の遺伝子を特定する能力が強化される。
【００２０】
本発明の一実施形態は、集団中の複数の生物によって示される複合形質の量的形質遺伝子座を特定する方法を提供する。複数の生物はある単一の種を含む。この方法においては、集団中の各生物を複数の亜集団中の一亜集団に分類する分類体系を用いて、集団を複数の亜集団に分割する。この分類体系は、各生物から得られる複数の細胞構成成分測定値を利用するものである。また、複数の亜集団中の各亜集団では、複合形質に対する量的形質遺伝子座を特定するために、亜集団の量的遺伝解析を実施する。一部の実施形態においては、各生物から得られる細胞構成成分測定値は、転写状態の測定値または翻訳状態の測定値である。別の実施形態においては、細胞構成成分は、複数の代謝産物を含み、複数の細胞構成成分測定値は、各生物における複数の代謝産物レベルを測定するメタボロミクス(metabolomic)技術などの細胞の表現型技術によって導出される。一部の実施形態においては、複数の細胞構成成分測定値は、遺伝子発現レベル、mRNA存在量、タンパク質発現レベルまたは代謝産物レベルを含む。
【００２１】
本発明の一部の実施形態においては、複合形質は、集団において不完全浸透率を示す対立遺伝子を特徴とする。一部の実施形態においては、複合形質は、集団内の生物が罹る疾患であり、その生物はその疾患の素因となる対立遺伝子を受け継いでいない。一部の実施形態においては、複合形質は、集団に相当する単一種のゲノム中の複数の異なる遺伝子のいずれかが突然変異するときに発生する。一部の実施形態においては、複合形質は、集団に相当する単一種のゲノム中の複数の遺伝子に突然変異が同時に起こる必要がある。一部の実施形態においては、複合形質は、集団における高頻度の疾患原因対立遺伝子に関連する。最後に、一部の実施形態においては、複合形質は、単一の遺伝子座に起因しうるメンデルの劣性遺伝または優性遺伝を示さない表現型である。
【００２２】
本発明のさらに別の実施形態においては、集団を分割するステップは、さらに、クラス予測変数(class predictor)を利用可能かどうかを決定するステップを含む。クラス予測変数が利用可能なときには、教師付き分類体系(supervised classification scheme)を使用して、集団中の各生物を複数の亜集団中の亜集団に分類する。クラス予測変数が利用不

可能なときには、教師なし分類体系を使用して集団中の各生物を複数の亜集団中の亜集団に分類する。

【００２３】

本発明の一部の実施形態においては、分類体系は教師付き分類体系であり、別の実施形態においては分類体系は教師なし分類体系である。本発明による教師付き分類体系は、これらだけに限定されないが、線形判別分析および線形回帰を含む技術を使用する。線形回帰は、これらだけに限定されないが、多重線形回帰、部分最小二乗回帰および主成分回帰を含む統計の広範なカテゴリーである。本発明による教師なし分類体系としては、階層型クラスター分析、非階層型クラスター分析、人工ニューラル・ネットワークおよび自己組織化地図が挙げられるが、これらだけに限定されない。

【００２４】

本発明の一部の実施形態においては、階層型クラスター分析は、最短距離アルゴリズム、最長距離(farthest-neighbor)法、平均連結(average linkage)法、重心(centroid)法または平方和アルゴリズムを用いて、(i)集団内の一生物から得られる複数の細胞構成成分測定値と、(ii)別の集団内の生物から得られる複数の細胞構成成分測定値との類似度を求める。一部の実施形態においては、階層型クラスター分析は、凝縮型(agglomerative)クラスタリング、多形質分割型(polythetic divisive)クラスタリングまたは単形質分割型(monthetic divisive)クラスタリングである。本発明の一部の実施形態においては、凝縮型クラスタリング手順は、ピアソン相関係数、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量または二乗ピアソン相関係数を使用して、(i)集団内の一生物からの複数の細胞構成成分測定値と(ii)別の集団内の生物からの複数の細胞構成成分測定との類似度を求める。本発明のさらに別の実施形態においては、非階層型クラスター分析は、K平均クラスタリング、ファジーk平均クラスタリングまたはJavis-Patrickクラスタリングである。本発明の一部の実施形態においては、人工ニューラル・ネットワークは、Kohonen人工ニューラル・ネットワークまたは自己連想ニューラル・ネットワークである。

【００２５】

本発明の一部の実施形態においては、連鎖解析、複数の細胞構成成分測定値を表現型形質として使用する量的形質遺伝子座(QTL)解析の一形式、または関連解析を用いて量的遺伝解析を実施する。QTL解析が複数の細胞構成成分測定値を表現型形質として使用する一部の実施形態においては、QTL解析は以下のステップを含む。第1に、複数のQTL解析から得られるQTLデータは、クラスター化されてQTL相互作用地図を形成する。複数のQTL解析の各QTL解析を、QTLデータを生成するために遺伝マーカー地図および量的形質を用いて複数の生物のゲノム中の複数の遺伝子中の1個の遺伝子Gに対して実施する。各QTL解析では、量的形質は、集団中の各生物のQTL解析が実施された遺伝子Gに対する発現統計量を含む。また、遺伝マーカー地図は、複数の生物に関連する1組の遺伝マーカーから構築される。第2に、QTL相互作用地図を解析して、量的形質に関連するQTLを特定する。

【００２６】

本発明の一部の実施形態は、コンピュータ・システムとともに使用されるコンピュータ・プログラム製品を提供する。コンピュータ・プログラム製品は、コンピュータ読み取り可能な記憶媒体、およびその中に埋め込まれたコンピュータ・プログラム機構を備える。コンピュータ・プログラム機構は、分類モジュールおよび量的遺伝解析モジュールを備える。分類モジュールは、集団内の複数の生物を複数の亜集団に分類体系を用いて分割する命令を含む。分類体系は、集団中の各生物を複数の亜集団中の一亜集団に分類する。複数の生物は単一種を含み、分類体系は集団中の各生物からの複数の細胞構成成分測定値を使用する。量的遺伝解析モジュールは、複数の生物中の1個または複数の生物によって示される複合形質に対する量的形質遺伝子座を特定するために、複数の亜集団中の各亜集団に対して、亜集団についての量的遺伝解析を実施する命令を含む。

【００２７】

本発明の一部の実施形態は、集団内の複数の生物によって示される複合形質に対する量

的形質遺伝子座を特定するコンピュータ・システムを提供する。コンピュータ・システムは、中央処理装置、および中央処理装置と接続されたメモリを備える。メモリには、分類モジュールおよび量的遺伝解析モジュールが保存される。分類モジュールは、集団内の複数の生物を複数の亜集団に分類体系を用いて分割する命令を含む。分類体系によって、集団中の各生物は複数の亜集団中の一亜集団に分類される。複数の生物は単一種を含み、分類体系は、集団中の各生物から得られる複数の細胞構成成分測定値を使用する。量的遺伝解析モジュールは、集団内の1個または複数の生物によって示される複合形質に対するQTLを特定するために、複数の亜集団中の各亜集団に対して、亜集団の量的遺伝解析を実施する命令を含む。

【００２８】

本発明の別の態様は、量的遺伝解析に使用する複数の亜集団を取り出すために、複数の生物Sを細分する方法を提供する。複数の生物Sは、単一種を含む。また、複合形質は、複数の生物S内の1個または複数の生物によって示される。2つ以上の生物群が、複合形質に関してそれぞれ個別の極値を示す複数の生物S内で特定される。単一種に関連する細胞構成成分Gのセットが決定される。細胞構成成分Gのセットは、複数の生物S内の2つ以上の生物群を識別することができる。細胞構成成分Gのセット中の各細胞構成成分iに対して、複数の生物Sの少なくとも一部の各生物からそれぞれ測定される細胞構成成分iの量を量的形質として用いて細胞構成成分iに対するQTL解析を実施する。その結果、複合形質に関連する第1のQTLと相互作用または重複するQTLを有する1個または複数の細胞構成成分が特定される。複数の生物Sを、特定された各細胞構成成分の測定量に基づいてクラスター化する。各細胞構成成分の測定量は、複数の生物Sの少なくとも一部の各生物からそれぞれ測定され、それによって複数の亜集団が取り出される。一部の実施形態においては、この方法は、さらに、複合形質のQTLを特定するために、複数の亜集団中の一亜集団について一連のQTL解析を実施するステップを含む。一連のQTL解析の各QTL解析は、細胞構成成分Gのセット中の細胞構成成分の測定量を量的形質として使用する。また、細胞構成成分の測定量は、複数の生物Sの少なくとも一部の各生物からそれぞれ測定される。

【００２９】

図面のいくつかの図を通して、同じ参照番号は対応する部分を示す。

【発明を実施するための最良の形態】

【００３０】

本発明は、コンピュータ・システム、コンピュータ・プログラム製品、および遺伝子を単一種の複数の生物内の1個または複数の生物によって示される形質と関連させる方法を提供する。例示的な生物としては、植物および動物が挙げられるが、これらだけに限定されない。具体的な実施形態においては、例示的な生物としては、トウモロコシ、マメ、イネ、タバコ、ジャガイモ、トマト、キュウリ、フルーツの木、キャベツ、レタス、コムギなどの植物が挙げられるが、これらだけに限定されない。具体的な実施形態においては、例示的な生物としては、哺乳動物、霊長類、ヒト、マウス、ネズミ、イヌ、ネコ、ヒヨコ、ウマ、ウシ、ブタ、サルなどの動物が挙げられるが、これらだけに限定されない。さらに他の具体的な実施形態においては、生物としては、ショウジョウバエ、酵母、ウイルスおよび線虫が挙げられるが、これらだけに限定されない。ある場合には、遺伝子は、遺伝子産物が関与する生物学的経路を特定することによって、形質に関連付けられる。本発明の一部の実施形態においては、目的形質は、ヒトの疾患などの複合形質である。目的とする例示的なヒト疾患および形質を以下の5.14セクションに記載する。

【００３１】

本発明は、量的形質を使用して、複合形質に影響を及ぼすゲノム領域の特定を支援する。一実施形態においては、量的形質は、細胞構成成分の発現レベルである。細胞構成成分としては、生体系における遺伝子発現レベル、mRNA存在量、タンパク質発現レベルおよび代謝産物レベルが挙げられるが、これらだけに限定されない。量的形質のより複雑な例を以下に示す。量的形質は、連続形質(例えば、脂質レベル)でも不連続形質(例えば、病態の有無)でもよい。本発明は、QTL解析を利用する。QTL解析の例は、セクション5.2に記載

した連鎖解析、セクション5.3に記載した細胞構成成分レベルを用いたQTL解析、およびセクション5.4に記載した関連解析であるが、これらだけに限定されない。

【００３２】

5.1. 本発明の概要

本発明の一態様によって取られる一般的ステップは、図1および図2を参照して理解することができる。図1は、本発明の一実施形態によって実施される処理ステップである。また、図2は、本発明の一実施形態によって操作されるシステム200である。システム200は、少なくとも1個のコンピュータ202を備える(図2)。コンピュータ202は、中央処理装置222、プログラム・モジュールおよびデータ構造を保存するメモリ224、ユーザー入/出力装置226、サーバー202を通信ネットワークを介して別のコンピュータに接続するネットワーク・インターフェース228(図示せず)、不揮発性ストレージ230を制御するディスク・コントローラ225、およびこれらの部品を相互接続する1個または複数のバス234を含めて、標準的な部品を備える。ユーザー入/出力装置226は、マウス236、ディスプレイ238、キーボード208などの1個または複数のユーザー入力/出力部品を備える。

【００３３】

メモリ224は、本発明によって使用されるいくつかのモジュールおよびデータ構造を含む。システム操作中のいかなる時点に於いても、メモリ224に保存されたモジュールおよび/またはデータ構造の一部は、モジュールおよび/またはデータ構造の別の一部が不揮発性ストレージ230に保存されている間は、ランダム・アクセス・メモリ224に保存されることを理解されたい。典型的な実施形態においては、メモリ224は、オペレーティング・システム240を含む。オペレーティング・システム240は、様々な基本的システム・サービスを取扱い、ハードウェアに依存するタスクを実施する手順を含む。メモリ224は、さらに、ファイル管理用のファイル・システム242を含む。一部の実施形態においては、ファイル・システム242は、オペレーティング・システム240の一部である。

【００３４】

図1および図2に示す実施形態においては、複雑性疾患などの複合形質を示す集団Pを細分する。次いで、細分された各集団を、有用な創薬標的である遺伝子を特定するために遺伝解析にかける。このプロセスは、集団P内の各生物から遺伝子発現測定値、タンパク質発現測定値などの細胞構成成分測定値を得るステップ(図1、ステップ102)から出発する。集団Pは、複雑性疾患などの複合形質を示す単一種の任意の集団として本明細書では定義される。一部の実施形態においては、この集団は、5、25、50、100、500、1000個またはそれ以上の生物を含む。細胞構成成分測定値は、転写状態測定値(セクション5.11参照)、翻訳データ測定値、(セクション5.12参照)、または細胞の表現型技術、異なる組織タイプ間の細胞構成成分レベル差の分析など別の測定値(セクション5.13参照)から導出することができる。細胞の表現型技術は、複数の代謝産物を測定するメタボロミクス技術を含む。細胞構成成分としては、生体系における遺伝子発現レベル、mRNA存在量、タンパク質発現レベルおよび代謝産物レベルが挙げられるが、これらだけに限定されない。遺伝子をコードするmRNA発現レベルおよび/またはタンパク質発現レベルなどの様々な細胞構成成分レベルは、薬物療法、および細胞の生物学的状態の別の撹乱(perturbation)に応じて変化することが知られている。したがって、複数のこのような「細胞構成成分」の測定値は、細胞の生物学的状態に対する撹乱の影響(affect)についての豊富な情報を含んでいる。このような測定値の集合は、一般に、細胞の生物学的状態の「プロファイル」と称する。別の非限定的な実施形態においては、細胞の生物学的状態のプロファイルを、細胞の生物学的状態の混在した態様で形成することができる。応答データは、例えば、あるmRNA存在量の変化、あるタンパク質存在量の変化、およびあるタンパク質活性の変化から作成することができる。

【００３５】

本発明の一部の実施形態においては、集団P内の各生物中の5個以上の細胞構成成分レベルをステップ102で測定する。本発明の別の実施形態においては、集団P内の各生物中の10、20、30、40、50、100、200個またはそれ以上の細胞構成成分レベルをステップ102で測

定する。本発明の一部の実施形態においては、集団P内の各生物中の300、400、500、800、1200、1500個またはそれ以上の細胞構成成分の細胞構成成分レベルをステップ102で測定する。本発明の一部の実施形態においては、集団P内の各生物中の1000、2000、5000、10000個またはそれ以上の細胞構成成分の細胞構成成分レベルをステップ102で測定する。一部の実施形態においては、集団P内の各生物から$2x10^4$、$3x10^4$、$4x10^4$、$5x10^4$、$6x10^4$個以上の細胞構成成分レベルを処理ステップ102で測定する。

【００３６】

一実施形態においては、処理ステップ102において得られる細胞構成成分測定値は、細胞構成成分データ244としてメモリ224に保存される(図2)。特に、集団P内の各生物は、データ構造244中のエントリー246を受け取る。各エントリー246は、複数の細胞構成成分248を含む。各エントリー246中の各細胞構成成分248では、ステップ102(図1)で測定する対応する細胞構成成分248の量を保存する量エントリー250がある。

【００３７】

ステップ104(図1)では、集団Pを部分群の集合に分割するのにクラス予測変数を利用できるかどうかを判定する。本明細書では、クラス予測変数を、標本(例えば、集団P内の患者または標本)を患者サブクラスに割り当てることができる構成体(construct)として定義する。一部の実施形態においては、クラス予測変数は、集団Pが示す複合形質に関連する基本的な生物学的プロセスに従って、定義済みの部分群にすでに細分された既知の標本(例えば、患者)の集合から、クラス予測変数作成モジュール260(図2)によって作成される。この実施形態を説明するために、集団Pのサブセットである集団P'を考える。集団P'は、集団Pが示す複合形質に関連する基本的な生物学的差異に基づく部分群「A」および「B」を含む。すなわち、P'の各メンバーは、基本的生態の差異に基づいて、部分群「A」または「B」に分類される。この状況においては、部分群「A」および「B」を使用して、一方の群(「A」または「B」)には多量に存在するがもう一方の群には多量に存在しない細胞構成成分を特定することによって、クラス予測変数が形成される。例えば、部分群「A」において高度に発現すが部分群「B」においては高度に発現しない遺伝子を、各部分群の各メンバーについて得られた細胞構成成分測定値から特定することができる。同様に、部分群「B」において高度に発現するが、部分群「A」においては高度に発現しない遺伝子を特定することができる。これらの示差的発現パターンを使用して1組の検定細胞構成成分を構築する。集団P'中に含まれない患者(標本)の細胞構成成分測定値を、患者(標本)を部分群「A」または「B」に分類するために、検定細胞構成成分のセットと比較することができる。このようにして、P'の解析から得られる検定細胞構成成分セットは、クラス予測変数として働く。この手法の説明は、Golub等 1999、Science 286:531を参照されたい。本発明の一部の実施形態においては、クラス予測変数作成モジュール260によって作成されるものなどのクラス予測変数は、クラス予測変数データ構造262として保存される(図2)。

【００３８】

一部の実施形態においては、クラス予測変数262は、集団Pが示す複合形質の様々な細区画(subdivision)において示差的に発現される情報価値のある遺伝子の集合から導出される。この状況においては、情報価値のある遺伝子は、目的とする複合形質の部分群の1つにおいて高度または低度に発現される遺伝子である。例えば、集団Pがトウモロコシであり、検討中の複合形質が平均収率である場合を考える。2個の亜集団をPから抽出することができる。一方の亜集団は平均収率が極めて高く、もう一方の亜集団は平均収率が極めて低い。2つの部分群の遺伝子発現データを解析することによって、2個の部分群において示差的に発現される遺伝子クラスを明らかにすることができる。次いで、集団P全体を、このクラスの遺伝子についてクラスター化することができる。このようなクラスタリングによって、このクラスの遺伝子を同様に発現する植物をクラスター(すなわち、部分群)に群分けする。次に、このような部分群を量的遺伝解析にかけることができる。部分群の性質がより均質であるために、部分群の量的遺伝解析によって、遺伝子を目的形質に結び付ける情報を、集団P全体の量的遺伝解析よりも多量に得ることができる。さらに別の実施形態においては、クラス予測変数は、複雑性疾患などの複合形質を示す集団を細分するため

に使用することができる任意の形の生物学的データから導出される。一部の実施形態においては、セクション5.16に記載した技術を用いてクラス予測変数を特定する。

【００３９】

クラス予測変数が利用可能な場合(図1、104-イエス)、以下に詳細に開示するように、教師付き分類方法106(図1)を使用して集団Pを細分する。クラス予測変数が利用不可能な場合(104-ノー)、以下に詳細に開示するように、教師なし分類方法108(図1)を使用して集団Pを細分する。本発明の一部の実施形態においては、教師なし分類方法108は、クラス予測変数が利用可能な場合でも使用されることを理解されたい。教師なし分類方法108を使用して、例えば、クラス予測変数262を検査することができる。本発明の一部の実施形態においては、システム200(図200)は、集団Pを1組のn個の部分群に分類するために、1個もしくは複数の教師付き分類モジュール106および/または1個もしくは複数の教師なし分類モジュール108を備える。ここで、nは、1、2、3、4、5、6、7、8またはそれよりも大きい整数である。

【００４０】

1つの教師付き分類方法106が、Golub等、1999、Science 286:531に記載されている。これらの研究者は、1個の複合形質サブクラスにおいて一様に高く別のサブクラスにおいて一様に低い遺伝子に対応する理想的な発現パターンcを定義した。次に、一連の標本における複数の遺伝子の発現パターンを調べて、偶然によって予想される以上に発現パターンcと相関がある遺伝子を特定した。具体的には、Golub等は、38個の急性白血病標本中の6817個の遺伝子の発現データを用いて、偶然によって予想される以上に特定の白血病タイプの特徴とより高い相関があるほぼ1100個の遺伝子を見出した。この相関は、分類が発現データに基づくことができることを示している。Golub等は、1100個の遺伝子のうち50個を使用して、所与の患者が急性骨髄性白血病(AML)と急性リンパ性白血病(ALL)のどちらであるかを識別することができるクラス予測変数を構築した。50個の遺伝子のうち25個はALL患者においてより高度に発現されるのに対して、残りの25遺伝子はALL患者においてより高度に発現される。Golub等は、50個の遺伝子セットが新しい標本をAMLまたはALLと認定する信頼できる予測変数として役立つことを示した。当業者は、Golub等の教師付き分類方法が、遺伝子発現データに限らず、実際に、ステップ102(図1)で得られるあらゆる形の細胞構成成分データに適用可能であることを理解されたい。

【００４１】

別の教師付き分類方法106は線形判別分析である。線形判別分析は、Ripley、1996、Pattern Recognition and Neural Networks、Cambridge University Press、New York、およびHastie等、1995、Penalized Discriminant Analysis、The Annals of Statistics 23:73～102に概説されている。この手法においては、その発現プロファイルが、あらかじめ指定された2個のカテゴリーのうちの1個に属する遺伝子のプロファイルと一致している程度に応じて、遺伝子にスコアが与えられる。正のスコアは1つのカテゴリーの遺伝子により類似している遺伝子に与えられ、負のスコアは別のカテゴリーの遺伝子により類似している遺伝子に与えられる。類似度を決定する際に、ある測定値は他の測定値よりも重要である。より重要な測定値にはより大きな重みが付けられる。この手法が、発現プロファイル中の関連のない多数の測定値を含むデータ・セットに使用される。特に、線形判別分析手法は、型別には役に立たない広範に発現されるある種の遺伝子の測定値を含む白血病データ・セットにうまく適用することができる。当業者は、線形判別分析分類方法が、ステップ102(図1)において得られるあらゆる形の細胞構成成分データに適用可能であることを理解されたい。さらなる教師付き分類方法を上記セクション5.15に記載する。

【００４２】

クラス予測変数262が集団Pを細分するのに利用不可能なときには、教師なし分類方法108を使用してPを細分するタスクを実行する。このような1つの教師なし分類方法108は、クラスター分析である。一般に、クラスター分析は、仮説を立てるために、細胞構成成分データの基本的構造を探索するのに使用することができる。クラスター分析は、観測データの内部構造編成(internal structural organization)を探索するものなので、各クラスタ

ーが複合形質の現実の細区画または有用な細区画である単一の「最適な」クラスター・セットではないこともある。クラスタリングの一形態は、すべての対象を相互に包括的に比較して、クラスターの系統タイプの階層木(Eisen等、1997、Proc Natl Acad Sci USA 998:14863～14868;Iyer等、1999、Science 283:83～87)、またはユークリッド距離、相関係数もしくは相互情報量を含めて様々な類似度もしくは距離計測量(distance metrics)を用いた関連性ネットワーク(relevance network)などのクラスターの別のグラフ表示を構築するアルゴリズムを使用する。例えば、Jain & Dubes、1988、Algorithms for Clustering Data、「Partitional Clustering」、Prentice Hall、New Jersey、89～133ページを参照されたい。処理ステップ108に関して、クラスター化される対象は、集団P内の各生物から得られる1個または複数の細胞構成成分測定値である。階層木クラスタリングは、対象またはセットを連結するしきい値を連続的に下げることによって、類似した対象(集団P内の個体から得られる細胞構成成分発現測定値のセット)を一緒に連結して、積み上げ方式で(すなわち、木の葉から根まで)連続的により大きなクラスターにする。関連性ネットワークは、逆の戦略をとる。これは、各対象を表す頂点(vertice)と関連性の指標となる辺を有する完全に連結されたグラフから出発し、次いで、リンクが徐々に削除されて、あるしきい値で「自然に出現する(naturally emerging)」クラスターが現れる。クラスタリング法を以下のセクション5.8.1に詳細に記載する。

【００４３】

教師なし分類方法108の別の形式としては、K平均分析(セクション5.8.2参照)、クラスター内散乱を最小にしまたはクラスター間散乱を最大にする最短距離クラスタリング、Jarvis-Patrickクラスタリング(セクション5.8.3参照)などの分割型クラスタリング・アルゴリズムなどが挙げられる。Jain & Dubes、1988、Algorithms for Clustering Data、「Partitional Clustering」、Prentice Hall、New Jersey、89～133ページも参照されたい。別の教師なし分類方法108としては、自己組織化地図などの人工ニューラル・ネットワーク学習アルゴリズムなどがある。例えば、Kohonen、1982、Biological Cybernetics 43:59～69を参照されたい。自己組織化地図(SOM)は、データ・セット中の少数のクラス(単純な形質)を特定するタスクに好適である。例えば、Tamayo等、1999、Proc Natl Acad Sci USA 96:2907を参照されたい。SOMを以下のセクション5.8.5に詳細に記載する。SOM手法においては、ユーザーは、特定するクラスターの数を指定する。SOMは、その付近にデータ・ポイントが集合していると考えられる「重心」の最適なセットを見つけるものである。次いで、SOMによってデータ・セットを分割する。クラスターを定義する各重心は最も近いデータ・ポイントからなる。Golub等は、2クラスターSOMを適用して、初期白血病標本38点を、各標本中の6817個の遺伝子の発現パターンに基づいて2個のクラス(A1およびA2)に分けた。38人の患者集団は、急性骨髄性白血病(AML)および急性リンパ性白血病(ALL)患者の異質な集団から構成されていた。すなわち、38人の患者集団内の各患者は、AMLまたはALLであった。SOMは、標本集合全体にわたって変動が5倍未満の遺伝子を除外する変動フィルターを有するGENECLUSTERソフトウェア(Tamayo等、1999、Proc Natl Acad Sci USA 96:2907)を用いて構築された。SOMクラスA1およびA2は解析され、クラスAMLおよびALLに相当することが判明した。したがって、Golub等は、複合形質(白血病)を示す患者から得られた遺伝子発現データを、実際の白血病のサブクラス(AMLおよびALL)に対応するサブクラスに患者集団を細分するために、SOMを使用してクラスター化できることを示した。別の教師なし分類方法としては、Kohonen人工ニューラル・ネットワーク(Kohonen、1989、Self-Organization and Associative Memory、Springer-Verlag、Berlin)、自己連想ニューラル・ネットワーク(Kramer、1992、Comput Chem Eng 16:313～328)などが挙げられる。

【００４４】

教師付き分類方法106または教師なし分類方法108を使用して患者集団Pを1組のn個のクラスに細分した後、n個のサブクラスの妥当性を、ステップ110(図1)において、クラス予測変数検証モジュール270を用いて場合によっては検証することができる。検証は、複雑性疾患の現実の構成成分に対応する「正しい」答えが不明な場合に特に重要である。検証

110は、多種多様な方法で実施することができる。一手法においては、方法106または108によって特定された推定サブクラスnを使用して、上記方法などの方法によってクラス予測変数262が作成される。複合形質に関連する基本的な生物学的プロセスの既知の差異に対応する正しい分類が知られている場合には、n個のサブクラスの各々に細分される標本を使用して、このクラス予測変数262を作成することができる。次いで、クラス予測変数を使用して、新しい標本を特定されたn個のサブクラスに分類することができる。各標本の真の分類はこの場合既知なので、新たに作成されたクラス予測変数が、遺伝子をn個のサブクラスに分類することができるか検証することができる。

【００４５】

真のサブクラスnが不明な(すなわち、教師なし技術を使用して集団Pをn個の新規なサブセットに分類する)場合には、新規なサブクラスnのクラス予測変数は、新しい標本について正確度を独立に評価することができない。なぜならば、個々の標本を分類する「正しい」方法が不明だからである。この場合、検証ステップ110を使用して、新しい標本に高い予測強度(prediction strength)が割り当てられるかどうか評価することができる。例えば、Golub等 1999、Science 286:531を参照されたい。高い予測強度は、初期のデータ・セットにおいて見られる構造が、個々のデータ・セットにおいても見られることを示唆している。方法106または108によって誘導されるn個のサブセットから導出されるクラス予測変数の予測強度を、無秩序に作成されたクラス予測変数と比較することができる。方法106または108によって見出されるn個のサブセットから導出されるクラス予測変数262(誘導クラス予測変数)は、誘導クラス予測変数の予測強度が、無秩序に作成されたクラス予測変数の予測強度よりもかなり大きい場合には意味があると考えられる。

【００４６】

処理ステップ112～120は、n個のサブクラスのセット中の各サブクラスiが量的遺伝解析にかけられる反復プロセスを提供する。一部の実施形態においては、処理ステップ112～120は、量的遺伝解析モジュール272(図2)によって実行される。処理ステップ112においては、サブクラスnのセットからのサブクラスiが量的遺伝解析に選択される。ステップ114においては、ステップ112におけるn個のサブクラスから選択されたサブクラスiが量的遺伝解析にかけられる。一部の実施形態においては、サブクラスiの量的遺伝解析114は、サブクラスi中の標本または患者の表現型形質の連鎖解析を含む。連鎖解析を以下のセクション5.2に記載する。別の実施形態においては、サブクラスiの量的遺伝解析114は、サブクラスi内の各患者または標本中の複数の遺伝子の転写レベルが表現型形質として扱われる新規な形式のQTL解析を含む。この新規な形式の量的遺伝解析を以下のセクション5.3に詳細に記載する。本発明のさらに別の実施形態においては、サブクラスiの量的遺伝解析は、以下のセクション5.4に記載する関連解析を含む。どの形式の量的遺伝解析を使用するにしても、ステップ114(図2)の目標は、各部分群nにおいて創薬標的を特定することである。

【００４７】

処理ステップ116(図1)においては、量的遺伝解析にかけられていない残余のサブクラスがあるかどうか質問される。部分群が残っている(116-イエス)場合、処理制御は、処理ステップ112に戻り、別のサブクラスiがサブクラスnのセットから選択される。しかし、サブクラスnのセット中の各サブクラスiが量的遺伝解析にかけられていた場合にはプロセスが終了する120。

【００４８】

本発明による一実施形態においては、量的遺伝解析モジュールによって特定される群を構成する各QTLが別のQTLとともに群内に属する程度は、多変量統計モデルをその群に適合させることによって検定される。本明細書で使用する量的形質遺伝子座(QTL)は、量的形質に影響を及ぼすゲノム領域である。多変量統計モデルは、複数の量的形質を同時に考慮し、QTL間のエピスタシス相互作用(epistatic interaction)をモデル化し、候補経路群中の遺伝子が同じ生物学的経路または関係する生物学的経路に属するかどうかを明らかにする別の変形形態を検定することが可能である。検討中の形質が同じQTLによって実際に制

御されるかどうか(多面発現効果)、またはそれらが独立しているかどうかを判定する具体的な検定を実施することができる。本発明によって使用することができる例示的な多変量統計モデルを以下のセクション5.9に示す。本発明の一部の実施形態においては、多変量QTL解析モジュール274(図2)を用いてこのような多変量統計モデルを実施する。

【００４９】

以上、本発明の一実施形態による処理ステップを開示したが、当業者は、本発明の技術が提供するいくつかの利点を理解したはずである。集団P全体ではなくサブクラスnに対して量的遺伝解析を実施することによって、不完全浸透率、表現型模写、遺伝的異質性、ポリジーン遺伝、高頻度の疾患原因対立遺伝子などの複合形質の量的遺伝解析に関する諸問題が最小限に抑えられる。また、集団P内の検討中の複合形質が癌などの複雑性疾患または1タイプの癌である場合、本発明の方法を使用して、集団を、各亜集団が特定の形態の癌を含む一連の亜集団nに分解することができる。例えば、白血病の場合を考える。AMLとALLの区別は十分に証明されているが、AMLとALLを診断するのに十分な単一の検定は現在ない。それどころか、臨床業務は、腫瘍形態学、組織化学、免疫型別、および細胞遺伝学的解析の経験を積んだ血液病理学者の解釈を必要とし、それぞれが高度に専門化した検査室で別個に実施される。それにもかかわらず、白血病分類は、不完全なままであり、誤りが多い。本発明の方法を使用して、細胞構成成分測定データを用いて、白血病集団をAML亜集団およびALL亜集団に分類することができる。実際に、本発明の方法を使用して、処理ステップ106/108のnを3に設定することによって、白血病集団を、AML、B系統ALL(B-lineage ALL)およびT系統ALLに分離することができる。腫瘍生物学の様々な態様を区別する多数の分類に集団Pを分類するために、白血病のさらなる細分を実施することができる。次いで、引き続き量的遺伝解析を使用して、基本的腫瘍生物学の様々な態様に関連する遺伝子を特定することができる。

【００５０】

より一般的な例では、本発明の方法を使用して、複合形質に関連する基本的生物学の特徴的な態様に基づいて集団Pを細分することができる。次いで、各細分された集団を、量的遺伝解析を用いて個々に検討することが有利である。集団を細分することによって、量的遺伝解析にかける集団が効率的に選別される。一般的な集団Pではなく、または一般的な集団Pに加えて、選別された亜集団を使用することによって、亜集団内に存在する形質に関連する遺伝子を特定する後続の量的遺伝解析の能力が改善される。後続の量的遺伝解析に対して必要な均一性レベルを得るために、所与の集団Pを2回以上細分することができることも理解されたい。

【００５１】

一部の実施形態においては、集団Pは、家系データが利用可能な分離集団である。このような実施形態においては、連鎖解析を使用して、集団Pを細分するのに使用することができる情報価値のある1組の遺伝子を作成することができる。このような実施形態は、目的形質の両極端の表現型を示す集団P内の部分群を識別することができる1組の遺伝子を特定することから出発する。次いで、予備的な連鎖解析を用いて、目的形質に対するQTLを特定する。予備的連鎖解析は、集団P全体に利用可能な家系情報、および集団全体に利用可能な表現型情報を使用して、量的形質に関連する1個または複数のQTLを特定する。連鎖解析を以下のセクション5.2に詳細に記載する。予備的連鎖解析に加えて、一連のQTL解析を実施する。一連の解析における各QTL解析においては、特徴的な遺伝子のセットから選択された遺伝子の発現レベルは、量的表現型として働く。この形式のQTL解析を以下のセクション5.3に詳細に記載する。目的形質に対して特定されたQTLと結び付けられる、または重複するQTLを生成するこれらの遺伝子は、特徴的な遺伝子のセット中に保持される。集団Pは、特徴的遺伝子セット中に残っている遺伝子の発現レベルに基づいてクラスター化される。以下のセクション5.8に記載するクラスタリング方法を含めて、任意の形式のクラスタリングを、このステップにおいて使用することができる。クラスタリングは、1組の部分群(クラスター)を生成する。次いで、セクション5.2に記載するように、目的形質を量的形質として使用して、各部分群に対して連鎖解析を実施する。集団を細分するこ

とによって、部分群の一部は、目的形質の遺伝的特徴がかなり強調されるはずである。目的形質のシグナルが増大したこのような部分群を、セクション5.3に従って一連のQTL解析にかける。この一連の解析における各QTL解析においては、特徴的遺伝子セットから選択された遺伝子が量的表現型として働く。一連のQTL解析から得られる連鎖データによって、どのQTLがセット特徴的遺伝子に関連するかが明らかになる。このようなQTLの多変量解析によって、目的形質に関連するQTLが特定される。この実施形態をセクション5.17に示す。家系情報が利用不可能な実施形態においては、セクション5.4に記載するように、予備的な関連解析を連鎖解析の変わりに使用して、集団Pを部分群にクラスター化するのに使用することができる1組の遺伝子を特定するのに役立てることができる。

【００５２】

5.1.1 クラスタリングを用いた細分

以下の方法においては、ある種を検定する。この種は、例えば、植物、動物、ヒトまたは細菌とすることができる。一部の実施形態においては、この種はヒト、ネコ、イヌ、マウス、ラット、サル、ブタ、ショウジョウバエまたはトウモロコシである。一部の実施形態においては、この種を代表する複数の生物を検定する。この種の中の生物数は、任意の数値とすることができる。一部の実施形態においては、検定される複数の生物は、5～100、50～200、100～500、または500個を超える生物である。一部の実施形態においては、複数の生物は、F_2雑種、(t-1世代に対してF_1を無作為に交配させて形成された)F_t集団、$F_{2:3}$設計(F_2個体の遺伝形質を決定し、次いで自殖(self)させる)、またはデザインIII(2つの近交系からのF_2を両方の親系統に戻し交雑する)である。したがって、本発明の一部の実施形態においては、生物246(図2)は、F_2集団、F_t集団、$F_{2:3}$集団、デザインIII集団などの集団である。

【００５３】

一部の実施形態においては、調べる生物の一部を撹乱させる。撹乱は、環境または遺伝的なものとすることができる。環境撹乱の例としては、検定化合物、アレルゲン、とう痛、および高温または低温への生物の暴露があるが、これらだけに限定されない。環境撹乱の別の例は、食餌(例えば、高脂肪食または低脂肪食)、睡眠遮断、隔離、および数量化自然環境影響(quantifying natural environmental influence)(例えば、喫煙、食餌、運動)である。遺伝的撹乱の例としては、遺伝子ノックアウトの使用、所定の遺伝子または遺伝子産物の阻害剤の導入、N-エチル-N-ニトロソ尿素(ENU)突然変異誘発、遺伝子のsiRNAノックダウン、またはある種の複数の生物が示す形質の定量化が挙げられるが、これらだけに限定されない。(RNA干渉または転写後遺伝子サイレンシングとも称される)様々なsiRNAノックアウト技術が、例えば、Xia等、2002、Nature Biotechnology 20、1006ページ;Hannon、2002、Nature 418ページ、244;Carthew、2001、Current Opinion in Cell Biology 13、244ページ;Paddison、2002、Genes & Development 16、948ページ;Paddison & Hannon、2002、Cancer Cell 2、17ページ;Jang等、2002、Proceedings National Academy of Science 99、1984ページ;Martinez等、2002、Proceedings National Academy of Science 99、14849ページに開示されている。

【００５４】

ステップ204。ステップ1504(図15)においては、遺伝子発現/細胞構成成分データ244を得るために、生物から選択された組織中の細胞構成成分レベルを複数の生物246から測定する。一部の実施形態においては、わずか1種の組織タイプから細胞構成成分データを収集する。別の実施形態においては、複数の組織タイプから細胞構成成分データを収集する。

【００５５】

一般に、複数の生物246は、ある目的形質に関して遺伝分散を示す。一部の実施形態においては、形質は数量化可能である。例えば、形質が疾患である場合には、形質をバイナリ形式で数量化可能である(例えば、生物が発症した場合を「1」、生物が発症しなかった場合を「0」とする)。一部の実施形態においては、形質を様々な数値として定量することができ、複数の生物46は、そのような様々ないくつかの値をとる。一部の実施形態におい

ては、複数の生物246は、未処置集団(例えば、未暴露集団、野生型集団など)および処置済み集団(例えば、暴露集団、遺伝的改変集団など)を含む。一部の実施形態においては、例えば、未処置集団は撹乱にかけられないが、処置済み集団は撹乱にかけられる。一部の実施形態においては、ステップ1504で測定する組織は血液、白色脂肪組織、または生物246から容易に得られるある別の組織である。

【００５６】

異なる実施形態においては、5個の細胞構成成分から100個の細胞構成成分、50個の細胞構成成分から100個の細胞構成成分、300～1000個の細胞構成成分、800～5000個の細胞構成成分、4000～15,000個の細胞構成成分、10,000～40,000個の細胞構成成分、40,000個を超える細胞構成成分のレベルを測定する。

【００５７】

一実施形態においては、遺伝子発現/細胞構成成分データ244は、検定集団内の各個体(生物)246に対するマイクロアレイの処理像を含む。一部の実施形態においては、このようなデータは、各個体246に対して、マイクロアレイ上で示される各遺伝子/細胞構成成分248の量(強度)情報250、任意選択のバックグラウンド・シグナル情報、および遺伝子プローブを表す付随注釈情報を含む。一部の実施形態においては、細胞構成成分データ244は、実際に、調べる生物246の特定組織における様々なタンパク質のタンパク質発現レベルである。

【００５８】

本発明の一態様においては、細胞構成成分レベルは、生物の所定の組織内の細胞構成成分量を測定することによってステップ1504で決定される。本明細書で使用する「細胞構成成分」という用語は、検討中の形質に影響を及ぼし得る個々の遺伝子、タンパク質、mRNA、代謝産物、および/または別の任意の細胞構成成分を含む。遺伝子以外の細胞構成成分のレベルを、多種多様な方法によって測定することができる。例えば、細胞構成成分レベルを、生物体、それらの活動、それらの改変状態(例えば、リン酸化)の量または濃度、あるいは検討中の形質に関連する別の測定値とすることができる。

【００５９】

一実施形態においては、ステップ1504は、生物246の1個または複数の組織内の細胞構成成分248の転写状態を測定するステップを含む。転写状態としては、構成成分RNA種、特にmRNAの本性および存在量などが挙げられる。この場合、細胞構成成分は、RNA、cRNA、cDNAなどである。細胞構成成分の転写状態は、核酸もしくは核酸模倣プローブのアレイへのハイブリッド形成技術、または別の遺伝子発現技術によって測定することができる。転写物アレイを以下のセクション5.11に記載する。

【００６０】

別の実施形態においては、ステップ1504は、組織内の細胞構成成分248の翻訳状態を測定するステップを含む。この場合、細胞構成成分はタンパク質である。翻訳状態としては、組織内のタンパク質の本性および存在量などがある。一実施形態においては、タンパク質のゲノム全体のモニタリング(例えば、「proteome」、Goffeau等、1996、Science 274、546ページ)を、マイクロアレイを構築することによって実施することができる。ここで、結合部位は、複数のタンパク質種に特異的な固定化抗体、好ましくは固定化モノクローナル抗体を含む。好ましくは、抗体は、コードされたタンパク質のかなりの部分(例えば、30%、40%、50%、60%またはそれ以上)に対して存在する。モノクローナル抗体を作製する方法は周知である。例えば、HarlowおよびLane、1998、Antibodies:A Laboratwy Manual、Cold Spring Harbor、N.Y.を参照されたい。一実施形態においては、モノクローナル抗体は、ゲノム配列に基づいて設計された合成ペプチド断片に対して産生される。このような抗体アレイを用いて、生物から得られたタンパク質をアレイに接触させ、その結合性を、当分野で既知のアッセイによって分析する。一部の実施形態においては、抗体-抗原相互作用のハイスループット・スクリーニング用の抗体アレイを使用する。例えば、Wildt等、Nature Biotechnology 18、989ページを参照されたい。

【００６１】

あるいは、大規模なタンパク質発現定量分析は、放射性(例えば、Gygi等、1999、Mol. Cell. Biol 19、1720ページ)および/または安定なイオストープ(iostope)(^{15}N)代謝標識(例えば、Oda等 Proc. Natl. Acad. Sci. USA 96、6591ページ)と、それに続く二次元(2D)ゲル分離、ならびに分離タンパク質のシンチレーション計数または質量分析法による定量分析によって実施することができる。二次元ゲル電気泳動は、当分野で周知であり、一般に、一次元での集束後に、二次元でのSDS-PAGE電気泳動を行うものである。例えば、Hames等、1990、Gel Electrophoresis of Proteins:A Practical Approach、IRL Press、New York;Shevchenko等、1996、Proc Nat'l Acad. Sci. USA 93、1440ページ;Sagliocco等、1996、Yeast 12、1519ページ;Lander 1996、Science 274、536ページ;およびNaaby-Haansen等、2001、TRENDS in Pharmacological Science 22、376ページを参照されたい。電気泳動図は、質量分析法、ポリクローナル抗体およびモノクローナル抗体を用いたウエスタン・ブロット法および免疫ブロット分析、内部およびN末端の微量配列分析を含めて、多数の技術によって分析することができる。例えば、Gygi等、1999、Nature Biotechnology 17、994ページを参照されたい。一部の実施形態においては、蛍光二次元差ゲル電気泳動(DIGS)を使用する。例えば、Beaumont等、Life Science News 7、2001を参照されたい。一部の実施形態においては、生物246の組織中のタンパク質量を、同位体によってコードされた親和性タグ(ICAT)と、それに続くタンデム型質量分析によって求める。例えば、Gygi等、1999、Nature Biotech 17、994ページを参照されたい。このような技術を用いて、生物246の所定の組織内で発現されるタンパク質のかなりの部分を同定することができる。

