(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6430998号
(P6430998)

(45)発行日　平成30年11月28日(2018.11.28)　(24)登録日　平成30年11月9日(2018.11.9)

(51)Int.Cl.		F I		
C12N 15/11	(2006.01)	C12N	15/11	Z
C12Q 1/6837	(2018.01)	C12Q	1/6837	
C12Q 1/6874	(2018.01)	C12Q	1/6874	
G06F 19/22	(2011.01)	G06F	19/22	

請求項の数 14　外国語出願　(全 118 頁)

(21)出願番号	特願2016-117074 (P2016-117074)
(22)出願日	平成28年6月13日 (2016.6.13)
(62)分割の表示	特願2008-542450 (P2008-542450)の分割
原出願日	平成18年11月22日 (2006.11.22)
(65)公開番号	特開2016-184429 (P2016-184429A)
(43)公開日	平成28年10月20日 (2016.10.20)
審査請求日	平成28年6月13日 (2016.6.13)
(31)優先権主張番号	60/739,882
(32)優先日	平成17年11月26日 (2005.11.26)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	60/742,305
(32)優先日	平成17年12月6日 (2005.12.6)
(33)優先権主張国	米国 (US)

(73)特許権者　513156537
ナテラ，　インコーポレイテッド
アメリカ合衆国　カリフォルニア　９４０７０，　サン　カルロス，　インダストリアル　ストリート　２０１，　スイート　４１０
(74)代理人　100099759
弁理士　青木　篤
(74)代理人　100077517
弁理士　石田　敬
(74)代理人　100087871
弁理士　福本　積
(74)代理人　100087413
弁理士　古賀　哲次

最終頁に続く

(54)【発明の名称】予測を行うための、遺伝子データを清浄化し、そして、そのデータを使用するためのシステムおよび方法

(57)【特許請求の範囲】

【請求項１】

標的個体の所与の染色体の所与のセグメント上の多数遺伝子座の測定値を用いて、標的個体のゲノム中の前記所与のセグメントの存在数を決定する方法であって、

（ｉ）標的個体のゲノムに存在する前記所与のセグメントの存在数に関する１以上の仮説のセットを創製し、

（ii）前記所与のセグメント上の複数の遺伝子座における可能な対立遺伝子の一部又は全部についての遺伝子データの量を測定し、ここで、前記遺伝子データは、遺伝子型及び／又は表現型のデータであり、

（iii）標的個体の遺伝子データの測定値、及び、関連個体の遺伝子データの測定値に基づいて、前記仮説の各々の相対的確率を決定し、ここで、前記仮説の各々の相対的確率の決定は、前記遺伝子座において予測される遺伝子データ及び観察される遺伝子データの統計学的な分布と、前記遺伝子座間の交差確率とに基づき、該仮説の各々の尤度を計算することにより行われ、ここで、前記関連個体は、前記標的個体の親から選択され、

（iv）各仮説に関連する相対的確率を用いて、最も高い確率を有する仮説を、標的個体の現実の遺伝物質の最も可能性の高い前記所与のセグメントの存在数として決定する

ことを含む方法。

【請求項２】

標的個体のゲノムに存在する染色体のセグメントの存在数の決定が、染色体異常のスクリーニングを通じて行われ、前記異常は、モノソミー、片親二染色体、トリソミー、異数

性、アンバランスなトランスロケーション、およびその組合せからなる群より選択される、請求項１記載の方法。

【請求項３】

各仮説の相対的確率の決定が、マッチドフィルタリングの概念を用いて行われる、請求項１又は２に記載の方法。

【請求項４】

各仮説の相対的確率の決定が、対立遺伝子要求を行わない定量的技術を用いて行われ、ここで、各遺伝子座の測定値の平均および標準偏差は、既知、未知、または均一のいずれかである、請求項１又は２に記載の方法。

【請求項５】

各仮説の相対的確率の決定が、対立遺伝子要求を用いる定性的技術を用いて行われる、請求項１又は２に記載の方法。

【請求項６】

各仮説の相対的確率の決定が、参照配列の公知の対立遺伝子と、対立遺伝子の定量的測定値とを用いて行われる、請求項１又は２に記載の方法。

【請求項７】

標的個体が、成人ヒト、若年ヒト、ヒト胎児、ヒト胚、非ヒト成体、非ヒト若年体、非ヒト胎児、および非ヒト胚よりなる群から選択される、請求項１～６の何れか一項に記載の方法。

【請求項８】

標的個体の遺伝子データが、ポリメラーゼ鎖反応（ＰＣＲ）、リガーゼ媒介ＰＣＲ、縮重オリゴヌクレオチドプライマーＰＣＲ、多重置換増幅、対立遺伝子－特異的増幅およびその組合せよりなる群から選択されるツールおよび／または技術を用いて増幅される、請求項１～７の何れか一項に記載の方法。

【請求項９】

標的個体の遺伝子データが、分子逆転プローブ（ＭＩＰ）、ゲノタイピングマイクロアレイ、Ｔａｑｍａｎ ＳＮＰゲノタイピングアッセイ、Ｉｌｌｕｍｉｎａゲノタイピングシステム、他のゲノタイピングアッセイ、蛍光イン－サイチュハイブリダイゼーション（ＦＩＳＨ）、およびその組合せよりなる群から選択されるツールおよび／または技術を用いて測定される、請求項１～８の何れか一項に記載の方法。

【請求項１０】

標的個体の遺伝子データが、標的個体のバルクジプロイド組織、標的個体から取られる１以上のジプロイド細胞、標的個体から取られる１以上の胚盤胞、標的個体上で見出された細胞外遺伝物質、母性血液で見出された標的個体からの細胞外遺伝物質、母性血液中の標的個体由来の細胞、標的個体に由来することが既知の遺伝物質、およびその組合せよりなる群から選択される物質を分析することにより測定される、請求項１～８の何れか一項に記載の方法。

【請求項１１】

標的個体の染色体または染色体セグメントの数の決定が、体外受精における胚選択の目的に用いられる、請求項１～１０の何れか一項に記載の方法。

【請求項１２】

標的個体の染色体または染色体セグメントの数の決定が、出生前遺伝子診断の目的に用いられる、請求項１～１０の何れか一項に記載の方法。

【請求項１３】

請求項１～１２の何れか一項に記載の方法を実現するように構成された、コンピューターにより実施されるシステム。

【請求項１４】

標的個体の遺伝子データの不完全な知識と、標的個体に遺伝的に関連する１人以上の関連個体の遺伝子データの知識とに基づいて、標的個体の遺伝子データと、標的個体のゲノムに存在する染色体または染色体セグメントの存在数とを決定する方法であって、

ここで、前記遺伝子データは、遺伝子型及び／又は表現型のデータであり、ここで、前記関連個体は、前記標的個体の親から選択され、
（ｉ）関連個体のどの染色体のどのセグメントが、標的個体のゲノムに見出されるセグメントに対応するかについて、１以上の仮説のセットを創製し、
（ii）標的個体のゲノムに存在する所与の染色体セグメントの数に関する１以上の仮説のセットを創製し、
（iii）前記所与のセグメント上の複数の遺伝子座における可能な対立遺伝子の各々について遺伝子データの量を測定し、
（iv）標的個体の遺伝子データの測定値の測定値と、関連個体の遺伝子データの測定値とに基づいて、前記仮説の各々の相対的確率を決定し、ここで、前記仮説の各々の相対的確率の決定は、前記複数の遺伝子座において予測される遺伝子データ及び観察される遺伝子データの統計学的な分布と、前記複数の遺伝子座間の交差確率とに基づき、該仮説の各々の尤度を計算することにより行われ、
（ｖ）各仮説に関連する相対的確率を用いて、最も高い確率を有する仮説を、標的個体の現実の遺伝物質の最も可能性が高い前記所与の染色体及びセグメントの存在数として決定する
ことを含む方法。
【発明の詳細な説明】
【技術分野】
【０００１】
（関連出願への相互参照）
本願は、米国特許法の下で、以下の米国仮特許出願の利益を主張する：２００５年１１月２６日出願の第６０／７３９，８８２号；２００５年１２月６日出願の第６０／７４２，３０５号；２００５年１２月９日出願の第６０／７５４，３９６号；２００６年2月２１日出願の第６０／７７４，９７６号；２００６年４月４日出願の第６０／７８９，５０６号；２００６年６月３０日出願の第６０／８１７，７４１号；２００６年７月３１日出願の第１１／４９６，９８２号；および２００６年９月２２日出願の第６０／８４６，６１０号；これらの開示は、その全体が本明細書中に参考として援用される。
【０００２】
（技術分野）
本発明は、一般には、医療的に予測される目的のための遺伝子データを獲得し、操作し、および用いる分野、具体的には、不完全に測定された遺伝子データを遺伝的に関連する個体の公知の遺伝子データを用いることによってより正確とし、それにより、種々の表現型結果をもたらす遺伝子不規則性のより効果的な同定を可能とするシステムに関する。また、本発明は、一般に、遺伝子、表現型および臨床的情報を分析し、管理し、それに作用させ、およびその情報を用いて、医療的決定の表現型結果を予測する分野に関する。さらに詳しくは、本発明は、対象の群からの一体化され、確証された遺伝子および表現型データを用いて、特定の対象に関して良好な決定を行う方法およびシステムに関する。
【背景技術】
【０００３】
（関連技術の背景）
出生前および着床前遺伝子診断
出生前診断の現行の方法は医師および親に対して成長する胎児における異常を警告することができる。出生前診断がなければ、５０人の赤ん坊の内１人は深刻な身体または精神的ハンディキャップを備えたまま誕生し、３０人の内１人のように多くの者は先天的奇形のいくつかの形態を有するであろう。あいにくと、標準的な方法は侵襲性テストを必要とし、流産の大まか１％の危険性を有している。これらの方法は羊水穿刺、絨毛膜絨毛バイオプシーおよび胎児血液サンプリングを含む。これらの内、羊水穿刺は最も普通の手法であり；２００３年において、それは全ての妊娠のほぼ３％で行われていたが、その使用頻度は過去１５年にわたって減少してきた。出生前診断の主な欠点は、限定された活動のコ

ースを仮定すれば、一旦異常が検出されれば、それは非常に深刻な欠陥についてテストするには価値がありかつ倫理的であるに過ぎない。結果として、出生前診断は、典型的には、高い危険性の妊娠の場合に試みられるに過ぎず、そこでは、潜在的異常の深刻性と組合わされた上昇した欠陥の確率が危険性を凌ぐ。これらの危険性を緩和する出生前診断の方法に対する要望が存在する。

【０００４】

最近、無細胞胎児ＤＮＡおよび無傷胎児細胞が母体血液循環に入ることができるのが発見された。結果として、これらの細胞の分析は、早期の非侵襲性出生前遺伝子診断（ＮＩＰＧＤ）を可能とすることができる。ＮＩＰＧＤを用いることにおける鍵となる挑戦は、母体血液から胎児の細胞または核酸を同定し、それを抽出する仕事である。母体血液における胎児細胞の濃度は胎児の妊娠の段階および状態に依存するが、見積もりは母体血液１ミリリットル毎に１ないし４０の胎児細胞、または１００，０００母体有核細胞当たり１未満の胎児細胞の範囲である。現在の技術は母親の血液から少量の胎児細胞を単離することができるが、胎児細胞をいずれかの量の純度まで豊富化するのは非常に困難である。この関係での最も効果的な技術はモノクローナル抗体の使用を含むが、胎児細胞を単離するのに用いられる他の技術は密度遠心、成人赤血球の選択的溶解、およびＦＡＣＳを含む。胎児ＤＮＡ単離は、胎児－特異的ＤＮＡ配列と共にプライマーを用いるＰＣＲ増幅を用いて示されてきた。各胚ＳＮＰの分子の１０がこれらの技術を通じて利用可能なのに過ぎないので、高い忠実度での胎児組織の下のタイピングは現在可能ではない。

【０００５】

正常なヒトはジプロイド細胞毎に２３染色体の２つの組を有し、１つのコピーは各親に由来する。異数性、余分なまたは失われた染色体を持つ細胞、および片親ニ染色体、一方の親に由来する２つの所与の染色体を持つ細胞は、着床の失敗、流産および遺伝病の大きなパーセンテージの原因であると考えられる。個体におけるある種の細胞のみが異数性である場合、該個体はモザイク現象を呈するといわれる。染色体異常の検出は成功した妊娠の確率の増大に加えて、とりわけ、ダウン症候群、クラインフェルター症候群およびターナー症候群のような疾患を持つ個体または胚を同定することができる。染色体異常についてのテストは母親の年齢のように特に重要であり；３５歳および４０歳の間では胚の４０％および５０％の間が異常であり、４０歳を超えると、胚の半分を超えて異常であると見積もられる。

【０００６】

異数性およびモザイク現象の予測で用いられる伝統的な方法である核型分析は、他のより高いスループットのよりコスト的に有利な方法に対する途を開く。最近多大な注目を集めてきた１つの方法はフローサイトメトリー（ＦＣ）および蛍光イン・サイチュハイブリダイゼーション（ＦＩＳＨ）であり、これを用いて、いずれかの相の細胞周期において異数性を検出できる。この方法の１つの利点は、それが核型分析よりも安価であるが、コストは、一般に、少し選択された染色体をテストするのでかなり十分である点である（通常、染色体１３、１８、２１、Ｘ、Ｙ；時々は８、９、１５、１６、１７、２２）；加えて、ＦＩＳＨは低いレベルの特異性を有する。１５細胞を分析するのにＦＩＳＨを用い、９５％信頼性を持って１９％のモザイク現象を検出することができる。テストの信頼性はモザイク現象のレベルが低くなるにつれ、および分析する細胞の数が減少するにつれかなり低くなる。テストが、対立の細胞を分析する場合、１５％と高い擬陽性率を有すると見積もられている。より高いスループット、より低いコスト、およびより大きな精度を有する方法に対する多大な要望が存在する。

【０００７】

遺伝病の古典的な出生前診断に対する代替法としての着床前遺伝子診断（ＰＧＤ）の使用に向けて多くの研究がなされてきた。ほとんどのＰＧＤは、今日、異数性のような高レベルの染色体異常、および成功した着床およびテイクーホームベイビーである主な結果を伴うバランスしたトランスロケーションに焦点を当てている。着床前段階における胚のより広範なゲノタイピングのための方法に対する要望が存在する。既知の病気に関連する対

立遺伝子の数は、現在、ＯＭＩＭによると３８９であり、常に上昇している。その結果、病気表現型に関連する多数の胚ＳＮＰを分析するのは益々重要となりつつある。出生前診断よりも優れた着床前遺伝子診断の明瞭な進歩は、それが、一旦望ましくない表現型が検出されたならば、作用の可能な選択に関して倫理的論争のいくつかを回避する点にある。

【０００８】

ゲノタイピング

単一の細胞を単離するための多くの技術が存在する。ＦＡＣＳマシーンは種々の適用を有し；１つの重要な適用は、サイズ、形状および総じてのＤＮＡ含有量に基づいて細胞間を区別することである。ＦＡＣＳマシーンは、単一細胞をいずれかの所望の容器に分類するように設定することができる。多くの異なるグループが、出生前遺伝子診断、組換え実験、および染色体不均衡の分析を含めた、多数の適用のために単一細胞ＤＮＡ分析を用いてきた。単一－精子ゲノタイピングは、従前、精子試料の法医学分析で用いて（混合試料から生起する問題を減少させ）、および単一－細胞組換え実験のために用いられてきた。

【０００９】

ヒト胚からの単一細胞の単離は、高度に技術的であるが、今日、体外受精クリニックにおいてルーチン的である。今日まで、出生前診断のほとんど大部分は、蛍光イン・サイチュハイブリダイゼーション（ＦＩＳＨ）を用いており、これは、（ダウン症候群、またはトリソミー２１のような）大きな染色体異常を決定することができ、およびＰＣＲ／電気泳動を用いてきており、これは少量のＳＮＰまたは他の対立遺伝子の要求を決定することができる。極体および胚盤胞は共に成功して単離されてきた。胚の一体性を危うくすることなく単一の胚盤胞を単離するのは非常に重要である。最も普通の技術は、３日胚（６または８細胞段階）から単一の胚盤胞を取り出すことである。胚を特殊な細胞培養基（カルシウムおよびマグネシウムを欠如する標準培養基）に移し、酸性溶液、レーザー、または機械的なドリリングを用いて穴を透明帯に導入する。技術者は、次いで、バイオプシーピペットを用いて、単一の目に見える核を取り出す。臨床的実験は、この目的は着床の成功を減少させないことを示している。というのは、この段階において、胚細胞は未分化だからである。

【００１０】

全ゲノム増幅（ＷＧＡ）に対して利用できる３つの主な方法がある：連結－媒介ＰＣＲ（ＬＭ－ＰＣＲ）、縮重オリゴヌクレオチドプライマーＰＣＲ（ＤＯＰ－ＰＣＲ）、および多数置換増幅（ＭＤＡ）。ＬＭ－ＰＣＲにおいては、アダプターと呼ばれる短いＤＮＡ配列をＤＮＡの平滑末端に連結する。これらのアダプターは普遍的増幅配列を含有し、これはＰＣＲによってＤＮＡを増幅するのに用いられる。ＤＯＰ－ＰＣＲにおいては、普遍的増幅配列をやはり含有するランダムプライマーを第一ラウンドのアニーリングおよびＰＣＲで用いる。次いで、第二ラウンドのＰＣＲを用いて、普遍的プライマー配列をさらに持つ配列を増幅する。最後に、ＭＤＡはｐｈｉ－２９ポリメラーゼを用い、これは、ＤＮＡを複製する高度にプロセッシング可能な非特異的酵素であり、単一－細胞分析で用いられてきた。これらの方法のうち、ＤＯＰ－ＰＣＲは、単一コピーの染色体を含めた、少量のＤＮＡから多量のＤＮＡを信頼性よく生産する。他方、ＭＤＡは最も速い方法であり、数時間以内にＤＮＡの１００折り畳み増幅を生産する。単一細胞からの増幅材料に対する主な制限は（１）極端に薄いＤＮＡ濃度または極端に小さな容量の反応混合物を用いる必要性、および（２）全ゲノムを横切って蛋白質からＤＮＡを信頼性よく解離させる困難性である。それにもかかわらず、単一－細胞全ゲノム増幅は、何年もの間種々の適用に対して成功して用いられてきた。

【００１１】

これらの関連でＤＮＡ増幅を用いるのに多数の困難がある。ＰＣＲによる単一－細胞ＤＮＡ（または少数の細胞からの、またはより少量のＤＮＡからのＤＮＡ）の増幅は、該ケースの５ないし１０％において報告されているように完全に失敗しかねない。これは、しばしば、ＤＮＡの汚染、細胞の喪失、そのＤＮＡ、またはＰＣＲ反応の間におけるＤＮＡの接近性である。増幅およびマイクロアレイ分析による胚ＤＮＡの測定で生じ得る誤差の

他の源は、特定のヌクレオチドがＰＣＲの間に誤ってコピーされるＤＮＡポリメラーゼによって導入される転写誤差、およびアレイ上での不完全なハイブリダイゼーションによるマイクロアレイのリーディング誤差を含む。しかしながら、最大の問題は、ヘテロ接合性細胞における２つの対立遺伝子のうちの一方を増幅できないことと定義される対立遺伝子ドロップ－アウト（ＡＤＯ）のままである。ＡＤＯは増幅の４０％を超えるまで影響しかねず、既に引き起こされたＰＧＤ誤診断を引き起こしてきた。ＡＤＯは特に優性病の症例において健康の論争となり、ここで、増幅できないことは侵された胚の着床に導きかねない。（ヘテロ接合体における）各マーカー当たり１を超えるプライマーの組に対する必要性はＰＣＲプロセスを複雑とする。従って、より信頼性があるＰＣＲアッセイがＡＤＯ起源の理解に基づいて開発されつつある。単一－細胞増幅のための反応容器は実験中である。アンプリコンのサイズ、ＤＮＡ分解の量、凍結および解凍およびＰＣＲプログラムおよび条件は、各々、ＡＤＯの速度に影響する。

【００１２】

しかしながら、全てのそれらの技術は、単一細胞における増幅で利用可能なＤＮＡの微量に依存する。このプロセスにはしばしば汚染が伴う。適当な滅菌条件およびマイクロサテライトサイジングは、汚染ＤＮＡの確率を排除することができる。というのは、出生前対立遺伝子においてのみ検出されるマイクロサテライト分析は汚染を排除するからである。対立－細胞レベルまで分子診断プロトコルを信頼性よく導入する研究は、最近、マイクロサテライトマーカーの第一ラウンド多重ＰＣＲ、続いての、リアル－タイムＰＣＲおよびマイクロサテライトサイジングを用いて追求されて、汚染の機会を排除してきた。多重ＰＣＲは単一－細胞ＤＮＡ分析における非常に重要な用件である単一反応における多数断片の増幅を可能とする。慣用的なＰＣＲはＰＧＤで用いられた最初の方法であるが、蛍光イン・サイチュハイブリダイゼーション（ＦＩＳＨ）は今日普通である。乱れていない細胞および組織構築物内での拡散の検出を可能とするのはデリケートなビジュアルアッセイである。それは、先ず、分析すべき細胞の固定に依拠する。その結果、試料の固定および貯蔵条件の最適化が、特に、単一－細胞懸濁液で求められる。

【００１３】

単一－細胞レベルでの多数の病気の診断を可能とする最新の技術は相間染色体変換、比較ゲノムハイブリダイゼーション（ＣＧＨ）、蛍光ＰＣＲ、および全ゲノム増幅を含む。これらの技術の全てによって得られたデータの信頼性は、ＤＮＡ調製の質に依拠する。ＰＧＤは高価でもあり、その結果、ミニ－配列決定のような安価なアプローチに対する要望が存在する。ほとんどの突然変異－検出技術とは異なり、ミニ－配列決定は低いＡＤＯ率での非常に小さなＤＮＡ断片の分析を可能とする。増幅およびＰＧＤについての単一－細胞ＤＮＡを調製する良好な方法が従って求められており、研究されている。より新規なマイクロアレイおよび比較ゲノムハイブリダイゼーション技術は、依然として結局は、分析されるＤＮＡの質に依拠する。

【００１４】

いくつかの技術が、少数の細胞、単一細胞（例えば、胚盤胞）、少数の染色体のＤＮＡについての、またはＤＮＡの断片からの多数ＳＮＰを測定するために開発されている。ポリメラーゼ鎖反応（ＰＣＲ）、続いてのマイクロアレイゲノタイピング分析を用いる技術がある。いくつかのＰＣＲ－ベースの技術は、多数置換増幅（ＭＤＡ）、および単一対のプライマーでのＰＣＲを用いて増幅することができる多数のタグドオリゴヌクレオチドを用いてゲノタイピングを行う分子逆転プローブ（ＭＩＰＳ）のような全ゲノム増幅（ＷＧＡ）技術を含む。非ＰＣＲベースの技術の例は蛍光イン・サイチュハイブリダイゼーション（ＦＩＳＨ）である。該技術は、対立遺伝子ドロップアウト、不完全なハイブリダイゼーション、および汚染のような効果のインパクトを亢進するであろう限定された量の遺伝物質によりひどく誤差の傾向があることが明らかである。

【００１５】

ゲノタイピングデータを供する多くの技術が存在する。ＴａｑｍａｎはＡｐｂｌｉｅｄ Ｂｉｏｓｙｓｔｅｍｓによって生産され、分配されるユニークなゲノタイピング技術で

ある。Ｔａｑｍａｎはポリメラーゼ鎖反応（ＰＣＲ）を用いて、注目する配列を増幅する。ＰＣＲサイクリングの間に、対立遺伝子特異的な従たる溝バインダー（ＭＧＢ）は増幅された配列にハイブリダイズする。ポリメラーゼ酵素によるストランド合成はＭＧＢプローブに連結されたレポーター色素を放出し、次いで、Ｔａｑｍａｎ光学リーダーは色素を検出する。このように、Ｔａｑｍａｎは定量的対立遺伝子区別を達成する。アレイベースのゲノタイピング技術と比較して、Ｔａｑｍａｎは反応当たりかなり高価であり、（～＄０．４０／反応）、およびスループットは比較的低い（実行当たり３８４遺伝子型）。反応当たり１ｎｇのＤＮＡが必要とされるに過ぎないが、Ｔａｑｍａｎによる数千の遺伝子型はマイクログラム量のＤＮＡを必要とし、従って、Ｔａｑｍａｎは必ずしもマイクロアレイよりも少ないＤＮＡを用いない。しかしながら、ＩＶＦ遺伝子型ワークフローに関しては、Ｔａｑｍａｎは最も容易に適用できる技術である。これはアッセイの高い信頼性および、最も重要なことには、アッセイのスピードおよび容易性のためである（実行当たりほぼ3時間、および最小の分子生物学工程）。また、（５００ｋ　Ａｆｆｙｍｅｔｒｉｘアレイのような）多くのアレイ技術とは異なり、Ｔａｑｍａｎは高度に慣用化でき、これは、ＩＶＦ市場で重要である。さらに、Ｔａｑｍａｎは高度に定量的であり、従って、異数性はこの技術単独で検出できよう。

【００１６】

Ｉｌｌｕｍｉｎａは、最近、高－スループットゲノタイピングにおけるリーダーとして出現した。Ａｆｆｙｍｅｔｒｉｘとは異なり、Ｉｌｌｕｍｉｎａゲノタイピングアレイはハイブリダイゼーションに専ら依拠しない。その代わり、Ｉｌｌｕｍｉｎａ技術が対立遺伝子－特異的ＤＮＡ延長工程を用い、これは、元の配列の決定について、ハイブリダイゼーション単独よりもかなり感受性であって、特異的である。従って、これらの対立遺伝子の全てはＰＣＲによって多重的に増幅され、次いで、これらの産物はビーズアレイにハイブリダイズされる。これらのアレイでのビーズはユニークな「アドレス」タグを含有し、天然配列を含有せず、従って、このハイブリダイゼーションは高度に特異的であって、感受性である。次いで、対立遺伝子がヘッドアレイの定量的スキャンニングによって呼ばれる。Ｉｌｌｌｕｍｉｎａ　Ｇｏｌｄｅｎ　Ｇａｔｅアッセイシステムは１５３６までの遺伝子座を同時に遺伝子型分けし、従って、スループットはＡａｑｍａｎよりも良好であるが、Ａｆｆｙｍｅｔｒｉｘ　５００ｋアレイほどは高くない。Ｉｌｌｕｍｉｎａ遺伝子型のコストはＴａｑｍａｎよりも低いが、Ａｆｆｙｍｅｔｒｉｘアレイよりも高い。また、Ｉｌｌｕｍｉｎａプラットフォームは５００ｋ　Ａｆｆｙｍｅｔｒｉｘアレイと同程度完全となるまでには長くを必要とし（７２時間まで）、これはＩＶＦゲノタイピングでは問題である。従って、Ｉｌｌｕｍｉｎａはかなり良好なコールレートを有し、アッセイが定量的であり、従って、異数性がこの技術で検出可能である。Ｉｌｌｕｍｉｎａ技術が５００ｋ　ＡｆｆｙｍｅｔｒｉｘアレイよりもＳＮＰの選択においてかなりフレキシブルである。