【００６２】

別の実施形態においては、ステップ1504は、複数の生物246の所定の組織内の細胞構成成分活性または翻訳後修飾を測定するステップを含む。例えば、ZhuおよびSnyder、Curr. Opin. Chem. Biol 5、40ページ;Martzen等、1999、Science 286、1153ページ;Zhu等、2000、Nature Genet. 26、283ページ;およびCaveman、2000、J. Cell. Sci. 113、3543ページを参照されたい。一部の実施形態においては、細胞構成成分活性の測定は、タンパク質マイクロアレイなどの技術を用いて容易になされる。例えば、MacBeathおよびSchreiber、2000、Science 289、1760ページ;およびZhu等、2001、Science 293、2101ページを参照されたい。一部の実施形態においては、翻訳後修飾または細胞構成成分状態の別の態様を質量分析法によって測定する。例えば、AebersoldおよびGoodlett、2001、Chem Rev 101、269ページ;Petricoin III、2002、The Lancet 359、572ページを参照されたい。

【００６３】

一部の実施形態においては、生物246から得られる組織のプロテオームをステップ1504において解析する。生物細胞のプロテオーム分析(例えば、すべてのタンパク質の定量化、およびそれらの翻訳後修飾の測定)は、一般に、マイクロアレイ技術などのハイスループット・タンパク質分析方法を使用するものである。例えば、Templin等、2002、TRENDS in Biotechnology 20、160ページ;AlbalaおよびHumphrey-Smith、1999、Curr. Opin. Mol. Ther. 1、680ページ;Cahill、2000、Proteomics:A Trends Guide、47～51ページ;EmiliおよびCagney、2000、Nat. Biotechnol.、18、393ページ;およびMitchell、Nature Biotechnology 20、225ページを参照されたい。

【００６４】

さらに別の実施形態においては、細胞構成成分量の「混合」態様をステップ1504で測定する。一例においては、生物246から得られる組織の1組の細胞構成成分の量または濃度を、そのような組織の別のある細胞構成成分の活性測定値とステップ1504において組み合わせる。

【００６５】

一部の実施形態においては、ステップ1504において、所与の生物中の異なる対立形質の細胞構成成分を検出し測定する。例えば、二倍体生物においては、「父」に由来するコピーと「母」に由来するコピーの、任意の所与の遺伝子の2個のコピーがある。ある場合には、所与の遺伝子の各コピーを異なるレベルで発現することができる。これは、このタイプの対立遺伝子の示差的発現が、検討中の形質に関連し、特に検討中の形質が複雑である

場合に関連し得るので極めて興味深い。

【００６６】

ステップ1506。遺伝子発現/細胞構成成分データ244が得られた後、このデータを発現統計データに変換する(図15、ステップ1506)。一部の実施形態においては、細胞構成成分データ244(図1)は、複数の細胞構成成分に対する転写データ、翻訳データ、活性データおよび/または代謝産物量を含む。一実施形態においては、複数の細胞構成成分は、少なくとも5個の細胞構成成分を含む。別の実施形態においては、複数の細胞構成成分は、少なくとも100個の細胞構成成分、少なくとも1000個の細胞構成成分、少なくとも20,000個の細胞構成成分、30,000個以上の細胞構成成分を含む。

【００６７】

本発明の一実施形態の分析において量的形質として一般に使用される発現統計データとしては、転写データから導出される平均対数比、対数強度およびバックグラウンド補正強度があるが、これらだけに限定されない。別の実施形態においては、別のタイプの発現統計データを量的形質として使用する。

【００６８】

一実施形態においては、各調べる生物における複数の遺伝子の各発現レベルが正規化される。任意の正規化ルーチンを使用してこの正規化を行うことができる。代表的な正規化ルーチンとしては、強度のZ-スコア、強度中央値、強度中央値の対数、強度のZ-スコア標準偏差対数、対数強度較正DNA遺伝子セットのZ-スコア平均絶対偏差、ユーザー正規化遺伝子セット、強度中央値の比率補正および強度バックグラウンド補正があるが、これらだけに限定されない。また、正規化ルーチンの組み合わせを実行することもできる。本発明による例示的な正規化ルーチンを以下のセクション5.6に詳細に開示する。

【００６９】

ステップ1508。ステップ1508においては、検討中の形質、および/または細胞構成成分測定前に集団に適用してもよい撹乱と関連する細胞構成成分レベル(例えば、遺伝子発現レベル、タンパク質存在量レベルなど)のパターンが特定される。ステップ1508を実施することができる方法はいくつかあり、このような方法はすべて本発明の範囲に含まれる。このような方法の1つは、形質を識別する細胞構成成分248をまず特定するものである。

【００７０】

一例においては、撹乱は、ステップ1504において細胞構成成分測定前に集団に適用される。撹乱は、例えば、生物をある化合物に曝すことである。ある化合物への生物の暴露は、投与、注射などを含めてこれらだけに限定されない様々な手段によって実施することができる。この例においては、生物246の集団を2つのクラスに分割する。化合物に暴露された生物246と化合物に暴露されなかった生物246である。この例においては、生物246におけるそのレベル(例えば、転写状態、翻訳状態、活性状態、翻訳後修飾状態など)によって投与群(生物に暴露された群)を対照群から識別する細胞構成成分(例えば、遺伝子、タンパク質、代謝産物など)を、対応のあるt検定、独立t検定、ウィルコクソン検定、符号付順位数検定などの統計手法を用いて、または形質と遺伝子発現値の相関を計算して特定する。集団に適用してもよい撹乱は複数の処理を含む場合もある。そのような場合には、T検定、およびAnova、クラスカル・ワリスなどの順位検定(ranks test)に対する一般化をこのステップにおいて使用する。

【００７１】

別の実施形態においては、検定集団に撹乱を適用しない。ある場合には、検定集団を、形質を示す生物246と形質を示さない生物に分割する。生物246におけるそのレベル(例えば、転写状態、翻訳状態、活性状態、翻訳後修飾状態など)によって罹患群を健常群から識別する細胞構成成分(例えば、遺伝子、タンパク質、代謝産物など)を、統計手法を用いて特定する。

【００７２】

さらに別の実施形態においては、検定集団を、検討中の形質に対する表現型の機能に基づいて群に分割する。生物246におけるそのレベルによって様々な群を識別する細胞構成

成分を、統計手法を用いて特定する。ステップ1508で使用することができる統計手法の詳細は以下のセクション5.19を参照されたい。

【００７３】

別の例においては、検定集団は、形質に対して広範な表現型を示す。次いで、生物246におけるそのレベルによってこれらの表現型の少なくともいくつかを識別することができる細胞構成成分を、統計手法を用いて特定する。一般に、このステップにおいては、一般に表現型が異なる群に集団を分割し、これらの表現型が異なる群を識別する細胞構成成分を、(2つの群に対する)t検定、(2つよりも多い群に対する)ANOVAなどの統計検定によって特定する。

【００７４】

様々な実施形態においては、ステップ1508において特定された細胞構成成分248セットは、5～100個の細胞構成成分、50～500個の細胞構成成分、400～1000個の細胞構成成分、800～4000個の細胞構成成分、3000～8000個の細胞構成成分、8000～15000個の細胞構成成分、さらに多い15000個の細胞構成成分、または30000個未満の細胞構成成分を含む。

【００７５】

一部の実施形態においては、集団内の両極端の表現型を特定する。例えば、ある場合には、目的形質は肥満である。このような例においては、極めて肥満した生物246と極めてやせた生物246を両極端の表現型としてこのステップで選択する。本発明の一実施形態においては、極端な表現型は、集団によって示される所与の表現型に関して集団の上位または下位40、30、20または10パーセンタイルとして定義される。一部の実施形態においては、(極端な表現型の生物において測定される)所与の細胞構成成分246の細胞構成成分レベル250をt検定または多変量検定などの別の検定にかけて、検定集団に対して特定される表現型群(例えば、処置と未処置)を、所与の細胞構成成分246によって識別できるかどうか判定する。各表現型群において細胞構成成分が特徴的に異なるレベルで存在するときには、細胞構成成分246によって表現型群を識別できるはずである。例えば、2つの表現型群がある場合、(極端な表現型の生物において測定された)細胞構成成分レベル250が、第1の表現型群中に第1のレベルで存在し、第2の表現型群中に第2のレベルで存在し、第1のレベルと第2のレベルが明確に異なる場合には、細胞構成成分によって2つの群が識別されるはずである。

【００７６】

ステップ1510。(例えば、両極端の表現型を示す集団中の生物を用いて)形質、または場合によっては撹乱を識別する細胞構成成分248セットを特定した後、それらをクラスター化することができる。本発明の一実施形態においては、2つ以上のクラス(例えば、罹患と健常、撹乱されたものと撹乱されていないもの)間の形質(またはステップ1504における測定前に集団に適用される撹乱)を識別する細胞構成成分セット中の各細胞構成成分248を細胞構成成分ベクトルとして扱う。例えば、2つ以上のクラス間の撹乱(例えば、複合形質)を識別する、細胞構成成分セット中のn番目の細胞構成成分248を

【数１】

$$C_n = (A_1^n, A_2^n, \ldots, A_m^n)$$

【００７７】

として表す。

【００７８】

式中、各Aは、複数の調べる生物内の生物246の組織中の細胞構成成分nのレベル(例えば、転写状態、翻訳状態、活性など)であり、mは、考慮される生物の数である。細胞構成成分ベクトルC_nは、各細胞構成成分ベクトルにおける対応するレベルAの値の類似性に基づいてクラスター化することができる。細胞構成成分ベクトルC_nは、そのような細胞構成成分ベクトルの対応するレベルと相関がある場合には、同じ群(細胞構成成分ベクトル・クラスター)にクラスター化される。説明のために、5つの異なる生物246において3個の異なる

細胞構成成分を測定することによって得られる仮定上の細胞構成成分ベクトルC_nを考える。すなわち、各細胞構成成分ベクトルは5個の値を有する。5個の各値は、5つの生物246のうちの1つの組織における対応する細胞構成成分nのレベル(例えば、活性、転写状態、翻訳状態など)である。

【００７９】

例示的な細胞構成成分ベクトルC_1:0、5、5.5、0、0

例示的な細胞構成成分ベクトルC_2:0、4.9、5.4、0、0

例示的な細胞構成成分ベクトルC_3:6、0、3、3、5

したがって、ベクトルC_1では、第1の生物における0任意単位(arbitrary unit)、第2の生物における5任意単位、第3の生物における5.5任意単位、ならびに第4および第5の生物における0任意単位の細胞構成成分「C_1」レベルがある。例示的な細胞構成成分ベクトルC_1、C_2およびC_3のクラスタリングによって2個のクラスター(細胞構成成分ベクトル・クラスター)が得られる。第1のクラスターは細胞構成成分ベクトルC_1およびC_2を含むはずである。というのは、各ベクトル内のレベルと相関があるからである(生物246-1における0と0、生物246-2における5と4.9、生物246-3における5.5と5.4、生物246-4における0と0、および生物246-5における0と0)。第2のクラスターは、例示的な細胞構成成分ベクトルC_3を含むはずである。というのは、ベクトルC_3におけるレベルのパターンが、C_1およびC_2におけるレベルのパターンに類似していないからである。この説明は、仮定上の細胞構成成分レベル・データを用いたクラスタリングのある態様を記述するのに役立つ。しかし、本発明においては、このステップに使用される細胞構成成分は、両極端の形質を識別するので選択されている。したがって、上述した仮定上のデータとは異なり、細胞構成成分レベルは、それらが両極端の表現型にわたって選択されたことを反映すべきである。この場合には、このステップにおけるクラスタリングは、両極端の形質を識別する細胞構成成分群内の細胞構成成分の部分群を特定するのに役立つ。この形式の二次元クラスタリングの例を以下のセクション5.20.2に示す。

【００８０】

本発明の一実施形態においては、凝縮階層型クラスタリングをステップ1510で細胞構成成分ベクトルに適用する。このようなクラスタリングにおいては、細胞構成成分ベクトル対のピアソン相関係数を用いて類似度を決定する。別の実施形態においては、細胞構成成分ベクトルのクラスタリングは、階層型クラスタリング法の適用、k平均法の適用、ファジーk平均法の適用、Jarvis-Patrickクラスタリング法の適用、自己組織化地図の適用、またはニューラル・ネットワークの適用を含む。一部の実施形態においては、階層型クラスタリング法は、凝縮型クラスタリング手順である。別の実施形態においては、凝縮型クラスタリング手順は、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである。さらに別の実施形態においては、階層型クラスタリング法は分割型クラスタリング手順である。遺伝子解析ベクトルをクラスター化するのに使用することができる例示的なクラスタリング法を以下のセクション5.8に記載する。好ましい実施形態においては、ノンパラメトリックなクラスタリング・アルゴリズムを細胞構成成分ベクトルに適用する。一部の実施形態においては、スピアマンのR、ケンドールのτ、またはガンマ係数を使用して、細胞構成成分ベクトルをクラスター化する。

【００８１】

ステップ1512。ステップ1512においては、集団は、ステップ1510からのクラスタリング情報を用いてサブタイプに再分類される。ステップ1512の目標は、これらのサブタイプを区別することができる細胞構成成分を含む分類子(classifier)を構築することである。一実施形態においては、集団中の各生物に対してそれぞれの表現型ベクトルを構築する。各表現型ベクトルは、ステップ1510において使用した細胞構成成分セットのすべてまたは一部に対する細胞構成成分レベルを含む。一部の実施形態においては、表現型ベクトルの成分の順序は、ステップ1510において得られたクラスタリング・パターンによって決定される。

【００８２】
表現型ベクトルは、セクション5.8に記載する技術のいずれかを用いてクラスター化される。各表現型ベクトルの成分の順序をステップ1510のクラスタリングに基づいて決定する実施形態においては、ステップ1512のクラスタリングによって二次元クラスターが生成する。一面では、細胞構成成分は、生物集団全体にわたるそれらの存在量の類似度に基づいてクラスター化される。例えば、2個の細胞構成成分が集団全体で類似したレベルで発現される場合には、それらは共にクラスター化される。別の面では、生物は、細胞構成成分セット全体の類似度に基づいてクラスター化される。例えば、2個の生物は、各生物中の対応する細胞構成成分が同等のレベルで発現する場合、共にクラスター化される。

【００８３】
本発明は、ステップ1510および1512に記載するクラスタリング法の代わりに使用することができる多数の他のパターン分類技術を提供する。これらの他のパターン分類技術を使用して、特徴的な細胞構成成分から分類子を構築することができる。次いで、このような分類子を使用して、一般的な集団を異なる部分群に分けることができる。このような別の技術をセクション5.18に記載する。

【００８４】
本質的に、ステップ1510および1512におけるクラスタリングは、集団を新しい部分群(例えば、表現型クラスター)に配列させる。各部分群(表現型クラスター)は、特有の細胞構成成分発現(またはレベル)パターンを特徴とする。説明のために、ステップ1510において実施されるクラスタリングによって3つの細胞構成成分群、すなわち群A、BおよびCが生成する場合を考える。次に、ステップ1512において、検定集団中の各生物に対して表現型ベクトルが構築される。表現型ベクトルの成分は、ステップ1510の細胞構成成分クラスタリング結果によって指定された順序で配置された、それぞれの生物に対して測定された細胞構成成分レベルである。説明のために、10個の細胞構成成分があると仮定する(1、2、3、4、5、6、7、8、9および10)。ここで、構成成分8～10は群A、構成成分4～7は群B、構成成分1～3は群Cに分類される。この場合、集団中の生物Mの表現型ベクトルV_Mは、
V_M=8、9、10、4、5、6、7、1、2、3
の形をとることができる。ここで、ベクトル中の各それぞれの細胞構成成分は、ベクトルによって示される生物中の細胞構成成分のレベルによって表される。各ベクトルV_Mは、これらのレベルに基づいてクラスター化される。そのような4つの生物に対する仮定上のベクトルを考える。ここで、細胞構成成分レベルは、単に、高レベルを「+」、低レベルを「-」として表す。

【００８５】
V_1=+、-、+、+、+、-、-、-、-、-
V_2=-、-、-、-、-、+、+、+、+、+
V_3=+、+、+、+、+、-、-、-、-、-
V_4=-、-、-、-、-、+、+、+、-、+
V_1からV_4までのクラスタリングによって、2つの群(IおよびII)が得られる。

【００８６】
群I: V_1=+、-、+、+、+、-、-、-、-、-
V_3=+、+、+、+、+、-、-、-、-、-
群II: V_2=-、-、-、-、-、+、+、+、+、+
V_4=-、-、-、-、-、+、+、+、-、+
群Iの各生物が類似した細胞構成成分発現(またはレベル)パターンを有することは明らかである。また、この類似パターンによって、群Iは群IIから区別される。同様に、群IIの各生物は、類似した細胞構成成分(またはレベル)パターンを有し、このパターンによって群IIは群Iから区別される。この例においては、ステップ1510から得られる順序付けられた細胞構成成分セットが、生物をサブタイプに再分類する分類子として働く。

【００８７】
一部の実施形態においては、ステップ1510のクラスタリングを実施せず、このような表

現型クラスターを特定するために表現型ベクトルのみをクラスター化する。しかし、上記例から、表現型クラスターを識別することができる細胞構成成分の特定は、ステップ1510のクラスタリングを実施した場合により容易に特定しうることが明らかである。というのは、ステップ1510のクラスタリングは、各表現型ベクトル内の特徴的な細胞構成成分を群化する傾向があるからである。

【００８８】

このステップにおいて得られる各サブタイプ(部分群)は、古典的な表現型観察を用いては得られないことに留意されたい。そうではなく、各サブタイプは、表現型を識別可能な群を識別する、順序付けられた細胞構成成分レベル・セットを用いて特定される。したがって、ステップ1512において特定された各サブタイプは、検討中の形質の異なる生化学的形態を示し得る。例えば、前のステップにおいて撹乱が適用された場合には、このステップにおいて特定された各サブタイプは、形質に関連する異なる生化学応答を示すことができる。

【００８９】

ステップ1512では、新たに特定された部分群(サブタイプ)を識別することができる細胞構成成分を決定する。例えば、以下のクラスターが得られた上記例を考える。

【００９０】

群I: V_1=+、-、+、+、+、-、-、-、-、-
V_3=+、+、+、+、+、-、-、-、-、-
群II: V_2=-、-、-、-、-、+、+、+、+、+
V_4=-、-、-、-、-、+、+、+、-、+
ここで、各ベクトルにおける成分の順序は
V_M=8、9、10、4、5、6、7、1、2、3
である。

【００９１】

細胞構成成分8、10、4、5、6、7、1および3は、群IとIIを識別するのに対して、細胞構成成分9および2は識別しないことがわかる。例えば、細胞構成成分9は、群Iにおける値が(-/+)であり、群IIにおける値が(-/-)である。細胞構成成分2は、群Iにおける値が(-/-)であり、群IIにおける値が(+/-)である。

【００９２】

ステップ1512において特定されたサブタイプ(部分群)を識別する細胞構成成分セットは、検定集団に対する分類子として働く。この分類子は、一般的な集団をサブタイプに分けることができる。前のステップでは、特徴的な細胞構成成分セット(分類子)を特定し順序付けるために選択生物(例えば、極端な表現型の生物)を使用したが、ステップ1512において特定された細胞構成成分によって、一般的な集団の生物すべてを部分群に分類することができる。

【００９３】

ステップ102。ステップ1512は、集団をサブタイプに分割するのに役立つ。ステップ1512の後、ステップ110～120を上述したように図1に示すように実施して、これらの集団サブタイプの各々を定量的な遺伝的方法を用いて解析する。

【００９４】

5.2. 連鎖解析

このセクションは、処理ステップ114(図1)に使用することができるいくつかの標準的な量的形質遺伝子座(QTL)連鎖解析アルゴリズムを説明する。このような連鎖解析はQTL解析とも呼ばれる。例えば、LynchおよびWalsch、1998、Genetics and Analysis of Quantitative Traits、Sinauer Associates、Sunderland、MAを参照されたい。連鎖解析の第1の目的は、複数の罹患生物を含むいくつかの家族の各々を通して、特定の遺伝モデルに一貫しており、かつ偶然だけでは起こりそうにないパターンで伝えられるゲノム片が存在するかどうかを明らかにすることである。換言すれば、これらのアルゴリズムの目的は、1個または複数の生物46によって示される表現型形質の遺伝子座(例えば、QTL)を特定すること

である。QTLは、検定種において表現型形質の変化割合の原因となる種のゲノム領域である。

【００９５】

組換え割合はθで表すことができ0～0.5である。2個の遺伝子座のθが0.5である場合には、2個の遺伝子座の対立遺伝子は独立して受け継がれ、2個の遺伝子座で配偶子の半分が組換えられ、半分が親のものである。この場合には、遺伝子座は連鎖していない。θ<0.5の場合には、対立遺伝子は独立に受け継がれず、2個の遺伝子座は連鎖している。極端なシナリオはθ=0のときであり、2個の遺伝子座は完全に連鎖しており、減数分裂中に2個の遺伝子座の組換えは起こらず、すなわちすべての配偶子が親のものである。連鎖解析は、既知の位置にあるマーカー遺伝子座が、検討中の表現型に影響を及ぼす未知の位置にある遺伝子座と連鎖しているかどうかを検定する。換言すれば、群内の生物の遺伝子型を、群が示す表現型と家系データを用いて比較することによってQTLを特定する。マーカー遺伝子型データによって作成される遺伝地図中の複数のマーカーの各マーカーにおける各生物の遺伝子型を、各生物の所与の表現型と比較する。遺伝地図は、遺伝マーカーを遺伝(線形)地図の順に配置することによって作成され、その結果、マーカー間の位置関係が理解される。マーカー地図によって与えられるマーカー間の関係を知ることから得られる情報によって、QTL効果とQTL位置の関係を扱う設定がなされる。

【００９６】

本発明の一部の実施形態においては、連鎖解析は、LynchおよびWalsch、1998、Genetics and Analyis of Quantitative Traits、Sinauer Associates, Inc.、Sunderland、MAに開示または参照されるQTL検出方法のいずれかに基づいている。

【００９７】

5.2.1. 使用する表現型データ

本発明は、QTL解析を実施するのに使用できる表現型データのタイプに制約がないことを理解されたい。表現型データは、例えば、生物集合中の定量可能な表現型形質の一連の測定値である。このような定量可能な表現型形質としては、例えば、尾の長さ、寿命、目の色、サイズおよび体重が挙げられる。あるいは、表現型データを、ある表現型形質の有無を追跡するバイナリ形式とすることができる。例えば、「1」は、特定の種の目的生物が所与の表現型形質を有することを示し、「0」は、特定の種の目的生物が所与の表現型形質を欠いていることを示すことができる。表現型形質は、検定集団中の各生物表現型に代表的な任意の形式の生物学的データとすることができる。一部の実施形態においては、表現型形質は数量化され、量的表現型と呼ばれることが多い。

【００９８】

5.2.2. 使用する遺伝子型データ

連鎖解析に必要な遺伝子型データを用意するために、遺伝マーカー地図中の各マーカーの遺伝子型を、検定集団内の各生物について決定する。遺伝子型情報は、集団内の生物のゲノムの多型から得られる。このような多型としては、一塩基多型、ミクロサテライト・マーカー、制限断片長多型、短鎖縦列反復、配列長多型、DNAメチル化パターンなどがあるが、これらだけに限定されない。

【００９９】

連鎖解析は、任意の所与の量的形質に対するQTLの位置の枠組みとして、マーカー遺伝子型データから得られる遺伝地図を使用する。一部の実施形態においては、順序付けられたマーカー対によって定義される区間を増分(例えば、2 cM)ごとに検索し、QTLが区間内の位置に存在し得るかどうかを統計方法によって検定する。一実施形態においては、連鎖解析は、単一のQTLについて遺伝地図中の順序付けられたマーカー全体にわたって各増分で統計的に検定する。検定結果はロッド・スコアとして表され、有望なQTLを位置付けるために帰無仮説(QTLなし)の下での尤度関数の評価結果を対立仮説(検定位置のQTL)と比較する。ロッド・スコアの詳細はセクション5.4、ならびにLanderおよびSchork、1994、Science 265、2037～2048ページにある。区間マッピングは、順序付けられた遺伝マーカーにわたって系統的に線形(一次元)で検索し、同じ帰無仮説を検定し、各増分で同じ形式の尤

度を使用する。

【０１００】

5.2.3. 使用する家系データ

連鎖解析には、マーカーの分離を統計的にモデル化するために、検定集団内の生物の家系データが必要である。様々な形式の連鎖解析を、家系データ(近交系と非近交系)を作成するために使用される集団のタイプによって分類することができる。

【０１０１】

いくつかの形式の連鎖解析は、近交系親系統から生じる集団の家系データを使用する。得られたF_1系統は、すべてのマーカーおよびQTLにおいて異型接合的である傾向にある。F_1集団から交雑種が作られる。例示的な交雑種としては、戻し交雑種、F_2雑種、(t-1世代に対してF_1を無作為に交配させて形成された)F_t集団、$F_{2:3}$設計(F_2個体の遺伝形質を決定し、次いで自殖させる)、デザインIII(2つの近交系からのF_2を両方の親系統に戻し交雑する)が挙げられる。したがって、本発明の一部の実施形態においては、生物はF_2集団などの集団であり、F_2集団の家系データは既知である。以下に詳細に考察するように、この家系データを使用してオッズ(ロッド)スコアの対数を計算する。

【０１０２】

ヒトを含めて多数の生物では、操作可能な近交系を利用することができず、連鎖解析を実施するために非近交系集団を使用しなければならない。非近交系集団を使用する連鎖解析は、集団内の変化の原因になるQTLを検出するのに対して、近交系集団を使用する連鎖解析は、系統間、さらには異なる種間の一定した差異の原因になるQTLを検出する。集団内の変化(非近交系集団)を使用すると、集団(近交系集団)間の一定した差異とは対照的に、QTL検出能力が低下する。近交系では、すべてのF_1親は(同じ連鎖相を含めて)同一の遺伝子型を有し、したがってすべての個体は情報価値があり、連鎖不平衡が最大になる。近交系同様、連鎖解析に必要な連鎖不平衡を有する標本を得る様々な設計が提案されている。一般に、血縁集合が頼りにされる。

【０１０３】

近交系交雑種を使用したQTL解析と非近交系集団との主要な差は、前者の親は遺伝的に均一であるが、後者の親は遺伝的に可変であることである。この違いは、いくつかの結果をもたらす。第1に、非近交系集団からの親の部分のみ情報価値がある。連鎖情報を提供する親は、マーカーと連鎖QTLの両方において異型接合的でなければならず、この状況においてのみマーカーと形質の関連性を子孫にもたらすことができる。非近交系集団の無秩序な親の部分のみが、そのような二重異型接合体(double heterozygote)である。近交系では、F_1はすべての遺伝子座において交雑種系統間で異なる異型接合的であり、したがってすべての親が十分情報価値がある。第2に、近交系交雑種設計中の遺伝子座において分離している対立遺伝子はわずか2個であるのに対して、非近交系集団は、任意の数の対立遺伝子を分離することができる。最後に、非近交系集団においては、各個体はマーカー-QTL連鎖相が異なることがあり、その結果、Mを有する配偶子が一方の親のQTL対立遺伝子Qに関連し、別の親のqと関連することがある。したがって、非近交系集団では、マーカーと形質の関連は、各親で個別に検定される可能性がある。近交系交雑種では、すべてのF_1親は(連鎖相を含めて)同一の遺伝子型を有するので、親に無関係にすべての子孫にわたってマーカーと形質の関連性を平均することができる。LynchおよびWalsh、Genetics and Analysis of Quantitative Traits、Sinauer Associates、Sunderland、Massachusettsを参照されたい。

【０１０４】

5.2.4. モデルのない連鎖解析とモデルに基づく連鎖解析

連鎖解析は、一般に、2つのクラス、すなわちモデルに基づく連鎖解析とモデルのない連鎖解析に分類される。モデルに基づく連鎖解析は、遺伝様式モデルを想定するのに対して、モデルのない連鎖解析は遺伝様式を想定しない。モデルのない連鎖解析は、対立遺伝子共有法(allele-sharing method)およびノンパラメトリック連鎖方法としても知られる。モデルに基づく連鎖解析は、「最大尤度」および「ロッド・スコア」方法としても知ら

れる。どちらの形式の連鎖解析でも本発明では使用することができる。

【０１０５】

モデルに基づく連鎖解析は、二分形質(dichotomous trait)に使用されることが最も多く、形質モデルを仮定する必要がある。これらの仮定には、疾患対立遺伝子頻度、浸透度関数などがある。疾患形質の場合、特に公衆衛生に関係する疾患形質の場合には、真の基本的モデルは複雑であり知られておらず、したがってこれらの手順を適用することはできない。もう一方の形式の連鎖解析(モデルのない連鎖解析)は、対立遺伝子共有を利用する。対立遺伝子共有法は、表現型が類似している血縁は、マーカーが目的遺伝子座と連鎖している場合およびその場合にのみ、マーカー遺伝子座における遺伝子型が類似しているはずであるという考えに依拠している。連鎖解析は、染色体の特異的領域に目的遺伝子座を限局化することができ、分解能範囲は一般に5 cMまたはほぼ5000 kb以上に限定される。モデルに基づく連鎖解析およびモデルのない連鎖解析についてのさらなる情報は、Olson等、1999、Statistics in Medicine 18、2961～2981ページ;LanderおよびSchork 1994、Science 265、2037ページ;およびElston、1998、Genetic Epidemiology 15、565ページ、ならびに以下のセクションを参照されたい。

【０１０６】

5.2.5. 既知の連鎖解析実施プログラム

本発明のこの態様にしたがって、多数の既知のプログラムを使用して連鎖解析を実施することができる。このようなプログラムの1つはMapMaker/QTLである。これは、MapMakerの同伴プログラムであり、最初のQTLマッピング・ソフトウェアである。MapMaker/QTLは、標準の区間マッピングを使用して、F_2または戻し交雑データを解析する。別のこのようなプログラムは、QTL Cartographerである。これは、単一マーカー回帰、区間マッピング(LanderおよびBotstein、同上)、複数の区間マッピングおよび複合区間マッピング(Zeng、1993、PNAS 90:10972～10976;およびZeng、1994、Genetics 136:1457～1468)を実施するものである。QTL Cartographerは、F_2または戻し交雑集団からの解析が可能である。QTL Cartographerは、http://statgen.ncsu.edu/gtlcart/Cartographer.html(North Carolina State University)から利用可能である。処理ステップ114によって使用することができる別のプログラムはQgeneである。これは、単一マーカー回帰または区間回帰によってQTLマッピングを実施するものである(MartinezおよびCumow 1994 Heredity 73:198～206)。Qgeneを用いて、(すべて同系交配に由来する)11個の異なる集団タイプを解析することができる。Qgeneは、http://www.qgene.org/から利用可能である。さらに別のプログラムはMapQTLである。これは、標準区間マッピング(LanderおよびBotstein、同上)、複数QTLマッピング(MQM)(Jansen、1993、Genetics 135:205～211;Jansen、1994、Genetics 138:871～881)、およびノンパラメトリック・マッピング(クラスカル・ワリス順位和検定)を実施するものである。MapQTLは、非近交系系統(他家受粉媒介者)を含めた様々な家系タイプを解析することができる。MapQTLは、Plant Research International、Plant Research International、P.O. Box 16、6700 AA. Wageningen、The Netherlands;http://www.plant.wageningen-ur.nl/default.asp?section=products)から利用可能である。処理ステップ210の一部の実施形態において使用することができるさらに別のプログラムは、QTLマッピング・プログラムのMap Manager QTである(ManlyおよびOlson、1999、Mamm Genome 10:327～334)。Map Manager QTは、単一マーカー回帰分析、回帰に基づく単純区間マッピング(simple interval mapping)(HaleyおよびKnott、1992、Heredity 69、315～324)、複合区間マッピング(Zeng 1993、PNAS 90:10972～10976)および並べ換え検定(permutation test)を行うものである。Map Manager QTの説明は、ManlyおよびOlson、1999、Overview of QTL mapping software and introduction to Map Manager QT、Mammalian Genome 10:327～334の参考文献にある。

【０１０７】

連鎖解析を実施することができるさらに別のプログラムはMultiCross QTLである。これは、近交系から生じる交雑種のQTLをマッピングするものである。MultiCross QTLは、線形回帰モデル手法を使用し、区間マッピング、全マーカー・マッピング(all-marker mapp

ing)、コファクターを含む複数QTLマッピングなどの異なる方法を扱う。このプログラムは、近交系種および非近交系種の多種多様な単純マッピング集団を扱うことができる。MultiCross QTLは、Unite de Biometrie et Intelligence Artificielle、INRA、31326 Castanet Tolosan、Franceから利用可能である。

【０１０８】

連鎖解析を実施するのに使用することができるさらに別のプログラムはQTL Cafeである。このプログラムは、F_2交雑種などの純系交雑種、戻し交雑種、リコンビナント近交系、および倍加半数体系統に由来するほとんどの集団を解析することができる。QTL Cafeは、Haley & Knottsのフランキング・マーカー回帰およびMarker回帰のJava実装が組み込まれ、複数のQTLを扱うことができる。このプログラムは、3つのタイプのQTL解析単一マーカーANOVA、マーカー回帰(KearseyおよびHyne、1994、Theor. Appl. Genet.、89:698～702)、および回帰による区間マッピング、(HaleyおよびKnott、1992、Heredity 69:315～324)が可能である。QTL Cafeは、http://web.bham.ac.uk/g.g.seaton/から利用可能である。

【０１０９】

連鎖解析を実施するのに使用することができるさらに別のプログラムはMAPLである。これは、区間マッピング(HayashiおよびUkai、1994、Theor. Appl. Genet. 87:1021～1027)または分散分析によってQTL解析を実施するものである。F_2、戻し交雑、F_2に由来するリコンビナント近交系、あるいは所与の自殖世代後の戻し交雑を含めて、様々な集団タイプを解析することができる。メトリック多次元尺度構成法による多数のマーカーの自動群化および順序付けが可能である。MAPLは、Institute of Statistical Genetics on Internet (ISGI)、Yasuo, UKAI、http://web.bham.ac.uk/g.g.seatol/から利用可能である。

【０１１０】

連鎖解析に使用することができる別のプログラムはR/qtlである。このプログラムは、実験交雑種においてQTLをマッピングする双方向環境を提供するものである。R/qtlは、欠測遺伝子型データを扱う隠れマルコフ・モデル(HMM)技術を使用する。R/qtlは、戻し交雑、雑種、および相既知の四系交雑(phase-known four-way cross)に対して、遺伝子型決定誤差を見込んだ多数のHMMアルゴリズムを実行する。R/gtlは、Haley-Knott回帰および多重代入法を用いた区間マッピングによる遺伝地図の推定機能、遺伝子型決定誤差特定機能、ならびに単一QTLゲノム走査および2個のQTLの二次元ゲノム走査の実施機能を含む。R/qtlは、Karl W. Broman、Johns Hopkins University、http://biosun01.biostat.jhsph.edu/~kbroman/gtl/から利用可能である。

【０１１１】

当業者は、量的遺伝解析を必要とする本発明の方法のステップにおいて使用することができるいくつかの別のプログラムやアルゴリズムが存在し、そのようなプログラムやアルゴリズムのすべてが本発明の範囲内にあることを理解されたい。

【０１１２】

5.2.6. モデルに基づくパラメトリック連鎖解析

モデルに基づく連鎖解析においては、(「ロッド・スコア」方法またはパラメトリック方法とも呼ばれる)、形質遺伝様式の詳細がモデル化されている。一般に、対立遺伝子頻度および浸透度関数の特定の値が指定される。

【０１１３】

5.2.6.1. 最大尤度/近交系集団による区間マッピング

本発明の一実施形態においては、連鎖解析は、LanderおよびBotstein、1989、「Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps」、Genetics 121:185～199によって初めて提案されたアルゴリズムから派生したアルゴリズムに従うQTL区間マッピングを含む。区間マッピングの原理は、マッピングされた2個のマーカー遺伝子座間の多数の位置におけるQTLの有無についてモデルを検定することである。モデルを適合させ、最尤法などの技術を用いてその適合度を検定する。最大尤度理論は、QTLが2個の両アレル・マーカー間にあるときに、遺伝子型(すなわち、倍加半数体子孫の場合はAABB、AAbb、aaBB、aabb)がそれぞれ量的形質遺伝子座(QTL)遺伝子型の混合物を含

むことを想定している。最大尤度は、各マーカー・クラスに対して観測される量的形質分布に最適近似を与えるQTLパラメータを検索するものである。モデルは、QTL効果を適合させ、またはQTL効果を適合させないで、観察された分布の尤度を計算することによって評価される。

【０１１４】

本発明の一部の実施形態においては、GeneHunterなどのプログラムにおいて実行されるLanderのアルゴリズムを使用して連鎖解析を実施する。例えば、Kruglyak等、1996、Parametric and Nonparametric Linkage Analysis:A Unified Multipoint Approach、American Journal of Human Genetics 58:1347～1363、KruglyakおよびLander、1998、Journal of Computational Biology 5:1～7;Kruglyak、1996、American Journal of Human Genetics 58、1347～1363を参照されたい。このような実施形態においては、マーカーを無制限に使用することができるが、家系サイズは、コンピュータ上の制限のために制約される。別の実施形態においては、MENDELソフトウエア・パッケージを使用する。(http://bimas.dcrt.nih.gov/linkage/ltools.html参照)。このような実施形態においては、家系サイズは無制限とすることができるが、使用することができるマーカー数はコンピュータ上の制限のために制約される。このセクションに記載する技術は、一般に、近交系集団を必要とする。

【０１１５】

5.2.6.2. 線形回帰/近交系集団を用いた区間マッピング

本発明の一部の実施形態においては、区間マッピングは、回帰法に基づき、最尤法によって得られるものに類似したQTL位置および効果の推定値が得られる。QTL遺伝子型は、回帰法に基づくマッピングにおいては未知であるので、遺伝子型は、最近接フランキング・マーカーの遺伝子型、または全連鎖マーカーの遺伝子型を用いて推定した確率で置換される。例えば、HaleyおよびKnott、1992、Heredity 69、315～324;JiangおよびZeng、1997、Genetica 101:47～58を参照されたい。このセクションに記載する技術は、一般に、近交系集団を必要とする。

【０１１６】

5.2.7. モデルのないノンパラメトリック連鎖解析

モデルに基づく連鎖解析(古典的連鎖解析)は、形質に対する特定の遺伝様式を仮定して、ゲノム中の所与の遺伝子座が形質に遺伝的に関連している確率を表すロッド・スコアを計算する。すなわち、対立遺伝子頻度および浸透度の値がパラメータとして含まれ、引き続き推定される。複雑性疾患の場合には、家族性集積性のすべての原因を確信をもってモデル化することは困難であることが多い。換言すれば、形質が非メンデル分離を示すときには、表現型模写リスクを含めた浸透度の値、および疾患突然変異の対立遺伝子頻度の信頼できる推定値を得ることは困難になり得る。実際、異なる遺伝子座における異なる突然変異は、感受性に異なる種類の効果を及ぼし、大きいこともあれば小さいこともあり、優性であることもあれば劣性であることもある。異なる家族において異なる遺伝様式が働く場合、または同じ家族において異なる遺伝子座が相互作用する場合には適切な遺伝モデルはない。連鎖解析の遺伝モデルの指定が不正確な場合には、得られる結果は妥当なものでも解釈できるものでもないと考えられる。

【０１１７】

上記難点の結果、遺伝モデルを定義するパラメータの値を指定する必要がない様々な連鎖検定方法が開発され、これらの方法は、モデルのない連鎖解析(これらが、真の遺伝モデルを顧慮せずに適用しうることを意味する)と称される。このような方法は、目的表現型が類似している血縁はマーカー遺伝子座が類似しており、表現型を生じる遺伝子座がマーカーと連鎖している場合にのみ、同一のマーカー対立遺伝子を共有するという前提に基づいている。

【０１１８】

モデルのない連鎖解析(対立遺伝子共有法)は、モデルを構築することではなく、モデルを拒否することに基づいている。具体的には、罹患した血縁が領域の同一コピーを偶然に

よって予想される以上に受け継いでいることを示すことによって、染色体領域の遺伝パターンが、無作為なメンデル分離と一致しないことを証明しようとするものである。罹患した血縁は、不完全浸透率、表現型模写、遺伝的異質性、および高頻度の疾患対立遺伝子が存在する場合でも、QTLに連鎖している領域における過剰な対立遺伝子共有を示すはずである。

【０１１９】

5.2.7.1. 同祖的罹患家系メンバー(IBD-APM)解析/非近交系集団

一実施形態においては、ノンパラメトリックな連鎖解析は、家系310内の罹患した血縁246(図1)を検定して、染色体領域の特定のコピーが同祖的(IBD)に共有される頻度、すなわち、系統内の共通の祖先から受け継がれる頻度を示す。次いで、遺伝子座におけるIBD共有の頻度を無作為な予想と比較することができる。同祖的罹患家系メンバー(IBD-APM)統計データは以下のように定義することができる。

【数２】

$$T(s) = \sum_{i,j} x_{ij}(s).$$

【０１２０】

式中、$x_{ij}(s)$は、染色体に沿った位置sにおいて同祖的に共有されるコピー数であり、合計は、系統310内の罹患血縁246の異なるすべての対(i,j)を取り込むものである。複数の家族から得られる結果を、重み付き合計T(s)として合算することができる。無作為な分離を想定すると、T(s)は、比較する血縁の親縁係数に基づいて計算することができる平均がμで分散がσの正規分布になる傾向にある。例えば、BlackwelderおよびElston、1985、Genet. Epidemiol. 2、85ページ;WhittemoreおよびHalpern、1994、Biometrics 50、118ページ;WeeksおよびLange、1988、Am. J. Hum. Genet. 42、315ページ;およびElston、1998、Genetic Epidemiology 15、565ページを参照されたい。無作為な分離からの偏差は、統計データ(T-μ)/σが臨界しきい値を超えるときに検出される。このセクションに記載する技術は、一般に、非近交系集団を使用する。

【０１２１】

5.2.7.2. 罹患同胞対解析/非近交系集団

罹患同胞対解析は、IBD-APM解析(セクション5.5.7.1)の一形式である。例えば、2つの同胞は、任意の遺伝子座の0、1または2個のコピーに対するIBD共有を示すことができる(無作為な分離の下では25%-50%-25%分布が予想される)。両方の親が利用可能である場合には、データを、母系染色体および父系染色体に対して別々のIBD共有に分割することができる(無作為な分離の下では50%-50%の分布が予想される0個または1個のコピー)。別の場合には、過剰の対立遺伝子共有をχ^2検定によって測定することができる。ASP手法においては、多数の小さな系統(罹患した同胞およびその親)を使用する。DNA標本を各生物から収集し、マーカー(例えば、ミクロサテライト、SNP)の大きな集合を用いて遺伝子型を決定する。次いで、機能的多型を調べる。例えば、Suarez等、1978、Ann. Hum. Genet. 42、87ページ;Weitkamp、1981、N. Engl. J. Med. 305、1301ページ;Knapp等、1994、Hum. Hered. 44、37ページ;Holmans、1993、Am. J. Hum. Genet. 52、362ページ;Rich等、1991、Diabetologica 34、350ページ;OwerbachおよびGabbay、1994、Am. J. Hum. Genet. 54、909ページ;およびBerrettini等、Proc. Natl. Acad. Sci. USA 91、5918ページを参照されたい。同胞対解析についての詳細な情報は、Hamer等、1993、Science 261、321ページを参照されたい。

【０１２２】

一部の実施形態においては、罹患同胞対が0.50を超える平均比率の同祖的マーカー遺伝子を有するかどうかを検定するASP統計量を計算した。例えば、BlackwelderおよびElston、1985、Genet. Epidemiol. 2、85ページを参照されたい。一部の実施形態においては、このような統計値は、SAGEパッケージのSIBPALプログラムを使用して計算される。例えば

、Tran等 1991、(SIB-PAL) Sib pair linkage program (Elston、New Orleans)、バージョン2.5を参照されたい。これらの統計値を、可能なすべての罹患対について計算する。一部の実施形態においては、t検定の自由度は、すべての可能な対の数の代わりに、標本中の(同胞群当たり罹患個体から1を引いた数として定義される)独立した罹患対の数に設定される。例えば、SuarezおよびEerdewegh、1984、Am. J. Med. Genet. 18、135ページを参照されたい。このセクションに記載する技術は、一般に、非近交系集団を使用する。

【０１２３】

5.2.7.3. 同質(identical by state)罹患家系メンバー(IBS-APM)解析/非近交系集団

一部の例においては、2つの血縁者が染色体領域IBDを受け継いでいるかどうかを見分けることは不可能であり、それらが領域中の遺伝マーカーに同じ対立遺伝子を有するかどうか、すなわち、同質(IBS)であるかどうかを見分けられるに過ぎない。IBDは、多型性の高いマーカーの密な集合を調べるときに、IBSから推測することができるが、遺伝解析の初期段階は、情報価値の少ないマーカーを含むより希薄な地図を伴うことがあり、その結果、IBD状態を正確に決定することができない。IBDをIBSから推測できない状況を扱う様々な方法が利用可能である。1つの方法は、マーカー・データに基づいてIBD共有を推定する(予想同祖的罹患系統メンバー;IBD-APM)。例えば、Suarez等、1978、Ann. Hum. Genet. 42、87ページ;およびAmos等、1990、Am J. Hum. Genet. 47、842ページを参照されたい。別の方法は、IBS共有に明確に基づく統計データを使用する(IBS-APM法)。例えば、WeeksおよびLange、1988、Am J. Hum. Genet. 42、315ページ;Lange、1986、Am. J. Hum. Genet. 39、148ページ;Jeunemaitre等、1992、Cell 71、169ページ;およびPericak-Vance等、1991、Am. J. Hum. Genet. 48、1034ページを参照されたい。

【０１２４】

一実施形態においては、WeeksおよびLange、1988、Am J. Hum. Genet. 42、315ページ;およびWeeksおよびLange、1992、Am. J. Hum. Genet. 50、859ページのIBS-APM技術を使用する。このような技術は、罹患個体のマーカー情報を使用して、家系内の各患者がマーカー遺伝子座において偶然によって予想される以上に互いに類似しているかどうかを検定する。一部の実施形態においては、マーカー類似度を同質の点から測定する。一部の実施形態においては、APM方法は、マーカー対立遺伝子頻度重み関数f(p)を使用する。ここで、pは対立遺伝子頻度であり、APM検定量は、3個の異なる重み関数、f(p)=1、$f(p)=1/\sqrt{p}$およびf(p)=1/pの各々に対して個別に存在する。第2および第3の関数は、患者間の稀な対立遺伝子の共有をより意味のある現象にするのに対して、第1の重み関数は、予想されるマーカー対立遺伝子共有度の計算においてのみ対立遺伝子頻度を使用する。第3の関数であるf(p)=1/pでは、(最初の2つの関数よりも頻繁に)検定量が非正規分布になり得る。第2の関数は、対立遺伝子頻度関数を取り入れつつ、検定量の正規分布を生成する妥当な折衷案である。一部の例では、APM検定量は、マーカー遺伝子座および対立遺伝子頻度の特定ミス(misspecification)に敏感である。例えば、Babron等、1993、Genet. Epidemiol. 10、389ページを参照されたい。一部の実施形態においては、Boehnke、1991、Am J. Hum. Genet. 48、22ページの方法を用いて、または対立遺伝子を検定することによって家系データから対立遺伝子頻度を推定する。例えば、Berrettini等、1994、Proc. Natl. Acad. Sci. USA 91、5918ページも参照されたい。

【０１２５】

一部の実施形態においては、APM検定量の有意性は、統計データの理論(正規)分布から計算される。また、マーカー対立遺伝子と疾患の独立遺伝(すなわち、非連鎖)を仮定して、これらのデータの多数の複製(例えば、10,000個)をシミュレートして偶然に実際の結果(または、より極端な統計データ)を観測する確率を評価する。この確率は、経験的P値である。各複製は、実際の系統を通して分離される非連鎖マーカーをシミュレートすることによって得られる。APM統計データは、実際のデータ・セットを解析するように、シミュレートされたデータ・セットを正確に解析することによって得られる。シミュレートされた統計データ分布における測定統計データの順位によって経験的P値が決まる。このセクションの技術は、一般に非近交系集団を使用する。

【０１２６】
5.2.7.4. 量的形質
モデルのない連鎖解析も量的形質に適用することができる。HasemanおよびElston、1972、Behav. Genet 2、3ページによって提案された手法は、2つの血縁間の表現型類似度は、形質の原因となる遺伝子座において共有される対立遺伝子の数と相関があるはずであるという考えに基づいている。正式には、2つの血縁間の形質の差の二乗Δ^2と、遺伝子座において同祖的に共有される対立遺伝子の数χとの回帰解析を実施する。この手法は、別の血縁(BlackwelderおよびElston、1982、Commun. Stat. Theor. Methods 11、449ページ)および多変量表現型(multivariate phenotype)(Amos等、1986、Genet. Epidemiol. 3、255ページ)にも適切に一般化することができる。Marsh等、1994、Science 264、1152ページ、およびMorrison等、1994、Nature 367、284ページ;Amos、1994、Am. J. Hum Genet. 54、535ページ;およびElston、Am J. Hum. Genet. 63、931ページも参照されたい。
【０１２７】
5.3. 細胞構成成分レベルを用いたQTL解析
このセクションでは、サブクラスiを新規形式のQTL解析を用いて解析する形式の量的遺伝解析114(図1)を用いる本発明の態様を説明する。この形式のQTL解析においては、サブクラスiの各患者または標本における複数の遺伝子の転写レベルそれぞれを表現型形質として扱う。サブクラスiの各生物における各遺伝子の発現レベルの測定値を、対応する発現統計量に変換する。遺伝子の「発現レベルの測定値」は、例えば、そのコードされたRNA(またはcDNA)またはタンパク質のレベルの測定値、あるいはコードされたタンパク質の活性レベルの測定値とすることができる。一部の実施形態においては、この変換は正規化ルーチンであり、未処理の遺伝子発現データを正規化して、平均対数比、対数強度およびバックグラウンド補正強度を得る。
【０１２８】
本発明のこの態様による量的遺伝解析114(図1)の態様をさらに説明するために、本発明のこの実施形態を容易にする特定のメモリ224中に存在するデータ構造およびモジュールを図3に示す。また、本発明のこの態様による処理ステップを図4に示す。このセクションに記載する本発明の態様による好ましい実施形態においては、細胞構成成分データ244は、遺伝子発現検定から得られる遺伝子発現データである。しかし、本発明のこの態様は、翻訳状態測定値(例えば、セクション5.12参照)などの別の形式の細胞構成成分データ、または別の態様の生物学的状態(セクション5.13参照)とともに使用することができる。本発明のこの態様は、検討中の実験交雑種またはヒト・コホートから得られる遺伝子型データ310(図3)を使用する。一部の実施形態においては、遺伝子型データ310には家系データが含まれる。一実施形態においては、細胞構成成分データ244(図3)は、検定集団内の各個体(生物)246に対するマイクロアレイ処理像からなる。このようなデータは、各個体246に対して、プロファイルされた各個体についてアレイ上で示される各遺伝子248の強度情報250、バックグラウンド・シグナル情報304、および遺伝子プローブを説明する付随注釈情報306を含む
遺伝子型および/または家系データ310は、各検定個体間の関係に加えて、これらの個体において分類された各マーカーに対する実際の対立遺伝子を含む。検定個体間の関係の程度は、F_2集団と同程度に単純な場合もあれば、広範なヒトの家系と同程度に複雑な場合もある。遺伝子型および家系データ310の例示的な出所を以下のセクション5.20.1に記載する。
【０１２９】
検定ゲノムにわたって規則的な間隔のマーカー・データ312(図3)、または目的遺伝子領域中のマーカー・データ312(図3)を使用して、分離をモニターし、または目的集団における関連性を見つける。マーカー・データ312は、検定集団において遺伝子型を評価するために使用されるマーカーを含む。一実施形態においては、マーカー・データ312は、マーカーの名前、マーカーのタイプ(例えば、SNP、ミクロサテライトなど)、ならびにゲノム配列中のマーカーの物理的および遺伝的位置を含む。また、一部の実施形態においては、

マーカー・データ312は、各マーカーに関連する異なる対立遺伝子を含む。例えば、「CA」の繰り返しからなる特定のミクロサテライト・マーカーは、検定集団における10個の異なる対立遺伝子であり、この10個の異なる対立遺伝子の各々がいくつかの繰り返しからなることができる。本発明の一実施形態による代表的なマーカー・データ312を以下のセクション5.5に示す。本発明の一実施形態においては、使用する遺伝マーカーは、一塩基多型(SNP)、ミクロサテライト・マーカー、制限断片長多型、短鎖縦列反復、DNAメチル化マーカーまたは配列長多型を含む。

【０１３０】

遺伝マーカー地図320(図3)は、複数の生物に関連する1組の遺伝マーカーから構築される。次いで、集団内の生物によって発現される複数の遺伝子中の各遺伝子Gに対して、QTLデータを得るために遺伝マーカー地図320を用いて量的形質遺伝子座(QTL)解析を実施する。1組の発現統計データは、各QTL解析に使用される量的形質である。QTL解析を、図4、要素410とともに以下にさらに詳細に説明する。任意の所与の遺伝子Gに対するこの発現統計データ・セットは、複数の生物中の各生物に対する遺伝子Gの発現統計量を含む。次に、各QTL解析から得られたQTLデータをクラスター化してQTL相互作用地図を作成する。QTL相互作用地図中の密にクラスター化されたQTLを特定することは、遺伝的に相互作用している遺伝子を特定するのに役立つ。この情報は、ヒト疾患などの複合形質の影響を受ける生物学的経路を解明するのに役立つ。本発明の一部の実施形態においては、QTL相互作用地図中の密にクラスター化されたQTLは、候補経路群とみなされる。これらの候補経路群を多変量解析にかけて、候補経路群中の遺伝子が特定の複合形質に影響を及ぼすかどうかを確認する。

【０１３１】

細胞構成成分レベルを用いたQTL解析の一実施形態による詳細な処理ステップを図4とともに説明する。この実施形態は、遺伝子発現検定から得られる細胞構成成分データ244、および検討中の実験交雑種またはヒト・コホートから得られる遺伝子型データ310から出発する(図3;図4、ステップ402)。一実施形態においては、データ310は、遺伝マーカー地図320中の各マーカーに対する検定集団中の各生物から得られる遺伝子型データを含む。一部の実施形態においては、遺伝子型データ310には表現型データが含まれる。しかし、連鎖解析を図4のステップのいずれかに使用するのでなければ、データ構造310が家系データを含む必要はない。出発データを収集した後、第1段階(図4、ステップ404)は、遺伝子発現データ244を、遺伝子発現データ244中の各遺伝子転写物量を量的形質として処理するために使用される発現統計データに変換することである。遺伝子発現データ244(図3)は、複数の遺伝子の遺伝子発現データを含む。一実施形態においては、複数の遺伝子は、少なくとも5個の遺伝子を含む。別の実施形態においては、複数の遺伝子は、少なくとも100個の遺伝子、少なくとも1000個の遺伝子、少なくとも20,000個の遺伝子、または30,000個を超える遺伝子を含む。本発明の一実施形態における解析において量的形質として一般に使用される発現統計データは、平均対数比、対数強度、バックグラウンド補正強度などであるが、これらだけに限定されない。別の実施形態においては、別のタイプの発現統計データを量的形質として使用する。一実施形態においては、この変換(図4、ステップ404)を正規化モジュール314(図3)を用いて実施する。このような実施形態においては、各調べる生物における複数の遺伝子の発現レベルが正規化される。

【０１３２】

任意の正規化ルーチンを、正規化モジュール314によって使用することができる。代表的な正規化ルーチンとしては、強度のZ-スコア、強度中央値、強度中央値の対数、強度のZ-スコア標準偏差対数、対数強度較正DNA遺伝子セットのZ-スコア平均絶対偏差、ユーザー正規化遺伝子セット、強度中央値の比率補正、強度バックグラウンド補正などがあるが、これらだけに限定されない。また、正規化ルーチンを併用して実行することもできる。本発明による例示的な正規化ルーチンを以下のセクション5.6により詳細に開示する。

【０１３３】

次いで、変換によって作成される発現統計データを、発現/遺伝子型ウェアハウス318に

保存する。ここで、発現統計データを、対応する遺伝子型情報と最終的に照合する。遺伝子発現データ244からの発現統計データの生成に加えて、遺伝マーカー地図320が遺伝マーカー312から作成される(図3;図4、ステップ406)。本発明の一実施形態においては、マーカー地図構築モジュール316(図3)を用いて遺伝マーカー地図320を作成する。検定個体の遺伝子型確率分布を計算することもできる。遺伝子型確率分布は、親のマーカー情報、マーカー間の既知の遺伝距離、マーカー間の推定遺伝距離などの情報を考慮に入れる。

【０１３４】

発現データを対応する発現統計データに変換し、遺伝マーカー地図320を構築した後、QTL解析ソフトウェアへ入力するために、このデータを、マーカー、遺伝子型および発現データすべてを関連付ける構造に変換する。この構造を発現/遺伝子型ウェアハウス318に保存する(図3;図4、ステップ408)。

【０１３５】

複数の遺伝子中の各遺伝子に対応するデータを量的形質として用いて、量的形質遺伝子座(QTL)解析を実施する(図4、ステップ410)。20,000個の遺伝子の場合には、20,000個の別々のQTL解析が行われることになる。一実施形態においては、各QTL解析を、QTL解析モジュール324によって実施する(図3)。一例においては、各QTL解析は、目的生物のゲノム中の各染色体を通して進む。検討中の量的形質との関連性を、染色体の長さ方向に沿って各ステップまたは各位置で検定する。このような実施形態においては、染色体の長さ方向に沿った各ステップまたは各位置は、規則的に規定された間隔ごとにある。一部の実施形態においては、これらの規則的に規定された間隔は、モルガン、またはより典型的にはセンチモルガン(cM)で定義される。モルガンは、染色体上のマーカー間の遺伝距離を表す単位である。モルガンは、一世代の1個の配偶子につき1回の組換え現象が起こると予想される染色体上の距離として定義される。一部の実施形態においては、各規則的に規定された間隔は100cM未満である。別の実施形態においては、各規則的に規定された間隔は10cM未満、5cM未満または2.5cM未満である。

【０１３６】

各QTL解析においては、複数の検定遺伝子から選択された遺伝子に対応するデータを量的形質として使用する。より具体的には、任意の所与の遺伝子に対してQTL解析に使用される量的形質は、セット504(図5)などの発現統計量セットである。発現統計量セット504(例えば、量的形質)は、検定集団内の各生物506に由来する遺伝子502に対応する発現統計量508を含む。したがって、本発明の一実施形態においては、各QTL解析(図4、ステップ410)は、(i)量的形質遺伝子座(QTL)解析に使用される染色体中の位置と量的形質(発現統計量セット504)との関連性を検定するステップと、(ii)ゲノム中の位置をある量だけ進めるステップと、(iii)ゲノムのすべてまたは一部が検定されるまでステップ(i)と(ii)を繰り返すステップとを含む。所与の染色体の遺伝子長がNcMであり、1cMのステップを使用する場合には、連鎖に対してN回の異なる検定を所与の染色体について実施する。

【０１３７】

一部の実施形態においては、各QTL解析から得られるQTLデータは、検定ゲノム中の各検定位置において計算されるオッズ・スコア(ロッド)の対数を含む。ロッド・スコアは、2個の遺伝子座が染色体上で互いの近傍にある可能性があるかどうか、したがって遺伝的に連鎖している可能性があるかどうかを統計的に推定するものである。この例では、ロッド・スコアは、検定ゲノム中の所与の位置が、所与の遺伝子に対応する量的形質と連鎖しているかどうかを統計的に推定するものである。以下のセクション5.7においてロッド・スコアをさらに説明する。ロッド・スコアを求めるには家系データが必要である。したがって、ロッド・スコアが得られる実施形態においては、処理ステップ410は、検討中の量的形質が目の色などの古典的な表現型ではなく細胞構成成分発現統計データなどのデータに由来する以外は、セクション5.2に記載するように本質的に連鎖解析である。

【０１３８】

家系データが利用不可能な状況においては、発現統計量セット504の差に関連する対立遺伝子型(allelic form)を有するゲノム領域を特定するために、セクション5.4に記載す

るように、遺伝マーカー地図320中の各マーカーについて、異なる生物246(図3)の各々から得られる遺伝子型データを各量的形質(発現統計量セット504)と関連解析によって比較することができる。命名法を統一するために、関連解析によって特定された領域を本明細書ではQTLと称する。

【０１３９】

連鎖解析や関連解析を処理ステップ410に使用したかどうかにかかわらず、各QTL解析結果はQTL結果データベース326に保存される(図3;図4、ステップ412)。QTL解析によって分析される複数の遺伝子中の各遺伝子に対応する量的形質328は、QTL結果データベース326中に存在する。各量的形質328(発現統計量セット504)に対して、QTL結果データベース326は、量的形質に対する関連性を検定した生物のゲノム中のすべての位置330を含む。位置330は、遺伝マーカー地図320から得られる。また、各位置330に対して、遺伝子型データ310は、複数の調べる生物中の各生物の位置330における遺伝子型を提供する。このようなデータを提供することによって、特定の量的形質328に関連する検定種のゲノム中のQTLを特定するように設計された統計解析が可能になる。QTL解析によって分析される各位置330に対して、位置と量的形質328の最大ロッド・スコアなどの統計測定値が一覧表示される。したがって、データ構造326は、検定した各量的形質328と遺伝的に関連する目的生物のゲノム中のすべての位置を含む。

【０１４０】

図6は、QTL結果データベース326をより詳細に説明するものである。各統計スコア332は、所与の位置330が、対応する量的形質328と関連する程度を測定したものである。任意の所与の量的形質328に対する統計スコア332セットは、QTLベクトルと考えることができる。したがって、本発明の一実施形態においては、調べる生物の染色体中の各検定遺伝子に対してQTLベクトルが作成される。より典型的な実施形態においては、QTLベクトルは、調べる生物のゲノム全体の各検定遺伝子に対して作成される。QTLベクトルは、遺伝子に対応して、QTL解析によって検定された各位置における統計スコア332を含む。QTLベクトルに加えて、ウェアハウス318に保存された変換遺伝子発現データから遺伝子発現ベクトルを構築することができる。各遺伝子発現ベクトルは、目的集団中の各生物から得られる遺伝子の変換発現レベルである。したがって、任意の所与の遺伝子発現ベクトルは、目的集団内の複数の異なる生物に由来する遺伝子の変換発現レベルを含む。

【０１４１】

QTLベクトルが得られると、本発明の次のステップは、QTLベクトルからQTL相互作用地図を作成することである(図4、ステップ414)。QTL相互作用地図を作成するためには、QTLベクトルを、QTLベクトル間の相互作用強度に基づいてQTL群にクラスター化する。本発明の一部の実施形態においては、QTL相互作用地図は、クラスタリング・モジュール340によって作成される。本発明の一実施形態においては、凝縮階層型クラスタリングをQTLベクトルに適用する。このクラスタリングにおいては、QTLベクトル対間のピアソン相関係数を用いて類似度が求められる。別の実施形態においては、各QTL解析から得られるQTLデータのクラスタリングは、階層型クラスタリング法の適用、k平均法の適用、ファジーk平均法の適用、Jarvis-Patrickクラスタリング法の適用、自己組織化地図の適用、またはニューラル・ネットワークの適用を含む。一部の実施形態においては、階層型クラスタリング法は、凝縮型クラスタリング手順である。別の実施形態においては、凝縮型クラスタリング手順は、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズムまたは平方和アルゴリズムである。さらに別の実施形態においては、階層型クラスタリング法は分割型クラスタリング手順である。QTLベクトルをクラスター化するのに使用することができる例示的なクラスタリング法を以下のセクション5.8に記載する。

【０１４２】

各QTLは目的集団内の複数の遺伝子中の所与の遺伝子に対応するので、QTL相互作用地図は、QTLが関連する情報を提供する。このような情報を遺伝子発現データと組み合わせて、複合形質に影響を及ぼす生物学的経路を解明するのに役立てることができる。本発明の

一実施形態においては、遺伝子発現クラスター地図を遺伝子発現統計データから構築する(図4、ステップ416)。複数の遺伝子発現ベクトルを作成する。複数の遺伝子発現ベクトル中の各遺伝子発現ベクトルは、目的集団内の複数の細胞構成成分中の遺伝子、遺伝子産物などの特定の細胞構成成分の発現レベル、活性または改変度を表す。次いで、複数の相関係数を計算する。複数の相関係数中の各相関係数を、複数の遺伝子発現ベクトル中の遺伝子発現ベクトル対間で計算する。次いで、複数の遺伝子発現ベクトルを、遺伝子発現クラスター地図を作成するために、複数の相関係数に基づいてクラスター化する。本発明の一実施形態においては、複数の相関係数中の各相関係数はピアソン相関係数である。本発明の別の実施形態においては、複数の遺伝子発現ベクトルのクラスタリングは、階層型クラスタリング法の適用、k平均法の適用、ファジーk平均法の適用、自己組織化地図の適用、またはニューラル・ネットワークの適用を含む。本発明の一実施形態においては、階層型クラスタリング法は、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズム、平方和アルゴリズムなどの凝縮型クラスタリング手順である。本発明の別の実施形態においては、階層型クラスタリング法は分割型クラスタリング手順である。遺伝子解析ベクトルをクラスター化するのに使用することができる例示的なクラスタリング法を以下のセクション5.8に記載する。

【０１４３】

この段階において、QTL相互作用地図は、遺伝子発現クラスター地図中に存在する遺伝子発現クラスター中の個々の遺伝子についての情報を提供する。遺伝子発現クラスター地図中に存在する遺伝子発現クラスターは、同じ候補経路群中にあるとみなすことができる。QTL相互作用を使用して、別の遺伝子よりも候補経路群に共に「近い」遺伝子を特定することができる。また、遺伝子発現地図中に存在し、遺伝的にまったく相互作用していない遺伝子発現クラスター中の遺伝子を、遺伝的に相互作用している遺伝子よりも軽視(down-weight)することができる。このようにして、QTL相互作用地図は、遺伝子発現クラスター地図中で特定される候補経路群を精緻化するのに役立つ。

【０１４４】

本発明の一実施形態においては、次のステップは、遺伝子発現データ244(図3)を作成するのに使用したすべてのプローブを、それぞれのゲノム座標および遺伝子座標にマッピングすることである。この情報は、所与の遺伝子が特定のQTLに直接対応する可能性(例えば、遺伝子が実際にQTLであったこと)を証明するのに役立つ。

【０１４５】

本発明の一実施形態においては、QTL相互作用地図から得られるQTL相互作用クラスター、および遺伝子発現クラスター地図から得られる遺伝子発現相互作用クラスターは、クラスター・データベース342中で表される(図3;図4、ステップ418)。クラスター・データベース342を使用して多変量QTL解析を与えるパターンを特定する。QTLおよび遺伝子発現クラスター情報に加えて、QTLおよび遺伝子の物理的位置がクラスター・データベース342中で表される。

【０１４６】

本発明の一部の実施形態においては、遺伝子は、候補経路群を得るためにQTL相互作用地図を選別することによって、QTL相互作用地図中で特定される。一実施形態においては、この選別は、QTL相互作用地図中の別のQTLと最も強く相互作用する候補経路群についてQTLを選択するステップを含む。一部の実施形態においては、QTL相互作用地図中の別のQTLと最も強く相互作用するQTLは、QTL相互作用地図中の各QTL間で計算されるすべての相関係数の75%、85%または95%よりも高い相関係数を、QTL相互作用地図中の別のQTLと共有する、QTL相互作用地図中のすべてのQTLである。

【０１４７】

本発明の一実施形態においては、クラスター・データベース342を使用して、遺伝子を形質と関連させる。代表的な形質としては、疾患状態、腫瘍段階、トリグリセリド・レベル、血圧、および/または診断検定結果などがあるがこれらだけに限定されない。この実施形態においては、クラスター・データベース342に保存されたQTL相互作用地図および/

またはデータは、候補経路群を得るために選別される(図4、ステップ420)。この選別は、遺伝子発現クラスター地図中の候補経路群におけるQTLを特定するステップを含む。本発明のこの実施形態による一例においては、QTL相互作用地図は、互いに密接に相互作用するQTL相互作用地図内のQTL群を特定することによって選別される。QTL相互作用地図において互いに密接に相互作用するQTL群中の各QTLに関連する遺伝子は、候補経路群と考えられる。一部の実施形態においては、選別は、遺伝子発現相互作用地図中の各候補経路群中の遺伝子を検索するステップをさらに含む。QTL相互作用地図において特定された候補経路群中の遺伝子が、遺伝子発現相互作用地図中で互いに密接に相互作用するかどうかは重要である。

【０１４８】

一般に、目的パターンは、クラスター・データベース342を照会することによって特定することができる。このような群は、遺伝的に最も強く相互作用する遺伝子を特定するQTL-QTL相互作用強度について選別し、次いでこの情報をこれらの群内で最も密にクラスター化している遺伝子と結び付けることによって特定することができる。これらの群のサイズは、相互作用するQTLおよび/または遺伝子を特定するために使用されるしきい値パラメータを動かすことによって容易に調節される。このような群はそれ自体を推定経路群と考えることができる。しかし、別の手法は、遺伝子が実際に同じ経路の一部であるかどうかを検定するために群を遺伝モデルに適合させることである。

【０１４９】

本発明による一実施形態においては、多変量統計モデルを候補経路群に適合させることによって、候補経路群を構成する各QTLが候補経路群内の別のQTLに属する程度を検定する(図4;ステップ422)。多変量統計モデルは、複数の量的形質を同時に考慮し、QTL間のエピスタシス相互作用をモデル化し、候補経路群の遺伝子が同じ生物学的経路または関係する生物学的経路に属するかどうかを判定する別の重要な変化を検定することが可能である。具体的な検定を実施して、検討中の形質が、実際に、同じQTLによって制御されるかどうか(多面発現効果)、あるいはそれらが独立かどうかを判断することができる。本発明に従って使用することができる例示的な多変量統計モデルを以下のセクション5.9に記載する。

【０１５０】

多変量QTL解析の結果を使用して候補経路群を「検証する」。次いで、これらの検証された群をデータベース化し、経路の再構築を含む最終解析段階で利用することができる。この段階では、ある種の一般的な遺伝的制御下にあり、発現レベルにおいてある程度相互作用し、これらの異なるレベルにおいて同じ経路または関係する経路におそらく属するのに十分な強さで相互作用することが示された遺伝子をデータベースは含む。したがって、いくつかの例においては、遺伝子と目的集団内の1つまたは複数の生物によって示される形質との関連性から、同じ経路または関係する経路の一部である遺伝子を含む経路群中に遺伝子が配置されることになる。

【０１５１】

次のステップは、所与の経路群内の経路を部分的に再構築しようとするものである。各候補経路群について、代表的なQTLベクトルと遺伝子発現ベクトルの相互作用を検討することができる。また、QTLおよびプローブ位置情報を使用して、原因経路の継ぎ合わせを開始することができる。また、グラフィカル・モデルを、相互作用強度、QTLオーバーラップ、および前のステップから蓄積された物理的位置情報を用いてデータに適合させて、候補経路群中の遺伝子を連結するエッジを重み付けし誘導することができる。このようなグラフィカル・モデルを適用して、候補経路群にさらに関連する遺伝子を明らかにし、その結果、経路のトポロジーに制約を加えるモデルが示唆される。したがって、相互作用、QTLオーバーラップおよび物理的QTL/プローブ位置による証拠が与えられたとして、このようなモデルによって、候補経路が特定の方向に進む可能性がより高いかどうかが検定される。発現データ、遺伝子型データおよびマーカー・データから出発した後、このプロセスの最終結果は、同じ経路または関係する経路の一部をなすものとして支持される遺伝子

からなる1組の経路群、および経路中の遺伝子(または、経路中の部分的な遺伝子セット)の正確な関係を示す因果情報である。

【０１５２】

5.4. 関連解析

このセクションでは、本発明において使用可能ないくつかの関連検定を説明する。関連検定は、系統標本または無関係な個体の標本を用いて実施することができる。また、関連検定は、二分形質(例えば、疾患)または量的形質に対して実施することができる。例えば、NepomおよびEhrlich、1991、Annu. Rev. Immunol. 9、493ページ;StrittmatterおよびRoses、1996、Annu. Rev. Neurosci. 19、53ページ;Vooberg等、1994、Lancet 343、1535ページ;Zoller等、Lancet 343、1536ページ;Bennet等、1995、Nature Genet. 9、284ページ;Grant等、1996、Nature Genet. 14、205ページ;およびSmith等、1997、Science 277、959ページを参照されたい。したがって、関連検定は、疾患と対立遺伝子が集団全体にわたって相関して出現するかどうかを検定するのに対して、連鎖検定は、系統内での遺伝に相関があるかどうかを判定する。

【０１５３】

連鎖解析は、ある世代から次の世代への配偶子の遺伝パターンに関係するのに対して、関連性は配偶子の集団の特性である。同じ配偶子内に存在する頻度が対立遺伝子頻度の積とは異なる場合には、2個の遺伝子座にある対立遺伝子間に関連性が存在する。この関連性が2個の連鎖遺伝子座間に存在する場合には、関連性の強さは数世代の家族内の組換えよりも歴史的な組換えに負うところが大きいので、関連性を利用することによって、精密に限局化することができる。最も簡単なシナリオでは、関連性は、疾患を引き起こす突然変異が、ある時間t_0においてある遺伝子座に出現するときに生じる。その際に、別のすべての遺伝子座における対立遺伝子で構成される特定の遺伝的背景で疾患突然変異が起こる。したがって、疾患突然変異は、この背景の対立遺伝子に完全に関連する。時間が進むにつれ、疾患遺伝子座と別のすべての遺伝子座との組換えが起こり、関連性が低下する。疾患遺伝子座により近い遺伝子座は、一般に、より高いレベルの関連性を有し、より遠方のマーカーでは関連性が急速に低下する。関連性が進化の歴史に依存することによって、50～75kbもの小さな領域への限局化が可能になる。関連性は、連鎖不平衡とも呼ばれる。関連性(連鎖不平衡)は、連鎖していない2個の遺伝子座にある対立遺伝子間に存在し得る。

【０１５４】

集団に基づく関連解析と家族に基づく関連解析の2つの形式の関連解析を以下のセクションで考察する。より一般的には、当業者は、いくつかの異なる形式の関連解析があり、そのような形式の関連解析のすべてを図1のステップ114および/または図4のステップ410において使用することができることを理解されたい。

【０１５５】

一部の実施形態においては、ゲノム全体の関連検定を本発明に従って実施する。「直接検定」法と「間接検定」法の2種類の方法を使用してゲノム全体の関連検定を実施することができる。直接検定法においては、所与の遺伝子の一般的な機能的変異体がすべて登録されており、所与の遺伝子のコード領域内において罹患個体中の特定の機能的変異体の蔓延率(関連性)が増加するかどうかを直接検定して判定する。「間接検定」法は、コード領域と非翻訳領域の両方にわたって並ぶ(例えば、図1のマーカー遺伝子型データ80から導かれる)極めて密なマーカー地図を使用する。そのような地図から得られる多型(例えば、SNP)の高密度パネルを対照中で(in controls)検定して、感受性遺伝子または耐性遺伝子の近傍の位置を狭めて決定する関連性を明らかにすることができる。この戦略は、疾患を引き起こす各配列変異体が過去のある時期に特定の個体で発生したに違いなく、したがって、その個体(生物246、図2)の変化した遺伝子の近傍の多型(ハプロタイプ)に特異的な対立遺伝子は、彼または彼女の子孫のすべてにおいて遺伝し得るという仮説に基づいている。したがって、認識可能な祖先のハプロタイプの有無は、疾患関連多型の指標になる。実際には、対立遺伝子には、関連性のあるものもあれば、突然変異体と別の多型との間で発生する組換えのために関連性のないものもある。

【０１５６】

5.4.1. 集団に基づく(モデルのない)関連解析

集団に基づく(モデルのない)関連検定においては、特定の対立遺伝子と複合形質に関連性があるかどうかを判定するために、罹患生物の対立遺伝子頻度を対照生物の対立遺伝子頻度と対比する。二分形質に対する集団に基づく関連検定は症例対照研究とも称される。症例対照研究は、集団の無関係な罹患個体と非罹患個体の比較に基づく。目的遺伝子の対立遺伝子Aは、罹患個体において対照個体よりもかなり高い頻度で存在する場合に量的表現型に関連すると言われている。統計的有意性は、ロジスティック回帰を含めて、これだけに限定されない数値方法(a number a methods)によって検定することができる。関連検定は、Lander、1996、Science 274、536;LanderおよびSchork、1994、Science 265、2037;RischおよびMerikangas、1996、Science 273、1516;およびCollins等、1997、Science 278、1533で考察されている。

【０１５７】

症例対照研究一般に該当することであるが、交絡は、集団に基づく関連解析を用いて、疾患と測定危険因子の原因関係を推測するのに問題になる。交絡に対処する一手法は、適合症例対照設計である。適合症例対照設計では、潜在的な交絡因子(例えば、年齢および性別)について個々の対照を症例と対比し、次いで適合した対の危険因子を個々に検討して、適合対照よりも症例において頻繁に発生するかどうかを確認する。一部の実施形態においては、症例と対照は民族的に同等である。換言すれば、均質な無作為交配集団を関連解析に使用する。一部の実施形態においては、以下に記載する家族に基づく関連検定を使用して、遺伝的に異質な集団による交絡の効果を最小限に抑える。例えば、Risch、2000、Nature 405、847ページを参照されたい。

【０１５８】

5.4.2. 家族に基づく関連解析

本発明の一部の実施形態においては、家族に基づく関連解析を使用する。一部の実施形態においては、各罹患生物を1個または複数の非罹患同胞(例えば、Curtis、1997、Ann. Hum. Genet. 61、319ページを参照されたい)または従兄弟(例えば、Witte等、1999、Am J. Epidemiol. 149、693ページを参照されたい)と対比し、適合症例対照研究の解析技術を使用して効果を推定し、仮説を検定する。例えば、BreslowおよびDay、1989、Statistical methods in cancer research I、The analysis of case-control studies 32、Lyon:IARC Scientific Publicationsを参照されたい。家族に基づく関連検定のいくつかの形態を以下のサブセクションで説明する。当業者は、家族に基づく関連検定には多数の形態があり、このような方法すべてが、ステップ114(図1)およびステップ410(図4)中を含めて、本発明に使用できることを理解すべきである。

【０１５９】

5.4.2.1. ハプロタイプ相対危険度検定

一部の実施形態においては、ハプロタイプ相対危険度検定を使用する。ハプロタイプ相対危険度方法においては、比較するすべてのマーカー対立遺伝子は同じ人から生じたものである。親から罹患子孫に遺伝したマーカー対立遺伝子(症例対立遺伝子)を、このような子孫に遺伝していないマーカー対立遺伝子(対照対立遺伝子)と比較する。遺伝した遺伝子型と遺伝していない遺伝子型を比較することもできる。n人の患者の2n人の親を考える。この集団は、遺伝対立遺伝子がマーカー対立遺伝子(M)であるか、別の対立遺伝子

【数３】

$$\overline{M}$$

【０１６０】

であるか、非遺伝対立遺伝子が同様にMであるか

【数４】

$\overline{M}$

【０１６１】
であるかによって2x2分割表(fourfold table)に分類することができる。
【表１】

	非遺伝対立遺伝子		
遺伝対立遺伝子	M	$\overline{M}$	合計
M	a	b	a+b
$\overline{M}$	c	d	c+d
	a+c	b+d	2n=a+b+c+d

【０１６２】
関連性を検定するために、遺伝したM対立遺伝子の割合a/(a+b)が、遺伝しないM対立遺伝子の割合a/(a+c)と大きく異なるかどうか判定する。これを判定するのに適切な1つの統計検定は、標本が大きいときに、$(b-c)^2/(b+c)$を自由度1のカイ二乗分布と比較するものである。
【０１６３】
上表の行の合計は、Mと
【数５】

$\overline{M}$

【０１６４】
である遺伝対立遺伝子の数であり、一方、列の合計は、Mと
【数６】

$\overline{M}$

【０１６５】
である非遺伝対立遺伝子の数である。これら4つの合計を、2n人の親ではなく、4n個の親対立遺伝子を分類する2x2分割表に記入することができる。

【表２】

マーカー対立遺伝子	遺伝	非遺伝	合計
M	a+b	a+c	2a+b+c
$\overline{M}$	c+d	b+d	b+c+2d
合計	2n	2n	4n

【０１６６】

ハプロタイプ相対危険度比は、(a+b)(c+d)/(a+c)(c+d)として定義される。自由度1のカイ二乗分布を使用して、ハプロタイプ相対危険度比が1と大きく異なるかどうかを判定することができる。例えば、Rudorfer等、1984、Br. J. Clin. Pharmacol. 17、433;MuellerおよびYoung、1997、Emery's Elements of Medical Genetics、Kalow編、169～175ページ、Churchill Livingstone、Edinburgh;およびRoses、2000、Nature 405、857ページ、Elson、1998、Genetic Epidemilogy、15、565ページを参照されたい。

【０１６７】

5.4.2.2. 遺伝平衡検定

一部の実施形態においては、遺伝平衡検定(TDT)を使用する。TDTは、対立遺伝子に異型接合的な親を考え、対立遺伝子が罹患子孫に遺伝する頻度を評価する。異型接合的な親に限定することによって、TDTは、多型マーカーの特定の対立遺伝子と疾患遺伝子座とを関連付ける別のモデルのない検定とは異なる。遺伝子座のパラメータ、サンプリングされた個体の遺伝子型、連鎖相、および組換え頻度は特定されない。それにもかかわらず、異型接合的な親のみを考慮することによって、TDTは、連鎖遺伝子座間の関連性に特異的である。

【０１６８】

TDTは、異質な集団に有効な連鎖および関連性の検定である。これは、本来、患児がいることによって確認された家族からなるデータに対して提案された。遺伝子データは、親と子供のマーカー遺伝子型からなる。TDTは、異型接合的な親、または遺伝子型が異なる対立遺伝子からなる親から患児への遺伝に基づいている。特に、対立遺伝子M_1およびM_2を有する両アレル・マーカーを考える。TDTは、マーカー対立遺伝子M_1がM_1M_2親から患児に遺伝する回数n_{12}、およびM_2が遺伝する回数n_{21}をカウントする。マーカーが疾患遺伝子座と連鎖していない場合、すなわちθ=0.5の場合、またはM_1と疾患突然変異に関連性がない場合には、異型接合的な親の数によって、かつ分離ひずみの非存在下で、n_{12}は二項分布B($n_{12}+n_{21}$,0.5)する。非連鎖または非関連性の帰無仮説は、統計データ

【数７】

$$T_{TDT} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

【０１６９】

によって検定することができる。統計的有意レベルは、自由度1のχ^2分布を用いて近似され、または二項分布を用いて正確に計算される。1家族当たり2人以上の患児からの遺伝がTDT統計データに含まれるときには、この検定は連鎖検定としてのみ有効である。

【０１７０】

TDT検定はいくつかの拡張検定が提案され、このような拡張検定のすべてが本発明の範

囲内にある。例えば、MortinおよびCollins、1998、Proc. Natl. Acad. Sci. USA 95、11389ページ;Terwilliger、1995、Am J Hum Genet 56、777ページを参照されたい。例えば、MuellerおよびYoung、1997、Emery's Elements of Medical Genetics、Kalow編、169～175ページ、Churchill Livingstone、Edinburgh;Zhao等、1998、Am. J. Hum. Genet. 63、225ページ;Roses、2000、Nature 405、857ページ;Spielman等、1993、Am J. Hum. Genet. 52、506ページ;およびEwensおよびSpielman;Am. H. Hum. Genet. 57、455ページも参照されたい。

【０１７１】

5.4.2.3. 同胞群に基づく検定

一部の実施形態においては、同胞群に基づく検定を使用する。例えば、Wiley、1998、Cur. Pharmaceut. Des. 4、417ページ;BlackstockおよびWeir、1999、Trends Biotechnol. 17、121ページ;KozianおよびKirschbaum、1999、Trends Biotechnol. 17、73ページ;Rockett等、Xenobiotica 29、655ページ;Roses、1994、J. Neuropathol. Exp. Neurol 53、429ページ;およびRoses、2000、Nature 405、857ページを参照されたい。

【０１７２】

5.5. マーカー・データの出所

マーカー遺伝子型データ312(マーカー地図)として使用されるいくつかの形式の遺伝マーカーが当分野で知られている。一般的な遺伝マーカーは、一塩基多型(SNP)である。SNPは、ゲノム中の600塩基対ごとに約1個存在する。例えば、KruglyakおよびNickerson、2001、Nature Genetics 27、235を参照されたい。本発明は、マーカー遺伝子型データ312源としてSNPデータベースなどの遺伝子型データベースの使用を企図する。物理的に近接したこのようなSNPのブロックを構成する各対立遺伝子は相関していることが多く、遺伝的変異性を減少させ、各々が単一の祖先の染色体からの遺伝を反映した限定された数の「SNPハプロタイプ」が定義される。Fullerton等、2000、Am. J. Hum. Genet. 67、881を参照されたい。このようなハプロタイプ構造は、解析に適切な遺伝的変異体を選択するのに有用である。Patil等は、一般的なハプロタイプ情報のすべてを取り込むために、極めて高密度のSNPセットが必要であることを見出した。一般的なハプロタイプ情報が利用可能になった後、それを使用して、包括的な全ゲノムの研究に有用なはるかに小さいSNPサブセットを特定することができる。Patil等、2001、Science 294、1719～1723を参照されたい。

【０１７３】

別の適切な遺伝マーカー源は、点状マイクロアレイ(マイクロアレイ)、高密度オリゴヌクレオチド・アレイ(HDA)、ハイブリッド形成フィルター(フィルター)、遺伝子発現の連続解析(SAGE)データなどのタイプのプラットフォームから得られる様々なタイプの遺伝子発現データを含むデータベースなどである。使用可能な遺伝子データベースの別の例は、DNAメチル化データベースである。代表的なDNAメチル化データベースの詳細については、Grunau等、MethDB- a public database for DNA methylation data、Nucleic Acids Research、印刷中、またはURL:http://genome.imb-jena.de/public.htmlを参照されたい。