【００１７】

一定時間において２５０，０００ＳＮＰまでの測定を可能とする最高スループット技術の内の１つはＡｆｆｙｍｅｔｒｉｘ　ＧｅｎｅＣｈｉｐ　５００Ｋゲノタイピングアレイである。この技術はＰＣＲをやはり用い、続いて、ハイブリダイゼーションによる分析、および水晶表面における異なる位置で化学的に合成されたＤＮＡプローブに対する増幅されたＤＮＡ配列の検出を用いる。これらのアレイの不利は低いフレキシビリティおよびより低い感度である。「完全なマッチ」および「ミスマッチプローブ」のような選択性を増加させることができる修飾されたアプローチがあるか、これらはアレイ当たりのＳＮＰコールの数を犠牲にしてそれを行う。

【００１８】

パイロ配列決定、または合成による配列決定もまたゲノタイピングおよびＳＮＰ分析で用いることもできる。パイロ配列決定に対する主な利点は、極端に速いターンアラウンドおよび曖昧でないＳＮＰコールを含むが、アッセイは、現在、高－スループット平行分析に導かれている。ＰＣＲ、続いての、ゲル電気泳動は、着床前診断においてほとんどの成

功に適合したかなり単純な技術である。この技術において、研究者はネステッドＰＣＲを用いて、注目する短い配列を増幅する。次いで、彼らは特殊なゲル上でこれらのＤＮＡ試料を実行して、ＰＣＲ産物を可視化する。異なる塩基は異なる分子量を有し、従って、どれぐらい速く産物がゲル中を泳動するかに基づいて塩基含有量を決定することができる。この技術は低－スループットであり、現行技術を用いる科学者による主題の分析を必要とするが、スピードの利点を有する（１ないし２時間のＰＣＲ、１時間のゲル電気泳動）。この理由で、それは、セラセミア、神経線維腫症２型、白血球接着欠乏症Ｉ型、アロポーシーメンス病、鎌状細胞貧血、網膜芽細胞腫、ペリツェーウス－メルツバッヒャー病、ドゥシェーヌ筋ジストロフィー、およびクラリノ症候群を含めた、膨大な病気についての出生前ゲノタイピングで従前用いられてきた。

【００１９】

非常に高い忠実度でもって少量の遺伝物質を遺伝子型分けするために開発されたもう１つの有望な技術は、Ａｆｆｙｍｅｔｒｉｘ’ｓ　Ｇｅｎｆｌｅｘアレイのような分子逆転プローブ（ＭＩＰ）である。この技術は、平行して多数のＳＮＰを測定する能力を有し；平行して測定された１０，０００を超えるＳＮＰＳが証明されている。少量の遺伝物質については、この技術についてのコールレートは概略９５％において確立されており、なされたコールの精度は９９％を超えることが確立されている。これまで、該技術は所与のＳＮＰについて１５０分子と小さなゲノムデータの量について実行されてきた。しかしながら、該技術は、着床前遺伝子診断について要求されるように、単一細胞、またはＤＮＡの単一ストランドからのゲノムデータで証明されてきた。

【００２０】

ＭＩＰ技術は、その２つの端部が、それらがＤＮＡの直ちに隣接する標的配列にハイブリダイズする場合に連結によって接合できる線状オリゴヌクレオチドであるパドロックプローブを用いる。プローブがゲノムＤＮＡにハイブリダイズされた後に、ギャップを満たす酵素をアッセイに加え、これは４つのヌクレオチドの内１つをギャップに加えることができる。もし加えられたヌクレオチド（Ａ，Ｃ，Ｔ，Ｇ）が測定下でＳＮＰに対して相補的であるならば、それはＤＮＡにハイブリダイズし、連結によってパドロックプローブの端部を接合するであろう。次いで、管状産物、または閉じたパドロックプローブをエキソヌクレオリシスによって線状プローブから区別される。エキソヌクレアーゼは、線状プローブを分解し、環状プローブを残すことによって、千倍以上だけ、閉じた－ｖｓ－閉じていないプローブの相対的濃度を変化させるであろう。次いで、残ったプローブをもう１つの酵素によって切断部位において開き、ＤＮＡから取り出し、ＰＣＲによって増幅する。各プローブは２０塩基タグよりなる異なるタグ配列が付され（１６，０００が作り出されている）、例えば、Ａｆｆｙｎｅｔｒｉｘ　ＧｅｎＦｌｅｘタグアレイによって検出することができる。特定のギャップを満たす酵素が加えられた反応からのタグドプローブからの存在は、関連ＳＮＰ上での相補的アミノ酸の存在を示す。

【００２１】

ＭＩＰＳの分子生物学利点は：（１）単一反応における多重ゲノタイピング、（２）遺伝子型「コール」はギャップを満たし連結することによって起こるが、ハイブリダイゼーションによっては起こらない、および（３）ユニバーサルタグのアレイへのハイブリダイゼーションは、ほとんどのアレイハイブリダイゼーションに固有な偽陽性を減少させることを含む。伝統的な５００ｋ、ＴａｑＭａｎおよび他のゲノタイピングアレイにおいて、全ゲノタイプ試料はアレイにハイブリダイズされ、これは種々の完全なマッチおよびミスマッチプローブを含有し、アルゴリズムはミスマッチおよび完全なマッチプローブの強度に基づく遺伝子型を要求するようである。しかしながら、ＤＮＡ試料の複雑性、およびアレイ上での膨大な数のプローブのため、ハイブリダイゼーションは固有にノイズがある。他方、ＭＩＰは、より長く、従って、より特異的であり、従って、プローブを環状化するのに頑強な連結工程を用いる多重プローブを用いる（すなわち、アレイ上にはない）。対立遺伝子ドロップアウトは（貧弱な実行プローブのため）高いであろうが、バックグラウンドは（特異性のため）このアッセイにおいてはかなり低い。

【００２２】
この技術を単一細胞（または少数の細胞）からのゲノムデータで用いる場合、それは、ＰＣＲベースのアプローチのように、一体性の争いに悩んでいる。例えば、パドロックプローブがゲノムＤＮＡにハイブリダイズできないことは、対立遺伝子ドロップアウトを引き起こすであろう。これは体外受精の関係で悪くなるであろう。というのは、ハイブリダイゼーション反応の効率は低く、かつそれは相対的に速く進行して、限定された時間内に胚を遺伝子型分けする必要があるからである。ハイブリダイゼーション反応は販売業者が推奨するレベルよりも十分低く減少でき、ミクロ－流動技術を用いて、ハイブリダイゼーション反応を加速することもできる。ハイブリダイゼーション反応のための時間を減少させることに対するこのアプローチは減少したデータの質を引き起こすであろう。

【００２３】
予測ゲノミックス
一旦遺伝子データが測定されれば、次の工程が予測目的でデータを用いることである。多くの研究が予測ゲノミックスにおいてなされ、これは、表現型予測を遺伝子型に基づいてなすことができるように、蛋白質、ＲＮＡおよびＤＮＡの正確な機能を理解することを試みる。カノニカル技術は単一－ヌクレオチド多形（ＳＮＰ）の機能に焦点を当てるが、より進歩した方法は多因子表現型特徴を担うようにされつつある。これらの方法は、遺伝子および表現型予測の組、および測定された結果の組の間の数学的関係を決定するように試みる、直線回帰および非直線神経ネットワークのような技術を含む。また、遺伝子データに典型的なように、結果の数に対して多くの潜在的プレディクターが存在し、データが過少決定される場合でさえパラメーターの重要な組を解決することができるように、さらなる制限を回帰パラメーターに適応するまばらなデータ組を収容するように設計されたＲｉｄｇｅ回帰、ｌｏｇ回帰および段階的選択のような回帰分析技術の組もある。他の技術は、未決定データ組から情報を抽出するために主な成分分析を適用する。決定ツリーおよび偶発性の表のような他の技術は、それらの独立した変数に基づいて主題を細分化して、主題を、表現型結果が同様であるカテゴリーまたはビンに入れるための戦略を用いる。論理的回帰といわれる最近の技術は、カテゴリー的に独立した変数の間の異なる論理的相互関係についてサーチして、遺伝子データに関連する多数の独立変数の間の相互作用に依存する変数をモデル化する方法を記載している。用いる方法に拘わらず、予測の質は、予測をなすのに用いる遺伝子データの質に自然に高度に依存する。

【００２４】
ＤＮＡ配列決定のコストは迅速に低下しており、近い将来において、個人の利益のための個々のゲノム配列決定はより普通になるであろう。個人的遺伝子データの知識は、広範な表現型予測が個人に対してなされるのを可能とするであろう。正確な表現型予測をなすためには、関係を問わず、高い質の遺伝子データが非常に重要である。出生前または着床前遺伝子診断の場合には、複雑化因子は入手可能な遺伝物質の相対的少量である。限定された遺伝物質をゲノタイピングで用いる場合に、測定された遺伝子データの性質に固有にノイズがあると仮定すれば、一次データの忠実度を増大させ、それをクリーンとできる方法に対する多大な要望が存在する。

【００２５】
臨床的決定がなされる現行の方法は、存在する情報の最良な可能な使用を行わない。医療的、生化学的および情報技術の進歩としては、増大した量のデータが作り出され、アカデミックおよび臨床的実験の関係においての個々の患者について双方を貯蔵する。分析で利用可能な遺伝子、表現型および臨床的情報の量における最近の急増に従い、臨床的に関連する相関関係を見出して、人々がより長く、より健康でかつよりエンジョイできる人生を送るのを助けるのに多大の努力が払われてきた。従前には臨床家および研究者は彼らの分析を少量の明らかな潜在的因子に焦点を当て、データの局所的貯蔵を用いるが、他の剤のスコアによって測定されたデータを活用することができ、および所与の遺伝子型または表現型に相関する従前に疑われていない因子を同定することができるより複雑なモデルを用いる潜在的利点がより明瞭になりつつある。この状況は、一旦個人的な遺伝子データが

病気の原因および治療、および対象の他の素因を理解するにおいてより抽象的な役割を占めれば、かなりより複雑になるであろう。次の１０年内に、臨床試験のために、または個人化された治療およびまたは薬物割当ての目的のために、患者の全ゲノムをスキャンし、ならびに膨大な表現型データ点を収集するのが可能であろう。

【００２６】

利用可能なデータの量が膨大となり、それが依然として迅速に増大するにつれ、問題の最も重要な点は、最も適当な関係が発見し、かつそれを用いて人々に役に立つのを可能とする設計および実行する良好な方法となった。分析するのに利用可能な変数の数が増大するにつれ、天文学的数の潜在的関係を会得でき、先見的にそれらのいずれかを除外しない方法を開発するのがより重要となった。同時に、それらの研究を同一プロトコルで実行しなかった場合でさえ、多数の研究の知見を総合し、それを利用することができる方法を開発するのが重要である。また、所与の分析において用いるために最適な方法を正しく同定することができるシステムを開発するために、研究されてきた非常に多数の予測モデルを仮定すれば、それは益々重要になりつつある。

【００２７】

ＨＩＶの関係におけるバイオインフォマティックス

ＨＩＶは三千万を超える人々が現在ＨＩＶに罹って生きているヒトにおいてＨＩＶは広域病と考えられ、毎年二百万を超える死亡がＨＩＶに帰せられている。ＨＩＶの主な特徴の１つはその速い複製サイクル、および逆転写酵素の高い誤差率および組換え原性の結果としてのその高い遺伝子可変性である。その結果、ＨＩＶウイルスの種々の株は異なるレベルの異なる薬物に対する耐性を示し、最適な治療養生法は感染性株の同一性およびその特別な罹患性を考慮することができる。

【００２８】

今日まで認可されたＡＲＴ薬物は１１のＲＴＩ：７のヌクレオシド、１つのヌクレオチド、および３つの非ヌクレオシド；７つのＰＩ；および１つの融合／エントリー阻害剤のリストよりなる。世界中でのＡＲＴ薬物が現在広く行きわたっていることを仮定すれば、ウイルスの耐性株の出現は、耐性に対する低い遺伝子バリア、および貧弱な薬物固執双方のため不可避的である。その結果、どのようにして突然変異したウイルスが抗－レトロウイルス療法に応答するかを予測する技術は益々重要となっている。というのは、それらはサルベージ療法についての結果に影響するだろうからである。ウイルス遺伝子配列決定の迅速に現象しているコスト－予備的に調製された配列については５ドルと低い容量価格－は、よりコストがかかりかつ関連するイン－ビトロ表現型測定よりはむしろ、ウイルス遺伝子配列データに基づく薬物の選択を魅力的なオプションとする。しかしながら、配列データの使用はウイルス遺伝子突然変異の出現に基づく、ウイルス薬物応答の正確な予測を必要とする。ウイルス突然変異の多くの異なる組合せは、全ての遺伝子補因子およびそれらの相互作用を含むモデルを設計し、限定されたデータでもってモデルを訓練するのを困難とする。後者の問題は、薬物養生法の多くの異なる組合せが、変数、すなわち、ベースライン臨床状態、処置履歴、臨床的結果および遺伝子配列を含有するいずれかの特定の養生法について十分に大きなデータ組を収集するのを困難とする場合に、イン－ビボ薬物応答をモデル化する関係が悪化した。

【００２９】

抗ウイルス薬物に対する耐性は、ＲＴまたはプロテアーゼ配列内の１つの突然変異、または複数の突然変異の組合せの結果であり得る。ＲＴ酵素は５６０コドンの鍵となる組によってコードされ；プロテアーゼ酵素は９９のコドンによってコードされる。アミノ酸を改変する突然変異のみをコードすることによって各アミノ酸遺伝子座は１９の可能な突然変異を有し；従って、ＲＴ酵素について野生型とは異なる合計１０，６４０の可能な突然変異、およびプロテアーゼ酵素についての１，９８１の可能な突然変異がある。単純な直線モデルを用い、データで総合した各突然変異（全ての突然変異が起こるのではない）が特定の重み付け、または直線回帰パラメーターと関連させる場合、数千のパラメーターが存在し得る。もし数百人の患者の試料のみが各薬物で利用できるならば、問題は過剰決定

的であるか、またはＨａｄａｍａｒｄの意味において不適切である。というのは、独立した方程式よりも評価するより多くのパラメーターがあるからである。不適切な問題のためにモデル構築する問題に適用することができる多くの技術が存在する。これらが先見的専門知識を観察と組み合わせて、専門家のルールに基づくシステム、ならびにｉ）リッジ回帰、ｉｉ）主要成分分析、ｉｉｉ）決定ツリー、ｉｖ）段系的選択技術、ｖ）神経ネットワーク、ｖｉ）最小絶対収縮および選択オペレーター（ＬＡＳＳＯ）およびｖｉｉ）Ｓｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅｓ（ＳＶＭ）を含めた統計的方法を作り出すことを含む。

【００３０】

３つの主な産業－標準専門家システムを典型的に用いて、ＡＲＴ薬物へのＨＩＶウイルスの罹患性：ＡＮＲＳ－ＡＣ１１システム、Ｒｅｇａシステム、およびＳｔａｎｆｏｒｄＨＩＶｄｂシステムを予測する。新しいアルゴリズムがこれらの専門家システムに対して評価されるのは文献において通常である。しかしながら、これらの専門家システムのいずれも、表現型応答の直接的予測を行うように設計されていないが、むしろ、異なる薬物をそれにより比較することができる数値スコアを供し、または感受性、中程度および耐性のような区別されるグループに薬物を分類するように設計されている。加えて、段階的選択でもって訓練された直線回帰モデルのような統計学的アルゴリズムは、表現型結果の予測において専門家システムを実質的に凌ぐことが明瞭に確立されている。結果として、統計学的技術の組のみが、文献に最近開示された方法を最良に実行することを含む詳細な記載中の新規な方法と比較される。

【００３１】

サルベージＡＲＴの臨床的結果の予測に対する現在のアプローチは、薬物養生法および遺伝子突然変異の多くの異なる順列と組み合せた、ほとんどは、統計学的に有意な結果のデータの欠如のため、良好な予測パワーを示さない。この分野は多数の不均一なデータ組の一体化、および薬物応答予測の増強の双方のための緊急の要望を有する。

【００３２】

癌の関係でのバイオインフォマティックス

見積って８０，０００の年次臨床試験のうち、２，１００は癌薬物のためである。癌療法のための危険性および利点をバランスさせることは、表現型および遺伝子型情報の組合せ使用についての臨床的先駆者を表す。過去数十年において化学療法で大きな進歩があったにもかかわらず、腫瘍学者は彼らの癌患者を、癌細胞について正常な細胞に対してしばしば毒性である原始的全身薬物で依然として治療している。かくして、化学の最大毒性用量および治療用量の間に微妙な線がある。さらに、用量－制限毒性は、他の患者ではなくある患者においてよりひどく、治療運動をより高くまたはより低くシフトさせ得る。例えば、乳癌治療で用いられるアントラサイクリンは有害な心血管事象を引き起こしかねない。現在、もし患者が心臓病に対して低い危険性であると決定できても、治療ウィンドウをより大きな用量のアントラサイクリン療法を可能とするようにシフトできたとしても、全ての患者はあたかも心血管毒性の危険性があるように治療される。

【００３３】

各患者についての化学療法の利点および危険性をバランスさせるために、副作用のプロフィール、および医薬介入の治療的有効性を予測することができる。癌療法は、しばしば、ユニークな宿主および腫瘍遺伝子型についての不適切な調整のため失敗する。単一の多形は、稀には、薬物応答において有意な変動を引き起こし；むしろ、マニフォールド多形の結果ユニークな生体分子組成物をもたらし、臨床的結果の予測を困難とする。「ファルマコゲネティックス」は、広く、遺伝子変異が薬物に対する患者の応答に影響する方法と定義される。例えば、肝臓酵素における天然の変異は薬物代謝に影響する。癌化学療法の将来は標的化医薬であり、これは、癌を、多数の遺伝子的、分子的、細胞的、および生化学的異常を含む病気プロセスとして理解する必要がある。酵素－特異的薬物の出現に伴い、腫瘍が特異的にまたは正常な組織よりも高いレベルで分子標的を発現することを確実とするために注意することができる。腫瘍細胞および健康な細胞の間の相互作用を考慮する

ことができる。というのは、患者の正常な細胞および酵素は腫瘍薬物の曝露を制限でき、または有害な事象をよりありそうにしかねないからである。

【００３４】

バイオインフォマティックスは癌治療に大変革を起こさせ、仕立てられた治療が利点を最大化し、有害な事象を最小化するのを可能とする。応答を予測するのに用いられる機能的マーカーはコンピュータアルゴリズムによって分析することができる。乳癌、結腸癌、肺癌および前立腺癌は４つの最も普通の癌である。これらの癌に対する２つの治療の例は乳癌を治療するのに用いられるタモキシフェン、および結腸癌患者において用いられるイリノテカンである。タモキシフェンまたはイリノテカンも、各々、乳癌または結腸癌を治療するのに必要でなく、または十分でない。癌および癌の治療は、患者の副作用のプロフィールおよび腫瘍応答に従って、療法の改正および、しばしば、組合せ療法を必要とする動的なプロセスである。もし癌治療を決定的なツリーとイメージして、他の療法の前、後またはそれと共にいずれかの１つの治療を与え、またはそれを差し控えるならば、このツリーは決定決断点のサブセットを含み、そこではツリーの多く（すなわち、他の治療）はブラックボックスと考えることができる。それにも拘わらず、医師を最も効果的な治療に部分的にガイドするためのデータを有することは有益であり、より多くのデータを集めるに従い、このデータに基づいて治療の決定を行うための効果的な方法は数千人の癌患者において平均余命および生活の質を有意に改善することができよう。

【００３５】

結腸または大腸は胃腸（ＧＩ）管の最後の６－フットのセクションである。合衆国癌協会は、結直腸癌の１４５，０００の症例が２００５年において診断され、５６，０００人が結果として死亡するであろうと見積もっている。結直腸癌はグレード、または細胞の異常、および段階について評価され、これは腫瘍のサイズ、リンパ節の関与、および遠い転移の存在または不存在に細分化される。結直腸癌の９５％は、結腸のルーメンをライニングする遺伝子的突然変異体上皮細胞から発生する腺癌である。症例の８０ないし９０％において、外科的処置単独が看護の標準であるが、転移の存在は化学療法を必要とする。転移性結直腸癌に対する多くの一次療法の１つは５－フルオロウラシル、ロイコボリン、およびイリノテカンの養生法である。

【００３６】

イリノテカンは、スーパーコイルドＤＮＡの絡みを解いて、ＤＮＡ複製が分裂細胞において進行するようにし、細胞をアポトーシスに対して感受性とするトポイソメラーゼを阻害するカンプトテシンアナログである。イリノテカンは生物学的経路において明確な役割を有さず、従って、臨床的結果は予測するのが困難である。用量－限定的毒性はひどい（グレードＩＩＩないしＩＶ）下痢および骨髄抑制を含む、その双方は直ちに医療的注意を必要とする。イリノテカンはウリジン二リン酸グルコロノシルトランスフェラーゼイソ形態１ａ１（ＵＧＴ１Ａ１）によって活性な代謝産物であるＳＮ－３８に代謝される。ＵＧＴ１Ａ１における多形はＧＩのひどさ、および骨髄副作用と相関する。

【００３７】

先行技術

本明細書中において、本発明の分野に関連する先行技術の組をリストする。この先行技術はいずれも、本発明の新規なエレメントを含まず、または断じてそれに言及しない。特許文献１において、Ｈａｒｔｌｅｙらは、作製された組換え部位および組換え蛋白質を用いてＤＮＡ分子のセグメントを移動させ、または交換する組換えクローニング方法を記載する。特許文献２において、Ｐａｒｒｏｔｔらは、生体活性脂質のレベルについて体外受精培養の培地検体を分析して、当該特徴を決定することによって、総じての胚の健康、着床性、および出産予定日まで成功して発生する増大した尤度を含めた体外受精胚の種々の生物学的特徴を決定する方法を提供する。特許文献３において、Ｔｈｒｅａｄｇｉｌｌらは、複数の単離された親細胞における部位－特異的有糸分裂組換えに関連するイン・ビトロフェノタイピングおよび遺伝子マッピングで言うようなホモ接合性細胞ライブラリーを調製する方法を記載する。特許文献４において、Ｓｔｅｗａｒｔらは、血清において直接

的に、またはＩＶＦ／ＥＴ手法の一部として患者から抽出された顆粒膜黄体細胞を培養することによって間接的にレラキシンを測定することによって成功する体外受精（ＩＶＦ）の確率を決定する方法を記載する。特許文献５において、Ｃｏｏｋｅらは、女性患者からの生物学的試料中の１１□－ヒドロキシステロイドデヒドロゲナーゼのレベルを測定することによってＩＶＦの結果を予測する方法を提供する。特許文献６において、Ｌａｒｄｅｒらは、神経ネットワークを用いて、療法剤に対する病気の抵抗性を予測する方法を記載する。特許文献７において、Ｖｉｎｇｅｒｈｏｅｔｓらは、所与のＨＩＶ株のインテグラーゼ遺伝子型を、関連表現型と共にＨＩＶインテグラーゼ遺伝子型の公知のデータベースと単純に比較して、マッチング遺伝子型を見出す方法を記載する。特許文献８において、Ｄｅｎｔｏｎらは、個人のハプロタイプを一般的集団におけるハプロタイプの公知のデータベースと比較して、治療に対する臨床的応答を予測する方法を記載する。特許文献９において、Ｓｃｈａｄｔらは、遺伝子マーカーのマップを構築し、個人の遺伝子および特性を分析して遺伝子－特性遺伝子座データを与え、次いで、これを遺伝子的に相互作用する経路を同定するための方法としてクラスター化し、これを多変数分析を用いて確証する方法を記載する。特許文献１０において、Ｖｅｌｔｒｉらは、パラメーターとしてバイオマーカーのコレクションを利用して、前立腺癌の再発の危険性を評価する神経ネットワークの使用を含む方法を記載する。特許文献１１において、Ｍａｓｃａｒｅｎｈａｓは、患者についての生化学的プロフィールを確立し、テストコフォルトのメンバーにおいて応答性を測定し、次いで、患者の生化学的プロフィールのパラメーターを個々にテストして、薬物応答性の尺度との相関性を見出すことによって薬物応答性を予測する方法を記載する。

【特許文献１】米国特許第６，７２０，１４０号明細書
【特許文献２】米国特許第６，４８９，１３５号明細書
【特許文献３】米国特許出願公開第２００４／００３３５９６号明細書
【特許文献４】米国特許第５，９９４，１４８号明細書
【特許文献５】米国特許第５，６３５，３６６号明細書
【特許文献６】米国特許第７，０５８，６１６号明細書
【特許文献７】米国特許第６，９５８，２１１号明細書
【特許文献８】米国特許第７，０５８，５１７号明細書
【特許文献９】米国特許第７，０３５，７３９号明細書
【特許文献１０】米国特許第６，０２５，１２８号明細書
【特許文献１１】米国特許第５，８２４，４６７号明細書

【発明の概要】
【課題を解決するための手段】
【００３８】
（発明の要旨）

開示するシステムは、情報の源として二次的遺伝子データを用い、またその遺伝子データを用いて、表現型および臨床的予測をする、不完全またはノイズがある遺伝子データの清浄化を可能とする。開示はヒト対象からの遺伝子データに焦点を当てているが、開示する方法は関連する範囲において生物の範囲の遺伝子データに適用されることは注意すべきである。遺伝子データを清浄化するために記載する技術は、体外受精の間の着床前診断、羊水穿刺と組み合わせた出生前診断、絨毛膜バイオプシー、および胎児血液サンプリング、および非侵襲性出生前診断との関係で最も関連し、ここで、少量の胎児遺伝物質は母体血液から単離される。診断は遺伝病、欠点または異常の増大した尤度、ならびに臨床的およびライフスタイルの決定を促進するための個体についての表現型予測を行うことに焦点をあてることができる。本発明は、先に議論された先行技術の欠点に取り組む。表現型および臨床的予測を行うための本明細書中に記載された技術は、着床前診断、出生前診断との関係、または医療的疾患、または罹患性を持つ個人の関係を含めた、多数の関係で関連する。本明細書中に開示される技術のある実施形態は、個体についての遺伝子、表現型および／または臨床的情報の組を仮定し、個体についての表現型結果または表現型罹患性の性格な予測を行うためのシステムを記載する。１つの態様において、遺伝子データに典型

的なように、測定された結果の数と比較して多くの潜在的予測が存在する場合に表現型を正確に予測することができる線形および非線形回帰モデルを形成するための技術が開示され；本発明のもう１つの態様において、該モデルは分割表に基づき、パブリックドメインで入手可能な情報から形成される。なおもう１つの発明において、システムが記載され、ここで、多数のモデルが関連データセットで訓練され、関連予測を行うのに最も正確なそのモデルを用いる。