【０１７４】

本発明の一実施形態においては、遺伝マーカー・セット(マーカー遺伝子型データ312)は、目的生物のゲノム変化を追跡する任意のタイプの遺伝子データベースから得られる。一般にこのようなデータベース中で示される情報は、目的生物のゲノム内の遺伝子座の集合である。各遺伝子座に対して、遺伝的変異情報が利用可能な系統が示される。示された各系統では、変化の情報が提供される。変化の情報は、任意のタイプの遺伝的変異情報である。代表的な遺伝的変異情報としては、一塩基多型、制限断片長多型、ミクロサテライト・マーカー、制限断片長多型、短鎖縦列反復などがあるが、これらだけに限定されない。したがって、適切な遺伝子型データベースとしては表1に開示したものが挙げられるが、これらだけに限定されない。

【表３】

表1:適切な遺伝子型データベース例

遺伝的変異タイプ	ユニフォーム・リソース・ロケーション
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.n1m.nih.gov/SNP/
マイクロサテライト・マーカー	http://www.informatics.jax.org/searches/polymorphism_form.shtml
制限断片長多型	http//www.informatics.jax.org/searches/polymorphism_form.shtml
短鎖縦列反復	http://www.cidr.jhmi.edu/mouse/mmset.html
配列長多型	http://mcbio.med.buffalo.edu/mit.html
DNAメチル化データベース	http://genome.imb-jena.de/public.html
短鎖縦列反復多型	Broman等, 1998, Comprehensive human genetic maps: Individual and sex-specific variation in recombination, American Journal of Human Genetics 63, 861-869
マイクロサテライト・マーカー	Kong等, 2002, A high-resolution recombination map of the human genome, Nat Genet 31, 241-247

【０１７５】

また、本発明の方法によって使用される遺伝的変異は、目的生物の実際に確認されたゲノム組成変化ではなく、遺伝子発現レベルの違いであってもよい。したがって、本発明の範囲内の遺伝子型データベースには、URL:http://www.ncbi.nlm.nih.gov/geo/にあるものなどの広範な発現プロファイル・データベースが含まれる。

【０１７６】

マーカー遺伝子型データ312として(例えば、マーカー地図として)使用することができる別の形式の遺伝マーカーは、制限断片長多型(RFLP)である。RFLPは、ヌクレオチド配列変異性によって生じるDNA制限断片間の対立遺伝子差による産物である。当業者には周知のように、RFLPは、一般に、ゲノムDNAの抽出および制限エンドヌクレアーゼによる消化によって検出される。一般に、得られた断片は、サイズに従って分離され、プローブとハイブリッド形成される。単一のコピー・プローブが好ましい。その結果、相同染色体からの制限断片が出現する。対立遺伝子間の断片サイズの差がRFLPである(例えば、Helentjaris等、1985、Plant Mol. Bio. 5:109～118、および米国特許第5,324,631号を参照されたい)。マーカー遺伝子型データ312として(例えば、マーカー地図として)使用することができる別の形式の遺伝マーカーは、無作為増幅多型DNA(RAPD)である。「無作為増幅多型DNA

」または「RAPD」という句は、DNAの対向する鎖の異なる部位に出現する単一のオリゴヌクレオチド・プライマーに相同なDNA配列間の増幅産物を意味する。結合部位における突然変異もしくは再配列、または結合部位間の突然変異もしくは再配列によって、増幅産物の有無によって検出される多型がもたらされる(例えば、WelshおよびMcClelland、1990、Nucleic Acids Res. 18:7213～7218;HuおよびQuiros、1991、Plant Cell Rep. 10:505～511を参照されたい)。マーカー遺伝子型データ312として使用することができるさらに別の形式の遺伝マーカー地図は、増幅断片長多型(AFLP)である。AFLP法は、多数の無秩序に分布した分子マーカーを生成するように設計されたプロセスである(例えば、欧州特許出願第0534858号A1を参照されたい)。マーカー地図を構築するために使用することができるさらに別の形式のマーカー遺伝子型データ312は「単純配列反復」または「SSR」である。SSRは、ゲノム内のジ-、トリ-またはテトラ-ヌクレオチド縦列反復である。反復領域は遺伝子型によって長さが変わり得るが、この反復に隣接するDNAは保存されるので、同じプライマーが複数の
遺伝子型で働く。2つの遺伝子型の多型は、2つの隣接する保存DNA配列間の長さの異なる反復である(例えば、Akagi等、1996、Theor. Appl. Genet. 93、1071～1077;Bligh等、1995、Euphytica 86:83～85;Struss等、1998、Theor. Appl. Genet. 97、308～315;Wu等、1993、Mol. Gen. Genet. 241、225～235;および米国特許第5,075,217号を参照されたい)。SSRは、サテライトまたはミクロサテライトとしても知られる。

【０１７７】

上述したように、本発明による使用に適切な多数の遺伝マーカーが公的に利用可能である。当業者は、適切なマーカーを調製することも容易にできる。分子マーカー法については、一般に、Genome Mapping in Plants (Andrew H. Paterson編)(第2章)中のAndrew H. Paterson、The DNA Revolution、1996、Academic Press/R. G. Landis Company、Austin、Tex.、7～21を参照されたい。

【０１７８】

5.6. 例示的な正規化ルーチン

正規化モジュール314(図3)によるいくつかの異なる正規化プロトコルを使用して、細胞構成成分データ248を正規化することができる。例示的な正規化プロトコルをこのセクションで説明する。一般に、正規化は、目的集団内の生物によって発現される複数の遺伝子中の各遺伝子の発現レベルの測定値を正規化することを含む。このセクションで説明する正規化プロトコルの多くが、マイクロアレイ・データを正規化するために使用される。本発明によって使用することができる多数の他の適切な正規化プロトコルがあることを理解されたい。そのようなプロトコルはすべて本発明の範囲内にある。このセクションにある正規化プロトコルの多くは、Microarray Explorer (Image Processing Section、Laboratory of Experimental and Computational Biology、National Cancer Institute、Frederick、MD 21702、USA.)などの公的に利用可能なソフトウェア中にある。

【０１７９】

正規化プロトコルの1つは、強度のZ-スコアである。このプロトコルにおいては、未処理発現強度を、標本中の全スポットの未処理強度の(平均強度)/(標準偏差)によって正規化する。マイクロアレイ・データの場合、強度のZ-スコア方法は、その標本中の全スポットの未処理強度の平均および標準偏差によって、ハイブリッド形成された各標本を正規化する。平均強度mnI_iおよび標準偏差sdI_iを、対照遺伝子の未処理強度に対して計算する。これは、平均を(0.0に)標準化し、ハイブリッド形成された標本間のデータ範囲を約-3.0～+3.0に標準化するために有用である。Z-スコアを使用するときには、比ではなくZ差(Z_{diff})を計算する。プローブi(ハイブリッド形成プローブ、タンパク質または他の結合要素)とスポットjの強度I_{ij}に対するZ-スコア強度(Z-スコア$_{ij}$)を以下のように計算する。

【０１８０】

$$Z\text{-スコア}_{ij} = (I_{ij} - mnI_i)/sdI_i$$

および

$$Zdiff_j(x,y) = Z\text{-スコア}_{xj} - Z\text{-スコア}_{yj}$$

式中、
xはxチャネルを表し、yはyチャネルを表す。
【０１８１】
別の正規化プロトコルは、各標本中の全スポットの未処理強度を、未処理強度中央値によって正規化する強度中央値正規化プロトコルである。マイクロアレイ・データでは、強度中央値正規化方法は、標本中の全スポットに対する対照遺伝子の未処理強度中央値(中央値I_i)によって、ハイブリッド形成された各標本を正規化する。したがって、強度中央値正規化方法によって正規化すると、プローブIとスポットjに対する未処理強度I_{ij}は、値Im_{ij}を有することになる。ここで、

$$Im_{ij} = (I_{ij}/\text{中央値}I_i)$$

である。
【０１８２】
別の正規化プロトコルは、対数強度中央値のプロトコルである。このプロトコルにおいては、標本中の全スポットに対する未処理発現強度を、代表的なスポットの中央値スケールの未処理強度の対数によって正規化する。マイクロアレイ・データの場合、対数強度中央値の方法では、各ハイブリッド形成標本を、対照遺伝子の中央値スケールの未処理強度(中央値I_i)の対数によって標本中の全スポットについて正規化する。本明細書で使用する対照遺伝子は、再現性のある正確に測定された発現値を有する1組の遺伝子である。強度がゼロであるときにlog(0.0)となるのを避けるために、強度値に値1.0を加算する。強度中央値正規化方法によって正規化すると、プローブIとスポットjに対する未処理強度I_{ij}は値Im_{ij}を有する。ここで、

$$Im_{ij} = \log(1.0 + (I_{ij}/\text{中央値}I_i))$$

である。
【０１８３】
さらに別の正規化プロトコルは、強度のZ-スコア標準偏差対数プロトコルである。このプロトコルにおいては、未処理の発現強度を、平均対数強度($mnLI_i$)および標準偏差対数強度($sdLI_i$)によって正規化する。マイクロアレイ・データの場合には、平均対数強度および標準偏差対数強度を、対照遺伝子の未処理強度の対数に対して計算する。したがって、プローブIとスポットjに対するZ-スコア強度$ZlogS_{ij}$は、

$$ZlogS_{ij} = (\log(I_{ij}) - mnLI_i)/sdLI_i$$

となる。
【０１８４】
さらに別の正規化プロトコルは、対数強度のZ-スコア平均絶対偏差プロトコルである。このプロトコルにおいては、未処理の発現強度を、式(log(強度) - 平均対数)/標準偏差対数による対数強度のZ-スコアによって正規化する。マイクロアレイ・データの場合、対数強度のZ-スコア平均絶対偏差プロトコルは、標本中の全スポットに対して、未処理強度の対数の平均および平均絶対偏差によって各結合標本を正規化する。平均対数強度$mnLI_i$および平均絶対偏差対数強度$madLI_i$を、対照遺伝子の未処理強度の対数に対して計算する。したがって、プローブIとスポットjに対するZ-スコア強度$ZlogA_{ij}$は、

$$ZlogA_{ij} = (\log(I_{ij}) - mnLI_i)/madLI_i$$

となる。
【０１８５】
別の正規化プロトコルは、ユーザー正規化遺伝子セット・プロトコルである。このプロトコルにおいては、未処理の発現強度を、各標本におけるユーザー定義遺伝子セット中の遺伝子の合計によって正規化する。この方法は、遺伝子サブセットが、1組の標本にわたって比較的一定して発現するように決定された場合に有用である。さらに別の正規化プロトコルは、各標本が較正DNA遺伝子の合計によって正規化される較正DNA遺伝子セット・プロトコルである。本明細書で使用する較正DNA遺伝子は、正確に測定される再現性のある発現値を与える遺伝子である。このような遺伝子は、いくつかの異なるマイクロアレイの各々で同じ発現値を有する傾向にある。このアルゴリズムは、上述したユーザー正規化遺

伝子セット・プロトコルと同じであるが、このセットは、較正DNAとして標識された遺伝子としてあらかじめ定義される。

【０１８６】

さらに別の正規化プロトコルは、強度中央値の比率補正プロトコルである。このプロトコルは、2色蛍光標識および検出スキームを使用する実施形態に有用である。(セクション5.11.1.5参照)。2色蛍光標識および検出スキームの2個の蛍光体がCy3およびCy5である場合には、比(Cy3/Cy5)に中央値Cy5/中央値Cy3強度を掛けて測定値を正規化する。バックグラウンド補正が可能である場合には、比(Cy3/Cy5)に(中央値Cy5 - 中央値BkgdCy5)/(中央値Cy3 - 中央値BkgdCy3)を掛けて測定値を正規化する。ここで、中央値Bkgdは、中央値バックグラウンド・レベルを意味する。

【０１８７】

一部の実施形態においては、強度バックグラウンド補正を使用して測定値を正規化する。スポット定量化プログラムから得られるバックグラウンド強度データを使用して、スポット強度を補正することができる。バックグラウンドは、全体値としても、スポット当たりでも指定することができる。アレイ像のバックグラウンドが低い場合、強度バックグラウンド補正が不要なこともある。

【０１８８】

5.7. オッズ・スコアの対数

すべての遺伝子型を受け継ぐ同時確率をP(g)、遺伝子型次第である全観測データx(形質およびマーカー種)の同時確率をP(x|g)とすると、1組のデータに対する尤度Lは、

$$L = \Sigma P(g)P(x|g)$$

となる。ここで、この合計は、すべての系統メンバーに対して可能な複合遺伝子型g(形質およびマーカー)の全体にわたる。この尤度における未知数は、P(g)が依存する組換え率θである。

【０１８９】

組換え率θは、減数分裂中に2個の遺伝子座が組み換わる確率である。組換え率θは、2個の遺伝子座間の距離と相関がある。定義によれば、遺伝距離は、異なる染色体上の遺伝子座間で無限大であると定義され(非シンテニック(nonsyntenic)遺伝子座)、そのような非連鎖遺伝子座ではθ = 0.5である。同じ染色体上の連鎖遺伝子座(シンテニック遺伝子座)ではθ < 0.5であり、遺伝距離はθの単調な関数となる。例えば、Ott、1985、Analysis of Human Genetic Linkage、第1版、Baltimore、MD、John Hopkins University Pressを参照されたい。セクション5.13に記載する連鎖解析の本質は、組換え率θを推定し、θ = 0.5であるかどうかを検定することである。ゲノム中の1個の遺伝子座の位置が判明すると、遺伝連鎖を利用して、第1の遺伝子座に対する第2の遺伝子座の染色体位置を推定することができる。セクション5.2に記載する連鎖解析においては、連鎖解析を使用して、遺伝地図中の多数のマーカー遺伝子座に対して、様々な量的表現型の素因となる遺伝子の未知の位置をマッピングする。組換え減数分裂および非組換え減数分裂を明確に数えることができる理想的な状況においては、大きな減数分裂標本における組換え減数分裂の頻度によってθを推定する。2個の遺伝子座が連鎖している場合、非組換え減数分裂数Nは、組換え減数分裂数Rよりも大きいと予想される。新しい遺伝子座と各マーカーの組換え率は、

【数８】

$$\hat{\theta} = \frac{R}{N+R}$$

【０１９０】

として推定することができる。

【０１９１】

その尤度は、

$$L = \Sigma P(g|\theta)P(x|g)$$

であり、検定組換え率θについての推定は、尤度比Λ = L(θ)/L(1/2)、または同じくその対数に基づく。

【０１９２】

したがって、典型的な臨床遺伝学検定においては、形質および単一マーカーの尤度は、1つまたは複数の関連する系統にわたって計算される。この尤度関数L(θ)は、形質(例えば、古典的形質または量的形質)とマーカー遺伝子座の組換え率θの関数である。標準化された対数尤度$Z(\theta) = \log_{10}[L(\theta)/L(1/2)]$をロッド・スコアと称する。ここで、「ロッド」は、「オッズの対数」の略語である。ロッド・スコアは、連鎖の証拠を可視化するものである。一般的な経験則として、ヒトの検定においては、その最大θにおいて区間[0,1/2]で

【数９】

$$Z(\hat{\theta}) \geq 3$$

【０１９３】

である場合には、遺伝学者は暫定的に連鎖を受け容れる。ここで、

【数１０】

$$\hat{\theta}$$

【０１９４】

は、その区間での最大θである。また、

【数１１】

$$Z(\hat{\theta}) \leq -2.$$

【０１９５】

であれば、連鎖は特定のθにおいて暫定的に棄却される。

【０１９６】

しかし、複合形質の場合には別の規則が提案された。例えば、LanderおよびKruglyak、1995、Nature Genetics 11、241ページを参照されたい。

【０１９７】

許容と却下は非対称的に扱われる。というのは、22対のヒト常染色体では、無秩序なマーカーが形質遺伝子座と同じ染色体上にはあり得ないからである。Lange、1997、Mathematical and Statistical Methods for Genetic Analysis、Springer-Verlag、New York;Olson、1999、Tutorial in Biostatistics:Genetic Mapping of Complex Traits、Statistics in Medicine 18、2961～2981を参照されたい。

【０１９８】

Lの値が大きいときには、既知の位置のマーカー遺伝子座に対する連鎖がない帰無仮説、L(1/2)を棄却し、量的形質に対応する遺伝子座の相対的位置を

【数１２】

$$\hat{\theta}$$

【０１９９】

によって推定することができる。したがって、ロッド・スコアによって、連鎖距離を計算する方法、ならびに2個の遺伝子(および/またはQTL)が連鎖している確率を推定する方法が提供される。

【０２００】
当業者は、ロッド・スコアの計算が種に依存することを理解されたい。例えば、マウスのロッド・スコアを計算する方法は、このセクションに記載した方法とは異なる。しかし、ロッド・スコアを計算する方法は当分野で知られており、このセクションに記載した方法は、単に説明のためのものであって、限定するためのものではない。
【０２０１】
5.8. クラスタリング技術
以下のサブセクションで、例示的なクラスタリング方法を説明する。このような技術を使用して、QTL相互作用地図を作成するためにQTLベクトルをクラスター化することができる。遺伝子発現クラスター地図を作成するために、同じ技術を遺伝子発現ベクトルに適用することができる。また、これらの技術を使用して、処理ステップ106および/またはステップ108(図2)に従って教師なし分類または教師付き分類を実施することができる。これらの技術においては、QTLベクトル、遺伝子発現ベクトル、または集団内の異なる生物から得られる細胞構成成分測定値セットは、データ(例えば、QTLベクトル、遺伝子発現ベクトルまたは細胞構成成分セット)間の相互作用強度に基づいてクラスター化される。クラスタリング技術についてのより詳細な情報は、KaufmanおよびRousseeuw、1990、Finding Groups in Data:An Introduction to Cluster Analysis、Wiley、New York、NY;Everitt、1993、Cluster analysis (3d ed.)、Wiley、New York、NY;Backer、1995、Computer-Assisted Reasoning in Cluster Analysis、Prentice Hall、Upper Saddle River、New Jersey;およびDuda等、2001、Pattern Classification、John Wiley & Sons、New York、NYを参照されたい。
【０２０２】
5.8.1. 階層型クラスタリング法
階層型クラスター分析は、測定した諸特性に基づいて成分の比較的均質なクラスターを見つける統計方法である。n個の標本からc個のクラスターへの一連の区分化を考える。最初は、各クラスターが正確に1個の標本を含むn個のクラスターへの区分化である。次は、n-1個のクラスターへの区分化、その次は、n-2個への区分化、そして、すべての標本が1個のクラスターを形成するn番目まで同様である。この一連の区分化におけるレベルkは、c = n - k + 1のときに生じる。したがって、レベル1はn個のクラスターに対応し、レベルnは1個のクラスターに対応する。任意の2個の標本xおよびx*が与えられたとして、これらはあるレベルで同じクラスターにともに群化される。2個の標本がレベルkの同じクラスターにあるときはいつでも、より高次のすべてのレベルでもそれらが一緒である特性をシーケンスが有する場合には、このシーケンスは階層型クラスタリングと呼ばれる。Duda等、2001、Pattern Classification、John Wiley & Sons、New York、2001、551を参照されたい。
【０２０３】
5.8.1.1. 凝縮型クラスタリング
一部の実施形態においては、遺伝子解析ベクトルをクラスター化するために使用される階層型クラスタリング法は、凝縮型クラスタリング手順である。凝縮型(ボトムアップ・クラスタリング)手順は、n個の単集合クラスター(singleton cluster)から出発し、クラスターを連続的に合体させることによって一連の区分を形成する。凝縮型クラスタリングにおける主要ステップは以下の手順に含まれる。ここで、cは所望の最終クラスター数であり、D_iおよびD_jはクラスターであり、x_iは遺伝子解析ベクトルであり、n個のそのようなベクトルが存在する。

【数１３】

1 初期化開始 $c, \hat{c} \leftarrow n, Di_{\leftarrow}\{x_i\}, i = 1, \ldots, n$
2 実行 $\hat{c} \leftarrow \hat{c} - 1$
3 最近接クラスター、例えばD_iおよびD_jを探索する
4 D_iとD_j合体させる
5 $c = \hat{c}$まで
6 c個のクラスターを戻す
7 終わり

【０２０４】

このアルゴリズムにおいては、用語a ←bは、変数aに新しい値bを割り当てることである。上述したように、この手順は、指定数のクラスターが得られたときに終了し、1組のポイントとしてクラスターを戻す。このアルゴリズムの要点は、2個のクラスターD_iとD_jの距離をどのように測定するかである。クラスターD_iとD_jの距離を定義するのに用いる方法によって、使用する凝縮型クラスタリング法のタイプが決まる。代表的な技術としては、最短距離アルゴリズム、最長距離アルゴリズム、平均連結アルゴリズム、重心アルゴリズム、平方和アルゴリズムなどがある。

【０２０５】

最短距離アルゴリズム。最短距離アルゴリズムは、クラスター間の距離を測定する以下の式を使用する。

【数１４】

$$d\ min(Di, Dj) = \min_{\substack{x \in Di \\ x' \in Dj}} \|x - x'\|.$$

【０２０６】

このアルゴリズムは、最小アルゴリズムとしても知られる。また、最近接クラスター間の距離が任意のしきい値を超えるときにアルゴリズムが終了する場合には、単連結アルゴリズムと称する。データ・ポイントがグラフのノードであり、エッジが同じサブセットD_i中のノード間に経路を形成する例を考える。dmin()を使用してサブセット間の距離を測定するときには、最近接ノードによって最近接サブセットが決まる。D_iとD_jの合体は、D_iとD_j中の最近接ノード対間にエッジを加えることに相当する。クラスターを連結するエッジは異なるクラスターを常に仲介しているので、得られるグラフはいかなる閉鎖ループも回路も決して含まない。グラフ理論用語では、この手順によってツリーが作成される。サブセットのすべてが連結されるまで続けることが可能な場合には全域木が得られる。全域木は、任意のノードから他の任意のノードまでの経路を含むツリーである。また、得られたツリーのエッジ長さの合計は、その標本セットに対する他のあらゆる全域木のエッジ長さの合計を超えないことがわかる。したがって、距離の尺度としてdmin()を使用することによって、凝縮型クラスタリング手順は、最小全域木を作成するアルゴリズムになる。Duda等、同上、553～554ページを参照されたい。

【０２０７】

最長距離アルゴリズム。最長距離アルゴリズムは、以下の式を使用してクラスター間距離を測定する。

【数１５】

$$d\ \max(Di, Dj) = \max_{\substack{x \in Di \\ x' \in Dj}} \|x - x'\|.$$

【０２０８】
このアルゴリズムは、最大アルゴリズムとしても知られる。最近接クラスター間距離が任意のしきい値を超えるときにクラスタリングが終了する場合には、完全連結アルゴリズムと称する。最長距離アルゴリズムによって、長いクラスターの成長が阻止される。エッジがクラスター中のノードのすべてを連結しているグラフを作成するときに、この手順の適用を考慮することができる。グラフ理論用語では、すべてのクラスターは完全な部分グラフを含む。2個のクラスター間の距離は、2個のクラスター中の最も遠いノードで終結する。最近接クラスターを合体させるときには、2個のクラスター中のノード対ごとにエッジを追加することによってグラフを変える。

【０２０９】
平均連結アルゴリズム。別の凝縮型クラスタリング法は、平均連結アルゴリズムである。平均連結アルゴリズムは、以下の式を使用してクラスター間の距離を測定する。

【数１６】

$$d_{avg}(Di, Dj) = \frac{1}{ninj} \sum_{x \in Di} \sum_{x' \in Dj} \|x - x'\|.$$

【０２１０】
階層型クラスター分析を、そのようなベクトル・セット中のすべての遺伝子解析ベクトルを対ごとに比較することによって開始する。セット中のすべての成分対の類似性を評価した後に距離行列を作成する。距離行列中で最短距離(すなわち、最も類似した値)にある一対のベクトルを選択する。次いで、平均連結アルゴリズムを使用するときには、2個のベクトルを平均することによって「ノード」(「クラスター」)を作成する。2個の結合成分を置換する新しい「ノード」(「クラスター」)で類似度行列を更新し、単一の成分のみが残るまでこのプロセスをn-1回繰り返す。以下の値を有する6つの成分A～Fを考える。

【０２１１】
A4.9、B8.2、C3.0、D5.2、E8.3、F2.3
第1の区分化においては、平均連結アルゴリズムを用いて、計算し得る1つの行列(解1)は、
(解1) A4.9、B-E8.25、C3.0、D5.2、F2.3
である。

【０２１２】
あるいは、平均連結アルゴリズムを用いた第1の区分化によって以下の行列、
(解2) A4.9、C3.0、D5.2、E-B8.25、F2.3
が得ることができる。

【０２１３】
解1が第1の区分化において求められたとして、平均連結アルゴリズムを用いた第2の区分化によって、
(解1-1) A-D5.05、B-E8.25、C3.0、F2.3
または
(解1-2) B-E8.25、C3.0、D-A5.05、F2.3
が得られる。

【０２１４】

解2が第1の区分化において求められたとして、平均連結アルゴリズムの第2の区分化によって、
(解2-1) A-D5.05、C3.0、E-B8.25、F2.3
または
(解2-2) C3.0、D-A5.05、E-B8.25、F2.3
が得られる。

【０２１５】

したがって、平均連結アルゴリズムにおける単なる2回の区分化の後に、すでに4つの行列が存在する。Duda等、Pattern Classification、John Wiley & Sons、New York、2001、551ページを参照されたい。

【０２１６】

5.8.1.2. ピアソン相関係数を用いたクラスタリング

本発明の一実施形態においては、QTLベクトルおよび/または遺伝子発現ベクトルを、ピアソン相関係数を用いた凝縮型階層型クラスタリングによってクラスター化する。この形式のクラスタリングにおいては、QTLベクトル対間、遺伝子発現対間、または細胞構成成分測定値セット間のピアソン相関係数を用いて類似度を求める。ピアソン相関係数に加えて、使用可能な他の尺度としては、ユークリッド距離、ユークリッド平方距離、ユークリッド平方和、マンハッタン計量、二乗ピアソン相関係数などがあるが、これらだけに限定されない。このような尺度は、SAS (Statistics Analysis Systems Institute、Cary、North Carolina)またはS-Plus (Statistical Sciences, Inc.、Seattle、Washington)を用いて計算することができる。

【０２１７】

5.8.1.3. 分割型クラスタリング

一部の実施形態においては、QTLベクトルおよび/または遺伝子発現ベクトルをクラスター化するために使用される階層型クラスタリング法は、分割型クラスタリング手順である。分割型(トップダウン・クラスタリング)手順は、1個のクラスター中の標本のすべてを用いて始まり、クラスターを首尾よく分割することによってシーケンスを形成させる。分割型クラスター化技術は、多形質的方法または単形質的方法のどちらかに分類される。多形質的手法は、クラスターを任意のサブセットに分割する。

【０２１８】

5.8.2. K平均クラスタリング

k平均クラスター化では、QTLベクトル、遺伝子発現ベクトルまたは細胞構成成分測定値セットが、K個のユーザー指定クラスターに無作為に割り当てられる。各クラスターの重心を、各クラスター中のベクトルの値を平均することによって計算する。次いで、各i=1、...、Nに対して、ベクトルx_iと各クラスター重心との距離を計算する。次いで、各ベクトルx_iを、重心が最も近いクラスターに再度割り当てる。次に、影響を受けた各クラスターの重心を再計算する。このプロセスを、それ以上再割り当てされなくなるまで繰り返す。Duda等、2001、Pattern Classification、John Wiley & Sons、New York、NY、526～528ページを参照されたい。関係する手法は、ファジーc平均アルゴリズムとしても知られるファジーk平均クラスタリング・アルゴリズムである。ファジーk平均クラスタリング・アルゴリズムにおいては、すべてのQTLベクトル、遺伝子発現ベクトルまたは細胞構成成分測定値セットが、任意の所与の時間、正確に1個のクラスター中にあるという仮定が緩和され、すべてのベクトル(またはセット)がクラスター中のある段階的または「ファジー」な帰属関係にある。Duda等、2001、Pattern Classification、John Wiley & Sons、New York、NY、528～530ページを参照されたい。

【０２１９】

5.8.3. JARVIS-PATRICKクラスタリング

Jarvis-Patrickクラスタリングは、1組のオブジェクトが、共有最近接数に基づいてクラスターに分割される最短距離非階層型クラスタリング法である。JarvisおよびPatrick、1973、IEEE Trans. Comput.、C-22:1025～1034が唱える標準的な方法では、データ・セ

ット中の各オブジェクトのK最近接を前処理段階で特定する。続くクラスター化段階においては、(i)iがjのK最近接の1つであり、(ii)jがiのK最近接の1つであり、(iii)iとjがK最近接の少なくともk_{min}を共有する場合には、2個のオブジェクトiとjが同じクラスターに加わる。ここで、Kおよびk_{min}は、ユーザー定義パラメータである。この方法は、断片デスクリプタに基づいてクラスタリング化学構造に広範に適用され、コンピュータ的要求が階層方法よりもはるかに厳しくない利点を有し、したがって、大きなデータベースにより適している。Jarvis-Patrickクラスタリングは、Jarvis-Patrick Clustering Package 3.0 (Barnard Chemical Information, Ltd.、Sheffield、United Kingdom)を用いて実施することができる。

【０２２０】

5.8.4. ニューラル・ネットワーク

ニューラル・ネットワークは、重みの層によって出力ユニット層に接続された入力ユニット(およびバイアス)層を含む層構造を有する。多層ニューラル・ネットワークには、入力ユニット、隠れユニットおよび出力ユニットがある。実際に、入力から出力までのあらゆる機能を、3層ネットワークとして実行することができる。そのようなネットワークにおいては、重みは、訓練パターンおよび所望の出力に基づいて設定される。多層ニューラル・ネットワークの教師あり訓練の一方法は逆伝播法である。逆伝播法によって、各隠れユニットの実効誤差の計算が可能になり、したがってニューラル・ネットワークの入力-隠れ重みに対する学習規則を導き出すことができる。

【０２２１】

ニューラル・ネットワークの基本的使用法は、訓練を受けていないネットワークから出発し、入力層に訓練パターンを与え、ネットを通してシグナルを送り、出力層で出力を測定するものである。次いで、これらの出力を標的値と比較する。あらゆる相違が誤差に対応する。この誤差または基準関数は、重みのスカラー関数であり、ネットワーク出力が所望の出力に一致したときに最小になる。したがって、重みを調整してこの誤差測定値を減少させる。一般に使用される3種類の訓練プロトコルは、確率、バッチおよびオンラインである。確率的訓練においては、パターンが訓練セットから無作為に選択され、ネットワーク重みが各パターン提示に対して更新される。確率的な逆伝播法などの勾配降下法によって訓練された多層非線形ネットワークは、ネットワーク・トポロジーによって定義されたモデルにおいて重み値の最尤推定法を実行する。バッチ訓練においては、すべてのパターンが、学習する前にネットワークに提示される。一般に、バッチ訓練においては、いくつかのパスが訓練データを通してなされる。オンライン訓練においては、各パターンがネットに1回のみ提示される。

【０２２２】

5.8.5. 自己組織化地図

自己組織化地図は、分割型クラスタリング手法に基づくニューラル・ネットワークである。その目的は、遺伝子を、それらの発現ベクトルと各区分に対して規定された参照ベクトルとの類似度に基づいて一連の区分に割り当てることである。2つの異なる実験から得られる2個のマイクロアレイが存在する例を考える。2つの実験における任意の所与の遺伝子の発現レベルにすべてのスポットが対応する二次元構築体を作製することが可能である。二次元グリッドを作製し、二次元構築体のいくつかの区分を得る。次に、遺伝子を無作為に選択し、その選択した遺伝子に最も近い参照ベクトル(ノード)の本性を距離行列に基づいて決定する。次いで、参照ベクトルを調節して、割り当てた遺伝子のベクトルにより近似させる。すなわち、参照ベクトルをx軸およびy軸上で1距離ユニット動かして、割り当てた遺伝子により近くする。他のノードをすべて、割り当てた遺伝子に合わせて調整する。ただし、1/2または1/4距離ユニットしか動かさない。このサイクルを数十万回繰り返して参照ベクトルを一定値に収束させ、そこでグリッドは安定する。このとき、すべての参照ベクトルは、遺伝子群の中心にある。最後に、最も類似している参照ベクトルに応じて、関連する区分に遺伝子をマッピングする。

【０２２３】

5.9. 多変量統計モデル

本発明の方法を用いて、QTL相互作用地図データおよび遺伝子発現クラスター地図の分析から候補経路群を特定する。各候補経路群はいくつかの遺伝子を含む。本発明の方法は、クラスタリング法を利用して、目的集団のゲノム中の数千個にもなり得る遺伝子を数個の候補経路群に選別するので有利である。典型的な例においては、候補経路群は、遺伝子発現クラスター地図中で密にクラスターを形成する遺伝子群である。候補経路群中の遺伝子も、QTL相互作用地図中で密にクラスターを形成する。QTL相互作用地図は、候補経路群中の遺伝子を規定する補足的手法として役立つ。例えば、遺伝子A、BおよびCが遺伝子発現クラスター地図中で密にクラスターを形成する場合を考える。また、遺伝子A、B、CおよびDは、対応するQTL相互作用地図中で密にクラスターを形成する。この例においては、遺伝子発現クラスター地図を分析しただけでは、遺伝子A、B、Cが候補経路群を形成することが示唆される。しかし、QTL相互作用地図と遺伝子発現クラスター地図の両方を分析すると、候補経路群が遺伝子A、B、CおよびDを含むことが示唆される。

【０２２４】

候補経路群が特定されると、多変量統計手法を使用して、候補経路群中の各遺伝子が複雑性疾患形質などの特定の形質に影響を及ぼすかどうかを明らかにすることができる。本発明の一部の実施形態において使用される多変量統計解析の形式は、利用可能な遺伝子型および/または家系データのタイプによって決まる。一般に、検定すべき集団が植物または動物の場合には、より多くの家系データが利用可能である。このような場合、JiangおよびZeng、1995、Nature Genetics 140:1111～1127ページのモデルなどの多変量統計モデル、ならびにQTL Cartographer(BastenおよびZeng、1994、Zmap-a QTL cartographer、Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software 22、Smith等編、65～66ページ、The Organizing Committee、5th World Congress on Genetics Applied to Livestock Production、Guelph、Ontario、Canada;Basten等、2001、QTL Cartographer、Version 1.15、Department of Statistics、North Carolina State University、Raleigh、North Carolinaにおいて実行される技術。また、マーカー回帰(結合マッピング、マーカー差回帰(marker-difference regression、MDR)、標識コファクターを含む区間マッピング、および複合区間マッピングを使用することができる。例えば、Lynch & Walsh、1998、Genetics and Analysis of Quantitative Traits、Sinauer Associates、Inc.、Sunderland、MAを参照されたい。

【０２２５】

JiangおよびZengは、複合区間マッピング(CIM)への複数形質拡張法(multiple-trait extension)を開発した。例えば、JiangおよびZeng、1995、Genetics 140、1111ページを参照されたい。CIMは、他の点では標準的な区間解析(例えば、線形モデルを用いた、または最大尤度によるQTL検出)にマーカー・コファクターを加える一般的手法である。CIMは、解析用コファクターとして追加のマーカーを含むように標準区間マッピングを改変して、生物から得られる多座マーカー情報を取り込むことによって複数のQTLを処理する。例えば、Jansen、1993、Genetics 135、205ページ;Zeng、1994、Genetics 136、1457ページを参照されたい。JiangおよびZengによって開発されたCIMへの複数形質拡張法は、本発明の方法を用いて構築される候補経路群を、これらの候補経路群中の遺伝子が同じ遺伝子領域と連鎖している場合に、検定する枠組みを提供するものである。JiangおよびZengの方法によって、同じ領域と連鎖している(候補経路群中の遺伝子の)発現値が、単一遺伝子の多面発現)または2個の密接に連鎖した遺伝子によって制御されるかどうかを判定することが可能になる。JiangおよびZengの方法が、複数の遺伝子が密接な連鎖遺伝子座(密接な連鎖遺伝子)によって実際に制御されていることを示唆する場合には、同じ領域と連鎖している遺伝子が同じ経路にあることは支持されない。さらに、経路群のサブセットを検定して、別の遺伝子に対して基本的な多面発現的関係を有する遺伝子を識別することによって、経路の構成成分(階層)を削減することができる。また、候補経路群中の他の遺伝子と多面発現的関係にない候補経路群中の特定の遺伝子を除去することによって、候補経路群の定義を精緻化することができる。この考えは、所与の領域と連鎖している遺伝子のうちどれ

が、それらの物理的位置と連鎖している他の遺伝子を有するかを明らかにし、階層および制御の順序を示すものである。

【０２２６】

現在、実用上の制約は、JiangおよびZengの方法などの多変量方法を用いて一度に10個以下の遺伝子しか処理することができないことである。理論的に、遺伝子数は、モデルに適合させるのに利用可能なデータ量によって制限されるが、特別な制限は、最適化技術が10次元を超えては有効でないことである。しかし、一部の実施形態においては、次元数削減技術(dimensionality reductions technique)を実施することによって10個を超える遺伝子を一度に処理することができる(主構成成分のように)。

【０２２７】

ヒトの遺伝子型と家系データの場合には、Amos等、1990、Am J. Hum. Genetics 47:247～254ページの方法を含めて、Allison、1998、Multiple Phenotype Modeling in Gene-Mapping Studies of Quantitative Traits:Power Advantages、Am J. Hum. Genetics 63:1190～1201ページに記載された方法が使用されるが、これらだけに限定されない。

【０２２８】

一部の実施形態においては、複数のタイプの組織の遺伝子発現データを収集する。そのような場合、多変量解析を使用して複雑性疾患の本質を明らかにすることができる。本発明のこの実施形態に使用される多変量技術は、Williams等、1999、Am J Hum Genet 65(4):1134～47;Amos等、1990、Am J Hum Genet 47(2):247～54、ならびにJiangおよびZeng、1995、Nature Genetics 140:1111～1127にある程度記載されている。

【０２２９】

喘息は、複数のタイプの組織から得られる発現データを用いて検定することができる複雑性疾患の一例である。喘息は、肺だけでなく血中の免疫系反応による影響をある程度受けると考えられる。肺および血中における遺伝子の発現を測定することによって、モデル・システム、例えば、F2マウス交雑種において共有される遺伝的効果を以下のモデルを用いて精査することができる。

【数１７】

$$y_{j1} = \alpha_1 + b_1 x_j + d_1 z_j + e_{j1}$$

$$y_{j2} = \alpha_2 + b_2 x_j + d_2 z_j + e_{j2}$$

$$\vdots$$

$$y_{jm} = \alpha_m + b_m x_j + d_m z_j + e_{jm}$$

【０２３０】

式中、個体jおよび推定QTLに対して、

$y_{j1},\ldots,y_{jm}$は、喘息に関連する表現型、肺における遺伝子発現の発現データ、および血中の遺伝子発現の発現データからなり、

x_jは、特定の親系統に由来するQTL対立遺伝子の数であり、

z_jは、個体がQTLに対して異型接合である場合は1であり、それ以外は0であり、

α_iは、表現型iの平均であり、

b_iおよびd_iは、表現型iに対するQTLの添加効果および優性効果であり、

e_{ji}は、個体jおよび表現型iに対する残余誤差である。

【０２３１】

一般に、残差は個体間で相関せず、個体内の残差間の相関は$Cov(e_{jk}, e_{jl}) = \rho_{kl}\sigma_k\sigma_l$としてモデル化されると考えられる。残差に対して多変量正規分布を想定すると、尤度分析を利用して形質ベクトルに対するQTLの結合連鎖(joint linkage)を検定し、多面発

現効果と近接連鎖(close linkage)を検定することができる。そのような情報を用いて、血中で発現する1組の遺伝子、および重複している可能性があるが、肺において発現する1組の遺伝子の遺伝子発現を変化させることによって、喘息の罹病性に影響を及ぼすQTLを検出することが可能なはずである。本発明によるこのような多変量解析を、複数の組織にわたる発現データを含む高品位表現型データと組み合わせて、複雑性疾患への罹患性に真に影響を及ぼす遺伝子の検出を改善することができる。

【０２３２】

5.10. 分析キットの使用

好ましい実施形態においては、キットを使用して本発明の方法を実施することができる。そのようなキットは、以下のサブセクションに記載するものなどのマイクロアレイを含む。そのようなキットに含まれるマイクロアレイは、固相の既知の位置においてプローブがハイブリッド形成または結合する固相、例えば、表面を備える。好ましくは、これらのプローブは、既知の異なる配列の核酸からなり、各核酸は、RNA種またはそれから誘導されるcDNA種にハイブリッド形成可能である。特定の実施形態においては、本発明のキットに含まれるプローブは、目的生物から収集される細胞中のRNA種に由来する核酸配列に特異的にハイブリッド形成可能な核酸である。

【０２３３】

好ましい実施形態においては、本発明のキットは、コンピュータ読み取り可能な媒体にエンコードされた、図2および/または図4の上述した1個または複数のデータベース、および/または遠隔のネットワーク・コンピュータから上述したデータベースを使用するアクセス許可も含む。

【０２３４】

別の実施形態においては、本発明のキットは、さらに、図2および/または図4に示した上述したものなどのコンピュータ・システムのメモリに読み込み可能なソフトウェアを含む。本発明のキットに含まれるソフトウェアは、図2および/または図4とともに上述したソフトウェアと本質的に同じものである。

【０２３５】

本発明の分析方法を実施する別のキットも、当業者には明らかであり、添付した特許請求の範囲に包含されるものである。

【０２３６】

5.11. 転写状態測定値

このセクションは、細胞構成成分の1タイプである遺伝子の発現レベルを測定するいくつかの例示的な方法を提供する。当業者は、本発明が、複数の生物中の各生物における細胞構成成分(例えば、遺伝子)の発現レベルを測定する以下の特定の方法に限定されないことを理解されたい。

【０２３７】

5.11.1. マイクロアレイを用いた転写物アッセイ

このセクションで説明する技術は、発現プロファイルをモニターすることによって、細胞または細胞型または他のあらゆる細胞標本の発現状態または転写状態を明らかにするのに特に有用である。これらの技術は、複数の遺伝子の発現レベルを同時に決定するのに使用することができるポリヌクレオチド・プローブ・アレイを用意することを含む。これらの技術は、さらに、そのようなポリヌクレオチド・プローブ・アレイを設計し作製する方法も提供する。

【０２３８】

遺伝子中のヌクレオチド配列の発現レベルを、任意のハイスループット技術によって測定することができる。その結果は、どう測定しても、転写物の絶対量もしくは相対量、または存在量もしくは存在割当量(abundance rations)を表す値を含めて、ただしこれらだけに限定されない応答データのどちらかである。発現プロファイルの測定は、このサブセクションに記載する転写物アレイへのハイブリッド形成によってなされることが好ましい。一実施形態においては、「転写物アレイ」または「プロファイリング・アレイ」を使用

する。転写物アレイは、細胞標本における発現プロファイルを分析するために使用することができ、特に、特定の組織タイプもしくは発生状態の細胞標本、または目的薬物に曝された細胞標本の発現プロファイルを測定するために使用することができる。