【００３９】

本発明の１つの態様において、方法は、減数分裂のメカニズムの知識、および胚ＤＮＡの不完全な測定と共に、母親および父親の遺伝子データの不完全な知識を用いて、高度な信頼性でもって鍵となるＳＮＰの位置において胚ＤＮＡをイン・シリコにて再構築する。親データは、貧弱に測定されたＳＮＰのみならず、挿入、欠失、およびＳＮＰ、または全く測定されなかったＤＮＡの全領域の再構築を可能とすることに注意するのは重要である。

【００４０】

開示された方法は体外受精との関係で適応でき、ここで、着床についてコードされる各胚からのゲノタイピングで利用できる。開示された方法は、少数の胎児細胞、または胎児ＤＮＡの断片のみが母親の血液から単離されている非侵襲性出生前診断（ＮＩＰＤ）の関係に等しく適応できる。開示された方法は、羊水穿刺の場合、および胎児の血液が直接的にサンプリングされる他の方法において等しく適応可能である。開示された方法は、限定された量の遺伝子データが標的個人から入手でき、およびさらなる遺伝子データが標的に遺伝的に関連する個体から入手できるいずれの場合においてもより一般的に適用可能である。

【００４１】

本発明の１つの態様において、再構築された胎児または胚ゲノムデータを用いて、細胞が異数性であるか、すなわち、少数の、または２を超える特定の染色体か細胞に存在するかを検出することができる。この疾患の普通の例はトリソリン－２１であり、これはダウン症候群を生起させる。再構築されたデータを用いて、所与の染色体の２つが存在し、その双方が１つの親に由来する疾患である片親二染色体についても検出することができる。これは、ＤＮＡの潜在的状態についての仮説の組を創製し、いずれの１つが測定されたデータを仮定して真実である最高の確率を有するかを見るためにテストすることによってなされる。異数性をスクリーニングするための高スループットゲノタイピングデータの使用は、各胚からの単一の胚盤胞が多数病気－関連遺伝子座を測定し、ならびに染色体異常についてスクリーニングする双方で用いられるのを可能とするのに注意されたし。

【００４２】

本発明のもう１つの態様において、複数の遺伝子座に存在する増幅されたまたは増幅されていない遺伝物質の量の直接的測定を用いて、異数性、または片親二染色体について検出することができる。この方法の背後にある考えは、単に、増幅の間に存在する遺伝物質の量は初期試料における遺伝子情報の量に比例し、多数の遺伝子座においてこれらのレベルを測定することは統計学的に有意な結果を与えることである。染色体異常についてスクリーニングするこの方法は、遺伝子データを清浄化するための本明細書中に記載された関連方法と組合せて用いることができる。

【００４３】

本発明のもう１つの態様において、開示された方法は、外来性遺伝物質によって生じたデータを同定することにより外来性ＤＮＡまたはＲＮＡに汚染されている個体の遺伝物質を清浄化できる。汚染ＤＮＡによって生じた偽シグナルは、異数性によって生じた染色体－幅特異的シグナルを検出できる方法と同様に認識することができる。

【００４４】

本発明のもう１つの態様において、標的細胞が単離され、これらの細胞に含有される遺伝子データが増幅され、以下の技術：ＰＣＲ－ベースの増幅技術、ＰＣＲ－ベースの測定技術、または分子逆転プローブに基づく検出技術、またはＧｅｎｅＣｈｉｐまたはＴａｑ

Ｍａｎシステムのようなマイクロアレイのうちの１以上の組合せを用いて多数ＳＮＰの測定を行う。次いで、この遺伝子データを本明細書中に記載されたシステムで用いる。

【００４５】

本発明のもう１つの態様において、双方の親からのジプロイドおよびハプロイドデータを用いて、個体の遺伝子データを清浄化できる。別法として、親からのハプロイドデータは、もし親のジプロイドおよびハプロイドデータを測定することができれば、シミュレートすることができる。もう１つの態様において、個体に対する公知の遺伝子関連のいずれかの個人からの遺伝子データを用いて、親、兄弟姉妹、祖父母、子孫、従兄弟、叔父、叔母などを含めた、個体のデータを清浄化することができる。

【００４６】

本発明のもう１つの態様において、標的および／または関連個体の遺伝子データはイン・シリコにて部分的にまたは全体的に知ることができ、いくつかの直接的測定の必要性を軽減する。遺伝子データの部分は、隠れたＭａｒｋｏｖモデルを利用するインフォーマティックスアプローチによってイン・シリコにて作り出すことができる。

【００４７】

本発明の１つの態様において、ＳＮＰの決定における信頼性を見積もることが可能である。

【００４８】

本明細書中に記載された技術は、１つの、または少数の細胞における遺伝物質の測定、ならびに非侵襲性出生前診断（ＮＩＰＤ）との関係で母親の血液から単離することができるもののようなより少量のＤＮＡについての測定の双方に関連することに注意されたし。また、この方法はイン・シリコでの、すなわち、遺伝物質から直接的に測定されないゲノムデータに等しく適応することができる。

【００４９】

本発明の１つの態様において、ＯＭＩＭ（男性におけるオンラインメンデル遺伝）データを介するように刊行物を介して、およびＨａｐＭａｐプロジェクトおよびヒトゲノムプロジェクトの他の態様から入手可能なデータを用いて入手可能なデータから構築することができる分割表に基づいてモデルを作り出すための技術が提供される。この技術のある実施形態は、モデルの予測的精度を改良するために、遺伝子の間の関連についての、および遺伝子および病気の間の関連についての出現する公のデータを用いる。

【００５０】

なおもう１つの態様において、最良のモデルを、特定の患者で利用できるデータで見出すことができる技術を開示する。この態様において、多くの異なるモデリング技術と共に、変数の多くの異なる組合せを調べることができ、他の対象からのテストデータと共に交差－確証に基づいて個々の対象についての最良の予測を生じるであろうその組合せを選択することができる。

【００５１】

いくつかの場合において、個体についての表現型の結果または表現型の感受性の正確な予測を行うにおいて最良のものを生じさせることができるモデルを、凸最適化技術を用いて訓練して、データの特定の組についての全体的に最適なパラメーターを見出すのが保証されるように、プレディクターの連続的サブセット選択を行う。この特徴は、モデルが複雑であり得、遺伝子突然変異または遺伝子発現レベルのような多くの潜在的プレディクターを含有することができる場合に特に有利である。さらに、いくつかの例においては、それらが単純な方法でデータを説明するように、凸最適化技術を用いて、モデルを希薄とすることができる。この特徴は、モデルにおける潜在的プレディクターの数が、訓練データにおける測定された結果の数と比較して大きい場合でさえ、訓練されたモデルが正確に一般化されるのを可能とする。同様な技術は学問的雑誌に公表されている（Ｒａｂｉｎｏｗｉｔｚ，Ｍ．ら，２００６，“Ａｃｃｕｒａｔｅ Ｐｒｅｄｉｃｔｉｏｎ ｏｆ ＨＩＶ－１ ｄｒｕｇ ｒｅｓｐｏｎｓｅ ｆｒｏｍ ｔｈｅ ｒｅｖｅｒｓｅ ｔｒａｎｓｃｒｉｐｔａｓｅ ａｎｄ ｐｒｏｔｅａｓｅ ａｍｉｎｏ ａｃｉｄ ｓｅｑｕｅｎｃｅ

ｓ ｕｓｉｎｇ ｓｐａｒｓｅ ｍｏｄｅｌｓ ｃｒｅａｔｅｄ ｂｙ ｃｏｎｖｅｘ ｏｐｔｉｍｉｚａｔｉｏｎ．”Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ ２２（５）：５４１－９）。この論文からの情報は背景および文脈のために本書類に含めてあることに注意されたし。

【００５２】

本明細書中に開示されたある説明的実施形態はヒト対象からの遺伝子データに焦点を当て、癌またはＨＩＶにかかった人々についての、またはアルツハイマー病または心筋梗塞のような病気に対する彼らの罹患性を理解したい人々についての特別な実施形態を提供するが、開示された方法は多数の異なる関係の範囲において生物の範囲の遺伝子データに適用されるのに注意すべきである。表現型予測および薬物応答予測について本明細書中に記載された技術は、種々の癌、遺伝子病、細菌、真菌またはウイルス感染の治療との関係で、並びに臨床的およびライフスタイルの決定を促進するために個体について表現型予測を行うにおいて関連し得る。さらに、該システムを用いて、遺伝子データ、具体的にはＩＶＦとの関係で胚（着床前）の、または羊水穿刺を含めた非侵襲性または侵襲性出生前診断との関係で胎児のＳＮＰ（単一ヌクレオチド多形）データを仮定し、特定の表現型結果の尤度を決定することができる。

【００５３】

１つの実施形態において、予測的モデルを、標準化された計算可能なフォーマットで貯蔵されている特定の個人についての遺伝子データに適用することができる。個人は、彼らに関連する特定の論点を記載することができ、あるいはシステムは、いずれの表現型罹患性がその個体が関連するかを自動的に決定することができる。新しい研究データが病気－遺伝子関連、治療、またはライフスタイルの嗜好性について入手できるようになるので、個体には、集合されたゲノムおよび臨床データから開発された予測的モデルに基づいて、彼らの決定および嗜好性についてのこの情報のインパクトを知らせることができる。別法として、該システムは新しい研究データを用いて、個体についての疑われていない危険性をここに検出することができ、その個体にはこの情報のインパクトを知らせることができる。

【００５４】

もう１つの実施形態において、遺伝子データ、表現型データおよび関連診断テストを含めた臨床記録のデータベースから一体化されたデータについて訓練された結果予測モデルを用いて臨床家のために増強された報告を作成することができる。このシステムは、限定されるものではないが、ＨＩＶ、癌、アルツハイマー病および心臓病を含めた、病気および／または病気素因を持つ個体についての増強された報告の創生を提供できる。この増強された報告は治療する医師に、いずれの病気－管理または予防的処置が与えられた個体についてより適当であるか、またはあまり適当でないであろうことを示すであろう。報告は、集合された対象データについて訓練されたモデルを用いるその個体についての鍵となる結果についての予測および信頼性限界を含むであろう。

【００５５】

もう１つの実施形態によると、特定の個体についてのデータを用いて、分割表に基づき、パブリックドメインで入手可能な情報から形成されたモデルを用いて該個体についての予測を行い、該データは該固体の遺伝子データ、該個体の表現型データ、および個体の臨床データ、およびその組合せよりなる群から取られ、ここで、該予測は該個体の表現型、表現型罹患性および可能な臨床的結果を含む群から取られたトピックスに関し、およびここで、該情報は、遺伝子型－表現型関連についての情報、ある遺伝子対立遺伝子の頻度についての情報、遺伝子対立遺伝子内のある関連の頻度についての情報、遺伝子対立遺伝子のある実施形態を仮定したある表現型の１以上の状態の確率についての情報、ある表現型の状態を仮定した遺伝子対立遺伝子のある組合せの確率についての情報、およびその組合せを含む群から取られるシステムおよび方法が開示される。

【００５６】

なおもう１つの実施形態によると、それにより、特定の個体についてのデータを用いて

、最良の精度を示すモデルを利用できるように集合データについて訓練された種々の数学的モデルを用い該個体についての予測を行うことができ、ここで、該個体のデータは該個体の遺伝子データ、該個体の表現型データ、および該個体の臨床的データよりなる群から取られ、およびここで、該予測は該個体の表現型、表現型罹患性、可能な臨床的結果、およびその組合せから取られるトピックスに関連するシステムおよび方法が提供される。ある実施形態において、該方法は、多数のモデルおよび多数のチューニングパラメーターを用いて、データの所与の組において異なる独立した変数および従属した変数の組合せの多くまたはすべてを調べることができ、次いで、最良の表現型予測を行う目的でテストデータにて最高の相関係数を達成した独立した変数および従属した変数およびその組合せ、そのモデルおよびそれらのチューニングパラメーターを選択する。

【００５７】

もう１つの実施形態によると、本明細書中に開示された方法のいずれも予測を用いて、該個体に関連する１以上のトピックスに関連する特定の個体についての報告を作成することができ、ここで、該トピックスはライフスタイルの決定、ダイエットの嗜好性、ホルモンサプリメント、病気についての可能な治療養生法、病原体に対する可能な治療養生法、薬物介入、およびその組合せを含む群から取られ、およびここで、該予測は該個体の遺伝子メイキャップ、該個体の表現型特徴、該個体の臨床的履歴およびその組合せに関連するデータに基づく。

【００５８】

他の実施形態によると、本明細書中に開示された方法のいずれも予測を用いて、医師または臨床家のような特定の個人の代理人のための報告を作成することができ、ここで、該予測は該個体に関連する情報を供することによって該代理人を助けることができ、およびここで、該情報の主題はライフスタイルの決定、ダイエットの嗜好性、ホルモンサプリメント、病気についての可能な治療養生法、病原体についての可能な治療養生法、薬物介入、他の治療的介入、およびその組合せを含むトピックスの群から取られ、およびここで、該予測は該個体の遺伝子メイキャップ、該個体の表現型特徴、該個体の臨床的履歴およびその組合せに関するデータに基づく。

【００５９】

もう１つの実施形態によると、本明細書中に開示された方法のいずれも予測を用いて、癌にかかった特定の個体に利益を与えることができ、およびここに該予測は、その個体および該個体の特定の癌に関連する情報を供することによって臨床家を助けることができ、およびここで、該情報の主題は治療養生法、ライフスタイルの決定、およびダイエットの嗜好性、薬物介入、他の治療的介入およびその組合せを含むトピックスの群から取られ、およびここで、該予測は該個体の遺伝子メイキャップ、該個体の表現型特徴、該個体の臨床的履歴、およびその組合せに関するデータに基づく。

【００６０】

１つの実施形態によると、本明細書中に開示された方法のいずれも、病原体に罹った特定の個体に利益を与えるために用いることができ、およびここで、該予測は、その個体、および該個体を感染する特定の病原体に関連する情報を供することによって臨床家を助けることができ、ここで、該病原体は細菌、ウイルス、微生物、アメーバー、真菌および他の寄生虫よりなる群から選択されるクラスのものであり、およびここで、該情報の主題は治療養生法、ライフスタイルの決定、およびダイエットの嗜好性、薬物介入、他の治療的介入、およびその組合せを含むトピックスの群から取られ、およびここで、該予測は該個体の遺伝子メイキャップ、該個体の表現型特徴、該個体の臨床的履歴、およびその組合せに関するデータに基づく。

【００６１】

もう１つの実施形態によると、本明細書中に開示された方法のいずれも、具体的な個体についての予測、新しい知識、およびデータを用いることができる。というのは、その知識およびデータは入手可能となるからであり、これを用いて、該個体に関連するトピックスについての、情報報告を自動的にまたは要求に応じて作成することができ、ここで、ト

ピックスはライフスタイルの決定、ダイエットの嗜好性、ホルモンサプリメント、病気についての可能な治療養生法、病原体についての可能な治療養生法、薬物の介入、他の治療的介入、およびその組合せを含む群から取られ、およびここで、新しい知識およびデータは性質において医療的であり、およびここで、該予測は、該個体の遺伝子のベーキャップ、該個体の表現型特徴、該個体の臨床的履歴、およびその組合せに関するデータに基づく。

【００６２】

もう１つの実施形態によると、本明細書中に開示された方法のいずれも、特定の胚からの遺伝子データを用いる予測を用いることができ、該予測を用いて、該胚のある表現型に対する予測された感受性に基づくＩＶＦの関係で胚の選択を助けることができる。

【００６３】

１つの実施形態によると、本明細書中に開示された方法のいずれも、特定の胎児からの遺伝子データを用いる予測を用いることができ、該予測を用いて、余命、乾癬の確率、または数学的能力の特定のレベルの確立のような、潜在的子孫についての特別な表現型の結果を見積もることができる。

【００６４】

この開示の利点を仮定すれば、他の態様、特徴および実施形態は本明細書中に開示された方法およびシステムの１以上を実施することができるのは当業者によって認識されるであろう。

例えば、本願発明は以下の項目を提供する。

（項目１）

（ｉ）関連個体からのいずれの染色体のいずれのセグメントが標的個体ゲノムで見出されるセグメントに対応するかに関する１以上の仮説のセットを創製し、

（ｉｉ）該標的個体遺伝子データの測定、および該関連個体遺伝子データの測定を仮定して該仮説の各々の確率を決定し、次いで、

（ｉｉｉ）各仮説に関連する確率を用いて、該標的個体の現実の遺伝物質の最もありそうな状態を決定する：

ことを含む、該標的個体の遺伝子データの不完全な知識、および該標的に遺伝的に関連する１以上の個体の遺伝子データの知識に基づいて該標的個体の遺伝子データを決定する方法。

（項目２）

前記方法が、前記標的の遺伝子データの測定、および親の遺伝子データを仮定した特定の測定の尤度の決定に基づいて、親染色体のいずれの領域が、標的個体に寄与した配偶子の形成に寄与した最大尤度を有するかを、決定することを含む、項目１記載の方法。

（項目３）

親の少なくとも１つのハプロタイプが、親のジプロイド試料から測定された遺伝子データ、およびジプロイド試料から測定されたいずれの対立遺伝子がいずれのハプロタイプに属するかを決定するのに用いられる親からのハプロイド試料から測定された遺伝子データを用いることによって決定されている、項目１記載の方法。

（項目４）

前記遺伝的に関連する個体からの遺伝子データが、ジプロイド母性試料、ハプロイド母性試料、ジプロイド父性試料およびハプロイド父性試料からの遺伝子データを含む群から選択される、項目１記載の方法。

（項目５）

清浄化された胚遺伝子データにおける個々のＳＮＰ要求の各々について信頼性が計算される、項目１記載の方法。

（項目６）

前記遺伝的に関連する個体からの遺伝子データが、ジプロイド母性細胞、ジプロイド父性細胞、ハプロイド父性細胞、母性祖父からのジプロイド細胞、および母性祖父からのハプロイド細胞からの遺伝子データを含む群から選択される、項目１記載の方法。

（項目７）
前記前記遺伝的に関連する個体からの遺伝子データが、ジプロイド母性細胞、ジプロイド父性細胞、および問題となる表現型のキャリアーであることが知られた関連個体からのジプロイド細胞からの遺伝子データを含む群から選択される、項目１記載の方法。
（項目８）
遺伝的に関連する個体が、父親、母親、息子、娘、兄弟、姉妹、祖父、祖母、叔父、叔母、甥、姪、孫息子、孫娘、従兄弟、クローン、前記標的に対する公知の遺伝的関係を持つ他の個体、およびその組合せよりなる群から選択される、項目１記載の方法。
（項目９）
前記標的個体が、成人ヒト、若年ヒト、ヒト胎児、ヒト胚、非ヒト成体、非ヒト若年体、非ヒト胎児、および非ヒト胚よりなる群から選択される、項目１記載の方法。
（項目１０）
前記個体の遺伝子データの１以上が、ポリメラーゼ鎖反応（ＰＣＲ）、リガンド媒介ＰＣＲ、縮重オリゴヌクレオチドプライマーＰＣＲ、多重置換増幅、対立遺伝子－特異的増幅技術、およびその組合せよりなる群から選択されるツールおよび／または技術を用いて増幅される、項目１記載の方法。
（項目１１）
前記個体の遺伝子データの１以上が、分子逆転プローブ（ＭＩＰ）、ゲノタイピングマイクロアレイ、Ｔａｑｍａｎ　ＳＮＰゲノタイピングアッセイ、Ｉｌｌｕｍｉｎａゲノタイピングシステム、他のゲノタイピングアッセイ、蛍光イン－サイチュハイブリダイゼーション（ＦＩＳＨ）、およびその組合せを含む群から選択されるツールおよび／または技術を用いて測定される、項目１記載の方法。
（項目１２）
前記個体の遺伝子データの１以上が、該個体のバルクジプロイド組織、該個体から取られた１以上のジプロイド細胞、該個体から取られた１以上の胚盤胞、該個体の精液、該個体の卵、該個体で見出される細胞外遺伝物質、母性血液で見出される該個体からの細胞外遺伝物質、母性血漿で見出される該個体からの細胞外遺伝物質、母性血液で見出される該個体からの細胞、該個体に由来することが知られている遺伝物質、およびその組合せを含む群から選択される物質を分析することによって測定される、項目１記載の方法。
（項目１３）
前記関連個体遺伝子データの１以上が、イン・シリコにて部分的にまたは全体的に知られているか、あるいは前記標的個体の遺伝子データを決定する以外の個人によって提供される、項目１記載の方法。
（項目１４）
前記個体の１以上のハプロイド遺伝子データが、ジプロイドデータからハプロイドデータをシミュレートするコンピュータアルゴリズムによってイン・シリコにて部分的にまたは全体的に創製される、項目１記載の方法。
（項目１５）
前記コンピュータアルゴリズムが隠れＭａｒｋｏｖモデルを含む項目１４記載の方法。
（項目１６）
前記標的遺伝子データの決定が、体外受精の関係で胚選択を目的として用いられる、項目１記載の方法。
（項目１７）
前記標的遺伝子データの決定が、出生前遺伝子診断の目的で用いられる、項目１記載の方法。
（項目１８）
前記標的遺伝子データの決定が、統計学的モデルおよび／または専門家則を用いて表現型罹患性の予測を行う目的で用いられる、項目１記載の方法。
（項目１９）
前記標的遺伝子データの決定が表現型予測を行う目的で用いられ、ここで、該表現型のい

くつかまたは全てを提示する尤度は、他の従前に知られた表現型情報によって影響される、項目１記載の方法。
（項目２０）
前記標的遺伝子データの決定が表現型予測を行う目的で用いられ、ここで、該予測は、該標的遺伝子データを、パブリックドメインで見出される公知の遺伝子マーカーと比較することによってなされる、項目１記載の方法。
（項目２１）
標的遺伝子データの決定が、臨床的決定を行う目的で用いられる、項目１記載の方法。
（項目２２）
標的遺伝子データの決定が、臨床的決定を行う目的で表現型マーカーと組合せて用いられる、項目１記載の方法。
（項目２３）
前記標的遺伝子データの決定が、１以上の病気に対する罹患性についてスクリーニングする目的で用いられ、ここで、家族の病歴が存在しない、項目１記載の方法。
（項目２４）
前記標的遺伝子データの決定が、１以上の表現型に対する罹患性についてスクリーニングする目的で用いられ、ここで、該表現型のいくつかまたは全てが多重遺伝子的である、項目１記載の方法。
（項目２５）
前記標的遺伝子データの知識が、汚染ＤＮＡまたはＲＮＡからの偽データを含有することが知られた、または含有することが疑われる、項目１記載の方法。
（項目２６）
前記個体の１以上の遺伝子データが、複数のＳＮＰについての対立遺伝子要求、および各ＳＮＰが知られている信頼性を含む、項目１記載の方法。
（項目２７）
前記標的個体のＳＮＰ要求における信頼性が、該ＳＮＰが正しくｖｓ正しくなく要求される確率のオッズ比を計算することによって決定される、項目１記載の方法。
（項目２８）
項目１記載の方法を達成するように構成されたシステム。
（項目２９）
項目１記載の方法を達成するように構成されたコンピュータ実施システム。
（項目３０）
（ｉ）標的個体のゲノムに存在する所与のセグメントの存在の数についての１以上の仮説のセットを創製し、
（ｉｉ）該所与のセグメント上の複数の遺伝子座における可能な対立遺伝子のいくつかまたは全てについての遺伝子データの量を測定し、
（ｉｉｉ）該標的個体の遺伝子データおよび、恐らくはまた、関連個体の遺伝子データの測定を仮定して該仮説の各々の相対的確率を決定し、次いで、
（ｉｖ）各仮説に関連する相対的確率を用いて、該標的個体の現実の遺伝物質の最もありそうな状態を決定する；
ことを含む、該標的個体の所与の染色体の所与のセグメント上の多数遺伝子座の測定を用いて、該標的個体のゲノム中の所与のセグメントの存在の数を決定する方法。
（項目３１）
該標的ゲノムに存在する染色体のセグメントの存在の数の決定が、染色体異常についてスクリーニングする関係で行われ、この異常は、モノソミー、片親ジソミー、トリソミー、他の異数性、アンバランスなトランスロケーション、およびその組合せを含むリストから選択される、項目３０記載の方法。
（項目３２）
各仮説の相対的確率の決定が、マッチドフィルタリングの概念を用いて行われる、項目３０記載の方法。

（項目３３）
各仮説の相対的確率の測定が、対立遺伝子要求を行わない定量的技術を用いてなされ、ここで、各遺伝子座の測定についての平均および標準偏差が既知、未知、または均一のいずれかである、項目３０記載の方法。
（項目３４）
各仮説の相対的確率の決定が、対立遺伝子要求を用いる定性的技術を用いてなされる、項目３０記載の方法。
（項目３５）
各仮説の相対的確率の決定が、参照配列の公知の対立遺伝子、および定量的対立遺伝子測定を用いることによってなされる、項目３０記載の方法。
（項目３６）
前記標的個体が、成人ヒト、若年ヒト、ヒト胎児、ヒト胚、非ヒト成体、非ヒト若年体、非ヒト胎児、および非ヒト胚よりなる群から選択される、項目３０記載の方法。
（項目３７）
前記標的個体の遺伝子データが、ポリメラーゼ鎖反応（ＰＣＲ）、リガーゼ媒介ＰＣＲ、縮重オリゴヌクレオチドプライマーＰＣＲ、多重置換増幅、対立遺伝子－特異的増幅およびその組合せを含む群から取られるツールおよび／または技術を用いて増幅される、項目３０記載の方法。
（項目３８）
前記標的個体の遺伝子データが、分子逆転プローブ（ＭＩＰ）、ゲノタイピングマイクロアレイ、Ｔａｑｍａｎ　ＳＮＰゲノタイピングアッセイ、Ｉｌｌｕｍｉｎａゲノタイピングシステム、他のゲノタイピングアッセイ、蛍光イン－サイチュハイブリダイゼーション（ＦＩＳＨ）、およびその組合せを含む群から選択されるツールおよび／または技術を用いて測定される、項目３０記載の方法。
（項目３９）
前記標的個体の遺伝子データが、該標的個体のバルクジプロイド組織、該標的個体から取られる１以上のジプロイド細胞、該標的個体から取られる１以上の胚盤胞、該標的個体上で見出された細胞外遺伝物質、母性血液で見出された該標的個体からの細胞外遺伝物質、母性血液で見出される該標的個体からの細胞、該標的個体に由来することが知られた遺伝物質、およびその組合せを含む群から取られる物質を分析することによって測定される、項目３０記載の方法。
（項目４０）
前記標的における染色体または染色体セグメントの数の決定が、体外受精の関係で胚選択を目的として用いられる、項目３０記載の方法。
（項目４１）
前記標的の染色体または染色体セグメントの数の決定が、出生前遺伝子診断の目的で用いられる、項目３０記載の方法。
（項目４２）
項目３０記載の方法を達成するように構成されたシステム。
（項目４３）
項目３０記載の方法を達成するように構成されたコンピュータ実施システム。
（項目４４）
（ｉ）関連個体からのいずれの染色体のいずれのセグメントが標的個体のゲノムで見出されるセグメントに対応するかについての１以上の仮説のセットを創製し、
（ｉｉ）該標的のゲノムに存在する所与の染色体セグメントの数についての１以上の仮説のセットを創製し、
（ｉｉｉ）該所与のセグメント上の複数の遺伝子座における可能な対立遺伝子の各々についてゲノムデータの量を測定し、
（ｉｖ）該標的個体の遺伝子データの測定、および該関連個体の遺伝子データの測定を仮定して仮説の各々の相対的確率を決定し、次いで、