【０２３９】

一実施形態においては、発現プロファイルは、細胞中に存在するmRNA転写物中のヌクレオチド配列である検出可能に標識されたポリヌクレオチド(例えば、全細胞mRNAから合成された蛍光標識cDNA)をマイクロアレイにハイブリッド形成させることによって得られる。マイクロアレイは、細胞または生物のゲノム、好ましくはほとんどの遺伝子またはほとんどすべての遺伝子中のヌクレオチド配列の多くを表す、担体上の位置的にアドレス指定可能な結合(例えば、ハイブリッド形成)部位のアレイである。そのような結合部位の各々は、担体上の所定領域に結合したポリヌクレオチド・プローブからなる。マイクロアレイは、いくつかの方法で作製することができ、そのいくつかを本明細書の以下で説明する。マイクロアレイは、どう作製してもある特性を共有する。アレイは再現性があり、所与のアレイの複数のコピーを作製することができ、互いに容易に比較することができる。マイクロアレイは、結合(例えば、核酸ハイブリッド形成)条件下で安定な材料で作製されることが好ましい。マイクロアレイは、小さいことが好ましく、例えば、約1 cm^2～25 cm^2、好ましくは約1～3 cm^2である。しかし、より大きなアレイもより小さなアレイも企図され、例えば、極めて多数または極めて少数の様々なプローブを同時に評価する場合に好ましいことがある。

【０２４０】

マイクロアレイ中の所与の結合部位または独特な結合部位セットは、細胞または生物から得られる単一遺伝子のヌクレオチド配列(例えば、特定のmRNAまたはそれに由来する特定のcDNAのエキソン)に特異的に結合する(例えば、ハイブリッド形成する)ことが好ましい。

【０２４１】

使用されるマイクロアレイは1個または複数の検定プローブを備えることができ、その各々が、検出しようとするRNAまたはDNAの部分配列に相補的であるポリヌクレオチド配列を有する。各プローブは一般に核酸配列が異なり、アレイの固体表面上の各プローブの位置は通常既知である。実際、マイクロアレイは、好ましくは、アドレス可能なアレイであり、より好ましくは、位置的にアドレス可能なアレイである。アレイの各プローブは固体担体上の既知の所定位置にあることが好ましく、各プローブの本性(すなわち、配列)をアレイ上(すなわち、担体上または表面)のその位置から決定することができる。一部の実施形態においては、アレイは順序付けられたアレイである。

【０２４２】

マイクロアレイまたは1組のマイクロアレイ上のプローブの密度は、様々な(すなわち、同一でない)プローブが約100個/cm^2以上であることが好ましい。より好ましくは、本発明の方法に使用されるマイクロアレイは、少なくとも550個/cm^2のプローブ、少なくとも1,000個/cm^2のプローブ、少なくとも1,500個/cm^2のプローブ、または少なくとも2,000個/cm^2のプローブを有する。特に好ましい実施形態においては、マイクロアレイは、異なるプローブが少なくとも約2,500個/cm^2の密度を好ましくは有する高密度アレイである。したがって、本発明に使用されるマイクロアレイは、好ましくは、少なくとも2,500個、少なくとも5,000個、少なくとも10,000個、少なくとも15,000個、少なくとも20,000個、少なくとも25,000個、少なくとも50,000個、または少なくとも55,000個の様々な(すなわち、同一でない)プローブを含む。

【０２４３】

一実施形態においては、マイクロアレイは、各位置が、遺伝子によってコードされる転写物のヌクレオチド配列に対して(例えば、mRNAまたはそれに由来するcDNAのエキソンに対して)別個の結合部位であるアレイ(例えば、行列)である。マイクロアレイ上の結合部位集団は、複数の遺伝子に対する結合部位セットを含む。例えば、様々な実施形態においては、本発明のマイクロアレイは、生物のゲノム中の50%未満の遺伝子によってコードさ

れる産物に対する結合部位を含むことができる。あるいは、本発明のマイクロアレイは、生物のゲノム中の少なくとも50%、少なくとも75%、少なくとも85%、少なくとも90%、少なくとも95%、少なくとも99%または100%の遺伝子によってコードされる産物に対する結合部位を有することができる。別の実施形態においては、本発明のマイクロアレイは、生物の細胞によって発現される遺伝子の50%未満、少なくとも50%、少なくとも75%、少なくとも85%、少なくとも90%、少なくとも95%、少なくとも99%または100%によってコードされる産物に対する結合部位を有することができる。この結合部位は、特定のRNAが特異的にハイブリッド形成することができるDNAまたはDNAアナログとすることができる。例えば、DNAまたはDNAアナログは、例えばエキソンに対応する合成オリゴマーまたは遺伝子断片とすることができる。

【０２４４】

本発明の一部の実施形態においては、遺伝子または遺伝子中のエキソンは、プロファイリング・アレイにおいて遺伝子またはエキソンの様々な配列セグメントに相補的である様々なポリヌクレオチドを有するプローブを含む1組の結合部位によって表される。そのようなポリヌクレオチドは、好ましくは15～200塩基長、より好ましくは20～100塩基長、最も好ましくは40～60塩基長である。各プローブ配列は、その標的配列に相補的である配列に加えてリンカー配列を含むこともできる。本明細書で使用するリンカー配列は、その標的配列に相補的である配列と担体表面との間の配列である。例えば、好ましい実施形態においては、本発明のプロファイリング・アレイは、各標的遺伝子またはエキソンに特異的である1個のプローブを含む。しかし、所望であれば、プロファイリング・アレイは、いくつかの標的遺伝子またはエキソンに特異的である少なくとも2、5、10、100、1000個またはそれ以上のプローブを含むことができる。例えば、アレイは、遺伝子の最長のmRNAアイソフォーム配列にわたって一塩基ステップで並べられたプローブを含むことができる。

【０２４５】

本発明の具体的な実施形態においては、エキソンが別のスプライスされた変異体を含むときには、エキソンの最長変異体を含むゲノム領域にわたって連続した重複配列、すなわち、並べられた配列の1組のポリヌクレオチド・プローブを、エキソン・プロファイリング・アレイに含めることができる。ポリヌクレオチド・プローブ・セットは、所定の塩基間隔のステップ、例えば1、5または10塩基間隔のステップで連続した重複配列を含むことができ、最長の変異体を含むmRNAの全体に及んでいるか、またはmRNAの全体にわたって並べられている。したがって、このようなプローブ・セットを用いて、すべてのエキソン変異体を含むゲノム領域を走査して、発現される変異体またはエキソン変異体を求めて、発現される変異体またはエキソン変異体を求めることができる。これとは別に、またはこれに加えて、エキソン特異的プローブおよび/または変異体接合プローブを含む1組のポリヌクレオチド・プローブをエキソン・プロファイリング・アレイに含めることができる。本明細書で使用する変異体接合プローブとは、特定のエキソン変異体および隣接エキソンの接合領域に特異的なプローブを意味する。いくつかの例では、プローブ・セットは、エキソンのすべての異なるスプライス接合配列の各々に特異的にハイブリッド形成可能な変異体接合プローブを含む。別の例では、プローブ・セットは、エキソンのすべての異なる変異体中の一般的な配列に特異的にハイブリッド形成可能なエキソン特異的プローブ、および/またはエキソンの異なるスプライス接合配列に特異的にハイブリッド形成可能な変異体接合プローブを含む。

【０２４６】

いくつかの例においては、エキソンは、エキソン・プロファイリング・アレイにおいて、完全長エキソンに相補的であるポリヌクレオチドを含むプローブによって表される。このような例においては、エキソンは、プロファイリング・アレイ上の単一結合部位によって表される。いくつかの好ましい例においては、エキソンは、プロファイリング・アレイ上の1個または複数の結合部位によって表され、結合部位の各々は、標的エキソンの重要な部分であるRNA断片に相補的であるポリヌクレオチド配列を有するプローブを含む。このようなプローブの長さは、通常、約15～600塩基、好ましくは約20～200塩基、より好ま

しくは約30～100塩基、最も好ましくは約40～80塩基である。エキソンの平均長さは、約200塩基である(例えば、Lewin、Genes V、Oxford University Press、Oxford、1994を参照されたい)。長さ約40～80のプローブは、それより長さの短いプローブよりもエキソンにより特異的に結合し、それによって標的エキソンに対するプローブの特異性が高くなる。ある種の遺伝子では、1個または複数の標的エキソンは、約40～80塩基未満の配列長さとすることができる。そのような場合、標的エキソンよりも長い配列を含むプローブを使用すべきときには、隣接する構成的スプライス・エキソン(constitutively splice exon)からの配列が隣接する標的エキソン全体を含む配列を含むプローブを、プローブ配列がmRNA中の対応する配列セグメントに相補的であるように設計することが望ましいことがある。ゲノムのフランキング配列、すなわち、イントロン配列ではなく、隣接する構成的にスプライスされたエキソンからのフランキング配列を用いることによって、同じ長さの別のプローブと同等のハイブリッド形成ストリンジェンシーが可能になる。使用するフランキング配列は、どんな代替経路にも関与しない、隣接する構成的にスプライスされたエキソンから得られることが好ましい。使用するフランキング配列は、隣接するエキソンの配列の重要な部分を含まず、交差ハイブリッド形成が最小限に抑えられることがより好ましい。一部の実施形態においては、所望のプローブ長よりも短い標的エキソンが選択的スプライシングに関与するときには、選択的にスプライシングされた様々なmRNA中のフランキング配列を含むプローブは、選択的にスプライシングされた様々なmRNA中で発現されるエキソンの発現レベルが測定できるように設計される。

【０２４７】

いくつかの例においては、選択的スプライシング経路および/または別個の遺伝子中のエキソン複製を区別しようとするとき、DNAアレイまたはアレイ・セットは、2個の隣接するエキソンの接合領域にまたがる配列に相補的であるプローブを含むこともできる。そのようなプローブは、各個々のエキソンに対するプローブと実質的に重複しない2個のエキソンからの配列を含み、交差ハイブリッド形成が最小限に抑えられることが好ましい。1個を超えるエキソンからの配列を含むプローブは、エキソンが、選択的にスプライシングされた1個もしくは複数のmRNAおよび/または複製されたエキソンを含む1個もしくは複数の別個の遺伝子中に存在するが、選択的にスプライシングされた他のmRNAおよび/または複製されたエキソンを含む他の遺伝子中には存在しない場合には、選択的スプライシング経路および/または別個の遺伝子中の複製されたエキソンの発現を識別するのに有用である。あるいは、別個の遺伝子中の複製エキソンでは、異なる遺伝子からの各エキソンの配列相同性がかなり違う場合には、異なる遺伝子からの各エキソンを識別できるように異なるプローブを含むことが好ましい。

【０２４８】

上記プローブ・スキームのいずれも、同じプロファイリング・アレイ上で、かつ/または同じセットのプロファイリング・アレイ内の異なるアレイ上で組み合わせて、複数の遺伝子の発現プロファイルをより正確に決定することができることは当業者には明白である。異なるプローブ・スキームを、プロファイリングにおける異なるレベルの正確度に対して使用できることも当業者には明白である。例えば、各エキソンに対する小セットのプローブを含むプロファイリング・アレイまたはアレイ・セットを使用して、ある特定の条件下で関連遺伝子および/またはRNAスプライシング経路を明らかにすることもできる。次いで、目的エキソンに対するより大きなプローブ・セットを含むアレイまたはアレイ・セットを使用して、そのような特定の条件下でエキソン発現プロファイルをより正確に求める。異なるプローブ・スキームをより有利に使用することができる別のDNAアレイ戦略も包含される。

【０２４９】

本発明に使用するマイクロアレイは、目的薬物の作用に関連する、または目的とする生物学的経路における1個もしくは複数の遺伝子に対するエキソン・セットの結合部位(すなわち、プローブ)を有することが好ましい。上述したように、「遺伝子」は、5'非翻訳領域(「UTR」)、イントロン、エキソンおよび3'UTRを含むことができる、RNAポリメラーゼ

によって転写されたDNAの一部として特定される。ゲノム中の遺伝子数は、細胞もしくは生物によって発現されるmRNAの数から推測することができ、または特性が十分明らかなゲノム部分から補外することによって推測することができる。目的生物のゲノム配列が決定されると、DNA配列を分析してORF数を求め、mRNAコード領域を明らかにすることができる。例えば、酵母(サッカロマイセス・セレビジエ)のゲノム配列は完全に決定されており、99アミノ酸残基長よりも長い配列をコードする約6275個のORFを有することが報告されている。これらのORFの分析から、タンパク質産物をコードしている可能性が高い5,885個のORFが存在することが示されている(Goffeau等、1996、Science 274:546～567)。これに対して、ヒト・ゲノムは、約30,000～130,000個の遺伝子を含むと推定される(Crollius等、2000、Nature Genetics 25:235～238;Ewing等、2000、Nature Genetics 25:232～234を参照されたい)。ショウジョウバエ、線虫、植物、例えば、イネおよびシロイヌナズナ、および哺乳動物、例えば、マウスおよびヒトを含めて、ただしこれらだけに限定されない別の生物のゲノム配列も完了またはほぼ完了している。したがって、本発明の好ましい実施形態においては、生物のゲノム中のすべての既知のエキソンまたは予想されるエキソンに対する全プローブを含むアレイ・セットを提供する。非限定的な例として、本発明は、ヒト・ゲノム中の既知の各エキソンまたは予想される各エキソンに対する1個または2個のプローブを含むアレイ・セットを提供する。

【０２５０】

細胞のRNAに相補的なcDNAを作製し、適切なハイブリッド形成条件下でマイクロアレイにハイブリッド形成させるとき、任意の特定の遺伝子のエキソンに対応するアレイ中の部位へのハイブリッド形成レベルは、その遺伝子から転写されるエキソンを含むmRNAの細胞における支配率を反映していることを理解されたい。例えば、全細胞mRNAに相補的である(例えば、蛍光団で)検出可能に標識されたcDNAをマイクロアレイにハイブリッド形成させるとき、細胞中で転写されず、またはRNAスプライシング中に除去される遺伝子のエキソンに対応する(すなわち、遺伝子発現の産物に特異的に結合可能である)アレイ上の部位は、ほとんどまたはまったくシグナル(例えば、蛍光性シグナル)を示さず、エキソンを発現するコードされたmRNAが優勢な遺伝子のエキソンは比較的強いシグナルを示す。次いで、選択的スプライシングによって同じ遺伝子から産生される様々なmRNAの相対存在量を、遺伝子に対して、モニターされるエキソンのセット全体にわたるシグナル強度パターンから決定する。

【０２５１】

一実施形態においては、2つの異なる条件から得られる細胞標本のcDNAを、2色プロトコル(two-color protocol)によってマイクロアレイの結合部位にハイブリッド形成させる。薬物応答の場合には、1個の細胞標本を薬物に曝し、同じタイプのもう1個の細胞標本を薬物に曝さない。経路応答の場合には、1個の細胞を経路の乱れに曝し、同じタイプのもう1個の細胞を経路の乱れに曝さない。2個の細胞型の各々に由来するcDNAは、(例えば、Cy3およびCy5で)異なって標識され、そのため識別することができる。一実施形態においては、例えば、薬物で処理した(または経路の乱れに曝した)細胞から得られるcDNAを蛍光標識dNTPを用いて合成し、薬物に曝していない第2の細胞から得られるcDNAをローダミン標識dNTPを用いて合成する。2個のcDNAを混合し、マイクロアレイにハイブリッド形成させるときに、各cDNAセットの相対シグナル強度をアレイ上の各部位で測定し、特定のエキソンの存在量の相対差を検出する。

【０２５２】

上述した例では、薬物処理した(または経路を撹乱させた)細胞から得られるcDNAは、蛍光団が刺激されると緑色の蛍光を発し、未処理細胞から得られるcDNAは赤色の蛍光を発する。その結果、細胞中の特定の遺伝子が転写および/または転写後スプライシングされると、直接的でも間接的でも薬物療法の効果がないときには両方の細胞においてエキソンの発現パターンを識別できず、逆転写されると、赤色標識されたcDNAも緑色標識されたcDNAも等しく優勢である。マイクロアレイにハイブリッド形成させると、そのRNA種に対する結合部位は、両方の蛍光団に特徴的な波長を放出する。これに対して、薬物に曝す細胞を

、細胞中の特定の遺伝子の転写および/または転写後のスプライシングを直接的でも間接的でも変える薬物で処理すると、各エキソン結合部位に対する緑色と赤色の蛍光比によって表されるエキソンの発現パターンが変化する。薬物がmRNAの優勢を強めると、mRNA中で発現する各エキソンの比が増加するのに対し、薬物がmRNAの優勢を弱めると、mRNA中で発現する各エキソンの比が減少する。

【０２５３】

遺伝子発現の変化を明確にする2色蛍光標識および検出スキームの使用については、例えば、Shena等、1995、Quantitative monitoring of gene expression patterns with a complementary DNA microarray、Science 270:467～470にmRNAの検出と関連して記載されている。その全体を参照により本明細書に援用する。このスキームは、エキソンの標識および検出にも等しく適用可能である。2個の異なる蛍光団で標識されたcDNAを使用する利点は、2つの細胞状態にある各遺伝子アレイに対応するmRNAまたはエキソン発現レベルの直接比較および内部対照比較を行うことができ、実験条件(例えば、ハイブリッド形成条件)のわずかな違いによる変化が後続の分析に影響を及ぼさない点にある。しかし、単一細胞に由来するcDNAを使用し、例えば、薬物処理細胞または経路撹乱細胞および未処理細胞中の特定のエキソンの絶対量を比較することも可能であることを認識されたい。また、3色以上で標識することも本発明では企図される。本発明の一部の実施形態においては、少なくとも5、10、20または100種の様々な色素を標識に使用することができる。そのような標識によって、識別可能に標識されたcDNA集団を同じアレイに同時にハイブリッド形成させることができ、したがって測定することができ、3個以上の標本から得られるmRNA分子の発現レベルを比較することもできる。使用可能な色素としては、フルオレセインおよびその誘導体、ローダミンおよびその誘導体、テキサス・レッド、5'カルボキシフルオレセイン(「FMA」)、2',7'-ジメトキシ-4',5'-ジクロロ-6-カルボキシフルオレセイン(「JOE」)、N,N,N',N'-テトラメチル-6-カルボキシローダミン(「TAMRA」)、6'カルボキシ-X-ローダミン(「ROX」)、HEX、TET、IRD40およびIRD41、Cy3、Cy3.5およびCy5を含めて、ただしこれらだけに限定されないシアミン色素、BODIPY-FL、BODIPY-TR、BODIPY-TMR、BODIPY-630/650およびBODIPY-650/670を含めて、ただしこれらだけに限定されないBODIPY色素、ALEXA-488、ALEXA-532、ALEXA-546、ALEXA-568およびALEXA-594を含めて、ただしこれらだけに限定されないALEXA色素、ならびに当業者に既知の他の蛍光色素などがあるが、これらだけに限定されない。

【０２５４】

本発明の一部の実施形態においては、ハイブリッド形成データを、複数の異なるハイブリッド形成時間で測定して、ハイブリッド形成レベルが平衡に達するのを確認することができる。そのような実施形態においては、ハイブリッド形成レベルは、最も好ましくは、0から、標識ポリヌクレオチドによって結合ポリヌクレオチド(すなわち、プローブ)をサンプリングするのに必要な時間以上にわたるハイブリッド形成時間で測定され、その結果、混合物は平衡に近いか実質的に平衡に達し、二本鎖は拡散ではなく親和性および存在量に依存する濃度になる。しかし、ハイブリッド形成時間は十分短く、標識ポリヌクレオチドとプローブおよび/または表面との不可逆結合相互作用が起こらない、または少なくとも限定されることが好ましい。例えば、ポリヌクレオチド・アレイを用いて、断片化されたポリヌクレオチドの複雑な混合物を精査する実施形態においては、典型的なハイブリッド形成時間を約0～72時間とすることができる。別の実施形態の適切なハイブリッド形成時間は、使用する特定のポリヌクレオチド配列およびプローブに依存し、当業者が決定することができる(例えば、Sambrook等編、1989、Molecular Cloning:A Laboratory Manual、第2版、1～3巻、Cold Spring Harbor Laboratory、Cold Spring Harbor、New Yorkを参照されたい)。

【０２５５】

一実施形態においては、異なるハイブリッド形成時間におけるハイブリッド形成レベルを、異なる同一のマイクロアレイで別個に測定する。そのような各測定では、ハイブリッド形成レベルを測定するハイブリッド形成時間において、好ましくは室温で、高濃度から

中濃度の塩(例えば、0.5～3 M塩濃度)の水性溶液中、結合またはハイブリッド形成したポリヌクレオチドのすべてが保持され、すべての未結合ポリヌクレオチドが除去される条件下でマイクロアレイを簡単に洗浄する。次いで、各プローブ上でハイブリッド形成した残留ポリヌクレオチド分子の検出可能な標識を、用いた特定の標識方法に適切な方法によって測定する。次いで、得られたハイブリッド形成レベルを組み合わせてハイブリッド形成曲線を作成する。別の実施形態においては、単一のマイクロアレイを用いてハイブリッド形成レベルを実時間で測定する。この実施形態においては、マイクロアレイは中断することなく標本とハイブリッド形成し、このマイクロアレイを各ハイブリッド形成時間において非侵襲的方法で調べる。さらに別の実施形態においては、1つのアレイを使用し、短時間ハイブリッド形成し、洗浄し、ハイブリッド形成レベルを測定し、同じ標本に戻し、別の時間ハイブリッド形成し、洗浄し、再度測定して、ハイブリッド形成時間曲線を得ることができる。

【０２５６】

好ましくは、2つの異なるハイブリッド形成時間で少なくとも2つのハイブリッド形成レベルを測定し、あるハイブリッド形成時間における第1のハイブリッド形成レベルは交差ハイブリッド形成平衡の時間スケールに近く、第2のハイブリッド形成レベルは第1のハイブリッド形成時間よりも長いハイブリッド形成時間で測定される。交差ハイブリッド形成平衡の時間スケールは、特に、標本組成およびプローブ配列に依存し、当業者が決定することができる。好ましい実施形態においては、第1のハイブリッド形成レベルを1～10時間で測定し、第2のハイブリッド形成時間を第1のハイブリッド形成時間の約2、4、6、10、12、16、18、48または72倍で測定する。

【０２５７】

5.11.1.1. マイクロアレイ用プローブの調製

上述したように、エキソンなどの特定のポリヌクレオチド分子が本発明によって特異的にハイブリッド形成する「プローブ」は、相補的ポリヌクレオチド配列である。1個または複数のプローブを、各標的エキソンに対して選択することが好ましい。例えば、最低数のプローブをエキソンの検出に使用するときには、プローブは、通常、約40塩基長を超えるヌクレオチド配列を含む。あるいは、大きな冗長プローブ(redundant probe)・セットをエキソンに使用するときには、プローブは、通常、約40～60塩基のヌクレオチド配列を含む。プローブは、完全長のエキソンに相補的な配列を含むこともできる。エキソンの長さは、50塩基未満から200塩基を超える範囲とすることができる。したがって、エキソンよりも長いプローブ長を使用するときには、プローブ配列が標的エキソンを含む連続mRNA断片に相補的であるように、隣接する構成的にスプライスされたエキソン配列でエキソン配列を補うことが好ましい。これによって、エキソン・プロファイリング・アレイの各プローブ間のハイブリッド形成ストリンジェンシーを同等にすることができる。各プローブ配列は、その標的配列に相補的である配列に加えて、リンカー配列も含むことができることを理解されたい。

【０２５８】

プローブは、生物のゲノム中の各遺伝子の各エキソンの一部に対応するDNAまたはDNA「模倣物」(例えば、誘導体およびアナログ)を含むことができる。一実施形態においては、マイクロアレイのプローブは、相補的RNAまたはRNA模倣物である。DNA模倣物は、DNAと特異的ワトソン-クリック様ハイブリッド形成が可能なサブユニット、またはRNAと特異的ハイブリッド形成が可能なサブユニットで構成されたポリマーである。核酸は、塩基部分、糖部分またはリン酸エステル骨格において修飾することができる。例示的なDNA模倣物としては、例えば、ホスホロチオエートがある。DNAは、例えば、ゲノムDNA、(例えば、RT-PCRによる)cDNAまたはクローン配列からのエキソン・セグメントをポリメラーゼ連鎖反応(PCR)で増幅して得ることができる。PCRプライマーは、一義的な断片(すなわち、マイクロアレイ上の他の任意の断片と、10塩基を超える隣接同一配列を共有しない断片)の増幅をもたらすエキソンまたはcDNAの既知の配列に基づいて選択されることが好ましい。Oligoバージョン5.0(National Biosciences)などの当分野で周知のコンピュータ・プログラム

が、必要な特異性および最適な増幅特性を有するプライマーを設計するのに有用である。マイクロアレイ上の各プローブは、一般には20塩基～600塩基であり、通常は30～200塩基長である。PCR方法は当分野で周知であり、例えば、Innis等編、1990、PCR Protocols:A Guide to Methods and Applications、Academic Press Inc.、San Diego、CAに記載されている。制御されたロボット・システムが核酸を単離し増幅するのに有用であることは、当業者には明らかであろう。

【０２５９】

マイクロアレイのポリヌクレオチド・プローブを作製する別の好ましい手段は、例えば、N-ホスホネートまたはホスホロアミダイト化学を用いて、合成ポリヌクレオチドまたはオリゴヌクレオチドを合成するものである(Froehler等、1986、Nucleic Acid Res. 14:5399～5407;McBride等、1983、Tetrahedron Lett. 24:246～248)。合成配列は、一般に、約15～約600塩基長であり、より典型的には約20～約100塩基であり、最も好ましくは約40～約70塩基長である。一部の実施形態においては、合成核酸としては、イノシンなど、ただしこれらだけに限定されない非天然塩基などがある。上述したように、核酸アナログを、ハイブリッド形成の結合部位として使用することができる。適切な核酸アナログの例は、ペプチド核酸である(例えば、Egholm等、1993、Nature 363:566～568;米国特許第5,539,083号を参照されたい)。

【０２６０】

別の実施形態においては、ハイブリッド形成部位(すなわち、プローブ)は、遺伝子のプラスミド・クローンもしくはファージ・クローン、cDNA(例えば、発現配列タグ)、またはそれらの挿入断片から作製される(Nguyen等、1995、Genomics 29:207～209)。

【０２６１】

5.11.1.2. 固体表面への核酸の付着

あらかじめ形成されたポリヌクレオチド・プローブを担体上に置いてアレイを作製することができる。あるいは、ポリヌクレオチド・プローブを、担体上で直接合成してアレイを作製することができる。プローブは、例えば、ガラス、プラスチック(例えば、ポリプロピレン、ナイロン)、ポリアクリルアミド、ニトロセルロース、ゲル、他の多孔質材料または非多孔質材料などでできた固体担体または表面に付着する。

【０２６２】

核酸を表面に付着させる好ましい方法は、Schena等、1995、Science 270:467～470に概略記載されているようにガラス・プレート上に印刷することによるものである。この方法は、cDNAのマイクロアレイを調製するのに特に有用である(DeRisi等、1996、Nature Genetics 14:457～460;Shalon等、1996、Genome Res. 6:639～645;およびSchena等、1995、Proc. Natl. Acad. Sci. U.S.A. 93:10539～11286も参照されたい)。

【０２６３】

マイクロアレイを作製する第2の好ましい方法は、高密度ポリヌクレオチド・アレイを作製することによるものである。表面の規定位置における規定配列に相補的である数千のオリゴヌクレオチドを含むアレイを、フォトリソグラフィ合成技術を用いてin situで生成する技術(Fodor等、1991、Science 251:767～773;Pease等、1994、Proc. Natl. Acad. Sci. U.S.A. 91:5022～5026;Lockhart等、1996、Nature Biotechnology 14:1675;米国特許第5,578,832号;同5,556,752号;および同5,510,270号を参照されたい)、または規定のオリゴヌクレオチドを迅速に合成し付着させる他の方法(Blanchard等、Biosensors & Bioelectronics 11:687～690)が知られている。これらの方法を使用するときには、既知配列のオリゴヌクレオチド(例えば、60量体)は、スライド・ガラス誘導体(derivatized glass slide)などの表面上で直接合成される。作製されたアレイは、エキソン1個当たりいくつかのポリヌクレオチド分子が重複していてもよい。

【０２６４】

マイクロアレイを作製する別の方法、例えば、マスキング(MaskosおよびSouthern、1992、Nucl. Acids. Res. 20:1679～1684)を使用する方法もできる。原則的には、上述したように、あらゆるタイプのアレイ、例えば、ナイロン・ハイブリッド形成膜上のドット・

ブロット(Sambrook等、同上参照)を使用することができる。しかし、当業者によって認識されているように、ハイブリッド形成体積がより小さくなるので、極めて小さなアレイが好ましいことが多い。

【０２６５】

特に好ましい実施形態においては、本発明のマイクロアレイを、例えば、Blanchard、1998年9月24日に公開された国際公開第98/41531号;Blanchard等、1996、Biosensors and Bioelectronics 11:687～690;Blanchard、1998、Synthetic DNA Arrays in Genetic Engineering、20巻、J.K. Setlow編、Plenum Press、New York、111～123ページ;およびBlanchard、米国特許第6,028,189号に記載された方法およびシステムを用いて、オリゴヌクレオチド合成用インクジェット式印刷装置によって製造する。具体的には、このようなマイクロアレイ中のポリヌクレオチド・プローブを、アレイ中、例えば、スライド・ガラス上で、炭酸プロピレンなどの表面張力の高い溶媒の「微小液滴」中の個々のヌクレオチド塩基を連続的に付着させることによって合成することが好ましい。微小液滴は、体積が小さく(例えば、100 pL以下、より好ましくは、50 pL以下)、マイクロアレイ上で(例えば、疎水性ドメインによって)互いに分離されて、アレイ・エレメント(すなわち、様々なプローブ)の位置を規定する表面張力円形ウェルを形成する。ポリヌクレオチド・プローブは、通常、ポリヌクレオチドの3'末端で表面に共有結合する。あるいは、ポリヌクレオチド・プローブは、ポリヌクレオチドの5'末端で表面に共有結合することができる(例えば、Blanchard、1998、Synthetic DNA Arrays in Genetic Engineering、20巻、J.K. Setlow編、Plenum Press、New York、111～123ページを参照されたい)。

【０２６６】

5.11.1.3. 標的ポリヌクレオチド分子

本発明の方法および組成物によって分析することができる標的ポリヌクレオチドとしては、メッセンジャーRNA(mRNA)分子、リボソームRNA(rRNA)分子、cRNA分子(すなわち、インビボで転写されるcDNA分子から調製されるRNA分子)およびそれらの断片など、ただしこれらだけに決して限定されないRNA分子などがある。やはり本発明の方法および組成物によって分析することができる標的ポリヌクレオチドとしては、ゲノムDNA分子、cDNA分子などのDNA分子、オリゴヌクレオチド、EST、STSなどを含めたそれらの断片などがあるが、これらだけに限定されない。

【０２６７】

標的ポリヌクレオチドは、あらゆる出所のものとすることができる。例えば、標的ポリヌクレオチド分子は、生物から単離されるゲノムDNA分子またはゲノム外DNA分子、生物から単離されるmRNA分子などのRNA分子などの天然核酸分子とすることができる。あるいは、例えば、cDNA分子などのインビボまたはインビトロで酵素的に合成される核酸分子、PCRによって合成されるポリヌクレオチド分子、インビトロでの転写によって合成されるRNA分子などを含めたポリヌクレオチド分子を合成することができる。標的ポリヌクレオチド標本は、例えば、DNA分子、RNA分子、またはDNAとRNAのコポリマー分子を含むことができる。好ましい実施形態においては、本発明の標的ポリヌクレオチドは、特定の遺伝子または特定の遺伝子転写物(例えば、細胞中で発現される特定のmRNA配列、またはそのようなmRNA配列から誘導される特定のcDNA配列)に相当する。しかし、多数の実施形態においては、特にポリヌクレオチド分子が哺乳動物細胞から得られる実施形態においては、標的ポリヌクレオチドは、遺伝子転写物の特定の断片に対応するものとすることができる。例えば、標的ポリヌクレオチドは、同じ遺伝子の異なるエキソンに対応し、その結果、例えば、その遺伝子の異なるスプライス変異体を検出し、かつ/または分析することができる。

【０２６８】

好ましい実施形態においては、分析する標的ポリヌクレオチドを、細胞から抽出される核酸からインビトロで調製する。例えば、一実施形態においては、RNAを細胞(例えば、全細胞RNA、ポリ$(A)^+$メッセンジャーRNA、それらの一部)から抽出し、メッセンジャーRNAを全抽出RNAから精製する。全RNAおよびポリ$(A)^+$RNAを調製する方法は当分野で周知であり、一般に、例えば、Sambrook等、同上に記載されている。一実施形態においては、本発明

において対象とする様々なタイプの細胞をチオシアン酸グアニジウムで溶解後、CsCl遠心分離し、オリゴdT精製してRNAを抽出する(Chirgwin等、1979、Biochemistry 18:5294～5299)。別の実施形態においては、細胞をチオシアン酸グアニジウム溶解後、RNeasyカラム(Qiagen)で精製してRNAを抽出する。次いで、例えば、オリゴdTまたはランダム・プライマーを用いて、精製したmRNAからcDNAを合成する。好ましい実施形態においては、標的ポリヌクレオチドは、細胞から抽出された精製メッセンジャーRNAから調製されるcRNAである。本明細書で使用するcRNAは、もとのRNAに相補的であるRNAとして定義される。RNAポリメラーゼ・プロモーターに結合したプライマーを用いて、アンチセンスRNAの転写を誘導することができる方向にRNAから二本鎖cDNAを合成するプロセスによって、抽出したRNAを増幅させる。次いで、RNAポリメラーゼを用いて、二本鎖cDNAの2番目の鎖からアンチセンスRNAまたはcRNAを転写する(例えば、米国特許第5,891,636号、同5,716,785号;同5,545,522号および同6,132,997号を参照されたい。また、米国特許第6,271,002号、およびZiman他によって2000年11月28日に出願された米国仮出願第60/253,641号も参照されたい)。RNAポリメラーゼ・プロモーターまたはその相補配列を含むオリゴdTプライマー(米国特許第5,545,522号および同6,132,997号)またはランダム・プライマー(Ziman他によって2000年11月28日に出願された米国仮出願第60/253,641号)を使用することができる。標的ポリヌクレオチドは、細胞の元の核酸集団に代表的な短鎖および/または断片ポリヌクレオチド分子であることが好ましい。

【０２６９】

本発明の方法および組成物によって分析される標的ポリヌクレオチドは、検出可能に標識されていることが好ましい。例えば、cDNAを、例えば、ヌクレオチド・アナログで直接標識することができ、または、例えば、第1の鎖を鋳型として使用して第2の標識cDNA鎖を作製することによって間接的に標識することができる。あるいは、二本鎖cDNAを転写してcRNAとし、標識することができる。

【０２７０】

検出可能な標識は、例えば、ヌクレオチド・アナログを組み込むことによる蛍光標識であることが好ましい。本発明における使用に適切な他の標識は、ビオチン、イミノビオチン、抗原、補因子、ジニトロフェノール、リポ酸、オレフィン化合物、検出可能なポリペプチド、電子に富む分子、基質に作用して検出可能なシグナルを発生可能な酵素、放射性同位体などであるが、これらだけに限定されない。好ましい放射性同位体としては、^{32}P、^{35}S、^{14}C、^{15}N、^{125}Iなどがある。本発明に適切な蛍光性分子としては、フルオレセインおよびその誘導体、ローダミンおよびその誘導体、テキサス・レッド、5'カルボキシフルオレセイン(「FMA」)、2',7'-ジメトキシ-4',5'-ジクロロ-6-カルボキシフルオレセイン(「JOE」)、N,N,N',N'-テトラメチル-6-カルボキシローダミン(「TAMRA」)、6'カルボキシ-X-ローダミン(「ROX」)、HEX、TET、IRD40、IRD41などがあるが、これらだけに限定されない。本発明に適切な蛍光性分子としては、さらに、Cy3、Cy3.5およびCy5を含めて、ただしこれらだけに限定されないシアミン色素、BODIPY-FL、BODIPY-TR、BODIPY-TMR、BODIPY-630/650およびBODIPY-650/670を含めて、ただしこれらだけに限定されないBODIPY色素、ALEXA-488、ALEXA-532、ALEXA-546、ALEXA-568およびALEXA-594を含めて、ただしこれらだけに限定されないALEXA色素、ならびに当業者に既知の他の蛍光色素などがある。本発明に適切な、電子に富む指示薬分子としては、フェリチン、ヘモシアニン、コロイド状金などがあるが、これらだけに限定されない。あるいは、これほどは好ましくない実施形態においては、標的ポリヌクレオチドを、第1の群をこのポリヌクレオチドに特異的に複合化することによって標識することができる。指示薬分子に共有結合し、第1の群に対して親和性を有する第2の群を使用して、標的ポリヌクレオチドを間接的に検出することができる。そのような実施形態においては、第1の群として使用するのに適切な化合物は、ビオチン、イミノビオチンなどであるが、これらだけに限定されない。第2の群として使用するのに適切な化合物は、アビジン、ストレプトアビジンなどであるが、これらだけに限定されない。

【０２７１】

5.11.1.4. マイクロアレイとのハイブリッド形成

上述したように、本発明によって分析されるポリヌクレオチド分子(本明細書では「標的ポリヌクレオチド分子と称する)が、アレイ、好ましくはその相補DNAが存在する特異的アレイ部位の相補的ポリヌクレオチド配列に特異的に結合または特異的にハイブリッド形成するように、核酸ハイブリッド形成および洗浄条件を選択する。

【０２７２】

その上に位置する二本鎖プローブDNAを含むアレイは、DNAを一本鎖にする変性条件にかけてから、標的ポリヌクレオチド分子と接触させることが好ましい。一本鎖プローブDNA(例えば、合成オリゴデオキシリボ核酸)を含むアレイは、標的ポリヌクレオチド分子と接触させる前に、例えば、自己相補的配列のために形成されるヘアピンまたは2量体を除去するために変性する必要がある場合がある。

【０２７３】

最適ハイブリッド形成条件は、プローブと標的核酸の長さ(例えば、オリゴマーと200塩基よりも大きなポリヌクレオチド)およびタイプ(例えば、RNAまたはDNA)によって決まる。核酸に対する特異的(すなわち、ストリンジェントな)ハイブリッド形成条件の一般的パラメータは、Sambrook等、(同上)、およびAusubel等、1987、Current Protocols in Molecular Biology、Greene Publishing and Wiley-Interscience、New Yorkに記載されている。Schena等のcDNAマイクロアレイを使用するときには、典型的なハイブリッド形成条件は、5 X SSCと0.2%SDS中で65℃で4時間のハイブリッド形成と、その後の低ストリンジェンシー洗浄緩衝剤(1 X SSCと0.2%SDS)による25℃での洗浄と、その後のより高いストリンジェンシーの洗浄緩衝剤(0.1 X SSCと0.2%SDS)による25℃で10分間の洗浄である(Shena等、1996、Proc. Natl. Acad Sci. U.S.A. 93:10614)。有用なハイブリッド形成条件は、例えば、Tijessen、1993、Hybridization With Nucleic Acid Probes、Elsevier Science Publishers B.V.およびKricka、1992、Nonisotopic DNA Probe Techniques、Academic Press、San Diego、CAにも記載されている。

【０２７４】

本発明のスクリーニングおよび/または情報伝達チップとともに使用される特に好ましいハイブリッド形成条件は、プローブの平均融解温度またはその近くの温度(例えば、5℃以内、より好ましくは2℃以内)における、1M NaCl、50 mM MES緩衝剤(pH 6.5)、0.5%サルコシンナトリウムおよび30%ホルムアミド中でのハイブリッド形成などである。

【０２７５】

5.11.1.5. シグナル検出およびデータ解析

細胞のRNAに相補的である標的配列、例えば、cDNAまたはcRNAを作製し、適切なハイブリッド形成条件下でマイクロアレイにハイブリッド形成させるときには、任意の特定の遺伝子のエキソンに対応するアレイ中の部位に対するハイブリッド形成のレベルは、その遺伝子から転写されるエキソンを含むmRNAの細胞における支配率を反映していることを理解されたい。例えば、全細胞のmRNAに相補的である(例えば、蛍光団で)検出可能に標識されたcDNAをマイクロアレイにハイブリッド形成させるときには、細胞中で転写されない、またはRNAスプライシング中に除去される遺伝子のエキソンに対応する(すなわち、遺伝子発現産物に特異的に結合可能である)アレイ上の部位は、ほとんどまたはまったくシグナル(例えば、蛍光性シグナル)を持たず、エキソンを発現するコードされたmRNAが優勢である遺伝子のエキソンは比較的強いシグナルを有する。次いで、選択的スプライシングによって同じ遺伝子から産生される様々なmRNAの相対存在量を、遺伝子に対して、モニターされるエキソンのセット全体にわたるシグナル強度パターンから決定する。

【０２７６】

好ましい実施形態においては、異なる2個の細胞からの標的配列、例えば、cDNAまたはcRNAを、マイクロアレイの結合部位にハイブリッド形成させる。薬物応答の場合には、1個の細胞標本を薬物に曝し、同じタイプのもう1個の細胞標本を薬物に曝さない。経路応答の場合には、1個の細胞を経路の乱れに曝し、同じタイプのもう1個の細胞を経路の乱れに曝さない。2つの細胞型の各々に由来するcDNAまたはcRNAは異なって標識され、そのため

識別することができる。一実施形態においては、例えば、薬物で処理した(または経路の乱れに曝した)細胞から得られるcDNAを蛍光標識dNTPを用いて合成し、薬物に曝されていない第2の細胞から得られるcDNAをローダミン標識dNTPを用いて合成する。2個のcDNAを混合し、マイクロアレイにハイブリッド形成させるときには、各cDNAセットからの相対シグナル強度をアレイ上の各部位で測定し、特定のエキソンの存在量の相対差を検出する。

【０２７７】

上述した例では、薬物処理した(または経路を撹乱させた)細胞から得られるcDNAは、蛍光団が刺激されると緑色の蛍光を発し、未処理細胞から得られるcDNAは赤色の蛍光を発する。その結果、細胞中の特定の遺伝子が転写および/または転写後スプライシングされると、直接的でも間接的でも薬物療法の効果がないときには両方の細胞においてエキソンの発現パターンを識別できず、逆転写されると、赤色標識されたcDNAも緑色標識されたcDNAも等しく優勢である。マイクロアレイにハイブリッド形成させると、そのRNA種に対する結合部位は、両方の蛍光団に特徴的な波長を放出する。これに対して、薬物に曝された細胞を、細胞中の特定の遺伝子の転写および/または転写後スプライシングを直接的でも間接的でも変える薬物で処理すると、各エキソン結合部位に対する緑色と赤色の蛍光比によって表されるエキソンの発現パターンが変化する。薬物がmRNAの優勢を強めると、mRNA中で発現する各エキソンの比が増加するのに対し、薬物がmRNAの優勢を弱めると、mRNA中で発現する各エキソンの比が減少する。

【０２７８】

遺伝子発現の変化を明確にする2色蛍光標識および検出方式の使用については、例えば、Shena等、1995、Quantitative monitoring of gene expression patterns with a complementary DNA microarray、Science 270:467～470にmRNAの検出と関連して記載されている。この全体を参照により本明細書に援用する。この方式は、エキソンの標識および検出に等しく適用可能である。異なる2個の蛍光団で標識された標的配列、例えば、cDNAまたはcRNAを使用する利点は、2つの細胞状態にある各遺伝子アレイに対応するmRNAまたはエキソン発現レベルの直接比較および内部対照比較を行うことができ、実験条件(例えば、ハイブリッド形成条件)のわずかな違いによる変化が後続の分析に影響を及ぼさない点にある。しかし、単一細胞のcDNAを使用し、例えば、薬物処理細胞または経路撹乱細胞および未処理細胞中の特定のエキソンの絶対量を比較することも可能であることを認識されたい。

【０２７９】

蛍光標識プローブを使用するときには、転写物アレイの各部位における蛍光放出を、走査型共焦点レーザー顕微鏡によって検出できることが好ましい。一実施形態においては、適切な励起ラインを用いた別々の走査を、使用する2個の蛍光団の各々に対して実施する。あるいは、2個の蛍光団に特有の波長で同時に標本を照射し、2個の蛍光団からの発光を同時に分析することができるレーザーを使用することができる(Shalon等、1996、Genome Res. 6:639～645を参照されたい)。好ましい実施形態においては、コンピュータ制御されたX-Yステージおよび顕微鏡対物レンズを備えたレーザー蛍光スキャナーでアレイを走査する。2個の蛍光団をマルチライン混合ガス・レーザーを用いて連続して励起し、放出光を波長によって分割し、2個の光電子増倍管で検出する。このような蛍光レーザー走査装置は、例えば、Schena等、1996、Genome Res. 6:639～645に記載されている。あるいは、Ferguson等、1996、Nature Biotech. 14:1681～1684に記載された光ファイバー束を使用して、mRNAの存在レベルを多数の部位において同時にモニターすることができる。

【０２８０】

シグナルを記録し、好ましい実施形態においては、コンピュータによって、例えば、12ビットのアナログ・デジタル変換ボードを用いて解析する。一実施形態においては、グラフィックス・プログラム(例えば、Hijaak Graphics Suite)を用いて走査イメージから斑を除去し、次いで、各部位において各波長で平均ハイブリッド形成のスプレッドシートを作成するイメージ・グリッディング・プログラムを用いて解析する。必要に応じて、2個の蛍光体に対するチャネル間の「クロストーク」(または重複)に対して実験的に決定され

た補正を行うことができる。転写物アレイ上の特定のハイブリッド形成部位に対して、2個の蛍光団の発光比を計算することができる。この比は、同族遺伝子の絶対的発現レベルには無関係であるが、薬物投与、遺伝子欠失、または他の任意の試験済みの現象によって発現がかなり調整される遺伝子に有用である。

【０２８１】

本発明による方法によれば、2個の細胞または細胞系中のmRNAおよび/またはmRNA中で発現されるエキソンの相対存在量は、撹乱された(すなわち、2つの検定mRNA源で存在量が異なる)ものとして、または撹乱されない(すなわち、相対存在量が同じ)ものとして記録される。少なくとも約25%(すなわち、一方の出所のRNAが、他方の出所のRNAよりも25%多い)、より一般的には約50%、さらにより一般的には約2倍(すなわち、2倍の量)、3倍(3倍の量)、または5倍(5倍の量)の本明細書で使用する2つのRNA源間の差が乱れとして記録される。本発明の検出方法は、約1.5倍～約3倍の差を信頼性よく検出できる。

【０２８２】

しかし、これは、2個の細胞または2個の細胞系におけるmRNAおよび/またはmRNA中で発現されるエキソンの存在量の相対差の程度を明らかにするのにも有利である。これは、上述したように、示差的標識に使用される2個の蛍光団の発光比を計算することによって、または当業者には容易に明らかな類似の方法によって実施することができる。

【０２８３】

5.11.2. 別の転写状態測定方法

細胞の転写状態を、当分野で既知の他の遺伝子発現技術によって測定することができる。2種類の制限酵素による消化を整相プライマー(phasing primer)と組み合わせる方法(例えば、1992年9月24日にZabeau他によって出願された欧州特許第0 534858号A1を参照されたい)、規定mRNA末端に最も近い部位を有する制限断片を選択する方法(例えば、Prashar等、1996、Proc. Natl. Acad. Sci. USA 93:659～663を参照されたい)など一部の技術は、電気泳動分析に対してあまり複雑でない制限断片プールを生じる。他の方法は、複数のcDNAの各々における十分な塩基(例えば、20～50塩基)の配列を決定して各cDNAを特定することによって、または規定mRNA末端に対して既知の位置において生成する短いタグ(例えば、9～10塩基)の配列を決定することなどによって、cDNAプールを統計的にサンプリングするものである(例えば、Velculescu、1995、Science 270:484～487を参照されたい)。

【０２８４】

5.12. 翻訳状態の測定

本発明の様々な実施形態においては、翻訳状態、活性状態または混合態様などの転写状態以外の生物学的状態の諸態様を測定することができる。したがって、このような実施形態においては、細胞構成成分データ44(図1)は、翻訳状態測定値、さらにはタンパク質発現測定値を含むことができる。実際、一部の実施形態においては、遺伝子発現に基づく遺伝子発現相互作用地図を使用するのではなく、タンパク質発現地図に基づくタンパク質発現相互作用地図を使用する。転写状態以外の生物学的状態の諸態様がこのセクションと以下のセクションに記載されている実施形態の詳細。

【０２８５】

翻訳状態の測定を、いくつかの方法によって実施することができる。例えば、タンパク質のゲノム全体(すなわち、「プロテオーム」、Goffeau等、同上)のモニタリングを、細胞ゲノムによってコードされる複数のタンパク質種に特異的である固定化抗体、好ましくは固定化モノクローナル抗体を結合部位が含むマイクロアレイを構築することによって実施することができる。抗体は、コードされたタンパク質の実質的部分、または少なくとも目的薬物の作用に関連するタンパク質に対して存在することが好ましい。モノクローナル抗体を作製する方法は周知である(例えば、その全体を援用する、HarlowおよびLane、1988、Antibodies:A Laboratory Manual、Cold Spring Harbor、New Yorkを参照されたい)。好ましい実施形態においては、モノクローナル抗体は、細胞のゲノム配列に基づいて設計された合成ペプチド断片に対して産生される。そのような抗体アレイを用いて、細胞から得られるタンパク質をアレイに接触させ、それらの結合を当分野で既知のアッセイによっ

て分析する。

【０２８６】

あるいは、タンパク質を、二次元ゲル電気泳動システムによって分離させることができる。二次元ゲル電気泳動は、当業者によく知られており、一般に、一次元に沿った等電点電気泳動と、その後の二次元に沿ったSDS-PAGE電気泳動を含む。例えば、Hames等、1990、Gel Electrophoresis of Proteins:A Practical Approach、IRL Press、New York;Shevchenko等、1996、Proc. Natl. Acad. Sci. USA 93:1440～1445;Sagliocco等、1996、Yeast 12:1519～1533;Lander、1996、Science 274:536～539を参照されたい。得られた電気泳動図は、質量分析法、ウエスタン・ブロット法、ポリクローナル抗体およびモノクローナル抗体を用いた免疫ブロット分析、内部およびN末端の微量配列分析を含めて、多数の技術によって分析することができる。これらの技術を用いて、薬物に曝された細胞中(例えば、酵母中)、または例えば、特定の遺伝子の欠失または過剰発現によって改変された細胞中を含めて、所与の生理学的条件下で産生される全タンパク質の実質的部分を同定することができる。

【０２８７】

5.13. 生物学的状態の別の態様の測定

本発明の方法は、モニター可能なあらゆる細胞構成成分に適用可能である。例えば、タンパク質の活性が測定可能である場合には、本発明の実施形態をそのような測定に使用することができる。活性の測定は、分析する特定の活性に適切な任意の機能的手段、生化学的手段または物理的手段によって実施することができる。活性が化学変換を伴う場合には、細胞のタンパク質を天然の基質に接触させ、変換速度を測定することができる。活性が、多量体ユニット中の会合、例えばDNAとの活性DNA結合複合体の会合を含む場合には、会合タンパク質の量、または転写されたmRNAの量などの会合の二次的結果を測定することができる。また、例えば、細胞周期制御におけるように、機能上の活性のみが既知の場合には、その機能の働きを観察することができる。タンパク質活性の変化は、既知のものでも測定されたものでも、上述した本発明の方法によって分析される応答データを形成する。

【０２８８】

本発明の一部の実施形態においては、細胞構成成分の測定は、細胞の表現型技術から派生する。そのような細胞の表現型技術の1つは、汎用レポーターとして細胞呼吸を使用する。一実施形態においては、各ウェルが独特の化学反応性を含む96ウェルのマイクロタイター・プレートが提供される。独特な各化学反応性は、特定の表現型を検定するように設計されている。目的生物46(図1)の細胞をピペットで各ウェルに移す。細胞が適切な表現型を示す場合には、細胞は呼吸し、テトラゾリウム色素を活発に還元して濃い紫色を呈する。表現型が弱いと色が薄くなる。無色は、細胞が特定の表現型を持たないことを意味する。色の変化は、毎時何回も記録することができる。1回のインキュベーション中、5,000個を超える表現型を検定することができる。例えば、Bochner等、2001、Genome Research 11、1246～55を参照されたい。

【０２８９】

本発明の一部の実施形態においては、測定される細胞構成成分(遺伝子発現データ44)は代謝産物である。代謝産物としては、アミノ酸、金属、可溶性糖、糖リン酸、複合糖質などがあるが、これらだけに限定されない。このような代謝産物は、例えば、熱分解質量分析法(Irwin、1982、Analytical Pyrolysis:A Comprehensive Guide、Marcel Dekker、New York;Meuzelaar 等、1982、Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials、Elsevier、Amsterdam)、フーリエ変換赤外分光法(Griffithsおよびde Haseth、1986、Fourier transform infrared spectrometry、John Wiley、New York;Helm等、1991、J. Gen. Microbiol. 137、69～79;Naumann等、1991、Nature 351、81～82;Naumann等、1991、Modern techniques for rapid microbiological analysis、43～96、Nelson, W.H.編、VCH Publishers、New York)、ラマン分光法、ガスクロマトグラフィー-質量分析法(GC-MS)(Fiehn等、2000、Nature Biotechnology 18、1157～1161、キャピラリー電気泳動法(CE)/MS、高圧液体クロマトグラフィー/質量分析法(HPLC/MS)、ならびに液体クロマトグ

ラフィー(LC)-エレクトロスプレー、cap-LC-タンデム-エレクトロスプレー質量分析法などの方法によって細胞全体のレベルで測定することができる。このような方法は、密接に関係する各標本を識別するために、人工ニューラル・ネットワークおよび遺伝プログラミングを利用する既成の計量化学法と組み合わせることができる。

【０２９０】

5.14. 例示的な疾患

ヒトにおける複雑性疾患の例としては、喘息、血管拡張性失調症(JaspersおよびBootsma、1982、Proc. Natl. Acad. Sci. U.S.A. 79:2641)、双極性障害、一般的な癌、一般的な遅発性アルツハイマー病、糖尿病、心疾患、遺伝性早期発症型アルツハイマー病(George-Hyslop等、1990、Nature 347:194)、遺伝性非腺腫性大腸癌、高血圧、感染、若年成人発症型糖尿病(Barbosa等、1976、Diabete Metab. 2:160)、真性糖尿病、片頭痛、非アルコール性脂肪肝(NAFL)(Younossi等、2002、Hepatology 35、746～752)、非アルコール性脂肪性肝炎(NASH)(James & Day、1998、J. Hepatol. 29:495～501)、インシュリン非依存性糖尿病、肥満、多発性嚢胞腎(Reeders等、1987、Human Genetics 76:348)、乾癬、精神分裂病、脂肪性肝炎、色素性乾皮症(De Weerd-Kastelein、Nat. New Biol. 238:80)などがある。遺伝的異質性は、染色体領域を疾患と同時分離することができる家族もあれば、同時分離することができない家族もあるので遺伝マッピングが妨げられる。

【０２９１】

5.15. 教師付き分類方法

多重線形回帰(MLR)、部分最小二乗回帰(PLS)、主構成成分回帰(PCR)などの線形回帰方法を含めて、様々な方法を使用して処理ステップ106(図1)に従って教師付き分類を実施することができる。このような方法は、例えば、(Brereton、1992、Multivariate Pattern Recognition in Chemometrics、Elsievier、Amsterdam;Brown等、1992、Chemometrics. Anal. Chem. 64、22R-49R;MartensおよびNaes、1989、Multivariate Calibration、John Wiley & Sons、New York;およびMeloun等、1992、Chemometrics for Analytical Chemistry 1巻、PC- aided Statistical Data Analysis、Ellis Horwood、Chichester、UK(1992)に記載されている。また、これらの技術の非線形版を処理ステップ106(図1)に使用することもできる。例えば、Frank等、1990、Chemom. Intell. Lab. Sys. 8:109～119;Hoskuldsson、1992、J. Chemom. 6:307～334;Kvalheim等、1985、Anal. Chem. 57:2858～2864;Wold、1992、Chemom. Intell. Lab. Sys. 14:71～84;およびWythoff、1993、Chemom. Intell. Lab. Sys. 20:129～148を参照されたい。処理ステップ106(図1)に使用することができる関係手法は人工ニューラル・ネットワーク(ANN)である。

【０２９２】

教師あり学習の目標は、入力を標的と正確に関連させるモデルまたはマッピングを見つけることである。したがって、これらの教師あり学習技術の基本的考え方は、検討すべき最低4個のデータ・セットがあることである。「訓練データ」は、(i)sがオブジェクトの数であり、nが変数の数であるs行n列の行列と、(ii)やはりs行と一般に1列または2列からなる第2の行列からなる。ここで、列は、求めようとする変数であり、訓練セットではある既存の「ベンチマーク」方法によって実際に決定される。この変数は、(i)の同じ行のパターンと組み合わせられる。「検定データ」は、上記(i)および(ii)に対応するやはり2個の行列(iii)および(iv)からなるが、検定セットは異なるオブジェクトを含む。その名前が示すように、この第2の対はシステムの正確度を検定するために使用される。あるいは、これらを使用してモデルをクロス確認することができる。すなわち、訓練セット(i、ii)を用いてモデルを構築した後に、検定データ(iii)(これらは、新しいスペクトルとすることができる)を結果のモデル予測を得るように較正モデルを「通過」させる。次いで、これらを既知の予想される応答(iv)と比較することができる。他のすべてのデータ解析技術と同様に、これらの教師あり学習方法は、選択の誤った初期データに対する感度の影響を受ける。例えば、ZupanおよびGasteiger、1993、Neural Networks for Chemists:An Introduction、VCH Verlagsgeesellschaft、Weinheimを参照されたい。したがって、訓練セットの模範は注意深く選択されなければならない。

【０２９３】
5.16. クラス予測変数を特定する例示的方法
このセクションは、目的とする所与の複合形質のクラス予測変数262(図2)を、本発明の一実施形態によって見つける方法を説明する。クラス予測変数を使用して、教師付き分類体系により集団Pを複数の亜集団に分割することが容易になる。このセクションに開示する技術の例を、2002年5月14日に出願された米国仮出願第60/380,710号「Diagnosis and Prognosis of Breast Cancer Patients」、代理人整理番号9301-175-888に見ることができる。このセクションで使用する「マーカー」は、その発現もしくはレベルがある病態間で変化する遺伝子全体、またはその遺伝子に由来するESTを意味する。ある病態と遺伝子の発現とに相関がある場合には、その遺伝子はその病態のマーカーである。

【０２９４】
このセクションで説明する技術は、乳癌のクラス予測変数262を特定するマーカー・セットを使用する。当業者は、これらの技術を使用して、別の複合形質のクラス予測変数262を見つけることができることを理解されたい。一般に、クラス予測変数262は、約25,000個のヒト・マーカーのうち複合形質に関連する諸条件または指標(例えば、表現型)と相関がある発現パターンを有するものを求めることによって特定される。

【０２９５】
一実施形態においては、クラス予測変数262を特定する技術は以下のとおりである。標的ポリヌクレオチドを抽出し標識した後に、標本X中の全マーカーの発現を標準または対照中の全マーカーの発現と比較する。一実施形態においては、標準または対照は、正常個体(例えば、乳癌に罹患していない個体)から得られた標本に由来する標的ポリヌクレオチド分子を含む。好ましい実施形態においては、標準または対照は標的ポリヌクレオチド分子のプールである。このプールは、いくつかの正常個体から収集された標本に由来してもよい。好ましい実施形態においては、このプールは、散発性腫瘍のいくつかの個体から採取された標本を含む。別の好ましい実施形態においては、このプールは、腫瘍標本から得られるマーカー由来核酸のプール中の各マーカーに由来する核酸のレベルに近似するように設計された人工核酸集団を含む。さらに別の実施形態においては、このプールは、正常細胞系もしくは細胞系標本、または乳癌細胞系もしくは細胞系標本に由来する。

【０２９６】
比較は、当分野で既知のあらゆる手段によって実施することができる。例えば、様々なマーカーの発現レベルを、マーカーに由来する標的ポリヌクレオチド分子(例えば、RNAまたはcDNA)をアガロースまたはポリアクリルアミド・ゲル中で分離し、続いてマーカー特異的オリゴヌクレオチド・プローブとハイブリッド形成させることによって評価することができる。あるいは、標的ポリヌクレオチド分子を標識し、続いて配列決定ゲル上で分離させることによって比較することができる。患者のポリヌクレオチドと対照または標準のポリヌクレオチドが隣接レーンになるようにポリヌクレオチド標本をゲル上に配置する。目視または濃度計によって発現レベルを比較する。好ましい実施形態においては、マイクロアレイにハイブリッド形成させることによって、全マーカーの発現を同時に評価する。各手法においては、ある判定基準を満たすマーカーを乳癌に関連するものとみなす。

【０２９７】
マーカーは、標本中の発現と標準条件または対照条件との有意差に基づいて選択される。選択は、患者標本中のマーカーの有効な上方制御または下方制御に基づいて行うことができる。選択は、マーカーの発現と条件または指標との相関の統計的有意性(すなわち、p値)を計算することによっても行うことができる。好ましくは、両方の選択判定基準を使用する。したがって、本発明の一実施形態においては、乳癌関連マーカーは、マーカーが標準の2倍を超える発現変化(増加または減少)を示し、かつ乳癌の存在とマーカー発現変化の相関に対するp値が0.01以下である(すなわち、統計的に有意である)場合に選択される。

【０２９８】
次いで、特定された乳癌関連マーカーの発現を使用して、腫瘍を臨床タイプに分化させ

ることができるマーカーを特定する。いくつかの腫瘍標本を用いる特定の実施形態においては、臨床カテゴリーまたは臨床パラメータと、個々の遺伝子に対する全標本にわたる発現比の線形変換、対数変換または任意の変換との相関係数を計算することによってマーカーを特定する。具体的には、相関係数を

【数１８】

$$\rho = (\vec{c} \bullet \vec{r}) / (\|\vec{c}\| \cdot \|\vec{r}\|$$

【０２９９】

として計算する。ここで、

【数１９】

$$\vec{c}$$

【０３００】

は臨床パラメータまたは臨床カテゴリーであり、

【数２０】

$$\vec{r}$$

【０３０１】

は標本と対照の発現比の線形変換、対数変換または任意の変換である。相関係数がカットオフを超えるマーカーを、特定の臨床タイプに特異的な乳癌関連マーカーとみなす。そのようなカットオフまたはしきい値は、モンテ・カルロ・シミュレーションによって得られる特徴的遺伝子のある種の有意性に対応する。しきい値は、使用する標本数によって決まり、

【数２１】

$$3 \times 1 / \sqrt{n-3}$$

【０３０２】

として計算することができる。ここで、

【数２２】

$$\sqrt{n-3}$$

【０３０３】

は分布幅であり、nは標本数である。特定の実施形態においては、マーカーは、相関係数が約0.3を超え、または約-0.3未満である場合に選択される。

【０３０４】

次に、相関の有意性を計算する。この有意性は、このような有意性を計算する任意の統計手段によって計算することができる。具体的な例においては、1組の相関データを、特定のマーカーの発現差と臨床カテゴリーの関連性を無作為化するモンテ・カルロ法を用いて作成する。相関係数を計算して判定基準を満たすマーカーの頻度分布を、モンテ・カルロ法によって作成されるデータの判定基準を満たすマーカー数と比較する。モンテ・カルロ法において判定基準を満たすマーカーの頻度分布を使用して、臨床データとの相関によって選択されたマーカー数が有意であるかどうかを判定する。

【０３０５】

マーカー・セットを特定した後に、マーカーを識別の有意性順に順序付けすることができる。順序付ける1つの手段は、マーカーの遺伝子発現変化と、識別される特定の病態との相関の大きさによるものである。別の好ましい手段は統計量を使用する。特定の実施形態においては、この量はフィッシャー様統計データ、すなわち

【数２３】

$$t = \frac{\left(\langle x_1 \rangle - \langle x_2 \rangle\right)}{\sqrt{\left[\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)\right]/(n_1 + n_2 - 1)/(1/n_1 + 1/n_2)}}$$

【０３０６】

である。式中、

【数２４】

$$\langle x_i \rangle$$

【０３０７】

は第1の診断群(例えば、ER(-)内の転写物発現測定値の対数比の誤差加重平均(error-weighted average)であり、

【数２５】

$$\langle x_2 \rangle$$

【０３０８】

は第2の関係する診断群(例えば、ER(+))内の対数比の誤差加重平均であり、σ_1はER(-)群内の対数比の分散であり、n_1は有効な対数比測定値が利用可能な標本数である。σ_2は第2の診断群(例えば、ER(+))内の対数比の分散であり、およびn_2は有効な対数比測定値が利用可能な標本数である。t値は、2つの手段間の分散補正された差である。

【０３０９】

順序付けられたマーカー・セットを使用して、識別に使用したセット中のマーカー数を最適化することができる。これは、一般に以下の「1個を残す(leave one out)」方法で実施される。第1の分析においては、サブセット、例えば順序付けられたリストの上位から5個のマーカーを使用して鋳型を作成する。ここで、X個の標本のうちX-1個を使用して鋳型を作成し、残りの標本の状態を予測する。X標本のすべてが1回予測されるまで全標本についてこのプロセスを繰り返す。第2の分析においては、追加のマーカー、例えば5個を追加して、鋳型を10個のマーカーで作成し、残りの標本の結果を予測する。このプロセスを、全マーカー・セットを使用して鋳型を作成するまで繰り返す。各分析について、1型誤り(偽陰性)および2型誤り(偽陽性)をカウントする。マーカーの最適数は、1型誤り率、または2型誤り率、または好ましくは1型誤り率と2型誤り率の合計が最低になる数である。

【０３１０】

予後マーカーの場合には、マーカー・セットの検証は、追加の統計データ、生存モデルによって実施することができる。この統計データは、最初の診断からの時間の関数として腫瘍遠隔転移の確率を生成するものである。ワイブル、正規、対数正規、対数ロジスティック、対数指数、または対数レイリー(12章「Life Testing」、S-PLUS 2000 GUIDE TO STATISTICS、2巻、368ページ、2000)を含めて、いくつかのモデルを使用することができる。「正規」モデルの場合には、時間tにおける遠隔転移確率Pは、

$$P = \alpha \times \exp(-t^2/\tau^2)$$

として計算される。ここで、αは一定で1に等しく、τは適合パラメータであり「予想寿命」の尺度となる。

【０３１１】

上記方法、特に上記統計方法は、乳癌関連マーカーの特定だけに限定されず、複合形質を含めて任意の表現型のクラス予測変数(例えば、1組のマーカー遺伝子)を特定するのにも使用できることは当業者には明らかなはずである。表現型は、例えば、癌などの疾患の有無であっても、その癌に関連する特定の臨床状態の有無であってもよい。疾患に関連しては、表現型は、生存時間、疾患状態の遠隔転移確率、治療療法または予防療法に対する特定の応答の尤度などの予後とすることができる。表現型は癌または疾患である必要はなく、健康な個体に関連する公称特性とすることができる。

【０３１２】

5.17. 特定の遺伝子座における遺伝的サインを強化する遺伝子発現パターンの使用

複合形質は、多数の遺伝子(すなわち、多遺伝子である)および複雑な環境相互作用を潜在的に含む。雑種強勢の研究に使用されるものなどの実験計画は、植物の2個の近交系に注意を向けることで遺伝的異質性の程度を最小限に抑えようとする。また、制御された類似の条件下で植物を成長させることによって、このような研究設計における環境変化を抑える。しかし、対照実験においては、ある種の表現型は遺伝的異質性の観点から依然として複雑である。これはフィールド・データ形質(field data trait)において証明されるだけでなく、遺伝子発現形質においても見られ、一部の遺伝子は複数のQTLを有し、遺伝子発現が2個以上の遺伝子の制御下にあることを示している。

【０３１３】

多面発現的経路効果を分析するために、所与の集団をより均質な部分群に細分する手段としての、主要な目的形質(例えば、雑種強勢、コレステロール・レベル、体重など)に関連する細胞構成成分存在パターン(例えば、遺伝子発現パターン)の使用を、図7を参照して説明する。図7は、亜集団について量的遺伝解析を実施するために、遺伝学と細胞構成成分存在データを組み合わせて集団を細分することができる方法の単なる一例であることを理解されたい。

【０３１４】

例示的プロセスはステップ702から始まる。このステップでは、細胞構成成分測定、プロテオメトリクス(proteometrics)、細胞表現型決定、様々な表現型の分析などの技術を用いて、分離集団S中の各生物から細胞構成成分データを収集する。例えば、上記セクション5.11、5.12および5.13を参照されたい。分離集団Sは、F_2植物または2個の近交系に由来するマウスから大きなヒト系統にまで及ぶ。ステップ702においては、特定の定量可能な表現型(例えば、複合形質)に関して集団の独立した両極端を特定する。一実施形態においては、生物は、その生物が示す特定の表現型の大きさが、検定集団(例えば、複数の生物S)内の生物の少なくとも70パーセント、75パーセント、80パーセント、85パーセントまたは90パーセントが示す特定の表現型の大きさよりも大きいときに、特定の表現型(例えば、複合形質)に関して独立した極端な群内にある。独立した両極端を特定した後に、極端な表現型群(独立した両極端)を妥当な正確度で識別することが可能な存在量の全細胞構成成分(例えば、遺伝子転写物)を特定する(ステップ706)。一部の実施形態においては、2個の独立した極端な表現型群が存在する。別の実施形態においては、2個を超える独立した極端表現型群が存在する。独立した極端表現型群を識別することができる細胞構成成分セットをこの実施形態では細胞構成成分セットCと称する。t検定などの多数のタイプの統計解析を使用してセットG中の細胞構成成分を特定することができる。

【０３１５】

ステップ708においては、主要な目的形質のQTLを、上記セクション5.2に記載するものなどの標準連鎖解析によって特定する。すなわち、集団Sの家系データ、目的形質の表現型データ、および検討中の種の遺伝マーカー地図を使用して、検討中の形質と関連するQTLを特定する。家系情報が利用不可能な実施形態においては、関連解析を使用して目的形質と関連する遺伝子座を特定することができる。関連解析を上記セクション5.4に記載す

る。図7は、単一のQTL(または関連解析の場合は遺伝子座)がステップ708において特定される例に単純化される。実際に、あらゆる数のQTLおよび/または遺伝子座を処理ステップ708で特定することができ、処理ステップ708で特定された各QTLおよび/または遺伝子座、あるいはそのようなQTLおよび/または遺伝子座の群に対して、処理ステップ710～718を繰り返すことができることを理解されたい。また、図708の処理ステップ710で特定した単一のQTL(または遺伝子座)を、後続の処理ステップで特定される別のQTL(または遺伝子座)すべてからそのQTL(または遺伝子座)を区別するために「所定のQTL」と呼ぶ。

【０３１６】

ステップ710においては、細胞構成成分セットC中の各細胞構成成分を用いて量的遺伝解析を実施する。各解析では、細胞構成成分セットCの中から選択された細胞構成成分の発現レベルは表現型形質として役立つ。各解析を、上記セクション5.3に記載する量的遺伝解析に従って実施する。集団S中の所与の細胞構成成分Cの存在データ(例えば、発現データ)を使用する各量的遺伝解析によって、その細胞構成成分に関連するQTL(遺伝子座)を特定する。ステップ712においては、このデータを使用して特徴的セットG中に残る細胞構成成分を選択する。一実施形態においては、所定のQTLと連鎖しているQTL(遺伝子座)、または所定のQTLと実際に重複しているQTL(遺伝子座)を有する細胞構成成分Cのみがセット G中に残留することができる。所定のQTLと連鎖しているQTLを持たず、所定のQTLと重複しているQTLを持たない細胞構成成分は棄却される。理解しやすいように、ステップ712で精緻化された細胞構成成分セットを「DG」と称する。

【０３１７】

ステップ714は、セットDG中の細胞構成成分数を増加させるのに使用することができる任意選択のステップである。この任意選択ステップでは、検定集団全体にわたって、調べる生物中のいくつかの細胞構成成分の存在パターンを、セットDG中の任意の細胞構成成分の存在パターンと比較する。集団S全体にわたってセットDG中の細胞構成成分の存在パターンと高い相関がある存在パターンを有する細胞構成成分をセットDGに加える。このタイプの相関を計算することができる方法についてのさらに詳細な情報は、2000年7月6日付け国際公開第00/39338号にある。

【０３１８】

ステップ716においては、細胞構成成分セットCの存在パターンに基づいて集団Sをクラスター化する。したがって、細胞構成成分セットC全体にわたって類似の存在パターンを有する集団S中の生物はクラスターを形成する。使用されるクラスタリング・タイプは、上記セクション5.1および/または5.8に記載する様々なクラスタリング方法のいずれでもよい。クラスタリングによって、細胞構成成分セットC全体にわたって類似の存在パターンを有する集団Sの1組のクラスター(例えば、部分群)が得られる。

【０３１９】

ステップ718においては、目的形質の連鎖解析(セクション5.2)または関連解析(セクション5.4)を、ステップ716で特定された異なる部分群を用いて実施する。目的QTLにおける目的形質に対するロッド・スコア、または連鎖を定量化するのに使用される別の計量形式をかなり増加させる部分群をさらに解析する。特に、そのような部分群を一連の量的遺伝解析にかける。一連の各量的遺伝解析においては、セットDG中の細胞構成成分の中から選択される細胞構成成分の発現レベルを量的形質として使用する。このような各量的遺伝解析を上記セクション5.3に従って実施する。この解析の最終結果は、セットDG中の細胞構成成分と関連するQTLを特定することである。例えば、上記セクション5.9に記載するものなどの多変量技術を用いたこれらの遺伝子の解析によって、検討中の複合形質に影響を及ぼす遺伝子を特定する。セットDG中の細胞構成成分の解析は、これらの細胞構成成分が、検討中の複合形質の両極端の表現型を識別することができたので特に興味深い。したがって、このような細胞構成成分と関連するQTLによって、複合形質に影響を及ぼす遺伝子がもたらされる。

【０３２０】

5.18. 一般的手法

このセクションでは、複雑性疾患を示す集団を亜集団に細分する別の方法を説明する。

【０３２１】

ステップ1202。ステップ1202(図12A)では、形質を選択して種を検定する。一部の実施形態においては、形質は複合形質である。種は、植物、動物、ヒトまたは細菌とすることができる。一部の実施形態においては、種は、ヒト、ネコ、イヌ、マウス、ラット、サル、ブタ、ショウジョウバエまたはトウモロコシである。一部の実施形態においては、種を代表する複数の生物を検定する。種の生物数は任意の数値とすることができる。一部の実施形態においては、検定される複数の生物は、5～100、50～200、100～500または500個を超える。

【０３２２】

一部の実施形態においては、調べる生物の一部を、形質に影響を及ぼす撹乱にかける。撹乱は、環境的なものでも遺伝的なものでもよい。環境撹乱の例としては、検定化合物、アレルゲン、とう痛、および高温または低温への生物の暴露があるが、これらだけに限定されない。環境撹乱の別の例は、食餌(例えば、高脂肪食または低脂肪食)、睡眠遮断、隔離、および自然環境影響(例えば、喫煙、食餌、運動)の定量である。遺伝的撹乱の例としては、遺伝子ノックアウトの使用、所定の遺伝子または遺伝子産物の阻害剤の導入、N-エチル-N-ニトロソ尿素(ENU)突然変異誘発、遺伝子のsiRNAノックダウン、またはある種の複数の生物が示す形質の定量化が挙げられるが、これらだけに限定されない。

【０３２３】

ステップ1202に使用してもよい撹乱は、撹乱と形質の間に何らかの関係があるために選択される。例えば、撹乱は、検討中の形質に影響を及ぼすと考えられる遺伝子のsiRNAノックダウンとすることができる。本発明のシステムおよび方法において検定することができる形質の例を上記セクション5.14に記載する。

【０３２４】

ステップ1204。ステップ1204(図12A)においては、遺伝子発現/細胞構成成分データ244を導出するために、細胞構成成分レベルを複数の生物246から測定する。このような測定値を構成する組織の本性は、検討中の形質について何が知られているかによって決まる。一部の実施形態においては、細胞構成成分測定値はいくつかの異なる組織で構成される。

【０３２５】

一般に、複数の生物246は、形質に関して遺伝分散を示す。一部の実施形態においては、形質は数量化可能である。例えば、形質が疾患である場合には、形質をバイナリ形式で数量化可能である(例えば、生物が発症した場合を「1」、生物が発症しなかった場合を「0」とする)。一部の実施形態においては、形質を様々な数値として定量することができ、複数の生物246は、そのような様々ないくつかの異なる値をとる。一部の実施形態においては、複数の生物246は、未処置集団(例えば、未暴露集団、野生型集団など)および処置済み集団(例えば、暴露集団、遺伝的改変集団など)を含む。一部の実施形態においては、例えば、未処置集団を撹乱にかけないが、処置済み集団を撹乱にかける。一部の実施形態においては、ステップ1204で測定する組織は、血液、白色脂肪組織、または生物246から容易に得られるある別の組織である。

【０３２６】

多様な実施形態においては、5個の細胞構成成分～100個の細胞構成成分、50個の細胞構成成分～100個の細胞構成成分、300個～1000個の細胞構成成分、800個～5000個の細胞構成成分、4000個～15,000個の細胞構成成分、10,000個～40,000個の細胞構成成分、または40,000個を超える細胞構成成分のレベルを測定する。

【０３２７】

一実施形態においては、遺伝子発現/細胞構成成分データ244は、検討集団内の各個体(生物)246のマイクロアレイ処理像を含む。一部の実施形態においては、このようなデータは、各個体246に対する、マイクロアレイ上の各遺伝子/細胞構成成分248の強度情報250を含む。一部の実施形態においては、細胞構成成分データ244は、実際に、調べる生物246における特定の組織中の様々なタンパク質発現レベルである。

【０３２８】
本発明の一態様においては、細胞構成成分レベルは、ステップ1204において生物の所定の組織中の細胞構成成分量を測定することによって決定される。本明細書で使用する「細胞構成成分」という用語は、検討中の形質に影響を及ぼし得る個々の遺伝子、タンパク質、mRNA、代謝産物、および/または任意の別の細胞構成成分を含む。細胞構成成分レベルは、多種多様な方法によって測定することができる。例えば、細胞構成成分レベルを、検討中の形質に関連する生物組織中の量または濃度、それらの活性度、それらの修飾状態(例えば、リン酸化)、または別の測定値とすることができる。
【０３２９】
一実施形態においては、ステップ1204は、生物246の組織中の細胞構成成分248の転写状態を測定するステップを含む。転写状態としては、組織内の構成成分RNA種、特にmRNAの本性および存在量などがある。この場合、細胞構成成分は、RNA、cRNA、cDNAなどである。細胞構成成分の転写状態は、核酸のアレイもしくは核酸模倣プローブとのハイブリッド形成技術、または別の遺伝子発現技術によって測定することができる。転写物アレイは上記セクション5.11で考察した。
【０３３０】
別の実施形態においては、ステップ1204は、細胞構成成分248の翻訳状態を測定するステップを含む。この場合、細胞構成成分はタンパク質である。翻訳状態としては、生物246中のタンパク質の本性および存在量などがある。一実施形態においては、タンパク質のゲノム全体のモニタリング(すなわち、「プロテオーム」、Goffeau等、1996、Science 274、546ページ)を、調べる生物の1個または複数の組織中に存在する複数のタンパク質種に特異的な固定化抗体、好ましくは固定化モノクローナル抗体を結合部位が含むマイクロアレイを構築することによって実施することができる。抗体は、コードされるタンパク質のかなりの部分に対して提示されることが好ましい。モノクローナル抗体を作製する方法は周知である。例えば、HarlowおよびLane、1998、Antibodies:A Laboratory Manual、Cold Spring Harbor、N.Y.を参照されたい。一実施形態においては、モノクローナル抗体は、ゲノム配列に基づいて設計された合成ペプチド断片に対して産生される。このような抗体アレイを用いて、生物から得られるタンパク質をアレイと接触させ、それらの結合を当分野で既知のアッセイによって分析する。一部の実施形態においては、抗体-抗原相互作用のハイスループット・スクリーニング用抗体アレイを使用する。例えば、Wildt等、Nature Biotechnology 18、989ページを参照されたい。
【０３３１】
あるいは、大規模なタンパク質発現定量分析を、放射性(例えば、Gygi等、1999、Mol. Cell. Biol 19、1720ページ)および/または安定なアイソトープ(^{15}N)代謝標識(例えば、Oda等 Proc. Natl. Acad. Sci. USA 96、6591ページ)と、それに続く二次元(2D)ゲル分離、ならびに分離タンパク質のシンチレーション計数または質量分析法による定量分析によって実施することができる。二次元ゲル電気泳動は、当分野で周知であり、一般に、一次元での集束後に、二次元でのSDS-PAGE電気泳動を行うものである。例えば、Hames等、1990、Gel Electrophoresis of Proteins:A Practical Approach、IRL Press、New York;Shevchenko等、1996、Proc Nat'l Acad. Sci. USA 93、1440ページ;Sagliocco等、1996、Yeast 12、1519ページ;Lander 1996、Science 274、536ページ;およびNaaby-Haansen等、2001、TRENDS in Pharmacological Science 22、376ページを参照されたい。電気泳動図は、質量分析法、ポリクローナル抗体およびモノクローナル抗体を用いたウエスタン・ブロット法および免疫ブロット分析、内部およびN末端の微量配列分析を含めて、多数の技術によって分析することができる。例えば、Gygi等、1999、Nature Biotechnology 17、994ページを参照されたい。一部の実施形態においては、蛍光二次元ディファレンス・ゲル電気泳動(DIGE)を使用する。例えば、Beaumont等、Life Science News 7、2001を参照されたい。一部の実施形態においては、生物246中のタンパク質量を、同位体によってコードされた親和性タグ(ICAT)と、それに続くタンデム型質量分析によって求める。例えば、Gygi等、1999、Nature Biotech 17、994ページを参照されたい。このような技術を用いて、生

物246の1個または複数の所定の組織内で発現されるタンパク質のかなりの部分を同定することができる。

【０３３２】

別の実施形態においては、ステップ1204は、複数の生物246中の細胞構成成分の活性または翻訳後修飾を測定するステップを含む。例えば、ZhuおよびSnyder、Curr. Opin. Chem. Biol 5、40ページ;Martzen等、1999、Science 286、1153ページ;Zhu等、2000、Nature Genet. 26、283ページ;およびCaveman、2000、J. Cell. Sci. 113、3543ページを参照されたい。一部の実施形態においては、細胞構成成分活性の測定は、タンパク質マイクロアレイなどの技術を用いて容易になされる。例えば、MacBeathおよびSchreiber、2000、Science 289、1760ページ;およびZhu等、2001、Science 293、2101ページを参照されたい。一部の実施形態においては、翻訳後修飾または細胞構成成分状態の別の態様を質量分析法によって分析する。例えば、AebersoldおよびGoodlett、2001、Chem Rev 101、269ページ;Petricoin III、2002、The Lancet 359、572ページを参照されたい。

【０３３３】

一部の実施形態においては、調べる生物246のプロテオームをステップ1204において解析する。プロテオーム分析(例えば、すべてのタンパク質の定量化、およびそれらの翻訳後修飾の測定)は、一般に、マイクロアレイ技術などのハイスループット・タンパク質分析方法を使用するものである。例えば、Templin等、2002、TRENDS in Biotechnology 20、160ページ;AlbalaおよびHumphrey-Smith、1999、Curr. Opin. Mol. Ther. 1、680ページ;Cahill、2000、Proteomics:A Trends Guide、47～51ページ;EmiliおよびCagney、2000、Nat. Biotechnol.、18、393ページ;およびMitchell、Nature Biotechnology 20、225ページを参照されたい。

【０３３４】

さらに別の実施形態においては、細胞構成成分量の「混合」態様をステップ1204で測定する。一例においては、調べる生物246中の1組の細胞構成成分の量または濃度を、そのような生物中のある別の細胞構成成分の活性測定値と組み合わせる。

【０３３５】

一部の実施形態においては、ステップ1204において、所与の生物中の細胞構成成分の異なる対立形質を検出し測定する。例えば、二倍体生物においては、任意の所与の遺伝子の2個のコピーがあり、一方が「父」に由来し、もう一方は「母」に由来する。所与の遺伝子の各コピーを異なるレベルで発現することができる場合もある。これは、このタイプの対立遺伝子の示差的発現が検討中の形質に関連し、特に検討中の形質が複雑である場合に関連し得るので極めて興味深い。

【０３３６】

ステップ1206。遺伝子発現/細胞構成成分データ244が得られた後に、このデータを発現統計データに変換する(図12A、ステップ1206)。一部の実施形態においては、細胞構成成分データ244(図2)は、複数の細胞構成成分の転写データ、翻訳データ、活性データおよび/または代謝産物量を含む。一実施形態においては、複数の細胞構成成分は、少なくとも5個の細胞構成成分を含む。別の実施形態においては、複数の細胞構成成分は、少なくとも100個の細胞構成成分、少なくとも1000個の細胞構成成分、少なくとも20,000個の細胞構成成分、または30,000個を超える細胞構成成分を含む。

【０３３７】

本発明の一実施形態の分析において量的形質として一般に使用される発現統計データとしては、転写データから導出される平均対数比、対数強度およびバックグラウンド補正強度があるが、これらだけに限定されない。別の実施形態においては、別のタイプの発現統計データを量的形質として使用する。

【０３３８】

一実施形態においては、この変換(図12A、ステップ1206)を正規化モジュール(図示せず)を用いて実施する。このような実施形態においては、各調べる生物における複数の遺伝子の各々の発現レベルを正規化する。任意の正規化ルーチンを正規化モジュールで使用す

ることができる。代表的な正規化ルーチンとしては、強度のZ-スコア、強度中央値、強度中央値の対数、強度のZ-スコア標準偏差対数、対数強度較正DNA遺伝子セットのZ-スコア平均絶対偏差、ユーザー正規化遺伝子セット、強度中央値の比率補正および強度バックグラウンド補正があるが、これらだけに限定されない。また、正規化ルーチンの組み合わせを実行することもできる。本発明による例示的な正規化ルーチンを上記セクション5.6に詳細に開示する。

【０３３９】

ステップ1250。先行ステップにおいては、形質を特定し、細胞構成成分レベル・データを測定し、細胞構成成分データを発現統計データに変換する。ステップ1250(図12A)においては、1個または複数の表現型を、検定集団内の生物246のすべてまたは一部について測定する。図13に、ステップ1202～1206および1250の結果測定されたデータをまとめる。検定集団内の各生物246では、少なくとも2クラスの収集データがある。第1のクラスの収集データは表現型情報1301である。表現型情報1301は、検討中の形質に関係するあらゆるものとすることができる。例えば、表現型情報1301を、特定の生物が表現型を示すかどうか(+/-)などのバイナリ・イベントとすることができる。表現型情報を、各生物246の肥満測定結果などのある量とすることができる。図13に示すように、1つの生物246につき2回以上の表現型測定を行うことができる。