（ｖ）各仮説に関連する相対的確率を用いて、該標的個体の現実の遺伝物質の最もありそうな状態を決定する；
ことを含む、該標的個体の遺伝子データの不完全な知識、および該標的に遺伝的に関連する１以上の個体の遺伝子データの知識に基づいて、該標的個体の遺伝子データ、および該標的ゲノムに存在する染色体、または染色体のセグメントの存在の数を決定する方法。
（項目４５）
（ｉ）遺伝子－病気関連についての公に入手可能な情報から形成された偶発事象表に基づいてモデルを構築し；次いで、
（ｉｉ）該モデルを適用して、個体に関連するデータに対して操作することによって予測を行う；
ことを含む、該個体に関連する予測を行う方法。
（項目４６）
多数の独立変数を使用する前記偶発事象表の精度が、結果データを用いて洗練することができ、ここで、独立変数のサブセットのみが測定される、項目４５記載の方法。
（項目４７）
多数の独立変数を使用する前記偶発事象表の精度が、前記独立変数の関連についてのデータを用いて洗練することができる、項目４５記載の方法。
（項目４８）
多数の独立変数を使用する前記偶発事象表の制度が、前記独立変数のある値の出現の頻度についてのデータを用いて洗練することができる、項目４５記載の方法。
（項目４９）
（ｉ）予測すべき結果が知られている個体の第二のセットからの集合データを用いて複数のモデルを創製し、それをテストし；
（ｉｉ）第一の個体で利用可能なデータを仮定した予測を行うための種々のモデルの相対的精度を計算し；次いで、
（ｉｉｉ）最も正確なものとして同定されるモデルを用いて、該第一の個体について予測を行う；
ことを含む、第一の個体に関する予測を行う方法。
（項目５０）
前記個体に関連するデータのタイプは、該個体の遺伝子型データ、該個体の表現型データ、該個体の臨床データ、および該個体の実験室データよりなる群から選択されるデータを含む、項目４５記載の方法。
（項目５１）
前記個体に関連するデータのタイプが、該個体の遺伝子型データ、該個体の表現型データ、および該個体の臨床データ、ならびに該個体の実験室データよりなる群から選択されるデータを含む、項目４９記載の方法。
（項目５２）
前記データのタイプが、また、前記個体を感染させる病原体のデータよりなる、項目４５記載の方法。
（項目５３）
前記データのタイプが、また、前記個体を感染させる病原体のデータよりなる、項目４９記載の方法。
（項目５４）
前記予測が、前記個体の表現型、表現型罹患性、可能な臨床的結果、ライフスタイルの決定、身体の運動、ダイエットの嗜好性、ホルモンサプリメント、栄養サプリメント、病気のための治療、病原体のための処理、望まない疾患についての治療、医薬での治療、およびその組合せよりなる群から選択されるトピックに関する、項目４５記載の方法。
（項目５５）
前記予測が、前記個体の表現型、表現型罹患性、可能な臨床的結果、ライフスタイルの決定、身体の運動、精神的運動、ダイエット嗜好性、ホルモンサプリメント、栄養サプリメ

ント、病気についての治療、病原体についての処理、望ましくない疾患についての治療、医薬での治療、およびその組合せよりなる群から選択されるトピックスに関する、項目４９記載の方法。

（項目５６）

前記予測を用いて、前記個体のための、または該個体の代理人のための報告を作成する、項目４５記載の方法。

（項目５７）

前記予測を用いて、前記個体のための、または該個体の代理人のための報告を作成する、項目４９記載の方法。

（項目５８）

前記操作の行為が、新しいデータについて操作して、前記個体の予測を更新することを含み、ここで、該データは新しい研究データ、または他の対象についての新しい集合データを含む群から選択される、項目４５記載の方法。

（項目５９）

前記操作の行為が、新しいデータについて操作して、前記個体の予測を更新することを含み、ここで、該データは新しい研究データまたは他の対象についての新しい集合データを含む群から選択される、項目４９記載の方法。

（項目６０）

項目４５記載の方法を達成するように構成されたシステム。

（項目６１）

項目４９記載の方法を達成するように構成されたシステム。

【図面の簡単な説明】

【００６５】

【図１】配偶子形成についての減数分裂における組換えの概念の説明図。

【図２】ヒト染色体１の１つの領域に沿っての組換えの可変速度の説明図。

【図３】異なる仮定に対する偽陰性および偽陽性の確率の決定。

【図４】混合された女性試料、全てのヘテロ遺伝子座からの結果。

【図５】混合された男性試料、全てのヘテロ遺伝子座からの結果。

【図６】女性試料についてのＣｔ測定とは異なる男性試料についてのＣｔ測定。

【図７】混合された女性試料からの結果；Ｔａｑｍａｎ単一色素。

【図８】混合された男性試料からの結果；Ｔａｑｍａｎ単一色素。

【図９】混合された男性試料についての反復測定の分布。

【図１０】混合された女性試料からの結果；ｑＰＣＲ尺度。

【図１１】混合された男性試料からの結果；ｑＰＣＲ尺度。

【図１２】女性試料についてのＣｔ測定とは異なる男性試料についてのＣｔ測定。

【図１３】第三の似ていない染色体での異数性の検出。

【図１４】定常対立遺伝子ドロップアウト速度での２つの増幅分布の説明図。

【図１５】アルファのガウス確率密度関数のグラフ。

【図１６】入力データ、データベースデータ、アルゴリズムおよび出力の一般的な関係のダイヤグラム。

【図１７】Ｐ（Ｈ｜Ｍ）をどのように駆動するかの視覚的概観。

【図１８】シミュレートされたデータについての清浄化アルゴリズムの有効性を示すのに用いられるアルゴリズムを記載するフローチャートの視覚的表示。

【図１９】ＩＶＦの間における胚の表現型予測の関係での、本明細書中に開示された方法を達成するように構成されたシステムの説明図。

【図２０】疎な解を生じるＬＡＳＳＯ傾向の説明図。Ｒｉｄｇｅ回帰解は２つの円の接合に存在し、ＬＡＳＳＯ解は円および四角形の接合に存在する。

【図２１】訓練およびテストデータの１０の異なる９：１スプリットにわたって平均し、次いで、各々、７つのＰＩまたは１０のＲＴＩにわたって平均した、測定したおよび予測した応答の相関係数（％で表したＲ）の表。

【図２２】ＰＩ応答を予測するためのプロテアーゼ酵素における突然変異に関連するＬＡＳＳＯモデルパラメーターの値のグラフ表示。最大の絶対的大きさを持つ４０のパラメーターのみを示す。
【図２３】ＮＲＴＩ薬物応答を予測するためのＲＴ酵素における突然変異に関連するＬＡＳＳＯモデルパラメーターの値のグラフ表示。最大の絶対的大きさを持つ４０のパラメーターのみを示す。
【図２４】ＮＮＲＴＩ薬物応答を予測するためのＲＴ酵素における突然変異に関連するＬＡＳＳＯモデルパラメーターの値のグラフ表示。最大の絶対的大きさを持つ４０のパラメーターのみを示す。
【図２５】記載なし。
【図２６】記載なし。
【図２７】記載なし。
【図２８】記載なし。
【００６６】
表１：ＯＭＩＭ／ＮＣＢＩに見出される病気遺伝子のまとめ。
表２：異なる異数性検出技術のまとめ。
表３：低度な共分離を持つＳＮＰを用いて記載された方法についての入力データの例。
表４：高度な共分離を持つＳＮＰを用いて記載された方法についての入力データの例。
表５：表２に示された入力データに代えての出力データの例。
表６：表４に示された入力データに代えての出力データの例。
表７：予備的シミュレーションの結果。
表８：方法の全シミュレーションの結果。
表９：アルツハイマー病の開始への影響におけるＡＰＯＥおよびＡＣＥにおける突然変異の役割を理解するためのＦａｒｒｅｒ（２００５）、Ｌａｂｅｒｔ（１９９８）、およびＡｌｖａｒｅｚ（１９９９）の結果を表す３つの分割表。
表１０：表７の実験のメター分析から生じた結果。
表１１：訓練およびテストデータの１０の異なる９：１スプリットにわたって平均した、種々の方法についてのプロテアーゼ阻害剤（ＰＩ）薬物に対する測定されたおよび予測された応答の相関係数（％で表したＲ）の表。結果の標準偏差（Ｓｔｄ．ｔｅｖ．）は灰色で示す；測定された薬物応答の数は最後の列に示す。
法１２：訓練およびテストデータ１０の異なる９：１スプリットにわたって平均された、種々の方法についての逆転写酵素阻害剤（ＲＴＩ）薬物に対する測定されたおよび予測された応答の相関係数（％で表したＲ）の表。結果の標準偏差（Ｓｔｄ．ｄｅｖ．）は灰色で示す；測定された薬物応答の数は最後の列に示す。
表１３：プロテアーゼ阻害剤（ＰＩ）薬物応答についてのプレディクターとしての最小絶対選択および収縮オペレーター（ＬＡＳＳＯ）によって選択された非ゼロ重みを持つ突然変異の数と共に、種々の回帰方法についての訓練で用いられる試料の数、および突然変異の合計数。
表１４：逆転写酵素阻害剤（ＲＴＩ）応答についてのプレディクターとしてのＬＡＳＳＯによって選択された非ゼロ重みを持つ突然変異の数と共に、種々の方法での訓練で用いられる試料の数、および突然変異の合計数。
表１５：イリノテカン実験についての表現型データ。
【発明を実施するための形態】
【００６７】
（好ましい実施形態の詳細な説明）
システムの概念的概観
開示されたシステムの１つの目標は、遺伝子診断の目的の高度に正確なゲノムデータを提供することである。個体の遺伝子データが有意な量のノイズ、またはエラーを含有する場合、開示されたシステムは、関連個体の遺伝子データ、およびその第二の遺伝子データに含まれる情報の間の同様性を用いて、標的ゲノムにおけるノイズを清浄化する。これは

、染色体のいずれのセグメントが配偶子形成に関与し、およびどこで減数分裂の間に交差が起こったか、従って、第二のゲノムのいずれのセグメントが標的ゲノムのセクションに対してほとんど同一であると予測されるかを決定することによってなされる。ある状況においては、この方法を用いてノイジーな塩基対測定を清浄化することができるが、それを用いて、測定されなかったＤＮＡの個々の塩基対または全領域の同一性を推定することもできる。加えて、なされた各再構成要求について信頼性を計算することができる。高度に単純化された説明を最初に示し、非現実的な仮定をなして、本発明の概念を説明する。今日の技術に適用することができる詳細な統計学的アプローチを以後示す。

【００６８】

システムのもう１つの目標は、染色体の異常な数、染色体のセクション、および染色体の起源を検出することにある。異数性であり、アンバランスなトランスロケーション、片親二染色体、または他の正味の染色体異常を有する一般的試料において、複数の遺伝子座に存在する遺伝物質の量を用いて、試料の染色体状態を決定することができる。この方法に対して多数のアプローチが存在し、それらのうちいくつかをここに記載する。いくつかのアプローチにおいて、試料に存在する遺伝物質の量は、異数性を直接的に検出するのに十分である。他のアプローチにおいて、遺伝物質を清浄化する方法を用いて、染色体不均衡の検出の効率を増強させることができる。なされた各染色体要求に対して信頼性を計算することができる。

【００６９】

該システムのもう１つの目標は、遺伝子データに関連する変数の効果をモデル化するように設計された項目の広いアレイを開発することによって、遺伝子データから最も単純かつ触知可能な統計学的モデルを抽出する有効かつ効果的手段を提供することにある。より具体的には、遺伝子データに基づいて表現型または表現型感受性をモデル化するための現在利用可能な方法のほとんどまたは全ては以下の欠点を有する：（ｉ）それらは凸最適化技術を用いず、かくして、所与の訓練データセットに対するモデルパラメーターについての局所的最小解を見出すことは保証されない；（ｉｉ）それらはモデルの複雑性を最小化する技術を用いず、かくして、それらは、独立した変数の数に対して少数の結果が存在する場合に十分に一般化されるモデルを形成しない；（ｉｉｉ）それらは、正規分布したデータの単純化仮定をなすことなく、論理的回帰の関係でデータからの最も単純な触知のルールの抽出を可能とせず；（ｉｖ）それらは遺伝子－遺伝子関連、遺伝子－表現型関連および遺伝子－病気関連についての先見的情報を活用して、表現型または表現型感受性の最良の可能な予測をしない；（ｖ）それらは１を超えるモデルを提供せず、かくして、訓練データに対する種々のモデルの交差－確証に基づいて最良の可能なデータを選択するための一般的アプローチを提供しない。これらの欠点は、遺伝子および表現型情報に関連する多量のデータクラスの分析に基づいて結果を予測する関係で臨界的である。まとめると、現在利用可能な方法は個体が遺伝子型が所与の特定の表現型特徴の尤度についての、または親の遺伝子型特徴を仮定した子孫における特定の表現型特徴の尤度についての質問に答えるのに効果的に力を与えない。

【００７０】

以下に掲げる説明のいくつかは、本書類の著者によって従前に公表された仕事を含むことに注意されたし。それは背景情報として提供されて、本明細書中に開示された材料の理解を容易とし、および該材料に対するより大きな関係を与える。

【００７１】

３つのカテゴリーにおいて遺伝子型－表現型予測モデルを考慮することができる：ｉ）遺伝子欠陥または対立遺伝子は１００％の確実性でもって病気表現型を引き起こすことが知られている；ｉｉ）病気表現型の確率を増加させる遺伝子欠陥および対立遺伝子、ここで、プレディクターの数は表現型確率を分割表でモデル化できるのに十分に小さい；およびｉｉｉ）多次元線形または非線形回帰モデルを用いて表現型を予測するのに用いることができる遺伝子マーカーの複雑な組合せ。オンラインメンデル遺伝データベース（Ｏｎｌｉｎｅ　Ｍｅｎｄｅｌｉａｎ　Ｉｎｈｅｒｉｔａｎｃｅ　Ｄａｔａｂａｓｅ（ＯＭＩＭ）

）における現在知られている配列および病気表現型を持つ３５９の遺伝子（表１、列２参照）のうち、大部分はカテゴリー（ｉ）に入り；残りは圧倒的にカテゴリー（ｉｉ）に入る。しかしながら、経時的に、多数の遺伝子型－表現型モデルがカテゴリー（ｉｉｉ）において生起していると予測され、ここで、多数の対立遺伝子または突然変異の相互作用は、特定の表現型の確率を見積もるためにモデル化される必要があろう。例えば、シナリオ（ｉｉｉ）は、確実に、今日、ＨＩＶウイルスの遺伝子データに基づいて抗－レトロウイルス療法に対するＨＩＶウイルスの応答を予測する関係で当てはまる。

【００７２】

シナリオ（ｉ）については、経験則に基づいて表現型の発生を予測するのは通常直接的である。１つの態様において、シナリオ（ｉｉ）について表現型の正確な予測をなすのに用いることができる統計的技術が記載されている。もう１つの態様において、シナリオ（ｉｉｉ）について正確な予測を行うのに用いることができる統計学的技術が記載されている。もう１つの態様において、特定の表現型、集合データの特定の組、および特定の個々のデータについて最良のモデルを選択することができる方法が記載されている。

【００７３】

本明細書中に開示された方法のある実施形態は、分割表を実行して、シナリオ（ｉｉ）において正確に予測を行う。これらの技術は遺伝子－遺伝子関連および遺伝子－病気関連についての先見的情報を活用して、表現型または表現型感受性の予測を改良する。これらの技術は、関連した独立変数の全てがサンプリングされるのではない従前の実験からのデータを活用するのを可能とする。それらが失われたデータを有するという理由でこれらの従前の結果を捨てる代わりに、概技術はＨａｐＭａｐプロジェクトおよびその他からのデータを活用して、関連する独立変数のサブセットのみが測定された従前の実験を用いる。このように、全ての関連する独立した変数が測定された対象からのデータを単純に集合させるよりはむしろ、予測モデルを全ての集合データに基づいて訓練することができる。

【００７４】

本明細書中に記載されたある方法は凸最適化を用いて、シナリオ（ｉｉｉ）において正確な予測をなすのに用いることができる疎なモデルを創製する。遺伝子型－表現型モデリングの問題はしばしば過剰決定系であるか、または不適切である。というのは、潜在的プレディクター－遺伝子、蛋白質、突然変異およびそれらの相互作用－の数は、測定された結果の数に対して大きいからである。そのようなデータのセットは、依然として、Ｏｃｃａｍ’ｓ　Ｒａｚｏｒと同様な原理を発見することによって正確に一般化される疎なパラメーターモデルを訓練するのに用いることができる。多くの可能な理論が観察を説明することができる場合、最も単純なのは最も正しいらしいものである。この哲学は、先に議論したシナリオ（ｉｉｉ）において遺伝子型－表現型モデルの形成に関連する１つの態様において具体化される。遺伝子データへの適用について本明細書中に記載された技術は、過少判断されたまたは誤って条件付けされた遺伝子型－表現型データセットについて疎なパラメーターモデルを創製することを含む。疎なパラメーターセットの選択はＯｃｃａｍ’ｓ　Ｒａｚｏｒと同様な原理を発揮し、結果として、潜在的プレディクターの数が測定された結果の数に対して大きい場合でさえ、正確なモデルが開発されるのを可能とする。加えて、シナリオ（ｉｉｉ）において遺伝子型－表現型モデルを形成するための本明細書中に記載された技術のある実施形態は、所与の訓練データセットについてのモデルパラメーターに対する全体的最小解を見出すことが保証された凸最適化技術を用いる。

【００７５】

集合データのセット、および個体についての入手可能なデータのセットを仮定すれば、その個体についての最良な表現型予測を行うために、いずれの予測アプローチが最も適当であるかは稀にしか明瞭でない。正確な表現型予測を行う傾向があるモデルのセットを記載することに加えて、本明細書中に開示された実施形態は、多数の方法をテストし、所与の表現型予測についての最適方法、集合データの所与のセット、および予測がなされるべき個体についての入手可能なデータの所与の組を選択するシステムを代表する。開示された方法およびシステムは、多重モデルおよび多重訓練パラメーターを用いるデータの所与

のセットにおける全ての異なる独立した変数および従属する変数の組合せを調べ、次いで、独立した変数、従属した変数、およびテストデータで測定された最良のモデリング精度を達成するチューニングパラメーターの組を選択する。シナリオ（ｉ）に対応する場合には、専門家則を立案することができ；カテゴリー（ｉｉ）におけるような少数の独立した変数での他の場合には、分割表は最良の表現型予測を提供し；およびシナリオ（ｉｉｉ）のような他の場合には、線形または非線形回帰技術を用いて、予測の最適な方法を提供することができる。本開示を読んだ後には、個体について予測をなすための最良のモデルを選択するアプローチをどのようにして用いて、本明細書中に開示されたものを超えて多くのモデリング技術から選択することができるかは当業者に明瞭であろうことを注記する。

【００７６】

技術のある実施形態はいくつかの関係で示されている。まず、それは、分割表、および遺伝子マーカーに基づいて、アルツハイマー病の予測に焦点を当てる多くの臨床的実験から一体化されたデータの不完全な組を用いてアルツハイマー病を発生する尤度を予測する関係で示されている。次に、該システムは、回帰分析、およびウイルスゲノムにおける遺伝子マーカーの知識を用いて１型ヒト免疫不全ウイルス（ＨＩＶ－１）の薬物応答をモデル化する関係で示されている。最後に、該システムは、各々、回帰分析、および個体についての双方の遺伝子マーカーの不完全なデータ、および癌に関連する実験質的および臨床的対象情報を用いる、乳癌および結腸癌の種々の症例の治療におけるタモキシフェンおよびイリノテカンの用法によって引き起こされる副作用の予測の点で示されている。

【００７７】

遺伝子型テストの減少する費用のため、信頼性よくウイルス薬物応答、癌薬物応答、および他の表現型応答または遺伝子データからの結果を予測する統計学的モデルは、それらが病気治療、ライフスタイルまたは嗜好性決定、または他の活動であるか否かを問わず適当な作用のコースの選択において重要なツールである。記載された最適化技術は、臨床的決定を増強させる目的で多くの遺伝子型－表現型モデリングの問題に応用を有するであろう。

【００７８】

システムの技術的記載

データの清浄化：単純化された例

図１は、親における配偶子の形成について減数分裂の間に起こる組換えのプロセスを説明する。個体の母親からの染色体１０１はオレンジ色（または灰色）で示す。個体の父親からの染色体１０２は白色で示す。減数分裂の前相Ｉの間の複糸期として知られたこの間隔の間に、４つの染色分体１０３のテトラドが目に見える。相同対の非姉妹染色分体の間の交差は組換え小節１０４として知られた地点で起こる。説明の目的で該例は単一の染色体、および３つの遺伝子の対立遺伝子を特徴付けると推定される３つの単一ヌクレオチド多形（ＳＮＰ）に焦点を当てる。この議論では、ＳＮＰは母性および父性染色体上で別々に測定することができると仮定する。この概念は多くのＳＮＰ、多数のＳＮＰによって特徴付けられる多くの対立遺伝子、多くの染色体、および母性および父性染色体をゲノタイピング前には個々に単離することができない現行のゲノタイピング技術に適用することができる。

【００７９】

注目するＳＮＰの間における潜在的交差の地点に注意を払わなければならない。３つの母性遺伝子の対立遺伝子のセットは、ＳＮＰ（SNP_1，SNP_2，SNP_3）に対応する（a_{m1}，a_{m2}，a_{m3}）として記載することができる。３つの父性遺伝子の対立遺伝子のセットは（a_{p1}，a_{p2}，a_{p3}）として記載することができる。図１において形成された組換え小節をコードし、組換え染色分体の各対についてちょうど１つの組換えがあると仮定する。このプロセスで形成された配偶子のセットは遺伝子対立遺伝子：（a_{m1}，a_{m2}，a_{p3}）、（a_{m1}，a_{p2}，a_{p3}）、（a_{p1}，a_{m2}，a_{p3}）、（a_{p1}，a_{p2}，a_{m3}）を有するであろう。染色分体の交差がない場合において、配偶子は対立遺伝子（a_{m1}，a_{m2}，a_{m3}）、（a_{p1}，a_{p2}，a_{p3}）を有するで

あろう。関連領域において交差の２つの地点がある場合において、配偶子は対立遺伝子（a_{m1}，a_{p2}，a_{m3}）、（a_{p1}，a_{m2}，a_{p3}）を有するであろう。対立遺伝子のこれらの８つの異なる組合せを、その特定の親について、対立遺伝子の仮説セットという。

【００８０】

胚ＤＮＡからの対立遺伝子の測定はノイジーであろう。この議論の目的では、胚ＤＮＡからの単一染色体を取り、それが、その減数分裂を図１で説明する親に由来すると仮定する。この染色体上の対立遺伝子の測定は、もし胚染色体における測定された対立遺伝子がa_{m1}であれば$A_1=1$であり、もし胚染色体における測定された対立遺伝子がa_{p1}であれば$A_1=-1$であって、もし測定された対立遺伝子a_{m1}またはa_{p1}でなければ$A_1=0$であるインジケーター変数のベクトルの項：$A=[A_1\ A_2\ A_3]^T$で記載することができる。推定親についての対立遺伝子の仮説セットに基づき、前記したすべての可能な配偶子に対応する８つのベクトルのセットを作り出すことができる。前記した対立遺伝子については、これらのベクトルは$a_1=[1\ \ 1\ \ 1]^T$、$a_2=[1\ \ 1\ \ -1]^T$、$a_3=[1\ \ -1\ \ 1]^T$、$a_4=[1\ \ -1\ \ -1]^T$、$a_5=[-1\ \ 1\ \ 1]^T$、$a_6=[-1\ \ 1\ \ -1]^T$、$a_7=[-1\ \ -1\ \ 1]^T$、$a_8=[-1\ \ -1\ \ -1]^T$となろう。システムのこの高度に単純化された適用において、胚のありそうな対立遺伝子は、仮説セットおよび測定されたベクトルの間の単純な相関分析を行うことによって決定することができる：

$$i^*=\arg\max_i A^T a_i,\quad i=1...8 \qquad (1)$$

一旦i^*が見出されれば、仮説

【００８１】

【数１】

$$a_{i^*}$$

が胚ＤＮＡにおける対立遺伝子の最もありそうなセットとして選択される。次いで、２つの異なる仮定、すなわち、胚染色体は母親または父親に由来するという仮定を立て、このプロセスを２回反復する。最大の相関

【００８２】

【数２】

$$A^T a_{i^*}$$

を生じるその過程は正しいと仮定されるであろう。各場合において、母親または父親の各ＤＮＡの測定に基づき、対立遺伝子の仮説セットを用いる。開示された方法の典型的な実施形態においては、特定の病気表現型とのその関連のため重要であるＳＮＰの間の多数のＳＮＰを測定し－これらは表現型－関連ＳＮＰまたはＰＳＮＰといわれるであろうことに注意されたし。ＰＳＮＰの間の非表現型－関連ＳＮＰ（ＮＳＮＰ）は、個体の間で実質的に異なる傾向があるＲｅｆＳＮＰをＮＣＢＩ　ｄｂＳＮＰデータベースから選択することによって、（例えば、特殊化されたゲノタイピングアレイを開発するための）先見的に選択することができる。別法として、ＰＳＮＰの間のＮＳＮＰは親の特定の対について選択することができる。なぜならばそれらは親の間で異なるからである。ＰＳＮＰの間のさらなるＳＮＰの使用は、交差がＰＳＮＰの間で起こるか否かをより高いレベルの信頼性でもって決定することを可能とする。異なる「対立遺伝子」をこの注記において言及するが、これは単に便宜的なものであり；ＳＮＰは蛋白質をコードする遺伝子には関連しないであろうことに注意するのは重要である。

【００８３】

現行の技術との関連でのシステム

もう１つのより複雑な実施形態において、特定の交差の確率を考慮して、対立遺伝子の事後確率を特定の測定を仮定して計算する。加えて、マイクロアレイに典型的なシナリオおよび他のゲノタイピング技術をアドレスし、ここで、ある時点で単一の染色体についてよりはむしろ染色体の対についてＳＮＰを測定する。胚、父性および母性染色体について