【０３４０】

検定集団内の各生物246の第2のクラスの収集データは、複数の細胞構成成分の細胞構成成分レベル250(例えば、量、存在量)である(ステップ1204～1206、図12A)。図13には示していないが、各生物に対していくつかの細胞構成成分測定値が存在し得る。これらのセットの各々を、検討中の形質に影響を及ぼす撹乱に生物246をかけた後に、それぞれの生物246において測定される細胞構成成分測定値とすることができる。代表的な撹乱としては、生物246をある量の化合物に曝すことが挙げられるが、これだけに限定されない。また、各生物246の各細胞構成成分セットを、生物の異なる組織から得られた測定値とすることができる。例えば、1組の細胞構成成分測定値を、各生物から採取した血液標本から得ることができ、別の細胞構成成分測定値セットを各生物の脂肪組織から得ることができる。

【０３４１】

ステップ1252。ステップ1252(図12A)においては、ステップ1250で収集した表現型データ1301(図13)を使用して集団を表現型群1410に分ける(図14)。ステップ1252を実施する方法は、ステップ1250で測定される表現型データのタイプによって決まる。例えば、唯一の表現型データが、生物246が特定の形質を示すかどうかである場合には、ステップ1252は単純明快である。形質を示す生物246を第1の群とし、形質を示さない生物246を第2の群とする。これよりもやや複雑な例は、量1301が、各生物246が示す量的形質の段階的変化である場合である。例えば、形質が肥満である場合には、各量1301は、各生物246の肥満度(例えば、肥満度指数など)に対応し得る。この第2の例では、生物246を、肥満度の関数として表現型群1410に分類する(bin)ことができる。

【０３４２】

本発明によるさらに別の例においては、複数の表現型測定値(例えば、2、3、4、5、8、10、20以上、10～20、20以上など)を所与の生物246から得ることができる。そのような実施形態においては、各生物246の各表現型測定値1301を、各生物246に対応する表現型ベクトルの成分として扱うことができる。次いで、表現型群1410を導出するために、これらの表現型ベクトルを、例えば、セクション5.8に開示するクラスタリング法のいずれかを用いてクラスター化することができる。説明上、一例においては、生物246はヒトであり、測定値1301は標準12リード心電図グラフ(ECG)から得られる。標準12リードECGは、体表の電極から記録される心臓の電気活動を表示するものである。ECGは、心拍数、心臓の鼓動、伝導、波形図、およびECG解釈(一般に、バイナリ・イベント、例えば、正常、異常)を含めて、ただしこれらだけに限定されない表現型データの宝庫である。これらの異なる表現型(心拍数、心臓の鼓動)の各々を、表現型ベクトルの成分として定量することができる。また、クラスタリング中に表現型ベクトルの一部の成分(例えば、ECG解釈)にさらに重

みを付けることができる。例えば、各生物246の表現型ベクトルを導出するために、血中コレステロール・レベル、血中トリグリセリド・レベル、性別、年齢などの追加の表現型によってECG測定値を増強することができる。適切な表現型ベクトルが構築された後に、表現型群1410を特定するために、セクション5.8の任意のクラスタリング・アルゴリズムを用いてそれらをクラスター化することができる。

【０３４３】

一部の実施形態においては、ステップ1252は、明確で特徴的な群を生成する表現型ベクトル形式が特定されるまで、様々な表現型ベクトルを構築しクラスター化する反復プロセスである。ある種の表現型(例えば、異常なECG/高コレステロール・部分群、正常ECG/低コレステロール・部分群)によって一義的に特徴付けられる表現型群1410を生成することができる表現型ベクトルは特に重要である。

【０３４４】

上記例を使用して、反復して検定することができる表現型ベクトルとしては、ECGデータのみを含むベクトル、血液測定値のみを含むベクトル、ECGデータと血液測定値を組み合わせたベクトル、選択ECGデータのみを含むベクトル、重み付けされたECGデータを含むベクトルなどがある。また、最適な表現型ベクトルを、確率論的検索技術(例えば、シミュレーテッド・アニーリング法、遺伝アルゴリズム)などの検索技術を用いて特定することができる。例えば、Duda等.、2001、Pattern Recognition、第2版、John Wiley & Sons、New Yorkを参照されたい。

【０３４５】

ステップ1254。ステップ1254においては、集団内の両極端の表現型が特定される。このような両極端の表現型を1組の極端生物(extreme organism)と称することができる。例えば、ある場合には、目的形質は肥満である。ステップ1254においては、極度に肥満した生物246と極度にやせた生物246を両極端の表現型としてこのステップで選択することができる。本発明の様々な実施形態においては、極端な表現型は、集団によって示される所与の表現型に関して集団の上位または下位40、30、20または10パーセンタイルとして定義される。一部の実施形態においては、両極端の表現型と称される極端生物セット中の5個を超える、10個を超える、20個を超える、100個を超える、1000個を超える、2～100、25～500、100個未満または1000個未満の生物が存在する。

【０３４６】

ステップ1256。ステップ1256においては、極端生物セットによって示される種の複数の細胞構成成分(レベル250、図13)が選別される。ステップ1254で選択された極端な表現型の生物246(極端生物セット)に対して測定されるレベル250のみがこの選別に使用される。図13を用いて説明するために、ある表現型に関して生物246-1と生物246-Nが両極端の表現型であり生物246-2が両極端の表現型ではない場合を考える。次いで、この場合においては、選別する際に生物246-6と246-Nの測定レベル250を考慮するのに対し、生物246-2の測定レベル250を考慮しない。

【０３４７】

一部の実施形態においては、所与の細胞構成成分248について(極端な表現型の生物において測定された)細胞構成成分レベル250をt検定(または多変量検定などのある別の検定にかけて、所与の細胞構成成分248が、上記ステップ1252で特定された各表現型群1410(図14)を識別することができるかどうかを判定する。細胞構成成分248は、細胞構成成分が表現型群1410の各々において特徴的に異なるレベルで存在するときに、各表現型群を識別する。例えば、2個の表現型群1410が存在する場合には、細胞構成成分は、(極端な表現型の生物において測定された)細胞構成成分レベル250が第1の表現型群においては第1のレベルで存在し、第2の表現型群においては第2のレベルで存在し、第1のレベルと第2のレベルが明確に異なる場合に、2個の群1410を識別する。

【０３４８】

好ましい実施形態においては、生物のもう一方の細胞構成成分を考慮せずに各細胞構成成分をt検定にかける。しかし、別の実施形態においては、表現型群1410を識別する細胞

構成成分を特定するために、細胞構成成分群をステップ1256において多変量解析で比較する。

【０３４９】

ステップ1258。一般に、ステップ1252で特定された各表現型群を識別すると考えられる極端な表現型の生物中で発現される細胞構成成分は、多数あるはずである。この細胞構成成分248の数は、検定に利用可能な生物246の数を超える場合もある。例えば、一部の実施形態においては、25,000個以上の遺伝子を前のステップで考慮する。したがって、極端な表現型の群を識別する遺伝子が数千はなくても数百あることがある。過剰決定体系をもたらすので生物以上に細胞構成成分を収容することができない多数の統計パラメータを含む統計モデルを用いて、これらの特徴的細胞構成成分を後続ステップで解析する場合もある。このような場合には、縮小アルゴリズムを用いて細胞構成成分数を減らすことが望ましい。しかし、検定細胞構成成分数を削減する必要のない別の形式の統計解析を使用する場合もある。

【０３５０】

ステップ1258において使用してもよい縮小アルゴリズムは、ステップ1256で特定された細胞構成成分セットの次元数を削減する基礎として、ステップ1256において各細胞構成成分に対して計算されるp値または別の形式の計測量を使用する。例示的な縮小アルゴリズムをいくつか考察する。しかし、当業者は、多数の縮小アルゴリズムが当分野で既知であり、このようなアルゴリズムのすべてをステップ1258に使用することができることを理解されたい。

【０３５１】

縮小アルゴリズムの1つは段階的回帰である。段階的回帰の基本手順は、(1)初期モデル(例えば、初期細胞構成成分セット)を特定するステップ、(2)「ステップ操作」を繰り返すこと、すなわち、「ステップ操作」に従って予測変数(細胞構成成分)を追加または除去することによって、前のステップでモデルを繰り返し変更するステップ、および(3)ステップ操作基準が所与の場合に、ステップ操作がもはや不可能であるとき、または指定した最大ステップ数に達したときに検索を終了するステップを含む。前方段階的回帰は、モデル項なしで(例えば、細胞構成成分なしで)開始される。各ステップにおいて、回帰は、何も残らなくなるまで最も統計的に有意な項を追加する。後方段階的回帰は、モデルのすべての項を用いて開始され、残余の細胞構成成分すべてが統計的に有意になるまで、最も有意でない細胞構成成分を除去していく。全細胞構成成分のサブセットを用いて開始し、次いで、所望の次元数還元が得られるまで、有意な細胞構成成分を追加し、または有意でない細胞構成成分を除去することも可能である。

【０３５２】

ステップ1258に使用することができる別の縮小アルゴリズムは、総当り(all-possible-subset)回帰である。実際には、総当り回帰を段階的回帰と併用することができる。段階的回帰検索手法は、細胞構成成分の単一の「最適」サブセットの存在を想定し、それを特定しようとするものである。総当り回帰手法では、有用であると考えることができるサブセット・サイズ範囲を作成する。次いで、このサブセット・サイズ範囲内の可能な全サブセットのうち「最適」なものだけを考慮する。いくつかの異なる判定基準を、複数の決定係数(R-square)、自由度修正済み決定係数(adjusted R-square)、MallowのCp統計などの「良さ(goodness)」の点からサブセットを順序付けるのに使用することができる。総当り回帰を段階的方法と併用するときには、サブセット複数決定係数統計データ(subset multiple R-square statistic)によって、各手法を用いて特定された各「最適」サブセットを直接比較することができる。

【０３５３】

本発明のステップ1258(図12A)に従って高次元の空間を低次元の空間に縮小させる別の手法は、細胞構成成分の一次結合を使用するものである。要するに、線形方法は、高次データを低次元空間に射影するものである。この射影を実施する2つの手法は、主成分分析法(PCA)および重判別分析(MDA)である。PCAは、データを最小二乗的に最適に表す射影を

求めるのに対して、MDAは、データを最小二乗的に最適に分離させる射影を求める。例えば、Duda等、2001、Pattern Classification、第3章および第10章を参照されたい。

【０３５４】

ステップ1258の最終目標は、ステップ1252で特定された表現型群1410に生物246を申し分なく分類する、ステップ256で特定された細胞構成成分セットから導出される分類子、またはステップ1256で特定された細胞構成成分サブセットを特定することである。本発明の一部の実施形態においては、シミュレーテッド・アニーリング法などの確率論的検索方法を使用してそのような分類子またはサブセットを特定することができる。シミュレーテッド・アニーリング法においては、例えば、ステップ1256で特定されて、ステップ1252で特定された表現型クラスに生物246を区別する細胞構成成分セットの全能力を評価する関数における重みを、各検定細胞構成成分に割り当てることができる。シミュレーテッド・アニーリング・アルゴリズム中に、これらの重みを調節することができる。実際に、一部の細胞構成成分にゼロの重みを割り当てることができ、したがって、アニール中に効果的にその細胞構成成分を除去することができ、それによって、後続ステップに使用される細胞構成成分の数を効果的に削減することができる。ステップ1258に使用することができる別の確率論的方法としては、遺伝アルゴリズムなどがあるが、これらだけに限定されない。例えば、Duda等、2001、Pattern Classification、第2版、John Wiley & Sons、New Yorkの第7章にある確率論的方法を参照されたい。

【０３５５】

ステップ1260。一部の実施形態においては、各表現型亜集団内の部分群をさらに特定するために、ステップ1256および/または1258で特定された細胞構成成分をクラスター化する。このようなクラスタリングを実施するために、各検定細胞構成成分に対する発現ベクトルを作成する。各細胞構成成分の発現ベクトルを作成するために、極端な表現型の各生物中の各細胞構成成分に対して測定されるレベル1301をベクトルの成分として使用する。例えば、細胞構成成分248-1の発現ベクトルを生物246-1、246-2および246-3から構築する場合を考える。レベル250-1-1、250-2-1および250-3-1は、細胞構成成分248-1を表す発現ベクトルの3個の成分として働く。次いで、各発現ベクトルを、例えば、上記セクション5.8に記載するクラスタリング法のいずれかによってクラスター化する。一実施形態においては、k平均クラスタリング(セクション5.8.2)を使用する。

【０３５６】

ステップ1260の利点は、このステップで実施されるクラスタリングによって、検討中の形質が、量1101(図11)などの(細胞構成成分レベル以外の)観察可能な全表現型データを用いても識別不可能な群1220(図12)に精緻化されることである。したがって、任意選択のステップ260は、実際に臨床形質を生じる細胞構成成分、またはその形質に対する多様な生化学応答をよく反映する細胞構成成分に焦点を絞ることによって、検討中の臨床形質の定義を改良する方法を提供するものである。しかし、ステップ260での改良は、一般検定集団の選択部分のみ、すなわち、両極端の表現型を示す生物のみに基づいているので不完全となり得る。このため、本発明の方法の後続ステップにおいてパターン分類技術を使用して、表現型レベル1101(図11)に依拠しない方法で一般的な集団を部分群に分類することができる堅牢な分類子を構築する。

【０３５７】

ステップ1264。ステップ1264においては、前のステップで特定された両極端の表現型の判別子(discriminator)として特定された細胞構成成分セット(またはこのような細胞構成成分から得られる主構成成分)を使用して分類子を構築する。この細胞構成成分セットは、実際に、検討中の臨床表現型の定義を改善するものである。ベイズの決定理論、最大尤度推定、線形判別関数、多層ニューラル・ネットワーク、および教師あり学習ならびに教師なし学習を含めて、ただしこれらだけに限定されないいくつかのパターン分類技術を使用してこのタスクを実施することができる。

【０３５８】

ステップ1264による一実施形態においては、極端な表現型の生物を表現型群に区別する

細胞構成成分セットを使用して、ニューラル・ネットワークを、例えば、逆伝播アルゴリズムによって訓練する。この実施形態においては、ニューラル・ネットワークは分類子として働く。まず、ニューラル・ネットワークを、極端な表現型の生物を表現型群に区別する細胞構成成分セットから得られる確率分布を用いて訓練する。例えば、一部の実施形態においては、確率分布は、ステップ1256で計算される各細胞構成成分t値または別の統計データを含む。ニューラル・ネットワークを訓練した後に、それを使用して一般的な集団を表現型群に分類する。一部の実施形態においては、訓練されるニューラル・ネットワークは多層ニューラル・ネットワークである。別の実施形態においては、射影追跡回帰、一般付加モデルまたは多変量適応回帰スプライン(multivariate adaptive regression spline)を使用する。例えば、Duda等、2001、Pattern Classification、第2版、John Wiley & Sons, Inc.、New Yorkの第6章に開示されたいずれかの技術を参照されたい。

【０３５９】

ステップ1264による別の実施形態においては、ベイズの決定理論を使用して分類子を構築することができる。ベイズの決定理論は、分類すべきものについての演繹的情報がいくつかあるときに役に立つ。ここで、極端な表現型の生物を表現型群に区別する細胞構成成分セットから得られる確率分布は、演繹的情報として役立つ。例えば、一部の実施形態においては、この確率分布は、ステップ1256で計算される各細胞構成成分p値または別の統計データを含む。ベイズの決定理論についてのより詳細な情報は、例えば、Duda等、2001、Pattern Classification.、第2版、John Wiley & Sons, Inc.、New Yorkの第2章および第3章に開示されたいずれかの技術を参照されたい。

【０３６０】

ステップ1264によるさらに別の実施形態においては、線形判別分析(関数)、線形プログラミング・アルゴリズムまたはサポート・ベクトル・マシンを使用して、一般的な生物246集団を表現型群1410に分類することができる分類子を構築する。この分類は、臨床の表現型の定義を改良する細胞構成成分248(すなわち、ステップ1256、1258および/または1260で選択された細胞構成成分)の細胞構成成分データ250に基づく。このクラスのパターン分類関数についてのより詳細な情報は、例えば、Duda等、2001、Pattern Classification.、第2版、John Wiley & Sons, Inc.、New Yorkの第5章に開示されたいずれかの技術を参照されたい。

【０３６１】

ステップ1266。ステップ1266においては、ステップ1264で得られる分類子を使用して、検定集団の全部またはかなりの(例えば、30%を超える、50%を超える、75%を超える)部分を分類する。本質的に、分類子は、残余の集団(両極端の表現型を含まない集団部分)を、それらの表現型(例えば、表現型量1301、図13)を考慮せずに分類する。分類子を用いて一般集団を分類するプロセスによって、表現型分類(表現型部分群)1450(図14)が作成される。表現型部分群1450は、検討中の形質を精密化したものと考えられ、検討中の形質を以下の技術を用いて群1250に区別する基本的生化学プロセスの解析に引き続き使用できる。

【０３６２】

ステップ1268。ステップ1260に至るステップおよびステップ1260を含むステップは、差次的に発現され、極端な表現型の生物に由来する細胞構成成分を特定するものである。ステップ1264においては、この細胞構成成分セットを使用して分類子を構築する。図12に示すように、ステップ1266においては、ステップ1264で構築された分類子によって、検討中の形質を、表現型データを考慮せずに部分群1250に分類する。部分群1250によって、検討中の形質の部分群が規定され、各部分群によって、検討中の形質の均質な生化学形態のある形態が規定されると予想される。各群1250における生化学的均質性は、以下に詳細に説明するように、検討中の形質に影響を及ぼす遺伝子および生化学経路を特定するために、定量的な遺伝的方法を用いて利用することができる。

【０３６３】

その形式にかかわらず、ステップ1264において形成された分類子は、ステップ1252で規定される表現型群1410、またはステップ1260で規定される部分群1420をさらに改良するの

に役立つ。このように、このセクションに開示する方法を使用して、検討中の形質を精緻化することができる。この精緻化を図14に示す。初めに、検討中の形質は、生物246のある集団1200によって示される。本方法のステップ1252においては、形質に関係する(細胞構成成分レベル以外の)全体の(目に見える、測定可能な)表現型の観察を使用して、一般的な集団1200を2個以上の表現型群1410(図14)に分割する。本方法のステップ1260においては、選択された細胞構成成分の任意選択のクラスタリングは、表現型群を改良して部分表現型群1420(図14)にするのに役立つ。

【０３６４】

ステップ1260の利点は、ステップ1260のクラスタリングによって、検討中の形質が、量1301(図13)などの(細胞構成成分レベル以外の)観察可能な全表現型データを用いても識別不可能な群1420(図14)に改良されることである。したがって、任意選択のステップ1260は、実際に臨床形質を生じる細胞構成成分、またはその形質に対する多様な生化学応答をよく反映する細胞構成成分に焦点を絞ることによって、検討中の臨床形質の定義を改良する強力な方法を提供するものである。しかし、ステップ1260で提供される改良は、一般検定集団の選択部分のみ、すなわち、両極端の表現型を示す生物のみに基づいているので不完全である。したがって、ステップ1264(図12)においては、出発点として両極端の表現型生物246に基づいて選択される初期細胞構成成分セットを用いて、より堅牢な分類子を構築する。図14に示すように、ステップ1266においては、ステップ1264で得られる分類子によって、検討中の形質を高度に精緻化された部分群1450に分類する。したがって、分類子を構築するために群1410または1420などの全カテゴリーのみを使用したが、分類子によって、集団は、群1410および/または1120に含まれ得るクラスターに分割される。これらのクラスターを図14では部分群1450と表す。これらの部分群1450の各々は、検討中の形質を精緻化するのに役立つ。換言すれば、部分群1450の各々は、検討中の全形質のより均質な形態である。分類子は、表現型データを考慮せずに一般集団を分類する(例えば、レベル1301、図13)。したがって、群1450が、群1420および/または1410に整然と含まれないこともある。

【０３６５】

このセクションに記載する方法を用いて作成される分類子は、目的形質の定義を改良するのに役立つ。したがって、分類子を用いて特定された図14の各群1450(亜集団)は、目的形質よりも均質な集団である。各群1450の生物から得られる細胞構成成分測定値を、連鎖解析(セクション5.2)、関連解析(セクション5.4)などの定量的な遺伝研究における量的形質として使用することができる。一般的な集団ではなく個々の群1450から得られるデータを用いた連鎖解析および/または関連解析によって、特に検討中の形質が複雑であり、かつ/または多数の異なる遺伝子によってもたらされる状況においては、結果が改善されると期待される。このような場合においては、個々の群1250を、より均質な集団または状態とすることができる。したがって、このような集団1450においてQTL(または遺伝子座)パターンをもたらす遺伝子または関連する遺伝子を、細胞構成成分データ形式全体集団をこのような研究において量的形質として使用する場合よりも容易に特定することができる。一般集団ではなく部分群についての量的遺伝解析を使用して目的形質に関連する遺伝子を特定する例が、Schadt等、2003、Nature 422、297ページに示されている。

【０３６６】

5.19. 形質に関連する細胞構成成分を特定する方法

上記セクション5.1.1のステップ1508においては、検討中の形質に関連する細胞構成成分レベル(例えば、遺伝子発現レベル、タンパク質存在量レベルなど)のパターンが特定される。このセクションは、セクション5.1.1のステップ1508を実施することができるいくつかの異なる方法を説明する。当業者は、ステップ1508を実施することができる方法がさらにいくつかあり、このような方法はすべて本発明の範囲に含まれることを理解されたい。

【０３６７】

5.19.1. 相関分析

相関分析を目的形質と細胞構成成分レベルに使用することができる。この手法の例は、Golub等、1999、Science 286:531に示されている。Golub等は、急性リンパ性白血病(ALL)患者と急性骨髄性白血病(AML)患者のクラス予測変数を作成した。37人の患者(ALL27人、AML11人)から6817個の遺伝子の発現データを得た。次に、37人の患者の6817個の遺伝子の発現パターンを近接性分析によって調査した。

【０３６８】

近接性分析においては、各細胞構成成分は、発現ベクトル$\nu(g) = (e_1, e_2, \ldots, e_n)$で表される。ここで、$e_i$は、複数の生物中のi番目の生物中の細胞構成成分gの発現レベル(または存在量)である。クラス・ベクトルは、理想発現パターン(存在量)$c = (c_1, c_2, \ldots, c_n)$で表される。ここで、$c_i$は、i番目の標本がクラス1(例えば、ALL)またはクラス2(例えば、AML)に属する患者から採取されるかどうかによって+1または0である。cと$\nu(g)$の相関は、細胞構成成分とクラス区分との間で様々な方法によって測定される。例えば、ピアソン相関係数またはユークリッド距離を使用することができる。Golub等は、予測変数として細胞構成成分を使用する際の「信号雑音」比を強調する相関尺度$P(g, c)$を使用した。$[\mu_1(g), \sigma_1(g)]$および$[\mu_2(g), \sigma_2(g)]$は、それぞれクラス1(例えば、ALL)およびクラス2(例えば、AML)の標本に対する細胞構成成分gの発現レベル(または存在量)の対数の平均および標準偏差であり、$P(g, c) = [\mu_1(g) - \mu_2(g)]/[\sigma_1(g) + \sigma_2(g)]$は、クラス内の標準偏差に対する各クラス間の差を示している。$|P(g, c)|$の大きな値は、細胞構成成分レベル(例えば、遺伝子発現)とクラス区分の相関が大きいことを示し、正または負である$P(g, c)$の符号は、gがクラス1またはクラス2で豊富であることに対応する。標準ピアソン相関係数と異なり、$P(g, c)$は範囲[-1、+1]に限定されない。クラス1およびクラス2の周りの半径rの近傍$N_1(c,r)$および$N_2(c,r)$は、それぞれ$P(g, c) = r$および$P(g, c) = -r$となる細胞構成成分セットであると定義される。近傍内の異常に大きな細胞構成成分数は、多数の細胞構成成分が、クラス・ベクトルと密接に相関する存在量(例えば、発現パターン)を有することを示している。

【０３６９】

近接性分析から、1組の情報価値のある細胞構成成分(クラス1とクラス2を識別する1組の細胞構成成分;形質を識別する1組の細胞構成成分)を選択することができる。Golub等によれば、例えば、情報価値のある細胞構成成分セットは、クラス1が高いクラス・ベクトルに最も近い[すなわち、できるだけ大きな$P(g, c)$]n/2個の遺伝子とクラス2に最も近い[すなわち、できるだけ大きな$-P(g, c)$]n/2個の遺伝子からなる。

【０３７０】

5.19.2. T検定

検討中の形質と関連する細胞構成成分レベル(例えば、遺伝子発現レベル、タンパク質存在量レベルなど)を特定するために使用することができる別の方法はt検定である。t検定は、2個の群の平均が互いに統計的に異なるかどうかを評価する。t検定を使用するときには、上記セクション5.1.1の処理ステップ1508は、生物246のクラスにおいて平均存在量がかなり異なる細胞構成成分を特定しようとするものである。例えば、複数の生物246を2個の群、すなわち、薬物で処理した群と薬物で処理していない群に分割する場合には、t検定を用いて、薬物で処理した生物と薬物で処理していない生物の平均発現レベルがかなり異なる細胞構成成分を見つける。例えば、Smith、1991、Statistical Reasoning、AllynおよびBacon、Needham Heights、Massachusetts、361～365ページを参照されたい。t検定は、以下の式で示される。

【数２６】

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

【０３７１】

式中、

分子は第1の群(T)と第2の群(C)の所与の細胞構成成分の平均レベルの差である分子であり、

var_Tは、群T中の所与の細胞構成成分レベルの分散(偏差の二乗)であり、

var_Cは、群C中の所与の細胞構成成分レベルの分散(偏差の二乗)であり、

n_Tは、群T中の生物46数であり、

n_Cは、群C中の生物46数である。

【０３７２】

t値は、第1の平均が第2の平均よりも大きい場合は正であり、小さい場合は負である。t値の有意性は、有意表中の値を調べて群間の差が偶然である可能性は低いと言えるほどその比が十分に大きいかどうかを検定することによって決定される。有意性を検定するために、リスク・レベル(アルファ・レベルと呼ばれる)を設定する。本発明の一部の実施形態においては、アルファ・レベルを0.05に設定する。これは、平均間の統計的有意差がない(すなわち、「偶然」)である場合でも、100回に5回は有意差があることを意味する。一部の実施形態においては、アルファ・レベルを0.025、0.01または0.005に設定する。また、有意性を検定するために、検定の自由度(df)を求める必要がある。t検定においては、自由度は、両方の群(TおよびC)の合計人数-2である。アルファ・レベル、dfおよびt値が与えられたとすると、標準有意表(例えば、FisherおよびYates、Statistical Tables for Biological, Agricultural, and Medical Research、Longman Group Ltd.、Londonの表IIIを参照されたい)中でt値を調べて、t値が十分に大きく有意であるかどうかを判定することができる。一部の実施形態においては、tが3以上、4以上、5以上、6以上または7以上のときに、細胞構成成分は、生物246の2群(例えば、化合物で処理した第1の群と化合物で処理していない第2の群)を識別すると考えられる。

【０３７３】

5.19.3. 対応のあるT検定

検討中の形質に関連する細胞構成成分レベル(例えば、遺伝子発現レベル、タンパク質存在量レベルなど)を特定するために使用することができる別の方法は対応のあるt検定である。対応のあるt検定は、2個の群の平均が互いに統計的に異なるかどうかを評価する。対応のあるt検定は、薬物の注射などのある撹乱の前後に同じ生物246から測定値が得られたときに一般に使用される。例えば、対応のあるt検定を、セクション5.1.1の処理ステップ1508の実施形態に使用して、血圧に影響を及ぼす化合物の投与前後の血圧差の有意性を判定することができる。対応のあるt検定は、以下の式で表される。

【数２７】

$$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

【０３７４】

式中、
分子は、標本対の平均であり、
Sdは、標本対の偏差であり、
nは、当該対の数である。

【０３７５】

5.19.4. 別のパラメトリック検定

統計を、データが正規分布などのある一般的な分布に従うという仮定の下に計算したときには、それらはパラメトリック統計と呼ばれる。これらのパラメトリック統計に基づく検定はパラメトリック検定と呼ばれることになる。したがって、データが正規分布を有するときには、任意の数の周知のパラメトリック検定をセクション5.1.1の処理ステップ1508に使用することができる。このような検定としては、上記t検定、分散解析(ANOVA)、反復測定ANOVA、ピアソン相関、単回帰、非線形回帰、多重線形回帰、多重非線形回帰などがあるが、これらだけに限定されない。例えば、回帰を使用して2個の変数(2個の異なる細胞構成成分)がともにどのように変化するかを見ることができる。

【０３７６】

5.19.5. ノンパラメトリック検定

集団分布についての仮定をしない検定はノンパラメトリック検定と呼ばれる。セクション5.1.1の処理ステップ1508の一部の実施形態においては、ノンパラメトリック検定を使用する。一部の実施形態においては、ウィルコクソンの符号付き順位検定、マン-ホイットニー検定、クラスカル・ワリス検定、フリードマン検定、スピアマンの順位相関係数、ケンドールのτ分析、またはノンパラメトリック回帰検定を使用する。

【０３７７】

5.20. 実施例

以下の実施例は、これまで記述してきた発明を説明するためのものであって、その記述を限定するためのものではない。

【０３７８】

5.20.1. 遺伝子型と家系データの例示的出所

マウス。本発明の方法は、遺伝的変異をたどることができるあらゆる生きた生物に適用可能である。したがって、例として、遺伝子型および/または家系データ63(図1)が、遺伝子型決定情報および関連する臨床形質情報が提供される実験交配またはヒト集団から得られる。複雑なヒトの疾患に対するマウス・モデルのそのような実験計画の1つを図9に示す。図9においては、交配されてF_1世代が得られる2種類の親近交系がある。F_1世代を異種交配してF_2世代が得られる。この時点で、F_2集団の遺伝子型が決定され、集団中の各F_2に対する生理的表現型が決定されて、遺伝子型と家系のデータ68が得られる。これらの同じ決定が、親ならびにF_1集団の採取標本に対してなされる。

【０３７９】

ヒト集団。本発明は、モデル・システムに制約されず、ヒト集団に直接適用することができる。例えば、Cephファミリーに対する家系および他の遺伝子型の情報が公的に利用可能であり(Center for Medical Genetics、Marshfield、Wisconsin)、これらのファミリー中の個体から得られるリンパ芽球状細胞系を、Coriell Institute for Medical Research (Camden、New Jersey)から購入することができ、本発明の発現プロファイリング実験に使用することができる。このセクションで考察する植物、マウスおよびヒト集団は、本発明に使用する遺伝子型および/または家系の非限定的な例である。

【０３８０】

5.20.2. 脂肪足蹠塊の例

以下の実施例は、図1に開示する本方法の一実施形態を説明するものである。

【０３８１】

ステップ102。F2雑種を、C57BL/6JとDBA/2Jマウス系統から構築した。すべてのマウスを、Association for Accreditation of Laboratory Animal Careの指針を満たす条件下で飼育した。マウスに12月齢までげっ歯類固形飼料を与え、その後アテローム生成的な高

脂肪、高コレステロール食餌に切り替えてさらに4ヶ月与えた。この交雑種の詳細はDrake等、2001、Physiol. Genomics 5、205ページに記載されている。親マウスおよびF_2マウスを16月齢で屠殺した。屠殺後、肝臓を素早く取り出し、液体窒素で急速冷凍し、-80℃で保存した。全細胞RNAを、Rneasy Miniキットを用いて製造者(Qiagen、Valencia、CA)の指示に従って25mgから精製した。競合ハイブリッド形成を、111個の雌F_2肝臓標本、5個のDBA/2J肝臓標本、および3個のC57BL/6J肝臓標本の各々から得られる蛍光標識cRNA(5mg)を、111個の分析済み肝臓標本の各々から得られる等量のcRNAからなる参照プールから得られる等量のcRNAと混合することによって実施した。

【０３８２】

上述した近交系C57BL/6JとDBA/2Jから構築されたF2マウスは、多数のマウスがアテローム硬化型病変部を生じ、他のマウスは同じ集団の他のマウスよりも脂肪足蹠塊がかなり大きく、コレステロール・レベルが高く、骨構造が大きく、天然集団における多様な疾患のモデルとなる。例えば、Drake、2001、J. Orthop Res 19、511ページ、およびDrake、2001、Physiol. Genomics 5、205ページを参照されたい。

【０３８３】

111匹のF2マウスの肝臓から得られる遺伝子の競合発現値を、23,574個の遺伝子を含むマイクロアレイを用いて測定した。アレイ像をHughes、2000、Cell 102、109ページに記載のようにして処理して、バックグラウンド・ノイズ、単一チャネル強度、および付随する測定誤差推定値を得た。肝臓標本と参照プールの発現変化を$\log_{10}$(発現比)として定量化した。ここで、「発現比」は、アレイ上の各スポットに対する2チャネル(赤と緑/肝臓標本と参照プール)の正規化バックグラウンド補正強度値の比である。対数比の誤りモデルを、Roberts、2000、Science 287、873ページに記載されたように適用して、肝臓標本と参照プールの発現変化の有意性を定量した。

【０３８４】

ステップ104-イエス。この実施例で使用するクラス予測変数/262は、複合形質皮下脂肪足蹠塊(FPM)の様々な細区画において示差的に発現される情報価値のある遺伝子の集合から得られる。FPMは、定量可能なマウス表現型形質である。例えば、Schadt等、2003、Nature. 422、297ページを参照されたい。そのため、280個の遺伝子を、皮下脂肪足蹠塊(FPM)形質の上下25パーセンタイルを含むマウスにおいて最も示差的に発現される遺伝子セットとして選択した。この遺伝子セット(FPMセット)を、FPM形質分布の両端にあるマウスの最も転写的に活性な遺伝子セットと考えることができる。両極端のFPM形質を識別できるかどうかに基づいて遺伝子を選択することによって、この遺伝子セットの選択の偏りをなくした。

【０３８５】

ステップ108。280個の遺伝子を、教師付き分類体系(図1、ステップ106)ではなく、教師なし分類体系で使用した。図1には示していないが、分類子が存在するのにもかかわらず教師なし分類体系を使用することは、セクション5.1に開示する本発明の一実施形態を実施するものである。

【０３８６】

ステップ108の場合には、280個の各遺伝子の発現ベクトルを構築した。各発現ベクトルは、F2集団内の全マウスにわたる280個の遺伝子のセット中の所与の遺伝子の発現値を含む。したがって、例えば、280個の遺伝子のセット中の所与の遺伝子iの発現ベクトルは111個の発現値を含み、各発現値はF2集団内の各マウス中の遺伝子iの発現を示す。

【０３８７】

図8は、二次元クラスター分析である。x軸上には、280個の各遺伝子の発現ベクトルがクラスター化されている。y軸上でクラスタリングを形成するために、111匹のマウスの各々に対してベクトルを構築した。このような各ベクトルは、ベクトルに関連するそれぞれのマウスにおいて考慮される280個の各遺伝子の発現値を含む。次いで、これらのベクトルをy軸に沿ってクラスター化する。したがって、図8において、x軸は、マウス集団にわたって同様に発現する遺伝子をクラスター化し、y軸は、280個の遺伝子のセットに対して

類似した遺伝子発現値を有するマウスをクラスター化する。二次元グラフの各x、y座標は、所与の生物における遺伝子の発現レベルである。図8では明確に示されていないが、二次元グラフの各x、y座標は色わけされており、参照プールに対する所与の生物における遺伝子の発現レベルが示されている。

【０３８８】

図8に示す二次元クラスター分析によって、集団内の部分群を求めることができる。このような亜集団がy軸上のクラスターによって定義されることは明らかである。しかし、x軸上のクラスタリングによって作成されるパターンは、y軸上の亜集団を定義する助けとなる。すなわち、y軸上の各部分群は、280メンバーの遺伝子セット全体にわたる発現に類似したパターンを有するはずである。図8の解析によって、そのような3個のセットが明らかになった。y軸は、臨床形質に基づいてクラスター化されなかった。それにもかかわらず、y軸上のマウスは、別個の表現型群にクラスター化された。第1のセットは低脂肪足蹠塊群である。低脂肪足蹠塊群は、2個の要因によって定義される。第1に、低脂肪足蹠塊群はy軸上のクラスターを定義する。第2に、セット802中の低脂肪足蹠塊群の遺伝子は、参照プールに対して緑色側にシフトする傾向にあるのに対し、セット804の遺伝子は参照プールに対して赤色側にシフトする傾向にある。y軸に沿った280メンバーのセットにおける遺伝子発現パターンは、低脂肪足蹠塊群が、実際に、2個以上の部分群の複合体ではないことを検証するのに役立つ。この形式の解析を継続して、以下の表に要約するように、別の2群(高脂肪足蹠塊1および高脂肪足蹠塊2)をy軸上で定義し、y軸に沿った発現パターンによって検証する。

【表４】

名称	Y軸	X軸遺伝子セット802	X軸遺伝子セット804
低FPM	クラスター810	緑	赤
高FPM2	クラスター812	緑	赤
高FPM1	クラスター814	緑/赤	緑

【０３８９】

ステップ112および114。図8で得られるパターンは、肥満形質、FPMを定義するのに役立つ。実際に、これらのパターンは、発現データなしで可能である以上にFPMの定義を改善するものである。図8に示される高FPMマウスに関連する2個の別個のパターンが存在することは明らかである(高FPM2および高FPM1)。臨床形質に関連する発現パターンの異質性が、臨床形質自体の異質性を示していることはほぼ間違いがない。

【０３９０】

この臨床形質をさらに解明するために、臨床および遺伝子発現データが存在する111匹のF2動物を、図8に示す3個の群の1個に分類した。続いて、2セットの動物、すなわち、1)高FPM群1または低FPMに分類される動物、および2)高FPM群2または低FPMに分類される動物について別個の連鎖解析(セクション5.2)を実施した。この連鎖解析においては、集団全体ではなく上で特定した亜集団を用いて量的形質FPMを解析した。

【０３９１】

図9および図10は、2個の染色体のこれらの解析結果である。2番染色体FPM QTL(図9)は、すべての動物を一緒に考慮したときに、FPMに対して最初に特定された4個のQTLのうち最大のものである。連鎖解析にすべてのマウスを用いた2番染色体のこの位置におけるQTLの大きさは曲線902で示される。しかし、このQTLは、高FPM群1を低FPM群とともに考慮したときに消滅するが(図9、曲線906)、高FPM群2を低FPM群とともに考慮したときに曲線902よりもほぼ2ロッド単位増加する(図9、曲線904)。

【０３９２】

図10は、マウス・セット全体についての初期の解析では19番染色体上にFPM形質に対し

て意味のあるQTLが生成されないが(図10、曲線1002)、低FPM群とともに考慮した高FPM群2によってかなりのロッド・スコアを有するQTLが生じ(図10、曲線1006)、低FPM群とともに考慮した高FPM群1が完全セットのそれよりも有意でない(図10、曲線1004)遺伝子座を示す。

【０３９３】

この実施例の結果は、2番染色体および19番染色体のQTLがそれぞれF2集団のサブセットのみに有意な影響を及ぼしているが、これは肥満などの形質の基礎をなす複雑さを直接示す一種の不均質である。また、19番染色体QTLは、高FPM群1/低FPMサブセットのFPM形質変化原因の19%を説明するが、部分表現型を定義するのに発現データを使用しなかった場合には完全に見落とされていた。図9および10に示したロッド・スコアが最も高いQTLの有意性を、F2動物の完全セットから繰り返し(10,000回)標本を採取することによって、高FPM群1/低FPMおよび高FPM群2/低FPM群とサイズが等しい群が各反復で得られるように評価した。10,000回の標本採取で、図9および10に示したQTLの有意性に近いQTLは得られなかった。

【０３９４】

臨床形質および図9に示した上記2番染色体遺伝子座と関連する遺伝子発現形質の一部の拡大図を図11に示す。FPM QTLと共存するのは、Drake等、2001、Physiol. Genomics 5、205ページに記載された肥満に関係する形質の別のQTLである。これらの形質としては、肥満症、脂肪足蹠塊、血しょう脂質レベル、骨密度などがある。図11に、肥満に関係する形質のうちの4つに対するロッド・スコア曲線を示す。そのうちの4個が肥満症または脂肪足蹠塊形質と共存する(すべてロッド・スコアが2.0を超える)7個の別の遺伝子座に加えて、主要な尿タンパク質遺伝子(MUP1、MUP4およびMUP5)の群が2番染色体遺伝子座と連鎖していることは興味深い。MUP1遺伝子は、すべてこれらの経路に関与することが知られているペルオキシソーム増殖因子活性化受容体(PPAR)ガンマ、RXR相互作用タンパク質およびLPR6のような他の遺伝子と共存するQTLを有することに加えて、レチノイドX受容体(RXR)ガンマ(R = 0.75/P値 << 1.0E<SUP>-15)、アシル補酵素Aオキシダーゼ1(R = 0.65/P値 = 3.78 E^{-15})およびレプチン受容体(R = -0.74/P値 << 1.0E<SUP>-15)を含めて、肥満に関係する経路に関与することが知られている他の多数の遺伝子と最も高い相関があるので目立つ。マウスおよびヒトにおけるレプチン受容体の突然変異は、過食症および極度の肥満を引き起こす。例えば、Chen等、1996、Cel 84、492ページ;Chua等、1996、Science 271、994ページ、Clement等、1998、Nature 392、398ページ、Montague等、1997、Nature 387、903ページ;Strobel等、1998、Nat. Genet. 18、213ページ;およびTsigos等、2002、J Pediatr Endocrinol Metab. 15、241ページを参照されたい。RXRは、脂質代謝、耐糖能およびインスリン感受性の制御に多数の面で関与するPPARαおよびPPARγを含めた多数の核内受容体の絶対パートナー(obligate partner)である。例えば、Chawla、2001、Science 294、1866ページを参照されたい。これは、2番染色体遺伝子座が、肥満症、脂肪足蹠塊、コレステロールおよびトリグリセリド・レベルをともに招き、肥満および糖尿病に確実な役割を有する遺伝子と連鎖していることを示している。また、MUP遺伝子は、リポカリン・タンパク質ファミリーのメンバーであり、マウスの生理および挙動に影響を及ぼすフェロモン結合プロセスに中心的役割を果たすことが知られている一方で(Timm等、2001、Protein Science 10、997ページ)、MUP発現の変化は体重および骨の長さ(Metcalf等、2000、Nature 405、1068ページ)ならびにVLDLレベル(Swift等、2001、J. Lipid Res. 42、218ページ)の変化に関連する。

【０３９５】

2番染色体遺伝子座を補助する領域は、ヒト肥満に関係する表現型とこれまで関連付けられてきた領域であるヒト染色体20q12-q13.12に相同である。例えば、Borecki等、1994、Obesity Research 2、213ページ;Lembertas、1997、J. Clin. Invest 100、1240ページ)を参照されたい。11において強調された遺伝子NM_025575およびNM_015731のヒト相同体はヒト20番染色体領域に存在し、完全には特徴付けられておらず、これまで肥満に関係する形質に結び付けられていない。メラノコルチン3受容体(MC3R)などの他の遺伝子は、こ