の遺伝子座ｉにおける遺伝子型の測定は、各々、ＳＮＰ測定の対を表すランダム変数（$e_{1,i}$，$e_{2,i}$）、（$p_{1,i}$，$p_{2,i}$）および（$m_{1,i}$，$m_{2,i}$）によって特徴付けることができる。もしすべての測定が対としてなされるならば、母性および父性染色体における交差の存在を決定することができないので、該方法は修飾される：受精胚および父性および母性ジプロイド組織を遺伝子型分けするに加えて、各親からの１つのハプロイド細胞、すなわち、精子細胞および卵細胞も遺伝子型分けする。精子細胞の測定された対立遺伝子は$p_{1,i}$，ｉ＝１．．．Ｎによって表され、父性ジプロイド組織から測定された相補的対立遺伝子は$p_{2,i}$によって表される。動揺に、卵細胞の測定された対立遺伝子は$m_{1,i}$によって表され、母親のジプロイド細胞におけるそれらの相補体は$m_{2,i}$によって表される。これらの測定は、どこで親染色体が測定された精子および卵細胞を生じるかにおいて交差したかについての情報を提供しない。しかしながら、卵または精子上のＮ個の対立遺伝子の配列は少数の交差によって、または交差なしによって、親染色体から作り出されたと仮定することができる。これは開示されたアルゴリズムを適用するための十分な情報である。あるエラーの確率は、父性および母性ＳＮＰの要求に関連する。このエラーの確率の見積もりは、なされた測定（$p_{1,i}$，$p_{2,i}$）および（$m_{1,i}$，$m_{2,i}$）、および用いる技術についてのシグナル－対－ノイズ比率に基づいて変化するであろう。これらのエラーの確率は、開示された方法に影響することなく各遺伝子座についてユニークに計算することができるが、父性および母性ＳＮＰを正しく要求する確立は、各々、p_pおよびp_mにおいて一定であると仮定することによってここでは代数は単純化される。

【００８４】

測定は、測定Ｍという胚ＤＮＡで行われると仮定する。加えて、Ａが今やセットであって、ベクトルではないように、表記法をわずかに修飾する：Ａとは、各親に由来する対立遺伝子の組合せ（またはセット）についての特定の仮説をいう。双方の親からの対立遺伝子Ａのすべての可能な実施形態のセットをS_Aとして示す。目標は、測定Ｍを与えて、最大の事後確率でもって、対立遺伝子の組合せ（またはその仮説）$A \in S_A$を決定することである：

$$A^* = arg\ max_A P(A|M),\ \forall A \in S_A \qquad (2)$$

条件付き確率の法則を用い、Ｐ（Ａ｜Ｍ）＝Ｐ（Ｍ｜Ａ）Ｐ（Ａ）／Ｐ（Ｍ）である。Ｐ（Ｍ）はすべての異なるＡについて共通するので、最適化サーチを：

$$A^* = arg\ max_A P(M|A)P(A),\ \forall A \in S_A \qquad (3)$$

として書き換えることができる。

【００８５】

今や、Ｐ（Ｍ／Ａ）の計算を考える。単一の遺伝子座ｉで開始し、胚上のこの遺伝子座は親ＳＮＰ $p_{t,1,i}$および$m_{t,1,i}$に由来すると仮定し、ここで、下付文字$_t$は、正しくても正しくなくてもよい行われた測定$p_{1,i}$および$m_{1,i}$とは反対に、これらの親ＳＮＰの真の値を示すのに用いられる。胚ＳＮＰの真の値は（$e_{t,1,i}$，$e_{t,2,i}$）として示される。もし仮説Ａが真であれば、（$e_{t,1,i}$，$e_{t,2,i}$）＝（$p_{t,1,i}$，$m_{t,1,i}$）または（$m_{t,1,i}$，$p_{t,1,i}$）である。測定（$e_{1,i}$，$e_{2,i}$）のいずれが、いずれの親に由来するかを区別できないので、双方の順番を考慮しなければならず、従って、仮説セットＡ＝［（$p_{t,1,i}$，$m_{t,1,i}$），（$m_{t,1,i}$，$p_{t,1,i}$）］となる。特定の測定Ｍの確率は、親ＳＮＰの真の値または基礎となる状態、すなわち、（$p_{t,1,i}$，$p_{t,2,i}$）および（$m_{t,1,i}$，$m_{t,2,i}$）に依存する。４つのＳＮＰ、$p_{t,1,i}$、$p_{t,2,i}$、$m_{t,1,i}$、$m_{t,2,i}$が存在し、かつこれらの各々は４つのヌクレオチド塩基Ａ、Ｃ、Ｔ、Ｇの値を取ることができるので、4^4または２５６の可能な状態が存在する。$p_{t,1,i} \neq p_{t,2,i} \neq m_{t,1,i} \neq m_{t,2,i}$であると仮定される１つの状態$s_1$についてアルゴリズムを説明する。この説明から、すべての２５６の可能な状態、s_k、ｋ＝１．．．２５６にどのようにして該方法を適用するかは明瞭であろう。胚ＳＮＰ（$e_{1,i}$，$e_{2,i}$）の測定Ｍを行い、結果$e_{1,i}=p_{1,i}$、$e_{2,i}=m_{1,I}$が

得られると仮定する。その仮説Ａおよび状態s_1が真実であるとしたこの測定についての事前確率を計算する：

【００８６】

【数３】

$$P(e_{1,i}=p_{1,i},e_{2,i}=m_{1,i}\mid A,s_1)=$$
$$P(e_{t,1,i}=p_{t,1,i},e_{t,2,i}=m_{t,1,i}\mid A,s_1)P(e_{1,i}=p_{1,i}\mid e_{t,1,i}=p_{t,1,i})P(e_{2,i}=m_{1,i}\mid e_{t,2,i}=m_{t,1,i})$$
$$+P(e_{t,1,i}=m_{t,1,i},e_{t,2,i}=p_{t,1,i}\mid A,s_1)P(e_{1,i}=p_{1,i}\mid e_{t,1,i}=m_{t,1,i},p_{t,1,i}\neq m_{t,1,i})P(e_{2,i}=m_{1,i}\mid e_{t,2,i}=p_{t,2,i},p_{t,2,i}\neq m_{t,1,i}) \quad (4)$$

第一項および第二項における最初の表現：Ｐ（$e_{1,i}=p_{1,i}$，$e_{2,i}=m_{1,i}$｜Ａ，s_1）＝Ｐ（$e_{1,i}=m_{1,i}$，$e_{2,i}=p_{1,i}$｜Ａ，s_1）＝０．５を考える。というのは、仮説Ａ＝［（$p_{t,1,i}$，$m_{t,1,i}$），（$m_{t,1,i}$，$p_{t,1,i}$）］は胚ＳＮＰについての２つの順序付けを等しくありそうとするからである。さて、第一項の第二の表現Ｐ（$e_{1,i}=p_{1,i}\mid e_{t,1,i}=p_{t,1,i}$）を考え、これは、胚ＳＮＰ $e_{t,1,i}$は現実には父性ＳＮＰ $p_{t,1,i}$.に由来すると仮定して$e_{1,i}=p_{1,i}$を測定する確率である。父性ＳＮＰ、母性ＳＮＰおよび胚ＳＮＰを正しく測定する確率はp_p，p_m，およびp_eである。仮定（$e_{t,1,i}=p_{t,1,i}$）を与えれば、測定（$e_{1,i}=p_{1,i}$）は、胚および父性ＳＮＰの双方が正しく測定されるか、あるいは双方は正しくなく測定され、それらは偶然に同一ヌクレオチド（Ａ，Ｃ，Ｔ，またはＧ）として正しくなく測定される、のいずれかを要求する。従って、Ｐ（$e_{1,i}=p_{1,i}\mid e_{t,1,i}=p_{t,1,i}$）＝$p_e p_p$＋（１－$p_e$）（１－$p_p$）／３であり、ここで、単純性のために、４つのヌクレオチドのすべてを正しくなく要求する確立は同等にありそうであると仮定される－該アルゴリズムは、もう１つの特定のヌクレオチドについての測定を与えて特定のヌクレオチド（Ａ，Ｃ，Ｔ，Ｇ）を要求する異なる確率を適合させるように容易に修飾することができる。同一アプローチを第一項中の３番目の表現に適用して、Ｐ（$e_{2,i}=m_{1,i}\mid e_{t,2,i}=m_{t,1,i}$）＝$p_e p_m$＋（１－$p_e$）（１－$p_m$）／３を得ることができる。さて、第二項の２番目の表現を考える。Ｐ（$e_{1,i}=p_{1,i}\mid e_{t,1,i}=m_{t,1,i}$，$m_{t,1,i}\neq p_{t,1,i}$）は、$e_{1,i}$または$p_{1,i}$が正しくない測定であるか、または双方が正しくない測定であるかのいずれかを要求し、従って、測定された値は偶然に等しい：Ｐ（$e_{1,i}=p_{1,i}\mid e_{t,1,i}=m_{t,1,I}$，$m_{t,1,i}\neq p_{t,1,i}$）＝$p_e$（１－$p_p$）／３＋（１－$p_e$）$p_p$／３＋（１－$p_e$）（１－$p_p$）２／９。同一の議論を第二項の最後の表現に適用して、Ｐ（$e_{2,i}=m_{1,i}\mid e_{t,2,i}=p_{t,2,i}$，$m_{t,1,i}\neq p_{t,2,i}$）＝$p_e$（１－$p_m$）／３＋（１－$p_e$）$p_m$／３＋（１－$p_e$）（１－$p_m$）２／９を得ることができる。さて、これらの項のすべてを組合せ、単に代数を単純化するために、$p_e=p_p=p_m=p$と仮定して、

【００８７】

【数４】

$$P(M(e_{1,i}=p_{1,i},e_{2,i}=m_{1,i})\mid A,s_1)=\frac{1}{162}(160p^4-160p^3+96p^2-28p+13) \quad (5)$$

を計算することができる。計算は変化するが、本明細書中に記載されたものに対して同様な概念的アプローチをすべての２５６の可能な状態、s_k、ｋ＝１．．．２５６で用いる。すべての２５６の状態s_iについてＰ（$e_{1,i}=p_{1,I}$，$e_{2,i}=m_{1,I}$｜Ａ，s_i）を計算し、各s_iの確率を合計し、Ｐ（$e_{1,i}=p_{1,I}$，$e_{2,i}=m_{1,i}$｜Ａ）を得る。言い換えれば：

【００８８】

【数５】

$$P(M \mid A) = \sum_{i=1...256} P(M \mid A, s_i) P(s_i)$$

(6)

である。各状態s_iの確率P（s_i）を計算するために、別々の事象としての状態をなすすべての別々の対立遺伝子を処理しなければならない。というのは、それらは別々の染色体上にあるからである、言い換えれば：P（s_i）＝P（$p_{t,1,i}$，$p_{t,2,i}$，$m_{t,1,i}$，$m_{t,2,i}$）＝P（$p_{t,1,i}$）P（$p_{t,2,i}$）P（$m_{t,1,i}$）P（$m_{t,2,i}$）である。ベイズ技術を適用して、個々の測定についての確率分布を見積もることができる。遺伝子座ｉにおける母性または父性染色体上の対立遺伝子の各測定をコイン投げ実験として処理して、特定の値（Ａ，Ｃ，Ｔ，またはＧ）であるこの対立遺伝子の確率を測定することができる。これらの測定を成人組織試料でなし、全く信頼性があるとして処理することができるが、対立遺伝子の対は各ＳＮＰについて測定し、いずれの対立遺伝子がいずれの染色体に由来かを決定するのは可能でない。$w_{p,1,i}$＝P（$p_{t,1,i}$）とし、これは、父親の染色体上のＳＮＰ　ｉの確率が値$p_{t,1,i}$であることに対応する。以下の説明において、$w_{p,1,i}$の代わりにｗを用いる。父親の染色体のＳＮＰ　ｉで行った測定は収集データとして特徴付けられるものとする。ｗについての確率分布ｐ（ｗ）を作り出し、データがベイズ理論：ｐ（ｗ｜Ｄ）＝ｐ（ｗ）ｐ（Ｄ｜ｗ）／ｐ（Ｄ）に従って測定した後これを更新することができる。ＳＮＰ　ｉのｎ個の対立遺伝子が観察され、ｗに対応する特定の対立遺伝子がｈ回出現する、言い換えれば、ヘッドはｈ回観察されると仮定する。この観察の確率は二項分布によって特徴づけることができる。

【００８９】

【数６】

$$p(D \mid w) = \binom{n}{h} w^h (1-w)^{n-h} \quad 、$$

(7)

データを収集する前に、０および１の間では均一である事前分布ｐ（ｗ）があると仮定する。ベイズ理論を適用することによって、直接的に、ｐ（ｗ｜Ｄ）についての得られた分布は形式：

【００９０】

【数７】

$$p(w \mid D) = \frac{1}{c} w^h (1-w)^{n-h} \text{ ここで } \quad c = \int_0^1 w^h (1-w)^{n-h} dw$$

(8)

のデータ分布であることを示し、ｃは正規化定数である。しかしながら、次いで、ベイズ理論および新しい測定を適用することによって、ｐ（ｗ｜Ｄ）を何回も更新し、それを、前記したデータ分布を有するように継続する。ｐ（ｗ）の見積もりは、新しい測定が収集されるごとに更新される。特定のＳＮＰにおける異なる対立遺伝子の確立は人種および性別のグループ分けに依存するので、Ｈａｐｍａｐプロジェクトで用いたのと同一のグループ分けを用いて、異なる人種および異なる性別について異なる関数ｐ（ｗ）があることに注意されたし。Ｐ（s_i）の計算では、各染色体上の各対立遺伝子は見積もられた確率分布、すなわち、$p_{p,1,i}$（$w_{p,1,i}$）、$p_{p,2,i}$（$w_{p,2,i}$）、$p_{m,1,i}$（$w_{m,1,i}$）および$p_{m,2,i}$（$w_{m,2,i}$）と関連するであろう。次いで、個々の分布の各々についてのＭＡＰ見積もりに従ってＰ（s_i）についての最大事後（ＭＡＰ）見積もりを計算することができる。例えば、$w_{p,1,i}{}^*$は、$p_{p,1,i}$

（$w_{p,1,i}$）を最大化する議論であるとする。Ｐ（s_i）のＭＡＰ見積もりは：
$P(s_i)_{MAP}=w_{p,1,i}*w_{p,2,i}*w_{m,1,i}*w_{m,2,i}*$
（９）
に従って見出すことができる。各ｗについて確率分布が存在するので、ＭＡＰ見積もりを単に用いるよりはむしろ、確率分布に渡って積分することによって、いずれかの特定の信頼性レベルまで値Ｐ（s_i）の保存的見積もりを計算することもできる。例えば、これを行って、ある信頼性レベル内まで保存的にＰ（Ｍ｜Ａ）を見積もることが可能である。保存的見積もりまたはＭＡＰ見積もりを用いるかに拘わらず、Ｐ（s_i）の見積もりはＰ（Ｍ｜Ａ）の計算のために継続的に洗練される。以下において、仮定された状態への言及をなくして、表記法を単純化し、状態s_1は詳細な計算のすべての説明のために仮定される。現実には、これらの計算は２５６の状態の各々について行われ、各々の確率に渡って合計することを銘記されたし。

【００９１】

Ｐ（Ｍ｜Ａ）を計算する方法は、今や、Ｍが胚上のＳＮＰのＮ個の対の測定のセット、Ｍ＝［M_1，．．．，M_N］を表すと仮定し、多数のＳＮＰ遺伝子座まで拡大される。また、Ａは、いずれの親染色体がそのＳＮＰに貢献したかについての各ＳＮＰに対する仮説のセットを表すと仮定する、Ａ＝［A_1，．．．，A_N］。ＳＡ’が、Ａとは異なる、またはセットＡ’に存在するというすべての他の可能な仮説のセットを表すものとする。Ｐ（Ｍ｜Ａ）およびＰ（Ｍ｜Ａ’）を計算することができる：

【００９２】

【数８】

$$P(M\mid A)=\prod_{i=1\ldots N}P(M_i\mid A_i),\quad P(M\mid A')=\sum_{A\in S_{A'}}P(A)\prod_{i=1\ldots N}P(M_i\mid A_i)\qquad (10)$$

Ｐ（Ａ）の計算を考える。本質的には、これは、胚を形成する配偶子の形成において起こる特定の交差の尤度に基づく。特定の対立遺伝子セットの確率は２つの因子、すなわち、胚染色体が母親または父親に由来する確率、および交差の特定の組合せの確率に依存する。異数性をこうむらない胚染色体の清浄なセットについては、胚染色体が母親または父親に由来する事前確率は～５０％であり、その結果、すべてのＡについて共通する。さて、組換え節の特定のセットの確率を考える。関連組換え部位Ｒの数は測定されたＳＮＰＳ：Ｒ＝Ｎ－１の数に依存する。注目するＰＳＮＰの回りのＮ個のＮＳＮＰを構成するＤＮＡセグメントは比較的短いので、交差干渉は、同一染色体上の２つの交差が１つの領域で起こり得ることをかなりありそうもなくする。計算の効率の理由で、この方法は、唯一の交差が各関連染色体についての各領域で起こると仮定し、これはＲ個の可能な部位で起こり得る。どのようにしてこの方法を拡大して、所与の領域に多数の交差がある確率を含めることができるかは当業者に明らかであろう。

【００９３】

ＳＮＰの間の各領域における交差をP_r，ｒ＝１．．．Ｎ－１で示すものとする。一次的には、２つのＳＮＰの間の領域ｒにおける組換え節の確率は、（ｃモルガンで測定された）それらのＳＮＰの間の遺伝子距離に比例する。しかしながら、多数の最近の研究は、２つのＳＮＰ遺伝子座の間の組換えの確率の正確なモデリングを可能とした。精子の実験からの観察、および遺伝子変異のパターンは、組換えの率はキロベーススケールに渡って広く変化し、および多数の組換えは組換えホットスポットで起こり、連鎖非平衡を引き起こして、ブロック－様構造を呈することを示す。ヒトゲノム上での組換え率についてのＮＣＢＩデータは、ＵＣＳＣ Ｇｅｎｏｍｅ Ａｎｎｏｔａｔｉｏｎ Ｄａｔａｂａｓｅを通じて公に入手可能である。

【００９４】

種々のデータセットを単独で、または組合せて用いることができる。最も普通のデータセットの内の２つはＨａｐｍａｐプロジェクトおよびＰｅｒｌｅｇｅｎヒトハプロタイププロジェクトからのものである。後者はより高い密度であり；前者はより高い質である。

ＨａｐＭａｐ相Ｉデータ、リリース１６ａに基づく、染色体１の位置１，０３８，４２３ないし４，４６７，７７５からの領域的組換え率については図２参照。これらの率は、パッケージＬＤＨａｔで入手可能な可逆的ジャンプＭａｒｋｏｖ Ｃｈａｉｎ Ｍｏｎｔｅ Ｃａｒｌｏ（ＭＣＭＣ）方法を用いて見積もられた。考えられる状態－空間は、ピース様定常組換え率マップの分布である。Ｍａｒｋｏｖ鎖は、各セグメント２０１についての率に加えて、率変更点の数および位置の分布を探索する。これらの結果を用いて、ＳＮＰＳの間の各定常セグメントの長さの組換え率倍に渡って積分することによってP_rの見積りを得ることができる。ヌクレオチド２０２に渡っての累積組換え率を赤色で図２に示す。

【００９５】

もし領域ｒおよびそうでなければ０で交差が起こったならば$c_r=1$であるように、Ｃをインジケーター変数c_rのセットとする。もし交差が起こらないか、そうでなければ０であれば、$c_0=1$である。ただ１つの交差がＮ個のＳＮＰの領域で起こり得ると仮定するので、セットＣのただ１つのエレメントは非０である。よって、セットＣによって表される交差の確率は：

【００９６】

【数９】

$$P_c=\left(1-\sum_{r=1\ldots N-1}P_r\right)^{c_0}\prod_{r=1}P_r^{c_r} \tag{11}$$

であることが判明する。ＳＮＰ １．．．Ｎについての仮説Ａにおいて、関連する４つの潜在的交差がある。すなわち、ｉ）（インジケーター変数のセットC_{pe}によって示される）胚を形成した父性染色体、ｉｉ）配列決定された精子を形成した父性染色体（セットC_{ps}）、ｉｉｉ）胚を形成した母性染色体（セットC_{me}）、およびｉ）配列決定された卵を形成した母性染色体（セットC_{ee}）。２つのさらなる仮定はｖ）第一の父性胚ＳＮＰが$p_{t,1,1}$または$p_{t,2,1}$に由来するか、およびｖｉ）第一の母性胚ＳＮＰが$m_{t,1,1}$または$m_{t,2,1}$に由来するかである。ＳＮＰの間の交差の確率は人種および性別の間で異なることが見出されるので、異なる交差確率は父性染色体については$p_{p,r}$として、および母性染色体については$p_{m,r}$として示されるであろう。従って、セットC_{pe}、C_{ps}、C_{me}、C_{ee}を包含する特定の仮説Ａの確率は；

【００９７】

【数１０】

$$P(A)=\frac{1}{4}\left(1-\sum_{r=1\ldots N-1}P_{p,r}\right)^{c_{ps,0}}\prod_{r=1\ldots N-1}P_{p,r}^{c_{ps,r}}\left(1-\sum_{r=1\ldots N-1}P_{p,r}\right)^{c_{ps,0}}\prod_{r=1\ldots N-1}P_{p,r}^{c_{ps,r}}\left(1-\sum_{r=1\ldots N-1}P_{m,r}\right)^{c_{me,0}}\prod_{r=1\ldots N-1}P_{m,r}^{c_{me,r}}\left(1-\sum_{r=1\ldots N-1}P_{m,r}\right)^{c_{ee,0}}\prod_{r=1\ldots N-1}P_{m,r}^{c_{ee,r}} \tag{12}$$

として表される。

【００９８】

さて、Ｐ（Ａ）およびＰ（Ｍ／Ａ）を決定するための方程式に関しては、前記方程式３についてのＡ＊を計算するのに必要な全ての要素は定義されている。よって、交差が起こった胚ＳＮＰの高度にエラー傾向の測定から決定し、および高度な信頼性でもって胚測定を結果的に清浄化することが可能である。最良の仮説Ａ＊における信頼性の低度を決定することが残っている。これを決定するためには、オッズ比Ｐ（Ａ＊｜Ｍ）／Ｐ（Ａ＊’｜Ｍ）を見出す必要がある。ツールは全てこの計算のために前記されている：

【００９９】

【数１１】

$$\frac{P(A^*|M)}{P(A^{*\prime}|M)}=\frac{P(A^*|M)}{1-P(A^*|M)}=\frac{P(A^*)P(M|A^*)}{P(A^{*\prime})P(M|A^{*\prime})}=\frac{P(A^*)P(M|A^*)}{(1-P(A^*))P(M|A^{*\prime})}=OR_{A^*} \tag{13}$$

次いで、Ａ＊における信頼性はＰ（Ａ＊｜Ｍ）＝OR_{A}＊／（１＋OR_{A}＊）として与え

られる。この計算は特定の仮説Ａ＊における信頼性を示すが、ＳＮＰの特定の決定における信頼性を示さない。胚ＰＳＮＰ　ｎの決定における信頼性を計算するためには、このＳＮＰの値を変化させない全ての仮説Ａのセットを作り出す必要がある。このセットは$S_{A*,n}$として示され、これは、仮説Ａ＊によって予測されるように、同一の値を有する胚にＰＳＮＰ　ｎをもたらす全ての仮説に対応する。同様に、仮説Ａ＊によって予測される異なる値を有するＰＳＮＰをもたらす全ての仮説に対応するセット$S_{A*'',n}$を作り出す。さて、ＳＮＰが正しく要求される確率－対－ＳＮＰが正しくなく要求される確率のオッズ比を計算することが可能である：

【０１００】

【数１２】

$$OR_{A^*,n}=\frac{\sum_{A\in S_{A^*,n}}P(A|M)}{\sum_{A\in S_{A^*,n}}P(A|M)}=\frac{\sum_{A\in S_{A^*,n}}P(A|M)}{1-\sum_{A\in S_{A^*,n}}P(A|M)}=\frac{\sum_{A\in S_{A^*,n}}P(A)P(M|A)}{\sum_{A\in S_{A^*,n}}P(A)P(M|A)} \quad (14)$$

オッズ比$OR_{A',n}$に基づく胚ＳＮＰの特定の要求における信頼性は：

【０１０１】

【数１３】

$$P(\text{正しく要求されたSNPn})=\sum_{A\in S_{A^*,n}}P(A|M)=\frac{OR_{A^*,n}}{1+OR_{A^*,n}} \quad (15)$$

として計算することができる。

【０１０２】

この技術を用いて、同一染色体の２つが同一の親からのものであり、他方、他の親からのその染色体のいずれも存在しない片親二染色体（ＵＰＤ）のような欠陥を検出することもできよう。親染色体における交差を推定しようと試みる際に、高い信頼性でもってデータを適切に説明する仮説はなく、もし複数のＵＰＤを含む別の仮説が許容されるならば、それらはよりありそうであることが判明するであろう。

【０１０３】

組換えラットにおける確実性の効果、およびＳＮＰ測定の信頼性のバウンディング

開示された方法は：特定のＳＮＰの間の組換えの確立についての仮定；胚、精子、卵、父性および母性染色体についての各ＳＮＰの正しい測定の確率についての仮定；および異なる集団群内のある対立遺伝子の尤度についての仮定に依存する。これらの仮定の各々を考慮し：組換えのメカニズムは完全には理解され、モデル化されておらず、交差確率は、個人の遺伝子型に基づいて変化することが確立されている。さらに、組換え率が測定される技術は実質的可変性を示す。例えば、可逆的－ジャンプＭａｒｋｏｖ　Ｃｈａｉｎ　Ｍｏｎｔｅ　Ｃａｒｌｏ（ＭＣＭＣ）方法を実行するパッケージＬＤＡａｔは、仮定のセットを作成し、組換えのメカニズムおよび特徴付けについてのユーザーの入力のセットを必要とする。これらの仮定は、種々の実験によって得られた異なる結果によって証明されているように、ＳＮＰの間の予測された組換え率に影響し得る。

【０１０４】

前記リストの全ての仮定のうち、組換え率についての仮定は方程式１５に対して最もインパクトを有するであろうと予測される。前記した計算は、ＳＮＰＳ、P_rの間の交差に対する確率の最良の見積もりに基づくべきである。その後、（正しくはＳＮＰ　ｎと呼ばれる）信頼性尺度Ｐを低下させる方向において、例えば、組換え率についての９５％信頼性範囲における値を用いてＰｒで用いることができる。９５％信頼性範囲は、組換え率の種々の実験によって生じた信頼性データに由来することができ、これは、異なる方法を用いて異なる群からの公表されたデータの間の不一致のレベルを見ることによって確証することができる。