の遺伝子座における肥満の候補として示唆されているが(Lembertas等、1997、J. Clin Invest. 100、1240ページ)、この実施例のデータは、遺伝子NM_025575およびNM_015731が、ネズミ2番染色体遺伝子座とかなり連鎖しているだけでなくやはり2番染色体遺伝子座と関連している脂肪足蹠塊形質のいくつかと大いに相互作用する基本的QTLをもたらしていることを示唆している。MC3Rの発現レベルは2番染色体遺伝子座と関連せず、Celera RefSNPデータベースの最近のビルド中のC57BL6系統とDBA/2J系統間でこの遺伝子のエキソンまたはイントロン中に注釈の付いたSNPはなかった。脳におけるMC3Rの多型発現が、2番染色体遺伝子座と連鎖する遺伝子の肝臓中での発現をある程度もたらすのでない限り、これらの事実は、この場合、MC3Rが2番染色体連鎖の基礎をなす遺伝子ではないことを示唆している。

【０３９６】

要約すると、F2動物を、本発明の方法によって3個の群(高FPM1、高FPM2および低FPM)の1個に分類した。次いで、これらの動物を、解析のためそれぞれ低FPM群と組み合わせた異なる高FPM群に適用されるQTL方法を用いて遺伝的に分析した。2番染色体の遠位端の結果を示した。2番染色体のこの領域のFPM QTLは、マウスの高FPM群の1個を考慮すると完全に消失するが、マウスのもう一方の高FPM群を考慮すると初期のロッド・スコアよりもほぼ2ロッド単位増加する。また、全マウスを同時に考慮したときには完全に見落とされていた別の興味深い遺伝子座を19番染色体上に発見した。この場合、2番染色体QTLの影響下にはないマウスの高FPM群は、かなりのロッド・スコアを有するQTLを生じたが、もう一方の高FPM群のロッド・スコアは、完全なセットに対して得られたスコアよりも小さかった。

【０３９７】

この実施例の結果は、遺伝子発現パターンを使用して、臨床形質の定義を、異なる遺伝子座の制御下にあるサブタイプに精緻化することができる初めての証拠である。創薬に対する潜在的重要性は大きく、複雑性疾患を精査する困難に直接関わるものである。FPM2番染色体QTLの基礎をなす遺伝子のみを標的にする化合物を開発することは、(この遺伝子座によって制御されていないので)高FPM群1の遺伝子にはまったく無効であるが、(この遺伝子座によって制御されているので)高FPM2群の遺伝子には極めて効果的なことは明らかである。すべての肥満個体を一緒に1群として処理することは、その他の方法でこの処理に応答するものを特定することによって得られるものよりもはるかに効果の少ない処理になる。また、すべての肥満患者の集団を構成する多数の亜集団の1個として所与の薬物療法に応答する可能性が最も高い亜集団を規定することによって、医薬品産業の薬剤開発と診断構成成分は、薬剤開発の可能な限り早い段階で集団を投与群によって層別化することによって各構成成分をより生産的にする自然な再構築に役立つ。この革新的戦略は、薬剤開発と診断学の2つの古典的に独立した世界をより密接に結び付けるものである。毒性を研究するのにも類似の議論をすることができる。というのは、薬物に対する有害応答も、上述したのと同様にして精査することができる複合形質だからである。

【０３９８】

6. 引用文献

本明細書で引用したすべての文献を、各個々の出版物または特許または特許出願が具体的かつ個別に参照によりその全体が本明細書に援用されるのと同じ程度に、その全体を参照により本明細書に援用する。

【０３９９】

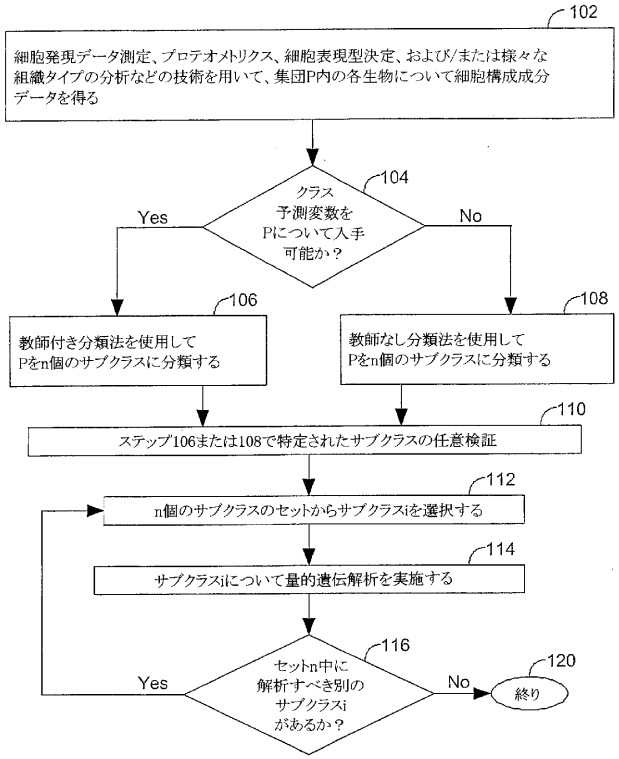
本発明は、コンピュータ読み取り可能な記憶媒体に埋め込まれたコンピュータ・プログラム機構を含むコンピュータ・プログラム製品として実施することができる。例えば、このコンピュータ・プログラム製品は、図1に示すプログラム・モジュールを含むことができる。これらのプログラム・モジュールは、CD-ROM、磁気ディスク記憶装置、または任意の他のコンピュータ読み取り可能なデータもしくはプログラム記憶装置に保存することができる。コンピュータ・プログラム製品中のソフトウェア・モジュールは、インターネット経由によって、または搬送波上で(ソフトウェア・モジュールが埋め込まれた)コンピュータ・データ信号を送信することによって電子的に配布することもできる。

【０４００】
当業者に明らかなように、本発明の精神および範囲から逸脱することなく、本発明の多数の改変形態および変更形態を実施することができる。本明細書に記載する具体的実施形態は、例としてのみ提供されるものであって、本発明は、このような特許請求の範囲の権利が与えられるあらゆる等価物とともに、添付した特許請求の範囲の用語によってのみ限定されるものである。

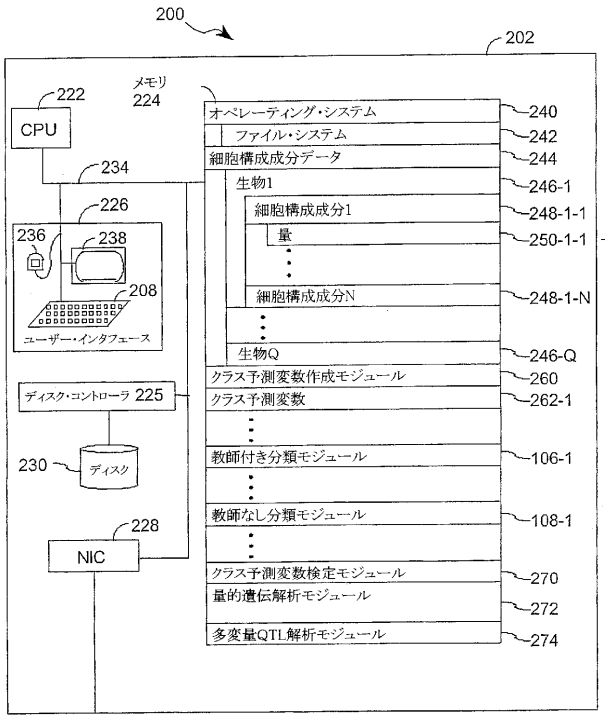
【図面の簡単な説明】

【０４０１】
【図１】集団Pをn個の部分群に細分し、次いで、そのn個の部分群の1個または複数を量的遺伝解析にかける、本発明の一実施形態による処理ステップを示す図である。
【図２】集団Pをn個の部分群に細分し、次いで、そのn個の部分群の1個または複数を量的遺伝解析にかける、本発明の好ましい実施形態によるコンピュータ・システムの図である。
【図３】細胞構成成分レベルを用いて、量的形質遺伝子座解析を容易にするデータ構造およびモジュールを含む、本発明の好ましい実施形態によるコンピュータ・システム中のメモリの図である。
【図４】本発明の好ましい実施形態による、細胞構成成分レベルを用いた量的形質遺伝子座解析アルゴリズム用の処理ステップを示す図である。
【図５】本発明の一実施形態による発現/遺伝子型ウェアハウスの図である。
【図６】本発明の一実施形態による量的形質遺伝子座結果データベースの図である。
【図７】疾患集団Pをn個の部分群に細分し、次いで、そのn個の部分群の1個または複数を量的遺伝解析にかける、本発明の別の実施形態による処理ステップを示す図である。
【図８】階層的にクラスター化された遺伝子および極端な脂肪足蹠塊のマウスを示す図である。
【図９】本発明の一実施形態によるマウス第2染色体の一部のQTL解析結果を示す図である。
【図１０】本発明の一実施形態によるマウス第19染色体の一部のQTL解析結果を示す図である。
【図１１】様々な肥満関連遺伝子のロッド・スコアを示すグラフである。
【図１２】疾患集団Pをn個の部分群に細分し、次いで、そのn個の部分群の1個または複数を量的遺伝解析にかける、本発明の好ましい実施形態による処理ステップを示す図である。
【図１３】検討中の形質を識別する細胞構成成分を特定するために使用されるデータを含むデータ構造を示す図である。
【図１４】本発明の一実施形態による、目的形質の部分形質(subtrait)への分類を示す図である。
【図１５】集団を部分群に細分する、本発明の一実施形態による処理ステップを示す図である。

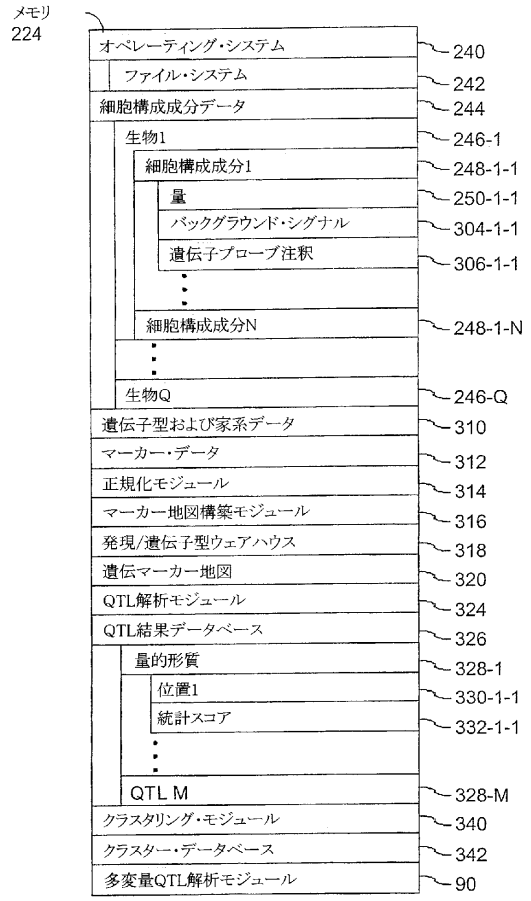
【図１】


102
細胞発現データ測定、プロテオメトリクス、細胞表現型決定、および/または様々な組織タイプの分析などの技術を用いて、集団P内の各生物について細胞構成成分データを得る
104
クラス予測変数をPについて入手可能か？
Yes
No
106
教師付き分類法を使用してPをn個のサブクラスに分類する
108
教師なし分類法を使用してPをn個のサブクラスに分類する
110
ステップ106または108で特定されたサブクラスの任意検証
112
n個のサブクラスのセットからサブクラスiを選択する
114
サブクラスiについて量的遺伝解析を実施する
116
セットn中に解析すべき別のサブクラスiがあるか？
Yes
No
120
終り

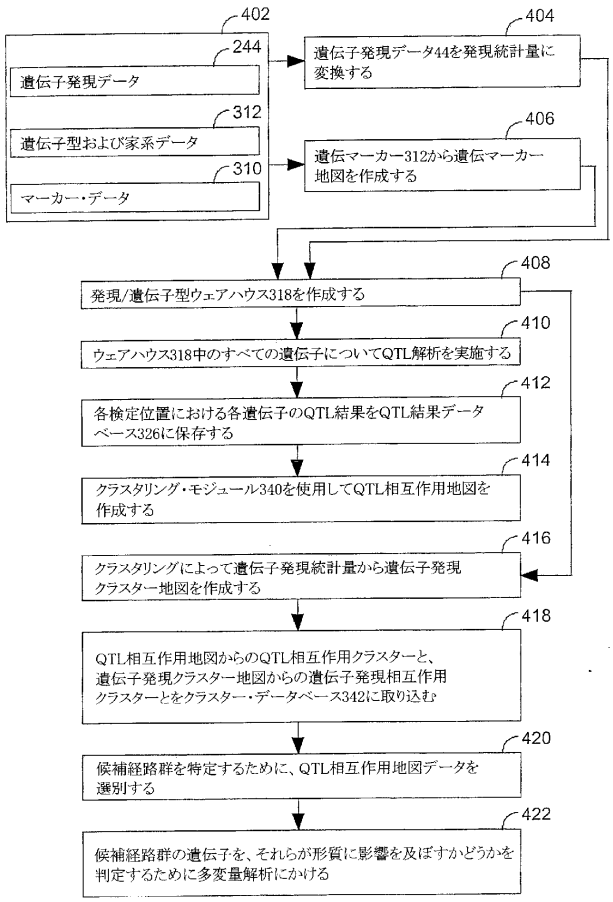
【図２】


200
202
メモリ
224
222
CPU
234
226
236
238
208
ユーザー・インタフェース
ディスク・コントローラ 225
230
ディスク
228
NIC
オペレーティング・システム 240
ファイル・システム 242
細胞構成成分データ 244
生物1 246-1
細胞構成成分1 248-1-1
量 250-1-1
細胞構成成分N 248-1-N
生物Q 246-Q
クラス予測変数作成モジュール 260
クラス予測変数 262-1
教師付き分類モジュール 106-1
教師なし分類モジュール 108-1
クラス予測変数検定モジュール 270
量的遺伝解析モジュール 272
多変量QTL解析モジュール 274

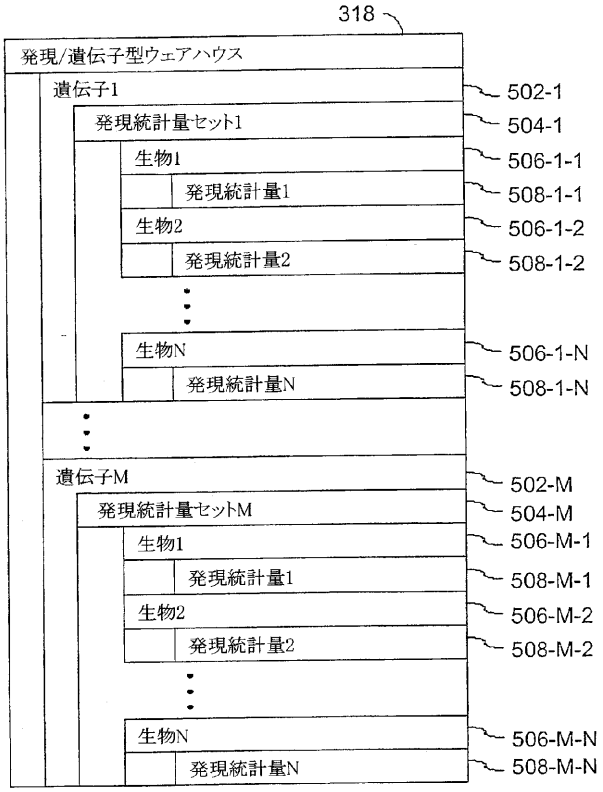
【図３】


メモリ
224
オペレーティング・システム 240
ファイル・システム 242
細胞構成成分データ 244
生物1 246-1
細胞構成成分1 248-1-1
量 250-1-1
バックグラウンド・シグナル 304-1-1
遺伝子プローブ注釈 306-1-1
細胞構成成分N 248-1-N
生物Q 246-Q
遺伝子型および家系データ 310
マーカー・データ 312
正規化モジュール 314
マーカー地図構築モジュール 316
発現/遺伝子型ウェアハウス 318
遺伝マーカー地図 320
QTL解析モジュール 324
QTL結果データベース 326
量的形質 328-1
位置1 330-1-1
統計スコア 332-1-1
QTL M 328-M
クラスタリング・モジュール 340
クラスター・データベース 342
多変量QTL解析モジュール 90

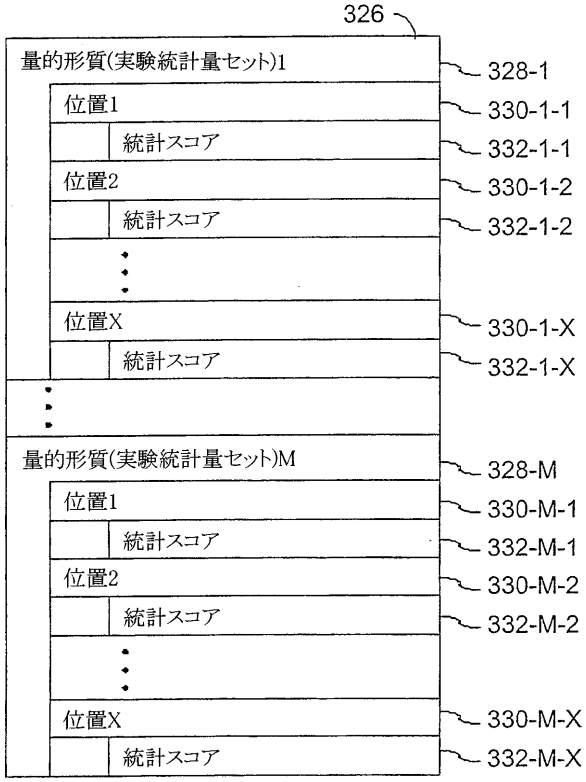
【図４】


402
244
遺伝子発現データ
312
遺伝子型および家系データ
310
マーカー・データ
404
遺伝子発現データ44を発現統計量に変換する
406
遺伝マーカー312から遺伝マーカー地図を作成する
408
発現/遺伝子型ウェアハウス318を作成する
410
ウェアハウス318中のすべての遺伝子についてQTL解析を実施する
412
各検定位置における各遺伝子のQTL結果をQTL結果データベース326に保存する
414
クラスタリング・モジュール340を使用してQTL相互作用地図を作成する
416
クラスタリングによって遺伝子発現統計量から遺伝子発現クラスター地図を作成する
418
QTL相互作用地図からのQTL相互作用クラスターと、遺伝子発現クラスター地図からの遺伝子発現相互作用クラスターとをクラスター・データベース342に取り込む
420
候補経路群を特定するために、QTL相互作用地図データを選別する
422
候補経路群の遺伝子を、それらが形質に影響を及ぼすかどうかを判定するために多変量解析にかける

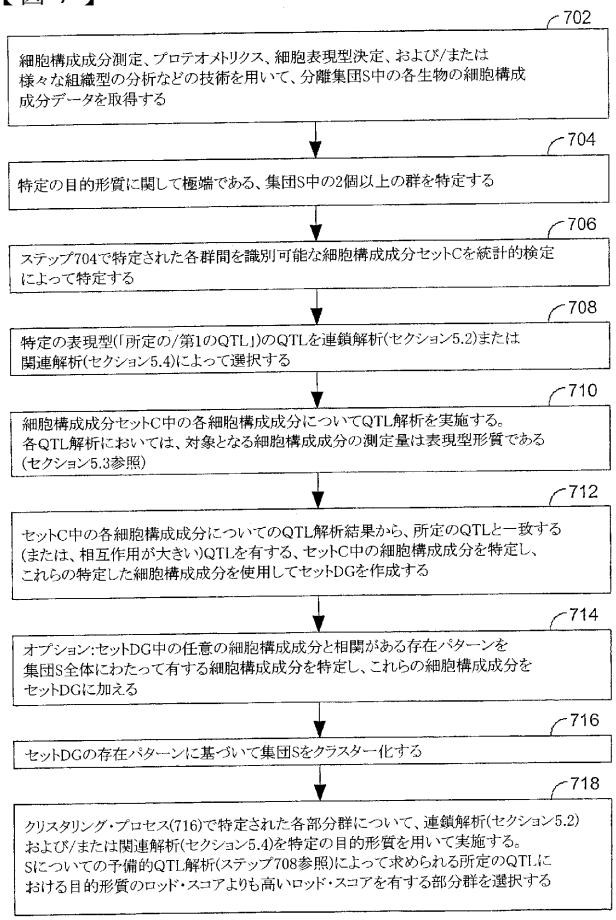
【図５】


318
発現/遺伝子型ウェアハウス
遺伝子1
502-1
発現統計量セット1
504-1
生物1
506-1-1
発現統計量1
508-1-1
生物2
506-1-2
発現統計量2
508-1-2
生物N
506-1-N
発現統計量N
508-1-N
遺伝子M
502-M
発現統計量セットM
504-M
生物1
506-M-1
発現統計量1
508-M-1
生物2
506-M-2
発現統計量2
508-M-2
生物N
506-M-N
発現統計量N
508-M-N

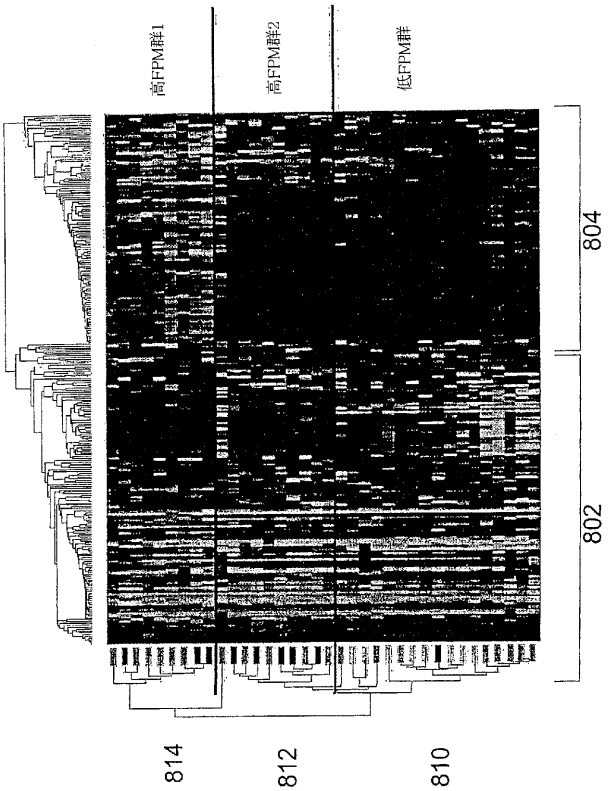
【図６】


326
量的形質(実験統計量セット)1
328-1
位置1
330-1-1
統計スコア
332-1-1
位置2
330-1-2
統計スコア
332-1-2
位置X
330-1-X
統計スコア
332-1-X
量的形質(実験統計量セット)M
328-M
位置1
330-M-1
統計スコア
332-M-1
位置2
330-M-2
統計スコア
332-M-2
位置X
330-M-X
統計スコア
332-M-X

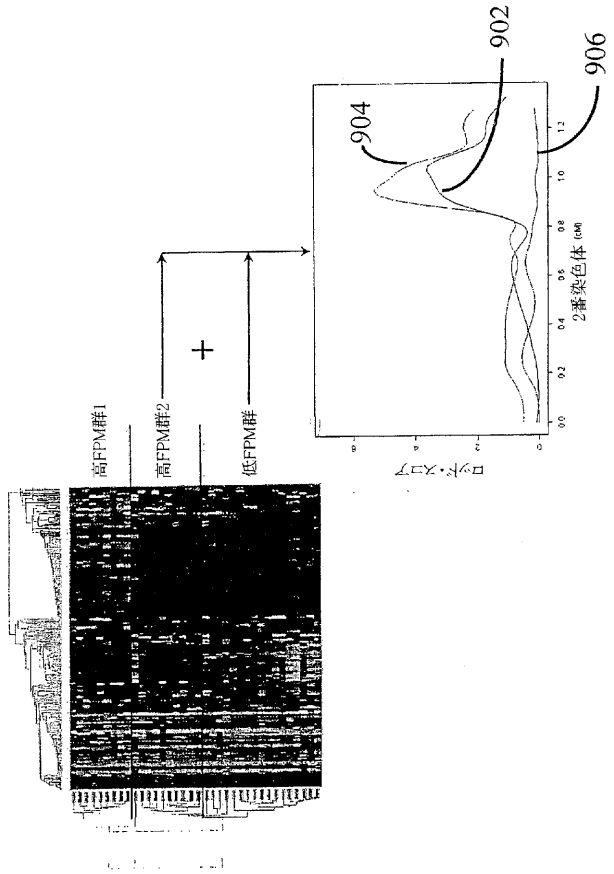
【図７】


702
細胞構成成分測定、プロテオメトリクス、細胞表現型決定、および/または
様々な組織型の分析などの技術を用いて、分離集団S中の各生物の細胞構成
成分データを取得する
704
特定の目的形質に関して極端である、集団S中の2個以上の群を特定する
706
ステップ704で特定された各群間を識別可能な細胞構成成分セットCを統計的検定
によって特定する
708
特定の表現型(「所定の/第1のQTL」)のQTLを連鎖解析(セクション5.2)または
関連解析(セクション5.4)によって選択する
710
細胞構成成分セットC中の各細胞構成成分についてQTL解析を実施する。
各QTL解析においては、対象となる細胞構成成分の測定量は表現型形質である
(セクション5.3参照)
712
セットC中の各細胞構成成分についてのQTL解析結果から、所定のQTLと一致する
(または、相互作用が大きい)QTLを有する、セットC中の細胞構成成分を特定し、
これらの特定した細胞構成成分を使用してセットDGを作成する
714
オプション:セットDG中の任意の細胞構成成分と相関がある存在パターンを
集団S全体にわたって有する細胞構成成分を特定し、これらの細胞構成成分を
セットDGに加える
716
セットDGの存在パターンに基づいて集団Sをクラスター化する
718
クリスタリング・プロセス(716)で特定された各部分群について、連鎖解析(セクション5.2)
および/または関連解析(セクション5.4)を特定の目的形質を用いて実施する。
Sについての予備的QTL解析(ステップ708参照)によって求められる所定のQTLに
おける目的形質のロッド・スコアよりも高いロッド・スコアを有する部分群を選択する

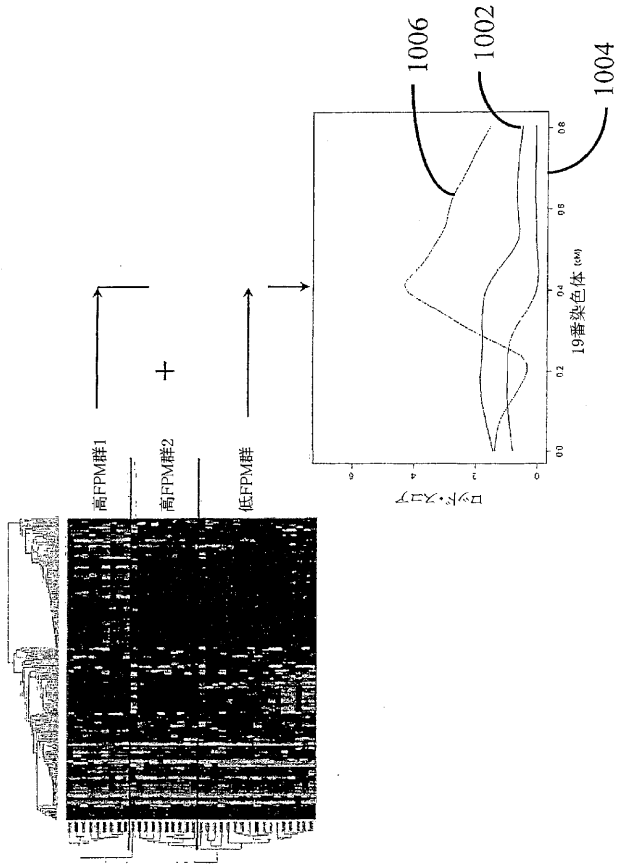
【図８】


高FPM群1
高FPM群2
低FPM群
804
802
814
812
810

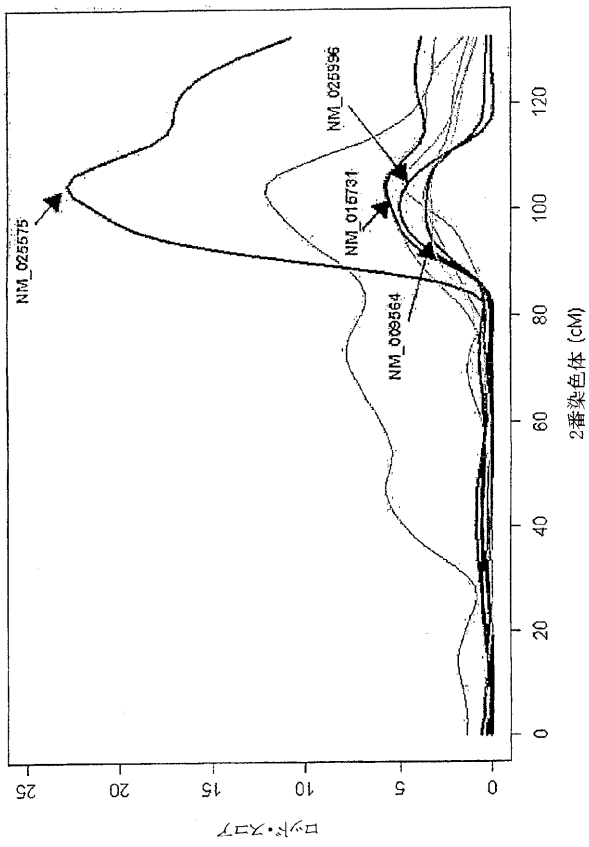
【図９】



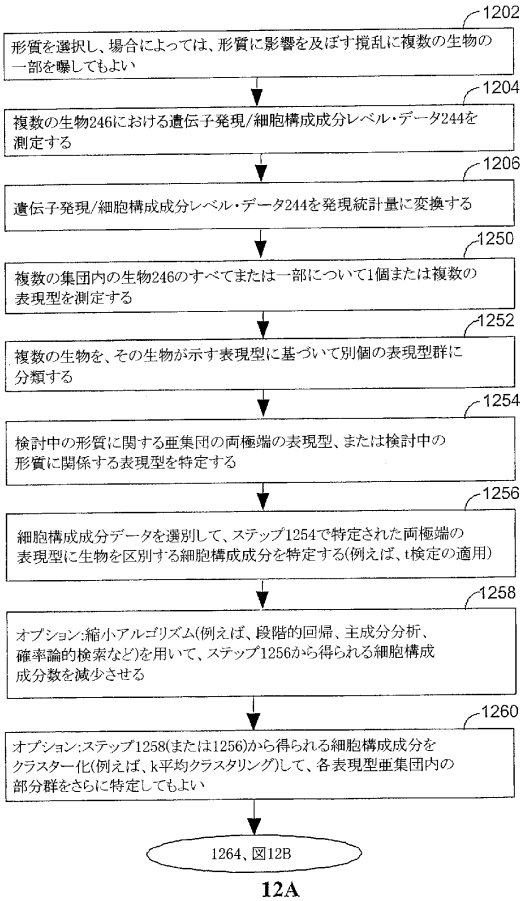
【図１０】



【図１１】

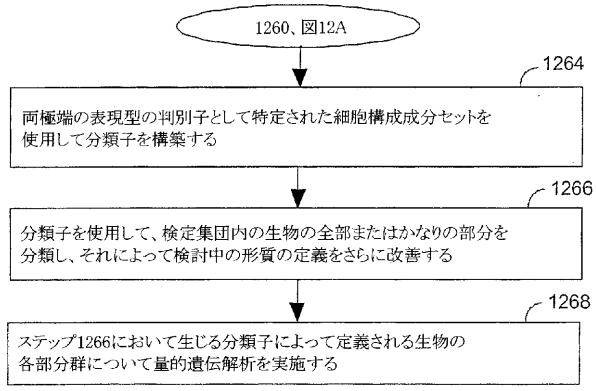


【図１２Ａ】



12A

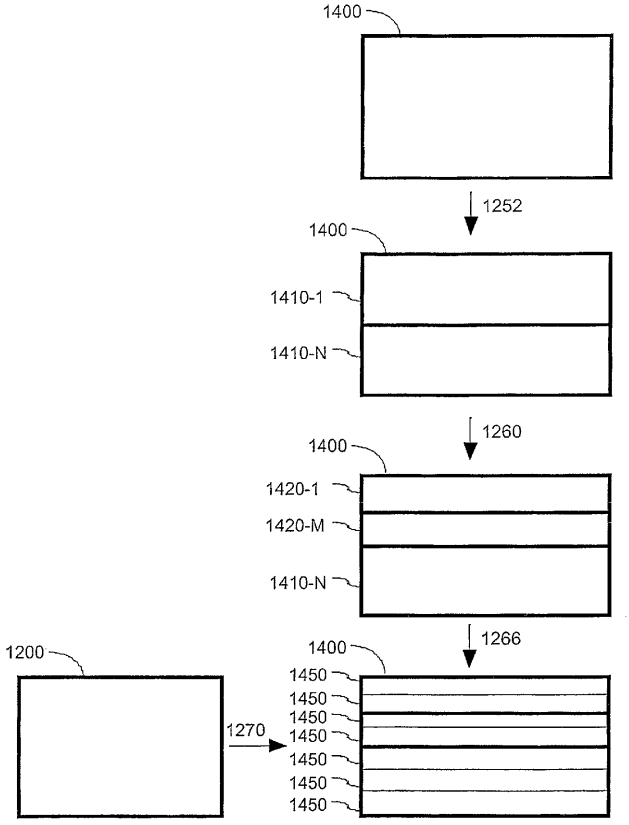
【図１２Ｂ】



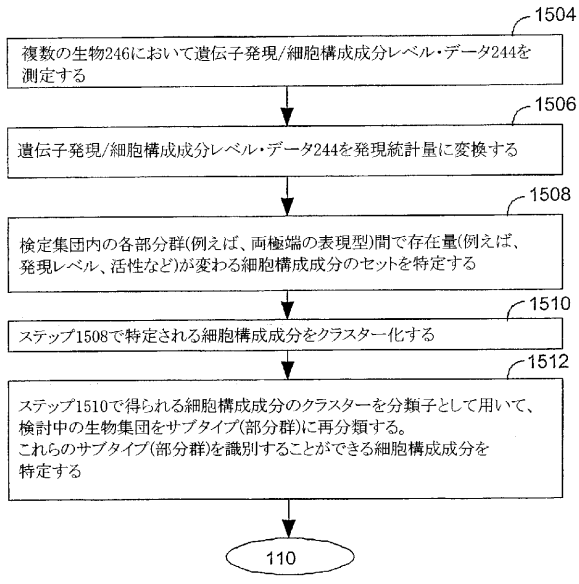
【図１３】

	表現型 1	・・・	表現型 M	CC 248-1	・・・	CC 248-Z
生物 246-1	量 1301-1-1	・・・	量 1301-1-M	レベル 250-1-1	・・・	レベル 250-1-Z
生物 246-2	量 1301-2-1	・・・	量 1301-2-M	レベル 250-2-1	・・・	レベル 250-2-Z
⋮	⋮	⋮	⋮	⋮	⋮	⋮
生物 246-N	量 1301-N-1	・・・	量 1301-N-M	レベル 250-N-1	・・・	レベル 250-N-Z

【図１４】



【図１５】



【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US03/15768

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) : G06F 19/00
US CL : 702/19
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. : 702/19

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 Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JANSEN et al. Genetical genomics: the added value from segregation. Trends in Genetics. July 2001, Vol. 17, No. 7, pages 388-391, especially page 388.	1, 2, 8-101, 106-126, 132-218, 223-245, 251-333
A	DOERGE Mapping and Analysis of Quantitative Trait Loci in Experimental Populations. Nature Reviews Genetics. January 2002, Vol. 31, pages 43-52, especially page 43.	1, 2, 8-101, 106-126, 132-218, 223-245, 251-333

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

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"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)

"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

"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

"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

"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

"&" document member of the same patent family

Date of the actual completion of the international search
20 November 2003 (20.11.2003)

Date of mailing of the international search report
09 DEC 2003

Name and mailing address of the ISA/US
Mail Stop PCT, Attn: ISA/US
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
Facsimile No. (703)305-3230

Authorized officer
John S. Brusca
Valerie Bell Harrison
Telephone No. 703 308-0196

Form PCT/ISA/210 (second sheet) (July 1998)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US03/15768

Box I Observations where certain claims were found unsearchable (Continuation of Item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claim Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claim Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claim Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of Item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
Please See Continuation Sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.: Please See Continuation Sheet

Remark on Protest
☐ The additional search fees were accompanied by the applicant's protest.
☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

PCT/US03/15768

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION IS LACKING

This application contains claims directed to more than one species of the generic invention. These species are deemed to lack unity of invention because they are not so linked as to form a single general inventive concept under PCT Rule 13.1.

In order for more than one species to be examined, the appropriate additional examination fees must be paid. The species are as follows:

The species of constituents are 1) transcription products, 2) translation products, and 3) metabolites.

The species of disease traits are: asthma, ataxia telangiectasia, bipolar disorder, cancer, common late-onset Alzheimer's disease, diabetes, heart disease, hereditary early-onset Alzheimer's disease, hereditary nonpolyposis colon cancer, hypertension, infection, maturity-onset diabetes of the young, mellitus, migraine, nonalcoholic fatty liver, nonalcoholic steatohepatitis, non-insulin-dependent diabetes mellitus, obesity, polycystic kidney disease, psoriasis, schizophrenia, and xeroderma pigmentosum.

The claims are deemed to correspond to the species listed above in the following manner:

For the species of constituents:
transcription product claims 2, 8, 31-68, 70, 71, 101, 106, 126, 132, 155-188, 218, 223, 245, 251, 274-307
translation product claims 2, 3, 8, 101, 106, 126, 132, 218, 223, 245, 246, 251
metabolite claims 4-8, 102-106, 128-132, 219-223, 247-251

For the species of disease traits: Markush claims 15, 113, 139, 230, 258

The following claim(s) are generic:
For the species of constituents: claims 1, 9-30, 69, 72-100, 107-125, 133-154, 189-217, 224-244, 252-273, 308-333

For the species of disease traits: claims 1-14, 16-112, 114-138, 140-229, 231-257, 259-333

The species listed above do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, the species lack the same or corresponding special technical features for the following reasons: each species of constituent is structurally different and has a different biological function. Each disease trait is a distinct and unrelated disease.

The total number of inventions was calculated based on the number of combinations that exist between the species. The formula is recited below:

Total Number of Inventions = number of species of constituents x number of disease trait species.

Continuation of Box II Item 4:

1, 2, 8-101, 106-126, 132-218, 223-245, 251-333, and transcript constituents and asthma trait species

Continuation of B. FIELDS SEARCHED Item 3:

INTERNATIONAL SEARCH REPORT

PCT/US03/15768

Medline, Biosis, US Patent publications and issued, Derwent WPI
search terms: quantitative trait locus, subpopulation

フロントページの続き

(51)Int.Cl.[7]		FI			テーマコード（参考）
Ｇ０６Ｆ	17/30	Ｇ０６Ｎ	3/00	５５０Ｃ	
Ｇ０６Ｎ	3/00	Ｇ０６Ｎ	3/00	５６０Ａ	
// Ｃ１２Ｑ	1/68	Ｃ１２Ｎ	15/00	Ｚ	
Ｇ０１Ｎ	27/447	Ｇ０１Ｎ	27/26	３１５Ｈ	
		Ｃ１２Ｑ	1/68	Ａ	

(81)指定国 AP(GH,GM,KE,LS,MW,MZ,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IT,LU,MC,NL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,ES,FI,GB,GD,GE,GH,GM,HR,HU,ID,IL,IN,IS,JP,KE,KG,KP,KR,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,MZ,NO,NZ,OM,PH,PL,PT,RO,RU,SC,SD,SE,SG,SK,SL,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,YU,ZA,ZM,ZW

（特許庁注：以下のものは登録商標）
ＪＡＶＡ

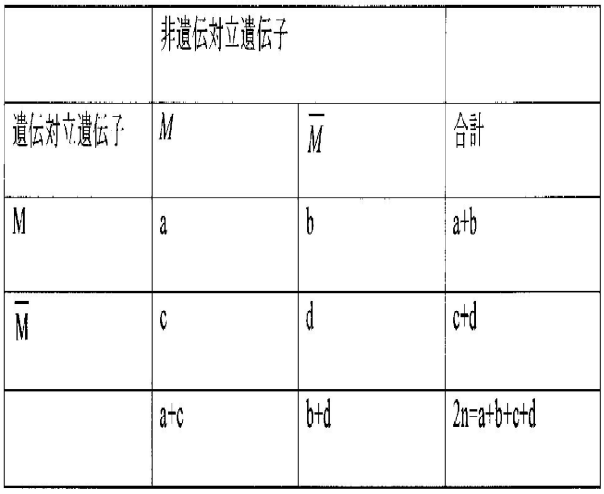
(72)発明者 シャット，エリック，イー.
アメリカ合衆国 ９８０３３ ワシントン州，カークランド，３アールディー プレイス １５１７
(72)発明者 モンクス，ステファニー，エー.
アメリカ合衆国 ９８１２５ ワシントン州，シアトル，エヌイー １２２エヌディー ストリート ９０６

Ｆターム(参考) 4B063 QA01 QA18 QQ52 QQ53 QR08 QR42 QR55 QR62 QR82 QS25 QS34 QS36 QX02
5B075 ND03 UU19

专利名称(译)	用于将复杂疾病细分为组成疾病的计算机系统和方法		
公开(公告)号	JP2005527904A	公开(公告)日	2005-09-15
申请号	JP2004507945	申请日	2003-05-20
[标]申请(专利权)人(译)	ROSETTA INPHARMACTIS		
申请(专利权)人(译)	罗塞塔中医药遗传学LLC		
[标]发明人	シャットエリックイー モンクスステファニーエー		
发明人	シャット,エリック,イー. モンクス,ステファニー,エー.		
IPC分类号	G01N33/53 C12N15/00 C12Q1/68 G01N27/447 G01N33/561 G01N37/00 G06F G06F17/30 G06F19/18 G06F19/24 G06N3/00 G06F19/00		
CPC分类号	G16B20/00 G16B25/00 G16B40/00		
FI分类号	G06F19/00.600 G01N33/53.D G01N33/561 G01N37/00.102 G06F17/30.170.F G06N3/00.550.C G06N3/00.560.A C12N15/00.Z G01N27/26.315.H C12Q1/68.A		
F-TERM分类号	4B063/QA01 4B063/QA18 4B063/QQ52 4B063/QQ53 4B063/QR08 4B063/QR42 4B063/QR55 4B063/QR62 4B063/QR82 4B063/QS25 4B063/QS34 4B063/QS36 4B063/QX02 5B075/ND03 5B075/UU19		
优先权	60/382036 2002-05-20 US 60/460304 2003-04-02 US		
外部链接	Espacenet		

摘要(译)

一种鉴定群体内多种生物表现出的复杂性状的数量性状基因座的方法。使用分类方案将该群体分成多个亚群。根据对该群体的了解，使用监督分类或无监督分类。该分类方案源自从群体中的每个生物获得的多个细胞成分测量。对亚群进行定量遗传分析以鉴定多个亚群中每个亚群的复杂性状的一个或多个数量性状基因座。



	非遺伝対立遺伝子		
遺伝対立遺伝子	M	$\overline{M}$	合計
M	a	b	a+b
$\overline{M}$	c	d	c+d
	a+c	b+d	2n=a+b+c+d