【０１０５】

同様に、９５％信頼性範囲を、各ＳＮＰが正しく要求される確率の見積もりで用いることができる：p_p、p_m、p_e。これらの数は、測定技術の信頼性についての経験的なデータと組み合わせた、ゲノタイピングアッセイ出力ファイルに含まれた現実の測定されたアレイ強度に基づいて計算することができる。これらのパラメーターp_p、p_mおよびp_eが確立されないＮＳＭＰは無視することができることを注記する。例えば、ジプロイド親データは信頼性よく測定されるので、親のハクロイド細胞、および親のジプロイド組織の関連ＳＮＰについての対立遺伝子のいずれにも対応しない胚についてのＮＳＮＰ測定を無視することができる。

【０１０６】

最後に、計算Ｐ（s_i）を生起する異なる集団群内のある対立遺伝子の尤度についての仮定を考える。これらの仮定もまた開示された方法に対して大きなインパクトを有しないであろう。というのは、親ジプロイドデータの測定は信頼性があり、すなわち、親試料からの状態s_iの直接的測定は、典型的には、高い信頼性を持つデータをもたらすからである。それにも拘わらず、方程式８に記載された各ｗについての確率分布を用いて、各状態Ｐ（s_i）の確率についての信頼性範囲を計算することが可能である。前記したように、（正しくはＳＮＰ　ｎと呼ばれる）信頼性尺度Ｐを低下させる保存的方向における各Ｐ（s_i）についての９５％信頼性範囲を計算することができる。

【０１０７】

（正しくはＳＮＰ　ｎと呼ばれる）Ｐの決定は、どのようにして多くのＮＳＮＰが各ＰＳＮＰについて測定される必要があるかについての決定を知らせて、所望のレベルの信頼性を達成するであろう。

【０１０８】

開示された方法の概念を実施する、すなわち、親のＤＮＡの測定、１以上の胚のＤＮＡの測定、および減数分裂のプロセスの事前知識を組合せて、胚ＳＮＰの良好な見積もりを得る異なるアプローチがあることを注記する。事前知識の異なるサブセットが知られており、または知られておらず、または大きなまたは小さな低度の確実性でもって知られている場合に、どのようにして同様な方法を適用することができるかは当業者に明らかであろう。例えば、多数の胚の測定を用いて、特定の胚のＳＮＰを要求する確実性を改良し、または親からの失われたデータを供給することができる。注目するＰＳＮＰを測定技術によって測定する必要がないことを注記する。たとえ測定システムによってＰＳＮＰが決定されなくても、それは、依然として、開示された方法によって高度な信頼性でもって再構築できる。

【０１０９】

一旦減数分裂の間に起こった交差の点が決定され、標的ゲノムの領域が親ＤＮＡの関連領域にマッピングされれば、注目する個体のＳＮＰの同一性のみならず、測定における対立遺伝子ドロップアウトまたは他のエラーによる測定された標的ゲノムで失われているであろうＤＮＡの全領域を推定することが可能である。または、親ＤＮＡにおける挿入および欠失を測定し、開示された方法を用いて、それらが標的ＤＮＡに存在すると推定することも可能である。

【０１１０】

種々の技術を用いて前記して開示アルゴリズムの計算の複雑性を改善することができる。例えば、母親および父親の間で異なるＮＳＮＰを選択することができるにすぎないか、または圧倒的に選択することができる。もう１つの考慮は、ＰＳＮＰの近くに間隔が設けられたＮＳＮＰを用いて、注目するＮＳＮＰおよびＰＳＮＰの間で起こる交差のチャンスを最小化することであろう。また、多数のＰＳＮＰの適用範囲を最大化するために染色体に沿って間隔を設けたＮＳＮＰを用いることもできる。もう１つの考慮は、最初に少数のＮＳＮＰのみを用いて、大まかにどこで交差が起こったかを、限定された程度の確率のみでもって決定することであろう。次いで、さらなるＮＳＮＰを用いて、交差モデルを洗練し、正しくＰＳＮＰを要求する確率を増加させることができる。考慮する交差組合せの数は、ＮがＳＮＰの数であって、Ｃが最大数の交差であるN^Cとして概略評価する。結果と

して、Ｃ＝４については、Ｐｅｎｔｉｕｍ（登録商標）－ＩＶプロセッサーに対して計算可能に御しやすくしつつ、各ＰＳＮＰについて概略Ｎ＝１００を供給することが可能である。前記したアプローチ、および増大した計算効率についての他のアプローチを用い、Ｎ＞１００、Ｃ＞４を容易に供給することができる。１つのそのようなアプローチを以下に記載する。

【０１１１】

基本となる概念を変化させることなく、胚データ、親データ、および用いるアルゴリズムの特定のセットに基づいて、ＰＳＮＰについての要求を行い、ＰＳＮＰが正しく決定された確率の見積もりを生じる多くの他のアプローチがあることを注記する。この確率は個人の決定をなすのに、およびＩＶＦまたはＮＩＰＧＤの関係で信頼性のよいサービスを実行するのに用いることができる。

【０１１２】

遺伝子データ清浄化アルゴリズムに対する帰納的解

直線的に範囲を定めるアルゴリズムに関連する本発明のもう１つの実施形態をここに記載する。計算パワーの限定された性質を仮定すると、計算の長さは開示された方法の使用において重要な因子であり得る。計算を実行する場合、必要とされる計算の数がＳＮＰの数と共に指数関数的に上昇するある値を計算しなければならないいずれのアルゴリズムも扱いにくくなり得る。ＳＮＰの数と共に直線的に増加する多数の計算を含む解は、常に、ＳＮＰの数が大きくなるにつれて時間の観点から好ましいであろう。以下に、このアプローチを記載する。

【０１１３】

全ての可能な仮説を考慮する単純なアプローチは、ＳＮＰの数が指数関数である実行時間と戦わなければならない。前記したように、ｋ個のＳＮＰについての測定された胚、父親および母親染色体の測定のコレクションであると仮定する。すなわち、Ｍ＝｛M_1，．．．，M_k｝であり、ここで、M_i＝（e_{1i}，e_{2i}，p_{1i}，p_{2i}，m_{1i}，m_{2i}，）である。前記したように、仮説空間はS_H＝｛H^1，．．．，H^q｝＝｛全ての仮説のセット｝であり、ここで、各仮説はフォーマットH^J＝｛$H^j{}_1$，．．．$H^j{}_k$｝のものであり、ここで、$H^J{}_I$はフォーマット$H^j{}_i$＝（$p_i{}^*$，$m_i{}^*$）のスニップｉについての「ミニ」仮説であり、ここで、$p_i{}^* \in \{p_{1i}, p_{2i}\}$および$m_i{}^* \in \{m_{1i}, m_{2i}\}$である。４つの異なる「ミニ」仮説$H^j{}_i$、特に：

$H^j{}_i1$：（e_{1i}，e_{2i}）＝｛（p_{1i}，m_{1i}）または（m_{1i}，p_{1i}）｝
$H^j{}_i2$：（e_{1i}，e_{2i}）＝｛（p_{1i}，m_{2i}）または（m_{2i}，p_{1i}）｝
$H^j{}_i3$：（e_{1i}，e_{2i}）＝｛（p_{2i}，m_{1i}）または（m_{1i}，p_{2i}）｝
$H^j{}_i4$：（e_{1i}，e_{2i}）＝｛（p_{2i}，m_{2i}）または（m_{2i}，p_{2i}）｝

がある。目標は、最もありそうな仮説H^*を：

【０１１４】

【数１４】

$$H^* = \arg\max_{H \in S_H} P(H \mid M) = \arg\max_{H \in S_H} F(M, H)$$

として選択することであり、ここで、関数Ｆ（Ｍ，Ｈ）＝Ｐ（Ｈ｜Ｍ）である。

【０１１５】

空間S^Hにおいて4^kの異なる仮設がある。全空間S^Hを専ら調べることによって最良の仮説を見出す試みによって、必要なアルゴリズムはｋ　Ｏ（ｅｘｐ（ｋ））における指数関数オーダーのものであり、ここで、ｋは関連するＳＮＰの数である。大きなｋ、ｋ＞５さえについても、これはかなり遅く、非現実的である。従って、一定時間内にサイズ（ｋ－１）の問題の関数としてサイズｋの問題を解く帰納的解に頼るのがより現実的である。本明細書中に示された解はｋ，Ｏ（ｋ）における直線オーダーのものである。

【０１１６】

ＳＮＰの数において直線的な帰納的解

Ｆ（Ｍ，Ｈ）＝Ｐ（Ｈ｜Ｍ）＝Ｐ（Ｍ｜Ｈ）＊Ｐ（Ｈ）／Ｐ（Ｍ）で始める。次いで、

$argmax_H F(M, H) = argmax_H P(M|H) * P(H)$であり、目標は直線的時間内にP（M｜H）＊P（H）を解くことである。$M_{(s,k)}$＝ＳＮＰ　ｓないしｋでの測定、$H_{(s,k)}$＝ＳＮＰ　ｓないしｋについての仮説とし、表現方法$M_{(k,k)} = M_k$，$H_{(k,k)} = H_k$＝ＳＮＰ　ｋについての測定および仮説を単純化する。先に示したように：

【０１１７】

【数１５】

$$P(M_{(1,k)} \mid H_{(1,k)}) = \prod_{i=1}^{k} P(M_i \mid H_i) = P(M_k \mid H_k) * \prod_{i=1}^{k-1} P(M_i \mid H_i) = P(M_k \mid H_k) * P(M_{(1,k-1)} \mid H_{(1,k-1)})$$

である。また、

【０１１８】

【数１６】

$$P(H_{(1,k)}) = 1/4 * \prod_{i=2}^{k} PF(H_{i-1}, H_i) = PF(H_{k-1}, H_k) * 1/4 * \prod_{i=2}^{k-1} PF(H_{i-1}, H_i) = PF(H_{k-1}, H_k) * P(H_{(1,k-1)})$$

であり、ここで、

【０１１９】

【数１７】

$$PF(H_{i-1}, H_i) = \begin{cases} 1 - PC(H_{i-1}, H_i) & H_{i-1} = H_i \\ PC(H_{i-1}, H_i) & H_{i-1} \neq H_i \end{cases}$$

であり、ＰＣ（H_{i-1}，H_j）＝H_{i-1}，H_iの間の交差の確率である。

【０１２０】

最後に、ｋ個のＳＮＰについては：

F（M，H）＝P（M｜H）＊P（H）＝P（$M_{(1,k)}$｜$H_{(1,k)}$）＊P（$H_{(1,k)}$）

＝P（$M_{(1,k-1)}$｜$H_{(1,k-1)}$）＊P（$H_{(1,k-1)}$）＊P（M_k｜H_k）＊PF（H_{k-1}｜H_k）

であり、従って、短くすると、

F（M，H）＝F（$M_{(1,k)}$，$H_{(1,k)}$））＝F（$M_{(1,k-1)}$，$H_{(1,k-1)}$）＊P（M_k｜H_k）＊PF（H_{k-1}，H_k）

であり、すなわち、ｋ個のＳＮＰについてのＦの計算をｋ－１個のＳＮＰについてのＦの計算に変えることができる。

【０１２１】

Ｈ＝（H_1，．．．H_k）については、ｋ個のＳＮＰについての仮説：

【０１２２】

【数１８】

$$\max_H F(M,H) = \max_{(H_{(1,k-1)}, H_k)} F(M,(H_{(1,k-1)}, H_k) = \max_{H_k} \max_{H_{(1,k-1)}} F(M,(H_{(1,k-1)}, H_k) = \max_{H_k} G(M_{(1,k)}, H_k)$$

であり、ここで、

【０１２３】

【数１９】

$$G(M_{(1,n)},H_n)=\max_{H_{(1,n-1)}}F(M_{(1,n)},(H_{(1,n-1)},H_n)=$$

$$\max_{H_{(1,n-1)}}F(M_{(1,n-1)},H_{(1,n-1)})*P(M_n\mid H_n)*PF(H_{n-1},H_n)=$$

$$P(M_n\mid H_n)*\max_{H_{(1,n-1)}}F(M_{(1,n-1)},H_{(1,n-1)})*PF(H_{n-1},H_n)=$$

$$P(M_n\mid H_n)*\max_{H_{n-1}}\max_{H_{(1,n-2)}}F(M_{(1,n-1)},(H_{(1,n-2)},H_{n-1})*PF(H_{n-1},H_n)=$$

$$P(M_n\mid H_n)*\max_{H_{n-1}}PF(H_{n-1},H_n)*G(M_{(1,n-1)},H_{n-1})$$

である。

【０１２４】

これをまとめると：

【０１２５】

【数２０】

$$\max_{H}F(M,H)=\max_{H_n}G(M_{(1,k)},H_k)$$

であり、ここで、Ｇが帰納的に見出すことができ：ｎ＝２，．．．，ｋについては、

【０１２６】

【数２１】

$$G(M_{(1,n)},H_n)=P(M_n\mid H_n)*\max_{H_{n-1}}\left[PF(H_{n-1},H_n)*G(M_{(1,n-1)},H_{n-1})\right]$$

およびＧ（Ｍ（１，１），Ｈ１）＝０．２５＊Ｐ（Ｍ１｜Ｈ１）である。

【０１２７】

該アルゴリズムは以下の通りである：

ｎ＝１については：４つの仮説Ｈ１ｉを作り出し、ｉ＝１，．．．，４についてＧ（Ｍ１｜Ｈ１ｉ）を計算する。

ｎ＝２については：Ｈ２ｉについて４つの仮説を作り出し、式：

【０１２８】

【数２２】

$$G(M_{(1,2)},H_2i)=P(M_2\mid H_2i)*\max_{j=1,\ldots,4}\left[PF(H_1,H_2i)*G(M_1,H_1j)\right]$$

を用い、一定時間内に、Ｇ（Ｍ（１，２）｜Ｈ２ｉ），ｉ＝１，．．．，４を計算する。

ｎ＝ｋについては：Ｈｋｉについて４つの仮説を作り出し、

【０１２９】

【数２３】

$$G(M_{(1,k)},H_ki)=P(M_k\mid H_ki)*\max_{j=1,\ldots,4}\left[PF(H_{k-1},H_ki)*G(M_{(1,k-1)},H_{k-1}j)\right]$$

によって、Ｇ（Ｍ（１，ｋ）｜Ｈｋｉ），ｉ＝１，．．．，４を作成する。

いずれの時点においても、覚えておくべき４つのみの仮説、および一定数の操作がある。

従って、アルゴリズムは、指数関数とは反対に、ＳＮＰの数ｋにおいて線形である。

【０１３０】

直線的時間内におけるＰ（Ｍ）の解

Ｐ（Ｍ）について解いて、最良の仮説を得る必要はない。というのは、それは全てのＨについて一定だからである。しかしながら、条件付確率Ｐ（Ｈ｜Ｍ）＝Ｐ（Ｍ｜Ｈ）＊Ｐ（Ｈ）／Ｐ（Ｍ）についての現実的な意味のある数を得るためには、Ｐ（Ｍ）を導く必要もある。前記したように、

【０１３１】

【数２４】

$$P(M)=P(M_{(1,k)})=\sum_{H_{(1,k)}}P(M_{(1,k)}\mid H_{(1,k)})*P(H_{(1,k)})$$
$$=\sum_{H_k}P(M_K\mid H_k)\sum_{H_{(1,k-1)}}P(M_{(1,k-1)}\mid H_{(1,k-1)})*P(H_{(1,k-1)})*PF(H_{k-1},H_k)$$
$$=\sum_{H_k}P(M_K\mid H_k)*W(M_{(1,k-1)}\mid H_k)$$

と書くことができ、ここで、

【０１３２】

【数２５】

$$W(M_{(1,k-1)}\mid H_k)=\sum_{H_{(1,k-1)}}P(M_{(1,k-1)}\mid H_{(1,k-1)})*P(H_{(1,k-1)})*PF(H_{k-1},H_k)$$

である。

帰納：

【０１３３】

【数２６】

$$W(M_{(1,k-1)}\mid H_k)=\sum_{H_{(1,k-1)}}P(M_{(1,k-1)}\mid H_{(1,k-1)})*P(H_{(1,k-1)})*PF(H_{k-1},H_k)$$
$$=\sum_{H_{k-1}}P(M_{k-1}\mid H_{k-1})\sum_{H_{(1,k-2)}}P(M_{(1,k-2)}\mid H_{(1,k-2)})*P(H_{(1,k-2)})*PF(H_{k-2},H_{k-1})*PF(H_{k-1},H_k)$$
$$=\sum_{H_{k-1}}P(M_{k-1}\mid H_{k-1})*PF(H_{k-1},H_k)*W(M_{(1,k-2)}\mid H_{k-1})$$

によってW（M，H）について解くことができ、従って、簡単に述べると、サイズｋの問題は、

【０１３４】

【数２７】

$$W(M_{(1,k-1)}\mid H_k)=\sum_{H_{k-1}}P(M_{k-1}\mid H_{k-1})*PF(H_{k-1},H_k)*W(M_{(1,k-2)}\mid H_{k-1})$$

および $W(M_{(1,1)}\mid H_2)=\sum_{H_1}P(M_1\mid H_1)*0.25*PF(H_1,H_2)$

によってサイズ（ｋ－１）の問題に変えられる。前記したように、ｎ＝２：ｋについては、最後に、

【０１３５】

【数２８】

$$P(M)=\sum_{H_k}P(M_K\mid H_k)*W(M_{(1,k-1)}\mid H_k)$$

を誘導することが可能となるまで、帰納的にW（２），．．．，W（Ｋ）＝W（$M_{(1,k-1)}$｜H_k）を得る。

【０１３６】

各レベルにおいて、４つの異なる仮設H_kがあるに過ぎず、従って、アルゴリズムは、再度、ＳＮＰ　ｋの数において線形である。

【０１３７】

直線的時間内における個々のＳＮＰ信頼性

一旦、最良の仮説H^*＝（H_1^*，．．．，H_k^*）が計算されたならば、今度は、各ＳＮＰについての最終的な解答における信頼性、すなわち、ｉ＝１，．．．，ｋについてのＰ（H_i^*｜Ｍ）誘導することが望まれるであろう。前記したように、Ｐ（H_i^*｜Ｍ

）＝Ｐ（Ｍ｜Ｈ$_{i}$＊）Ｐ（Ｈ$_{i}$＊）／Ｐ（Ｍ）＝Ｗ（Ｈ$_{i}$＊，Ｍ）／Ｐ（Ｍ）であり、ここで、Ｐ（Ｍ）は既に知られている。

【０１３８】

【数２９】

$$W(M,H_i^*)=\sum_{H,H_i=H_i^*}P(M\mid H)*P(H)=\sum_{H=\{H_{(1,i-1)},H_i^*,H_{(i+1,k)}\}}P(M\mid H)*P(H)$$

であり、すなわち、仮説Ｈは最初のｉ－１のＳＮＰ、ｉ番目のＳＮＰについての仮説、およびｉ＋１ないしｋ番目のＳＮＰについての仮説まで終えている。前記したように：

【０１３９】

【数３０】

$$P(M_{(1,k)}\mid H_{(1,k)})=\prod_{j=1}^{k}P(M_j\mid H_j)=\prod_{j=1}^{i-1}P(M_j\mid H_j)*P(M_i\mid H_i^*)*\prod_{j=i+1}^{k}P(M_j\mid H_j)$$
$$=P(M_{(1,i-1)}\mid H_{(1,i-1)})*P(M_i\mid H_i^*)*P(M_{(i+1,k)}\mid H_{(i+1,k)})$$

および

【０１４０】

【数３１】

$$P(H_{(1,k)})=1/4*\prod_{j=2}^{k}PF(H_{j-1},H_j)$$
$$=1/4*\prod_{j=2}^{i-1}PF(H_{j-1},H_j)*PF(H_{i-1},H_i^*)*PF(H_{i-1},H_i^*)*\prod_{j=j+2}^{k}PF(H_{j-1},H_j)$$
$$=1/4*T(H_{(1,i-1)})*PF(H_{i-1},H_i^*)*PF(H_{i-1},H_i^*)*T(H_{(i+1,k)})$$

であり、したがって、

【０１４１】

【数３２】

$$P(H_{(1,k)})=1/4*T(H_{(1,k)})=1/4*T(H_{(1,i-1)})*PF(H_{i-1},H_i^*)*PF(H_{i-1},H_i^*)*T(H_{(i+1,k)})$$

であり、ここで、

【０１４２】

【数３３】

$$T(H_{(1,k)})=\prod_{j=2}^{k}PF(H_{j-1},H_j)$$

である。これから、

【０１４３】

【数３４】

$$W(M_{(1,k)},H_i^*)=\sum_{H,H_i=H_i^*}P(M\mid H)*P(H)=\sum_{H,H_i=H_i^*}P(M\mid H)*1/4*T(H)$$

$$=\sum_{H=(H_{(1,i-1)},H_i^*,H_{(i+1,k)})}P(M_{(1,i-1)}\mid H_{(1,i-1)})*P(M_i\mid H_i^*)*P(M_{(i+1,k)}\mid H_{(i+1,k)})*1/4*T(H_{(1,i-1)})*PF(H_{i-1},H_i^*)*PF(H_{i-1},H_i^*)*T(H_{(i+1,k)})$$

$$=4*P(M_i\mid H_i^*)*\left(\sum_{H_{i-1}}P(M_{(1,i-1)}\mid H_{(1,i-1)})*1/4*T(H_{(1,i-1)})*PF(H_{i-1},H_i^*)\right)$$

$$*\left(\sum_{H_{i+1}}P(M_{(i+1,k)}\mid H_{(i+1,k)})*1/4*T(H_{(i+1,k)})*PF(H_i^*,H_{i+1})\right)$$

$$=4*P(M_i\mid H_i^*)*\left(\sum_{H_{i-1}}W(M_{(1,i-1)},H_{i-1})*PF(H_{i-1},H_i^*)\right)*\left(\sum_{H_{i+1}}W(M_{(i+,k)},H_{i+1})*PF(H_i^*,H_{i+1})\right)$$

を示すのは可能である。

【０１４４】

再度、サイズｋの場合は、前記したよりも複雑なビットであるにもかかわらず、より小さなサイズの２つのピースに変えられている。ピースの各々は

【０１４５】

【数３５】

$$W(M_{(1,n)},H_n)=P(M_n\mid H_n)*\left(\sum_{H_{n-1}}W(M_{(1,n-1)},H_{n-1})*PF(H_{n-1},H_n)\right)$$

$$W(M_{(m,k)},H_m)=P(M_m\mid H_m)*\left(\sum_{H_{m+1}}W(M_{(m+1,k)},H_{m+1})*PF(H_m,H_{m+1})\right)$$

として計算することができる。従って、アルゴリズムは、４つの異なるH_n、H_mの各々についてｎ＝１，．．，ｋ、ｍ＝ｋ，．．１について、Ｗ（$M_{(1,n)}$，H_n）、Ｗ（$M_{(m,k)}$，H_m）を計算し、次いで、必要に応じてそれらを組合せてｉ＝１，．．．，ｋについてＷ（$M_{(1,k)}$，H_i^*）を計算する。操作の数は依然としてｋについて直線的である。

【０１４６】

データの小さなまたは異なるセットが利用可能である場合の、胚データへの開示された方法の適用

システムの１つの実施形態において、親のいずれかまたは双方からのハプロイドデータの有りまたは無しにて、かつそのデータがより高いまたはより低い程度の確実性まで知られている場合、１人の親（恐らくは母親）からのジプロイドデータを利用する必要があるに過ぎない。例えば、卵の供与の厳しい性質を仮定すれば、母性ハプロイドデータが容易に入手できない場合があると予測される。この明細書を読んだ後に、どのようにして、特定のＳＮＰの尤度を計算するための統計学的方法を限定されたデータセットを仮定して修飾できるかは当業者に明らかとなろう。

【０１４７】

別のアプローチが、一方または双方の親の失われたジプロイドまたはハプロイドデータを補うためにより距離がある親族からのデータを用いる。例えば、個人の染色体の１つのセットは彼または彼女の親の各々に由来することが知られているので、母方祖父母からのジプロイドデータを用いて、失われたまたは貧弱にしか測定されていない母性ハプロイドデータを部分的に再構築できよう。

【０１４８】

この方法の帰納的性質に注意し：適当な祖父母のジプロイドおよび／またはハプロイドデータと共に、単一細胞親ハプロイドデータの天然ではノイジーな測定を仮定し、開示さ

れた方法を用いて、親ハプロイドデータを清浄化することができ、これは、今度は、胚のより正確なゲノタイピングを供するであろう。これらの場合に用いる方法をどのようにして修飾するかは当業者に明らかなはずである。

【０１４９】

より少ないよりはむしろより多くの情報を用いるのが好ましい。というのはこれは所与のＳＮＰにおいて正しい要求を行うチャンスを増大させることができ、かつそれらの要求において信頼性を増加させることができるからである。これは、システムの増大する複雑性とバランスしなければならない。というのは、データのさらなる技術および源を用いるからである。データを増大させるのに情報を用いるために利用できるさらなる情報、ならびに技術の多くの源がある。例えば、Ｈａｐｍａｐデータ、またはゲノムデータの他のレパートリーで見出すことができる相関関係を利用するインフォマティックスを基礎としたアプローチがある。加えて、そうでなければイン・シリコにて再度作り出す必要がある遺伝子データの直接的測定を可能とできる生物学的アプローチがある。例えば、そうでなければ利用できないハプロイドデータは、フローサイトメトリー技術を用いてジプロイド細胞から個々の染色体を抽出して、蛍光タグド染色体を単離することによって測定可能であろう。別法として、細胞融合を用いて、一対立遺伝子ハイブリッド細胞を作り出して、ジプロイドからハプロイドへの変換を行うことができる。

【０１５０】

いずれの胚が着床するようであるかを選択することへの開示された方法の適用

１つの実施形態において、システムを用いて、母親に着床し、ベイビーまで発生する胚の尤度を決定することができる。胚着床の尤度が胚のＳＮＰ、および／または母親のＳＮＰに対するそれらの関係によって決定される程度まで、開示された方法は、いずれがクリーンなＳＮＰデータに基づいて成功して着床するかの信頼性ある予測をなすことをベースとして、胚の選択を助けるにおいて重要であろう。尤度を最良に予測するためには、胚における遺伝子発現のレベル、母親における遺伝子発現のレベル、および／または母親の決定された遺伝子型と恐らくは組合された胚の決定された遺伝子型を考慮する必要があろう。

【０１５１】

加えて、異数性胚はあまり着床しないようであり、成功した妊娠をもたらさないようであり、健康な子供をもたらさないようであることはよく知られている。結果として、異数体についてのスクリーニングは、成功した結果を最ももたらすようである胚の選択に対して重要な面である。このアプローチについてのより詳細は以下に掲げる。

【０１５２】

親ハプロイドデータの推定

該方法の１つの実施形態において、親のジプロイドデータの詳細な知識を仮定し、親はプロイドを推定する必要があろう。これを行うことができる多数の方法がある。最も単純な場合において、ハプロタイプは、直接的関係（母親、父親、息子または娘）の単一ハプロイド細胞の分子アレイによって既に推定されている。この場合、分子からアッセイによって測定されたジプロイド遺伝子型からの公知のハプロイドを差し引くことによって姉妹ハプロイドを推定するのは当業者にとってたやすいことである。例えば、もし特定の遺伝子座がヘテロ接合性であれば、未知の親ハプロイドは公知の親ハプロタイプからの反対の対立遺伝子である。

【０１５３】

もう１つの場合において、親のノイジーなハプロイドデータは、精子のような個々の親ハプロイド細胞の分子生物学的ハプロタイピングから、または磁性ビーズおよびフローサートメトリーを含めた種々の方法によって単離することができる個々の染色体から知ることができる。この場合、決定されたハプロタイプが測定されたハプロタイプと同程度にノイジーであることを除いて、同一手法を前記したように用いることができる。

【０１５４】

また、（公のＨａｐｍａｐプロジェクトで作り出されたもののような）一般的集団にお

ける公知のハプロタイプブロックを利用する統計学的方法を用い、ジプロイドデータから直接的にハプロイドデータセットを推定する方法もある。ハプロタイプブロックは、本質的には、種々の集団において反復して起こる一連の関連する対立遺伝子である。これらのハプロタイプブロックはしばしば古くかつ共通するので、それらを用いて、ジプロイド遺伝子型からハプロイドを予測することができる。次いで、親の推定されたハプロイドブロックを本明細書中に記載された方法のために入力として用いて、胚からのノイジーなデータを清浄化することができる。この仕事を達成する公に入手可能なアルゴリズムは、不完全な系統発生アプローチ、共役事前分布、および集団遺伝学からの事前分布に基づくベイズアプローチを含む。これらのアルゴリズムのいくつかは隠れたＭａｒｋｏｖモデルを用いる。１つの研究は、公のトリオおよび無関係な個々のデータを用いて、これらのアルゴリズムが１ＭＢの配列にわたって０．０５％と低い誤差率にて実行されることを示した。しかしながら、予測されるように、精度は稀なハプロタイプブロックを持つ個人についてより低い。１つの見積もりにおいて、計算方法は、２０％のわずかな対立遺伝子頻度にて５．１％の遺伝子座と多くを同調できなかった。

【０１５５】

本発明の１つの実施形態において、ＩＶＦサイクルの間に異なる胚から取られた多数の胚盤胞からの遺伝子データを用いて、より大きな信頼性でもって親のハプロタイプブロックを推定する。

【０１５６】

高および中程度スループットのゲノタイピングを用いて異数性をスクリーニングするための技術

システムの１つの実施形態において、測定された遺伝子データを用いて、個体において異数体および／またはモザイク現象の存在について検出することができる。本明細書中に開示するのは、これらの試料からの増幅されたまたは増幅されていないＤＮＡからの染色体の数またはＤＮＡセグメントコピー数を検出するための中程度または高－スループットゲノタイピングを用いるいくつかの方法である。目標は、Ｉｌｌｕｍｉｎａ、ＡｇｉｌｅｎｔおよびＡｆｆｙｍｅｔｒｉｘからのＡＢＩ　Ｔａｑｍａｎ、ＭＩＰＳ、またはマイクロアレイのような異なる定量的および／または定性的ゲノタイピングプラットフォームを用いて異数性のあるタイプおよびモザイク現象のレベルの検出において達成することができる信頼性を見積もることである。これらの場合の多くにおいて、遺伝物質はゲノタイピングアレイ上のプローブへのＰＣＲによって増幅して、特定の対立遺伝子の存在を検出する。これらのアッセイをゲノタイピングでどのようにして用いるかは本開示において他の箇所に記載されている。

【０１５７】

以下に記載するのは、欠失、異数体および／またはモザイク現象から生起するかに拘らず、異常な数のＤＮＡセグメントについてスクリーニングするいくつかの方法である。該方法は以下のようにグループ分けされる：（ｉ）対立遺伝子要求を行うことのない定量的技術；（ｉｉ）対立遺伝子要求を活用する定性的技術；（ｉｉｉ）対立遺伝子要求を活用する定量的技術；（ｉｖ）各遺伝子座における遺伝子データの増幅についての確率分布関数を用いる技術。全ての方法は、標的個体のゲノムにおける所与のセグメントの存在の数を決定するための、所与の染色体の所与のセグメント上の多数の遺伝子座の測定を含む。加えて、該方法は、所与のセグメントの存在の数についての１以上の仮説のセットを作り出し；所与のセグメント上の多数の遺伝子座における遺伝子データの量を測定し；標的個体遺伝子データの測定を仮定して、仮説の各々の相対的確率を決定し；次いで、所与のセグメントの存在の数を決定するために、各仮説に関連する相対的確率を用いることを含む。さらに、該方法は、全て、多数の遺伝子座における遺伝子データの量の測定の計算された関数である組合せ測定Ｍを作り出すことを含む。全ての方法において、閾値は、測定Ｍに基づいて各仮説H_iの選択について決定され、測定すべき遺伝子座の数を見積もって、仮説の各々の偽検出の特定のレベルを有するようにする。

【０１５８】

測定Mを仮定して各仮説の確率はP（H_i｜M）＝P（M｜H_i）P（H_i）／P（M）である。P（M）はHiから独立しているので、P（M｜H_i）P（H_i）のみを考慮することによってMを仮定した仮説の相対的確率を決定することができる。以下において、技術の分析および比較を単純化するために、我々は、我々がP（M｜H_i）のみを考慮することによって全てのP（H_i｜M）の相対的確率を計算できるように、P（H_i）は全ての{H_i}について同一である。その結果、閾値、および測定すべき遺伝子座の数の我々の決定は、P（H_i）が全ての{H_i}について同一であるという仮定の下で偽仮説を選択する特定の確率を有することに基づく。この開示を読んだ後に、どのようにしてアプローチを修飾して、B（H_i）がセット{H_i}において異なる仮説で変化するという事実を受け入れるであろうかは当業者に明瞭である。いくつかの実施形態において、全てのiにわたってP（H_i｜M）を最大化する仮説H_i*が選択されるように閾値を設定する。しかしながら、閾値はP（H_i｜M）を最大化するように必ずしも設定される必要はないが、むしろ、セット{H_i}における異なる仮説の間の偽検出の確率の特定の比率を達成するように設定される必要がある。

【０１５９】

異数体を検出するための本明細書中で言及する技術は、片親二染色体、バランスしないトランスロケーションについて、および染色体の性別分け（男性または女性；ＸＹまたはＸＸ）について検出するのに等しく良く用いることができることに注意するのは重要である。概念の全ては、所与の試料に存在する染色体（または染色体のセグメント）の同一性および数を検出することに関連し、かくして、全ては、本処理に記載された方法によって取り組まれる。どのようにして、本明細書中に記載された方法のいずれかの方法を拡大して、これらの異常性のいずれかについて検出するのは当業者に明らかなはずである。

【０１６０】

マッチドフィルタリングの概念

ここに適用される方法は、デジタルシグナルの最適検出において適用されるのに同様である。正常に分布したノイズの存在下においてシグナル－ノイズ比率（ＳＮＲ）を最大化する最適アプローチは、可能なノイズ－フリーシグナルの各々に対応する、理想化されたマッチングシグナル、またはマッチドフィルターを形成すること、およびこのマッチドシグナルを受け取られたノイジーなシグナルと相関させることは、Ｓｃｈｗａｒｔｚ不均衡を用いて示すことができる。このアプローチは、可能なシグナルのセット、ノイズの統計学的分布－平均および標準偏差（ＳＤ）が公知であることを必要とする。ここで、染色体、またはＤＮＡのセグメントが試料中に存在するかまたは存在しないことを検出する一般的アプローチを記載する。全染色体を調べること、または挿入されたまたは欠失された染色体セグメントを調べることの間に差を設けない。この記載を呼んだ後に、どのようにして、該技術を異数性および性別決定の多くのシナリオ、または胚、胎児または産まれた子供の染色体における挿入および欠失の検出まで拡大できるかは明らかなはずである。このアプローチは、Ｔａｑｍａｎ、ｑＰＣＲ、Ｉｌｌｕｍｉｎａアレイ、Ａｆｆｙｍｅｔｒｉｘアレイ、Ａｇｉｌｅｎｔアレイ、ＭＩＰＳキット等を含めた広い範囲の定量的および定性的ゲノタイピングプラットフォームに適用することができる。

【０１６１】

一般的問題の公式化

２つの対立遺伝子変異が起こる（ｘおよびｙ）ＳＮＰにおいてプローブがあると仮定する。各遺伝子座ｉ、ｉ＝１．．．Ｎにおいて、２つの対立遺伝子からの遺伝物質の量に対応するデータを収集する。Ｔａｑｍａｎアッセイにおいて、これらの尺度は、例えば、各対立遺伝子－特異的色素のレベルが閾値を交差するサイクル時間C_tであろう。どのようにして、このアプローチを、各遺伝子座における、または遺伝子座における各対立遺伝子に対応する遺伝物質の量の異なる測定まで拡大できるかは明らかであろう。遺伝物質の量の定量的測定は非線形であり、その場合、注目するセグメントの存在によって引き起こされた特定の遺伝子座の測定の変化は、どのようにして、その遺伝子座の多くの他のコピーが他のＤＮＡセグメントからの試料に存在するかに依存するであろう。いくつかの場合に

おいて、技術が、注目するセグメントの存在によって引き起こされた特定遺伝子座の測定の変化が、どのようにしてその遺伝子座の多くの他のコピーが他のＤＮＡセグメントからの試料に存在するかに依存しないように、線形測定を必要とするであろう。アプローチを、どのようにしてＴａｑａｍａｎまたはｑＰＣＲアッセイからの測定を線形化することができるかについて記載するが、異なるアッセイについて適応できる非線形測定を線形化するための多くの他の技術がある。

【０１６２】

遺伝子座１．．．Ｎにおける対立遺伝子ｘの遺伝物質の量の測定は、データ$d_x=[d_{x1}...d_{xN}]$によって与えられる。同様に、対立遺伝子ｙについては、データ$d_y=[d_{y1}...d_{yN}]$によって与えられる。各セグメントｊは、各要素a_{ji}がｘまたはｙいずれかである対立遺伝子$a_j=[a_{j1}....a_{jN}]$を有すると仮定する。対立遺伝子ｘの遺伝物質の量の測定データを、ｓｘがシグナルであって、υ_xが擾乱である$d_x=s_x+\upsilon_x$として記載する。該シグナルは$s_x=[f_x(a_{11},...,a_{J1})...f_x(a_{JN},...,a_{JN})]$であり、ここで、ｆｘは測定に対する対立遺伝子からのセットのマッピングであり、ＪがＤＮＡセグメントコピーの数である。擾乱ベクトルυ_xは測定誤差によって引き起こされ、非線形測定の場合においては、注目するＤＮＡセグメント以外の他の遺伝子物質の存在によって引き起こされる。測定誤差は通常に正規分布し、それらは、非線形によって引き起こされた擾乱に対して大きく（線形化測定についてのセクション参照）、従って、$\upsilon_{xi}\ n_{xi}$であり、ここで、非n_{xi}が偏差$\sigma_{xi}2$を有し、ベクトルn_xは正規分布する～Ｎ（０、Ｒ），Ｒ＝Ｅ（$n_x n_x^T$）と仮定する。さて、いくつかのフィルターｈをこのデータに適応して、測定$m_x=h^T d_x=h^T s_x+h^T \upsilon_x$を行うと仮定する。ノイズに対するシグナルの比率（$h^T s_x/h^T n_x$）を最大化するためには、ｈはマッチドフィルター$h=\mu R^{-1}s_x$によって与えられ、ここで、μはスケーリング定数であることを示すことができる。対立遺伝子ｘについての議論は対立遺伝子ｙについて反復することができる。

【０１６３】

方法１ａ：各遺伝子座についての平均および標準偏差が知られている場合に、対立遺伝子要求を行わない定量的技術による異数性または性別の測定

このセクションでは、データは、（例えば、ｑＰＣＲを用いる）対立遺伝子値に拘わらず遺伝子座における遺伝物質の量に関係し、またはデータは、集団において１００％浸透度を有する対立遺伝子についてのみであると仮定し、あるいはデータは、各遺伝子座における多数の対立遺伝子において組合せて（線形化測定についてのセクション参照）、その遺伝子に座における遺伝物質の量を測定すると仮定する。その結果、このセクションにおいては、データd_xに言及でき、d_yを無視することができる。また、２つの仮説：ＤＮＡセグメントの２つのコピーがあるh_0（これらは、典型的には同一のコピーではない）およびただ１つのコピーがあるh_1、があると仮定する。各仮説については、データは、各々、$d_{xi}(h_0)=s_{xi}(h_0)+n_{xi}$および$d_{xi}(h_1)=s_{xi}(h_1)+n_{xi}$として記載でき、ここで、$s_{xi}(h_0)$は、２つのＤＮＡセグメントが存在する場合に、遺伝子座ｉにおける遺伝物質の予測される測定（予測されるシグナル）であり、$s_{xi}(h_1)$は１つのセグメントについて予測されるデータである。仮説h_0：$m_{xi}=d_{xi}-s_{xi}(h_0)$についての予測されるシグナルを差分することによって各遺伝子座についての測定を構築する。もしh_1が真であれば、測定の予測される値は$E(m_{xi})=s_{xi}(h_1)-s_{xi}(h_0)$である。先に議論したマッチドフィルターを用い、$h=(1/N)R^{-1}(s_{xi}(h_1)-s_{xi}(h_0))$を設定する。測定は$m=h^T d_x=(1/N)\Sigma_{i=1...N}((s_{xi}(h_1)-s_{xi}(h_0))/\sigma_{xi}^2)m_{xi}$と記載される。

【０１６４】

もしh_1が真であれば、$E(m|h_1)=m_1=(1/N)\Sigma_{i=1...N}(s_{xi}(h_1)-s_{xi}(h_0))^2/\sigma_{xi}^2$の予測される値、およびｍの標準偏差は$\sigma_{m|h1}^2=(1/N^2)\Sigma_{i=1...N}((s_{xi}(h_1)-s_{xi}(h_0))^2/\sigma_x$

$_{i}{}^{4}$）$\sigma_{x_i}{}^{2}$＝（１／Ｎ^{2}）$\Sigma_{i=1...N}$（$s_{x_i}(h_1)-s_{x_i}(h_0)$）^{2}／$\sigma_{x_i}{}^{2}$である。

【０１６５】

h_0が真であれば、ｍの予測される値はＥ（ｍ｜h_0）＝m_0＝０であって、ｍの標準偏差は、再度、$\sigma_{m|h0}{}^{2}$＝（１／Ｎ^{2}）$\Sigma_{i=1...N}$（$s_{x_i}(h_1)-s_{x_i}(h_0)$）^{2}／$\sigma_{x_i}{}^{2}$である。

【０１６６】

図３は、どのようにして、偽陰性および偽陽性検出の確率を決定するかを説明する。閾値ｔは、偽陰性および偽陽性の確率を等しくするために、m_1およびm_0の間に設定すると仮定する（これは、後に記載するように、当てはまる必要はない）。偽陰性の確率は、（m_1-t）／$\sigma_{m|h1}$＝（m_1-m_0）／（２$\sigma_{m|h1}$）の比率によって決定される。「５－シグマ」統計学は、偽陰性の確率が１－ｎｏｒｍｃｄｆ（５，０，１）＝２．８７ｅ－７であるように用いることができる。この場合、目標は（m_1-m_0）／（２$\sigma_{m|h0}$）＞５または１０ｓｑｒｔ（（１／Ｎ^{2}）$\Sigma_{i=1...N}$（$s_{x_i}(h_1)-s_{x_i}(h_0)$）^{2}／$\sigma_{x_i}{}^{2}$）＜（１／Ｎ）$\Sigma_{i=1...N}$（$s_{x_i}(h_1)-s_{x_i}(h_0)$）^{2}／$\sigma_{x_i}{}^{2}$またはｓｑｒｔ（$\Sigma_{i=1...N}$（$s_{x_i}(h_1)-s_{x_i}(h_0)$）^{2}／$\sigma_{x_i}{}^{2}$）＞１０に対するものである。Ｎのサイズを計算するために、平均シグナル－ノイズ比率が、集合データから計算することができるＭＳＮＲ＝（１／Ｎ）$\Sigma_{i=1...N}$（$s_{x_i}(h_1)-s_{x_i}(h_0)$）^{2}／$\sigma_{x_i}{}^{2}$。次いで、前記不均衡からＭを見出すことができる：ｓｑｒｔ（Ｎ）．ｓｑｒｔ（ＭＳＮＲ）＞１０またはＮ＞１００／ＭＳＮＲ。

【０１６７】

このアプローチは、Ｘ染色体上の４８ＳＮＰを用いてＡｐｐｌｉｅｄ　ＢｉｏＳｙｓｔｅｍｓからのＴａｑｍａｎアッセイで測定したデータに適用した。各遺伝子座についての測定は、この遺伝子座に対応するウェルに放出された色素が閾値を超えるのに要する時間C_tである。試料０は、対象が２つのＸ染色体を有した混合女性起源のウェル当たりおよそ０．３ｎｇ（５０細胞）の合計ＤＮＡよりなり；試料１は、対象が１つのＸ染色体を有した混合男性起源のウェル当たりおよそ０．３ｎｇのＤＮＡよりなるものであった。図４および図５は、試料１および０についての測定のヒストグラムを示す。これらの試料の分布は、m_0＝２９．９７；SD_0＝１．３２、m_1＝３１．４４、SD_1＝１．５９２によって特徴付けられる。このデータは混合男性および女性試料に由来するので、観察されたＳＤのいくつかは、混合試料中の各ＳＮＰにおける異なる対立遺伝子頻度によるものである。加えて、観察されたＳＤのいくつかは、各ＳＮＰにおける異なるアッセイの変化する効率、および各ウェルにピペットで入れられる色素の異なる量によるものであろう。図６は、男性および女性試料についての各遺伝子座における測定の差のヒストグラムを提供する。男性および女性試料の間の平均差は１．４７であって、差のＳＤは０．９９である。このＳＤは、依然として、混合男性および女性試料における異なる対立遺伝子頻度に従うが、それは、もはや、各遺伝子座における各アッセイの異なる効率に影響しないであろう。目標は、各々がおよそに同様なＳＤを持つ２つの測定を区別することであり、調整されたＳＤは、全ての遺伝子座についての各測定に対して０．９９／ｓｑｒｔ（２）＝０．７０と近似することができる。２つの実行を各遺伝子座について行って、マッチドフィルターを適用することができるように、その遺伝子座におけるアッセイについてσ_{x_i}を見積もった。σ_{x_i}の下限を０．２に設定して、σ_{x_i}を計算するためのただ２つの実行から得られる統計学的異常を回避した。双方の対立遺伝子にわたり、双方の実験実行にわたり、および男性および女性双方の試料にわたって、対立遺伝子ドロップアウトがなかった遺伝子座（ナンバリング３７）のみを、プロットおよび計算で用いた。前記したアプローチをこのデータに適用し、ＭＳＮＲ＝２．２６、よって、Ｎ＝２^{2}５^{2}／２．２６＾２＝１７遺伝子座であることが判明した。

【０１６８】

方法１ｂ：平均および標準偏差が知られていないか、または均一である場合に、対立遺

伝子要求を行わない定量的技術による異数性または性別の測定

各遺伝子座の特徴がよく知られていない場合、各遺伝子座における全てのアッセイが同様に挙動し、すなわち、その代わり、E（m_x）およびσ_xのみに言及するのが可能であるように、E（m_{xi}）およびσ_{xi}は全ての遺伝子座eにわたって一定であるという単純化仮定をすることができる。この場合、マッチドフィルタリングアプローチ$m=h^T d_x$はd_xの分布の平均を見出すことに変えられる。このアプローチは平均の比較といい、それは、真実のデータを用いる異なる種類の検出で必要とされる遺伝子座の数を見積もるのに用いられるであろう。

【０１６９】

前記したように、試料に存在する２つの染色体（仮説h_0）または存在する１つの染色体（h_1）がある場合のシナリオを考える。h_0では、分布は$N(\mu_0, \sigma_0^2)$であり、h_1については、分布は$N(\mu_1, \sigma_1^2)$である。各々、測定された試料平均およびＳＤ：m_1、m_0、s_1およびs_0を持つN_0およびN_1試料を用いて分布の各々を測定する。平均は、$M_0 \sim N(\mu_0, \sigma_0^2/N_0)$および$M_1 \sim N(\mu_1, \sigma_1^2/N_1)$として正規分布するランダム変数$M_0$、$M_1$としてモデル化することができる。$M_1 \sim N(m_1, s_1^2/N_1)$および$M_0 \sim N(m_0, s_0^2/N_0)$と仮定することができるように、$N_1$および$N_0$は十分に大きい（＞３０）と仮定する。分布が異なるか否かを検定するために、平均検定の差を用いることができ、ここで、$D=m_1-m_0$である。ランダム変数Ｄの偏差は$\sigma_d^2=\sigma_1^2/N_1+\sigma_0^2/N_0$であり、これは$\sigma_d^2=s_1^2/N_1+s_0^2/N_0$と近似することができる。$h_0$を与えると、E（ｄ）＝０となり；$h_1$を与えると、E（ｄ）＝$\mu_1-\mu_0$となる。$h_1$および$h_0$の間の要求を行うための異なる技術をここに議論する。

【０１７０】

Ｘ染色体上の４８ＳＮＰを用いるＴａｑｍａｎアッセイの異なる実行で測定されたデータを用いて、性能をキャリブレートした。試料１は、１つのＸ染色体を含有する混合男性起源のウェル当たりおよそ０．３ｎｇのＤＮＡよりなり；試料０は、２つのＸ染色体を含有する混合女性起源のウェル当たりおよそ０．３ｎｇのＤＮＡよりなるものであった。$N_1=42$および$N_0=45$。図７および図８は、試料１および０についてのヒストグラムを示す。これらの試料についての分布は$m_1=32.259$、$s_1=1.460$、$\sigma_{m1}=s_1/sqrt(N_1)=0.225$；$m_0=30.75$；$s_0=1.202$、$\sigma_{m0}=s_0/sqrt(N_0)=0.179$によって特徴付けられる。これらの試料では、$d=1.509$および$\delta_d=0.2879$である。

【０１７１】

このデータは混合男性および女性試料に由来するので、標準偏差の多くは、混合試料中の各ＳＮＰにおける異なる対立遺伝子頻度によるものである。ＳＤは、多数の実行にわたり、一定時刻における１つのＳＮＰについてのC_tにおける変動を考慮することによって見積もられる。このデータを図９に示す。ヒストグラムは０の周りに対称である。というのは、各ＳＮＰについてのC_tは２つの実行または実験で測定され、各ＳＮＰについてのC_tの平均値は差し引かれるからである。２つの実行を用いる混合男性試料中の２０のＳＮＰにわたる平均標準偏差はｓ＝０．５９７である。このＳＤは男性および女性双方の試料で保存的に用いられる。というのは、女性試料についてのＳＤは男性試料についてよりも小さいだろうからである。加えて、混合試料は全てのＳＮＰについてヘテロ接合性であると推定されるので、ただ１つの色素からの測定が用いられていることを注記する。双方の色素の使用は、遺伝子座における各対立遺伝子の測定が組み合わされることを必要とし、これはより複雑である（線形化測定についてのセクション参照）。双方の色素についての測定の組合せはシグナルの振幅を２倍とし、およそにｓｑｒｔ（２）によってノイズ振幅を増大させ、その結果、およそｓｑｒｔ（２）または３ｄＢのＳＮＲ改良がもたらされる。

【０１７２】

モザイク現象なしおよび参照試料なしを仮定する検出

m_0が多くの実験から完全に知られており、かつかく実験の実行は、m_1を計算してm０と比較するのにただ１つの試料を実行すると仮定する。ｎ１はアッセイの数であり、各アッセイは異なるＳＮＰ遺伝子座であると仮定する。閾値ｔはm_0およびm_1の間に設定して、偽陰性の尤度を偽陰性の数と等しくすることができ、もしそれが閾値を超えれば、試料は異常であると記される。$s_1=s_2=s=0.597$であると仮定し、偽陰性または陽性の確率が１－ｎｏｒｍｃｄｆ（５，０，１）＝２．８７ｅ－７となるように５－シグマアプローチを用いる。目標は$5s_1/sqrt(N_1)<(m_1-m_0)/2$、従って、$N_1=100s_1^2/(m_1-m_0)^2=16$についてのものである。さて、有害なシナリオである、偽陽性の偽陰性の確率よりも高くされるアプローチを用いることもできる。もし陽性を測定すれば、実験は再度行うことができる。その結果、偽陰性の確率は偽陽性の確率の平方と等しいはずであるということが可能である。図３を考え、ｔ＝閾値とし、シグマ０＝シグマ１＝ｓと仮定する。かくして、１－ｎｏｒｍｃｄｆ（（ｔ－m_0）／ｓ，０，１））^{2}＝１－ｎｏｒｍｃｄｆ（（m_1－ｔ）／ｓ，０，１）である。これを解き、$t=m_0+0.32(m_1-m_0)$であることを示すことができる。よって、目標は$5s/sqrt(N_1)<m_1-m_0-0.32(m_1-m_0)=(m_1-m_0)/1.47$、よって、$N_1=(5^2)(1.47^2)s^2/(m_1-m_0)^2=9$についてのものである。

【０１７３】

参照試料を実行することのないモザイク現象での検出

目標は９７．７％の確率でモザイク現象を検出することである（すなわち、２－シグマアプローチ）以外は前記したのと同一状況を仮定する。これは、およそ２０の細胞を抽出し、それらの写真を撮る羊水穿刺に対する標準アプローチよりも良好である。もし２０細胞のうち１が異数体であって、これは１００％の信頼性でもって検出されると仮定するならば、標準アプローチを用いる異数体である群の少なくとも１つを有する確率は$1-0.95^{20}=64\%$である。もし細胞の０．０５％が異数体であれば（この試料３を要求する、$m_3=0.95m_0+0.05m_1$および$var(m_3)=(0.95s_0^2+0.05s_1^2)/N_1$である。かくして、$std(m_3)2<(m_3-m_0)/2=>sqrt(0.95s_0^2+0.05s_1^2)/sqrt(N_1)<0.05(m_1-m_2)/4=>N_1=16(0.95s_2^2+0.05s_1^2)/(0.05^2(m_1-m_2)^2)=1001$である。慣用的アプローチを用いて達成することができるよりも依然として良好な（すなわち、８４．１％確率での検出）１－シグマ統計学の目標を用い、同様にして$N_1=250$であると示すことができる。

【０１７４】

モザイク現象がなく、参照試料を用いる検出

このアプローチは必要でないかもしれないが、各実験は２つの試料を実行して、m_1を真実の試料m_2と比較すると仮定する。$N=N_1=N_0$と仮定する。$d=m_1-m_0$を計算し、$\sigma_1=\sigma_0$と仮定し、閾値$t=(m_0+m_1)/2$を設定し、従って偽陽性および偽陰性の確率は等しい。偽陰性の確率を２．８７ｅ－７とし、それは、$(m_1-m_2)/2>5sqrt(s_1^2/N+s_2^2/N)=>N=100(s_1^2+s_2^2)/(m_1-m_2)^2=32$があてはまらなければならない。

【０１７５】

モザイク現象での検出および参照試料の実行

前記したように、偽陰性の確率は２．３％であると仮定する（すなわち、２－シグマアプローチ）もし細胞の０．０５％が異数体であれば（これを試料３と呼ぶ）、$m_3=0.95m_0+0.05m_1$および$var(m_3)=(0.95s_0^2+0.05s_1^2)/N_1$である。$d=m_3-m_2$および$\sigma_d^2=(1.95s_0^2+0.05s_1^2)/N$である。$std(m_3)2<(m_0-m_2)/2=>sqrt(1.95s_2^2+0.05s_1^2)/sqrt(N)<0.05(m_1-m_2)/4=>N=16(1.95s_2^2+0.05s_1^2)/(0.05^2(m_1-m_2)^2)=2002$でなければならない。再度１－シグマアプローチを用い、Ｎ＝５００であることが同様にして示すことができる

。

【０１７６】

目標が、現在の技術水準におけるように、６４％の確率でもって５％モザイク現象を検出するにすぎない場合を考える。従って、偽陰性の確率は３６％となろう。換言すれば、１－ｎｏｒｍｃｄｆ（ｘ，０，１）＝３６％となるようなｘを見出す必要があろう。かくして、２－シグマアプローチについてはＮ＝４（０．３６＾２）（１．９５s_2^2＋０．０５s_1^2）／（０．０５^2（m_1-m_2）^{2}）＝６５であり、または１－シグマアプローチについてはＮ＝３３である。この結果、取り組むことが必要な、非常に高いレベルの偽陽性がもたらされることに注意されたし。というのは、偽陽性のそのようなレベルは現在実行可能な代替ではないからである。

【０１７７】

また、もしＮが３８４に限定され、（すなわち、染色体当たり３８４ウェルＴａｑｍａｎプレート）、かつ目標が９７．７２％の確率でモザイク現象を検出することにあれば、１－シグマアプローチを用いて８．１％のモザイク現象を検出することが可能であろうことを注記する。８４．１％の確率でもって（または１５．９％偽陰性率でもって）モザイク現象を検出するには、１－シグマアプローチを用いて５．８％のモザイク現象を検出するのが可能である。９７．７２％の信頼性でもって１９％のモザイク現象を検出するには、およそ７０の遺伝子座を必要とするであろう。かくして、単一プレート上で５つの染色体についてスクリーニングできよう。

【０１７８】

これらの異なるシナリオの各々のまとめを表２に供する。また、この表２は、ｑＰＣＲから得られた結果、およびＳＹＢＲアッセイを含める。前記した方法を用い、各遺伝子座についてのｑＰＣＲアッセイの性能は同一であるという単純化した仮定を行った。図１０および図１１は、前記したような、試料１および０についてのヒストグラムを示す。$N_0=N_1=47$。これらの試料についての測定の分布は、$m_1=27.65$、$s_1=1.40$，$\sigma_{m1}=s_1/sqrt(N_1)=0.204$；$m_0=26.64$；$s_0=1.146$、$\sigma_{m0}=s_0/sqrt(N_0)=0.167$によって特徴付けられる。これらの試料について、ｄ＝１．０１および$\sigma_d=0.2636$である。図１２は、０．７５の全ての遺伝子座にわたる差の標準偏差での各遺伝子座についての男性および女性試料に対するC_tの間の差を示す。ＳＤは、男性または女性試料での各遺伝子座の各測定について０．７５／ｓｑｒｔ（２）＝０．５３と近似した。

【０１７９】

方法２：対立遺伝子要求を用いる定性的技術

このセクションにおいては、アッセイは定量的であるという仮定をしない。その代わり、仮定は、対立遺伝子要求は定性的であって、アッセイに由来する意味のある定量的データはないというものである。このアプローチは、対立遺伝子要求を行ういずれのアッセイについても適当である。図１３は、どのようにして、異なるハプロイド配偶子が減数分裂の間に形成されるか、およびそれを用いて、このセクションに関連する異なる種類の異数性を記載するのに用いる。最良のアルゴリズムは、検出されるべき異数性のタイプに依存する。

【０１８０】

異数性が、他の２つのセグメントのいずれかのコピーであるセクションを有しない第３のセグメントによって引き起こされる状況を考える。図１３より、例えば、もしp_1およびp_4、またはp_2およびp_3の双方が、他の親からの１つのセグメントに加えて、子供の細胞中で生起するならば、該状況は起きるであろう。これは、異数性を引き起こすメカニズムを仮定すれば、非常に普通である。１つのアプローチは、細胞中に２つのセグメントがある仮説h_0、およびこれらの２つのセグメントは何であるかでもって開始することである。説明の目的で、h_0は図１３からのp_3およびm_4についてのものであると仮定する。好ましい実施形態において、この仮説は本書類中の他の箇所に記載されたアルゴリズムに由来する。仮説h_1は、他のセグメントのコピーであるセクションを有しないさら

なるセグメントがあるというものである。これは、例えば、もしこのp_2またはm_1もまた存在するならば正直であろう。p_3およびm_4においてホモ接合性である全ての遺伝子座を同定することが可能である。異数性は、ホモ接合性であると予測される遺伝子座におけるヘテロ接合性遺伝子型要求をサーチすることによって検出することができる。

【０１８１】

各遺伝子座は２つの可能な対立遺伝子ｘおよびｙを有すると仮定する。各々、対立遺伝子ｘおよびｙの確率は一般にp_xおよびp_yであり、および$p_x+p_y=1$であるとする。もしｈ１が真であれば、それについてp_3およびm_4がホモ接合性である各遺伝子座ｉについて、非ホモ接合性要求の確率は、遺伝子座が、各々、ｘまたはｙにおいてホモ接合であるかに依存してp_yまたはp_xである。注意：親データ、すなわち、p_1、p_2、p_4およびm_1、m_2、m_3の知識に基づいて、各遺伝子座において非ホモ接合性対立遺伝子ｘまたはｙを有する確率をさらに改良することが可能である。これは、同一数のＳＮＰでの各仮説についてより信頼性のある測定を可能とするが、標記方法を複雑化し、従って、この延長は明示的には取り扱わない。どのようにしてこの情報を用いて、仮説の信頼性を増加させるかは当業者に明らかなはずである。

【０１８２】

対立遺伝子ドロップアウトの確率はp_dである。遺伝子座ｉにおいてヘテロ接合性遺伝子型を見出す確率は仮説h_0を仮定すれば、p_{0i}であり、仮説h_1を仮定すればp_{1i}である。

【０１８３】

h_0：$p_{0i}=0$とする。

【０１８４】

遺伝子座がｘまたはｙに対してホモ接合性であるかに依存して、h_1：$p_{1i}=p_x(1-p_d)$、または$p_{1i}=p_y(1-p_d)$とする。

【０１８５】

測定$m=1/N_h\Sigma_{i=1...N_h}I_i$を作り出し、ここで、$I_i$はインジケーター変数であり、もしヘテロ接合性要求がなされたならば、１であって、その他の場合は０である。N_hはホモ接合性遺伝子座の数である。$p_x=p_y$であって、全ての遺伝子座についてp_{0i}、p_{1i}が同一の２つの値p_0およびp_1であると仮定することによって、説明を簡略化することができる。h_0を与えて、$E(m)=p_0=0$、および$\sigma^2_{m|h0}=p_0(1-p_0)/N_h$となる。$h_1$を与えて、$E(m)=p_1$および$\sigma^2_{m|h1}=p_1(1-p_1)/N_h$となる。５シグマ－統計学を用い、偽陽性の確率を偽陰性の確率と等しくし、$(p_1-p_0)/2>5\sigma_{m|h1}$、よって、$N_h=100(p_0(1-p_0)+p_1(1-p_1))/(p_1-p_0)^2$と示すことができる。５－シグマ信頼性の代わりに２－シグマ信頼性では、$N_h=4.2^2(p_0(1-p_0)+p_1(1-p_1))/(p_1-p_0)^2$と示すことができる。

【０１８６】

信頼性が少なくとも９７．７％であるように（２－シグマ）十分な入手可能なホモ接合性遺伝子座$N_{h-avail}$があることは、十分な遺伝子座Ｎをサンプリングするのに必要である。$N_{h-avail}=\Sigma_{i=1...N}J_i$を特徴付け、ここで、$J_i$は、もし遺伝子座がホモ接合性であれば値１のインジケーター変数であり、そうでなければ、０である。ホモ接合性である遺伝子座の確率は$p_x^2+p_y^2$である。その結果、」$E(N_{h-avail})=N(p_x^2+p_y^2)$、および$\sigma_{Nh-avail}^2=N(p_x^2+p_y^2)(1-p_x^2+p_y^2)$となる。Ｎが９７．７％信頼性でもって十分に大きいことを補償するためには、$E(N_{h-avail})-2\sigma_{Nh-avail}=N_h$でなければならず、ここで、$N_h$は前記から見出される。

【０１８７】

例えば、もし$p_d=0.3$、$p_x=p_y=0.5$を仮定するならば、５－シグマ信頼性について、$N_h=186$および$N=391$を見出すことができる。同様に、２－シグマ信頼性、すなわち、偽陰性および偽陽性における９７．７％信頼性について、$N_h=30$で

あって、Ｎ＝６８であることを示すのは可能である。

【０１８８】

同様なアプローチを、h_0が２つの公知の染色体セグメントが存在する仮説であって、h_1が染色体セグメントの一方が失われている仮説である場合、セグメントの欠失を探すことに適応することができることを注記する。例えば、前記でなされたように、対立遺伝子ドロップアウトの効果をコードし、ヘテロ接合性であるが、ホモ接合性である遺伝子座を探すことが可能である。

【０１８９】

また、アッセイが定性的であったとしても対立遺伝子ドロップアウト率を用いて、存在するＤＮＡセグメントの数についての定量的尺度のタイプを供することができることを注記する。

【０１９０】

方法３：参照配列の公知の対立遺伝子、および定量的対立遺伝子測定の使用

ここで、セグメントの清浄なまたは予測されるセットは知られていると仮定する。これらの染色体についてチェックするためには、各染色体の２つを仮定して、第一の工程はデータを正常化することである。本発明の好ましい実施形態において、第一の工程におけるデータ正常化は、本書類の他の箇所に記載された方法を用いてなされる。次いで、予測される２つのセグメントに関連するシグナルは、測定されたデータから差し引かれる。次いで、残りのシグナル中のさらなるセグメントを探すことができる。マッチドフィルタリングアプローチを用いて、さらなるセグメントを特徴付けるシグナルは、存在すると信じられるセグメントの各々、ならびにそれらの相補的染色体に基づく。例えば、図１３をコードし、もしＰＳの結果が、セグメントｐ２およびｍ１が存在することを示すならば、本明細書中に記載された技術を用いて、さらなる染色体上でのｐ２、ｐ３、ｍ１、およびｍ４の存在をチェックすることができる。もし存在するさらなるセグメントがあれば、それは、これらのテストシグナルの少なくとも１つと共通する、５０％を超える対立遺伝子を有することが保証される。ここに詳細に記載されていないもう１つのアプローチは、染色体の異常な番号、すなわち、１、３、４、および５染色体を仮定し、書類の他の箇所に記載されたアルゴリズムを用いて、データを正常化し、次いで、本明細書中で議論した方法を適用することができる。このアプローチの詳細は、本書類を読んだ後に当業者に明瞭なはずである。

【０１９１】

仮説h_0は、対立遺伝子ベクトルa_1、a_2をもつ２つの染色体があるというものである。仮説a_1は、対立遺伝子ベクトルa_3を持つ第三の染色体があるというものである。遺伝子データを正常化するために本書類に記載した方法、またはもう１つの技術を用い、各要素a_{ji}がｘまたはｙいずれかである$h_0：a_1=[a_{11}...a_{1N}]$および$a_2=[a_{21}...a_{2N}]$によって予測される２つのセグメントの対立遺伝子を決定することが可能である。予測されるシグナルは、f_x、f_yが各対立遺伝子の測定に対する対立遺伝子のセットからのマッピングを記載する仮説$h_0：s_{0x}=[f_{0x}(a_{11},a_{21})...f_{x0}(a_{1N},a_{2N})]$、$s_{0y}=[f_y(a_{11},a_{21})...f_y(a_{1N},a_{2N})]$について作り出される。$h_0$を仮定すれば、データは$d_{xi}=s_{0xi}+n_{xi}$、$n_{xi}\sim N(0,\sigma_{xi}^2)$；$d_{yi}=s_{0yi}+n_{yi}$、$n_{yi}\sim N(0,\sigma_{yi}^2)$と記載することができる。データおよび参照シグナルを差分することによって測定を作り出す：$m_{xi}=d_{xi}-s_{xi}$；$m_{yi}=y_i-s_{yi}$。十分な測定ベクトルは$m=[m_x^T m_y^T]^T$である。

【０１９２】

さて、注目するセグメント、その存在が疑われるセグメントについてのシグナルを作り出し、それを、このセグメントの推定される対立遺伝子に基づいて求める：$a_3=[a_{31}...a_{3N}]$。残りについてのシグナルを：$s_r=[s_{rx}^T s_{ry}^T]^T$と記載し、ここで、$s_{rx}=[f_{rx}(a_{31})...f_{rx}(a_{3N})]$、$s_{ry}=[f_{ry}(a_{31})...f_{ry}(a_{3N})]$であり、ここで、もし$a_{3i}=x$であれば、$f_{rx}($

a_{3i}）＝δ_{xi}であって、その他の場合は０であり、もしa_{3i}＝ｙであればf_{ry}（a_{3i}）＝δ_{yi}であり、そうでなければ０である。この分析は、遺伝子座ｉにおける対立遺伝子ｘの１つのコピーの存在がデータ$\delta_{xi}+m_{xi}$を作り出し、遺伝子座ｉにおける対立遺伝子ｘのκ_xコピーの存在はデータ$\kappa_x\delta_{xi}+n_{xi}$を作り出すように、測定は線形化されている（後記セクション参照）と仮定する。しかしながら、この仮定は本明細書中に記載された一般的なアプローチでは必要ないことに注意されたし。ｈ１を仮定すれば、もし対立遺伝子a_{3i}＝ｘであれば、$m_{xi}=\delta_{xi}+n_{xi}$、$m_{yi}=n_{yi}$であり、もし$a_{3i}$＝ｙであれば、$m_{xi}=n_{xi}$、$m_{yi}=\delta_{yi}+n_{yi}$である。その結果、マッチドフィルターｈ＝（１／Ｎ）$R^{-1}s_r$を作り出すことができ、ここで、Ｒ＝ｄｉａｇ（［$\sigma_{x1}\ldots\sigma_{xN}^2\sigma_{y1}^2\ldots\sigma_{yN}^2$］）である。測定はｍ＝$h^T$ｄである。

h_0：$m=(1/N)\Sigma_{i=1\ldots N}s_{rxi}n_{xi}/\sigma_{xi}^2+s_{ryi}n_{yi}/\sigma_{yi}^2$

h_1：$m=(1/N)\Sigma_{i=1\ldots N}s_{rxi}(\delta_{xi}+n_{xi})/\sigma_{xi}^2+s_{ryi}(\delta_{yi}+n_{yi})/\sigma_{yi}^2$

必要なＳＮＰの数を見積もるためには、全ての対立遺伝子および全ての遺伝子座についての全てのアッセイが同様な特徴を有し、すなわち、ｉ＝１．．．Ｎについて$\delta_{xi}=\delta_{yi}=\delta$および$\sigma_{xi}=\sigma_{yi}=\sigma$であるという単純化仮定を行う。次いで、平均および標準編纂は以下のように見出すことができる

h_0：$E(m)=m_0=0$；$\sigma_{m|h0}^2=(1/N^2\sigma^4)(N/2)(\sigma^2\delta^2+\sigma^2\delta^2)=\delta^2/(N\sigma^2)$

h_1：$E(m)=m_1=(1/N)(N/2\sigma^2)(\delta^2+\delta^2)=\delta^2/\sigma^2$；$\sigma_{m|h1}^2=(1/N^2\sigma^4)(N)(\sigma^2\delta^2)=\delta^2/(N\sigma^2)$。

さて、h_1－対－h_0のこのテストについてシグナル－対－ノイズ比率（ＳＮＲ）を計算する。シグナルは$m_1-m_0=\delta^2/\sigma^2$であって、この測定のノイズの偏差は$\sigma_{m|h0}^2+\sigma_{m|h1}^2=2\delta^2/(N\sigma^2)$である。その結果、このテストについてのＳＮＲは$(\delta^4/\sigma^4)/(2\delta^2/(N\sigma^2))=N\delta^2/(2\sigma^2)$である。

【０１９３】

このＳＮＲを、対立遺伝子要求に基づいてマッチドフィルタリングを行うことなく、遺伝子情報を各遺伝子座において単純に合計するシナリオと比較する。

【０１９４】

【数３６】

$$h=(1/N)\bar{1}$$

と仮定し、ここで、

【０１９５】

【数３７】

$$\bar{1}$$

はＮのそれのベクトルであり、ｉ＝１．．．Ｎについて$\delta_{xi}=\delta_{yi}=\delta$および$\sigma_{xi}=\sigma_{yi}=\sigma$であると前記したように単純化仮定をする。このシナリオについては、もしｍ＝h^Tｄであれば：

h_0：$E(m)=m_0=0$；$\sigma_{m|h0}^2=N\sigma^2/N^2+N\sigma^2/N^2=2\sigma^2/N$

h_1：$E(m)=m_1=(1/N)(N\delta/2+N\delta/2)=\delta$；$\sigma_{m|h1}^2=(1/N^2)(N\sigma^2+N\sigma^2)=2\sigma^2/N$

であることを直接的に示すことができる。その結果、このテストについてのＳＮＲは$N\delta^2/(4\sigma^2)$である。言い換えれば、セグメントa_3について予測される対立遺伝子測定を単に合計するマッチドフィルターを用いることによって必要なＳＮＰの数は２倍だけ低下する。これは、各遺伝子座におけるアッセイの異なる効率を説明するためにマッチドフィルタリングを用いることによって達成されたＳＮＲ利得を無視する。

【０１９６】

もし参照シグナルs_{xi}およびs_{yi}を正しく特徴付けなければ、得られた測定シグナルm_{xi}およびm_{yi}についてのノイズまたは擾乱のＳＤは増加するであろう。これはもしδ＜＜σであれば有意でなく、そうでなければそれは偽検出の確率を増加させるであろう。その結果、この技術は、３つのセグメントが存在し、２つのセグメントは相互の正確なコピーであると推定される仮説をテストするのによく適合している。この場合、s_{xi}およびs_{yi}は、他の箇所に記載された定性的対立遺伝子要求に基づくデータ正常化の技術を用いて信頼性よく知られるであろう。１つの実施形態において、方法３は、定性的ゲノタイピングを用い、対立遺伝子ドロップアウトからの定量的測定とは別に、セグメントの第二の正確なコピーの存在を検出することができない方法２と組合せて用いられる。

【０１９７】

さて、対立遺伝子要求を用いるもう１つの定量的技術を記載する。該方法は、所与の対立遺伝子についての４つの登録の各々におけるシグナルの相対的量を比較することを含む。ホモ接合性増幅が起こる、（または増幅の相対的量が正規化される）、単一の正常な細胞を含む理想化された場合において、４つの可能な状況が起こり得ると想像することができる：（ｉ）ヘテロ接合性対立遺伝子の場合には４つの登録の相対的強度はほぼ１：１：０：０であり、シグナルの絶対的強度は１つの塩基対に対応し；（ｉｉ）ホモ接合性対立遺伝子の場合には、相対的強度はほぼ１：０：０：０であり、シグナルの絶対的強度は２つの塩基対に対応する；（ｉｉｉ）ＡＤＯが対立遺伝子のうち１つについて起こる対立遺伝子の場合において、相対的強度はほぼ１：０：０：０であって、シグナルの絶対強度は１つの塩基対に対応し；および（ｉｘ）ＡＤＯが対立遺伝子の双方について起こる対立遺伝子の場合において、相対強度はほぼ０：０：０：０であって、シグナルの絶対的強度は塩基対に対応しないであろう。

【０１９８】

しかしながら、異数体の場合には、異なる状況が観察されるであろう。例えば、トリソミーの場合には、ＡＤＯはなく、３つの状況の１つが起こり：（ｉ）三重にヘテロ接合性である対立遺伝子の場合には、４つの登録の相対的強度はほぼ１：１：１：０であり、シグナルの絶対的強度は１つの塩基対に対応し；（ｉｉ）対立遺伝子の２つがホモ接合性である場合には相対的強度はほぼ２：１：０：０であり、シグナルの絶対的強度は、各々、２つおよび１つの塩基対に対応し；（ｉｉｉ）対立遺伝子がホモ接合性である場合には、相対的強度はほぼ１：０：０：０であって、シグナルの絶対的強度は３つの塩基対に対応するであろう。もし対立遺伝子ドロップアウトがトリソミーを持つ細胞における対立遺伝子の場合で起こるならば、正常な細胞で期待される状況のうちの１つが観察されるであろう。モノソミーの場合には、４つの登録の相対的強度はほぼ１：０：０：０であって、シグナルの絶対的強度は１つの塩基対に対応するであろう。この状況は、ＡＤＯにおける対立遺伝子の１つが起こった正常な細胞の場合に対応するが、正常な細胞の場合には、これは対立遺伝子のいくらかのパーセンテージで観察されるのに過ぎないであろう。２つの同一の染色体が存在する片親二染色体の場合には、４つの登録の相対的強度はほぼ１：０：０：０であって、シグナルの絶対的強度は２つの塩基対に対応するであろう。１つの親からの２つの異なる染色体が存在するＵＰＤの場合には、この方法は、本特許に記載された他の方法を用いるデータのさらなる分析はこれを明らかにするであろうが、細胞は正常であることを示す。

【０１９９】

これらの場合の全てにおいて、正常であり、異数体またはＵＰＤを有する細胞いずれかにおいて、１つのＳＮＰからのデータは、細胞の状態について決定するのに適切ではないであろう。しかしながら、もし前記仮説の各々の確率を計算し、それらの確率を所与の染色体上の十分な数のＳＮＰと組み合わせるならば、１つの仮説が支配的であり、高い信頼性でもって染色体の状態を決定することが可能であろう。

【０２００】

定量的測定を線形化するための方法

多くのアプローチを採用して、異なる対立遺伝子からのデータを容易に合計し、または

差分できるように、特定の遺伝子座における遺伝物質の量の測定を線形化することができる。まず、上位概念的なアプローチを議論し、次いで、特定のタイプのアッセイについて設計されるアプローチを議論する。

【０２０１】

データd_{xi}は遺伝子座$_i$における対立遺伝子$_x$の遺伝物質の量の非線形測定をいうと仮定する。Ｎの測定を用いてデータの訓練セットを作り出し、ここに各測定については、データd_{xi}に対応する遺伝物質の量はβ_{xi}であると見積もられ、またはそのように知られている。この訓練セットβ_{xi}、ｉ＝１．．．Ｎは、現実に遭遇するであろう全ての異なる量の遺伝物質にわたるように選択される。標準回帰技術を用いて、線形測定Ｅ（β_{xi}）を期待して、非線形測定d_{xi}からマップされる関数を訓練することができる。例えば、線形回帰を用いて、ｃが係数$c=[c_0 c_1 . . . c_P]^T$のベクトルであるＥ（β_{xi}）＝［１　d_{xi} d_{xi}^2．．．d_{xi}^P］ｃであるように、次元Ｐの多項関数を訓練することができる。この線形化関数を訓練するために、Ｎの測定$\beta_x=[\beta_{x1}...\beta_{xN}]^T$についての遺伝物質の量のベクトル、およびパワーに生起された測定されたデータのマトリックス０．．．Ｐ：Ｄ＝［［１　d_{x1} d_{x1}^2．．．d_{x1}^P］^{T}［１　d_{x2} d_{x2}^2．．．d_{x2}^P］^{T}．．．［１　d_{xN} d_{xN}^2．．．d_{xN}^P］^{T}］^{T}を作り出す。次いで、最小二乗フィット$c=(D^TD)^{-1}D^T\beta_x$を用いて係数を見出すことができる。

【０２０２】

フィットした多項式のような上位概念的関数に依存するよりはむしろ、特定のアッセイの特徴について特殊化された関数を作り出すことができる。例えば、ＴａｑｍａｎアッセイまたはｑＰＣＲアッセイを考える。いくつかの閾値と交差する点までの時間の関数としての、対立遺伝子ｘおよびいくつかの遺伝子座ｉについてのダイの量を、α_{xi}がバイアスオフセットであり、γ_{xi}が指数関数的成長速度であって、β_{xi}が遺伝物質の量に対応するバイアスオフセット：$g_{xi}(t)=\alpha_{xi}+\beta_{xi}\exp(\gamma_{xi}t)$を持つ指数関数曲線として記載することができる。$\beta_{xi}$の項における測定をキャストするためには、曲線の漸近限界$g_{xi}(-\infty)$を探すことによってパラメーターα_{xi}を計算し、次いで、曲線のＬＯＧを取って、$\log(g_{xi}(t)-\alpha_{xi})=\log(\beta_{xi})+\gamma_{xi}t$が得られ、標準的な線形回帰を行うことによって$\beta_{xi}$および$\gamma_{xi}$を見出すことができる。一旦$\alpha_{xi}$および$\gamma_{xi}$についての値を有すれば、もう１つのアプローチは、閾値g_xがその時点で超過する時間t_xからβ_{xi}を計算することである。$\beta_{xi}=(g_x-\alpha_{xi})\exp(-\gamma_{xi}t_x)$。これは、特定の対立遺伝子の遺伝子データの真実の量のノイジーな測定であろう。

【０２０３】

どのような技術を用いても、線形化測定を$\beta_{xi}=\kappa_x\delta_{xi}+n_{xi}g_{xi}(-\infty)$としてモデル化することができ、ここに$\kappa_x$は対立遺伝子ｘのコピーの数であり、$\delta_{xi}$は対立遺伝子ｘおよび遺伝子座ｉについての定数であり、$n_{xi}\sim N(0,\sigma_x^2)$であり、ここで、$\sigma_x^2$は経験的に測定することができる。

【０２０４】

方法４：各遺伝子座における遺伝子データの増幅のための確率分布の関数の使用

特定のＳＮＰについての物質の量は、その上にそのＳＮＰが存在する細胞中の初期セグメントの数に依存するであろう。しかしながら、増幅およびハイブリダイゼーションプロセスのランダムな性質のため、特定のＳＮＰからの遺伝物質の量は、セグメントの出発数に直接的に比例しないであろう。$q_{s,A}$，$q_{s,G}$，$q_{s,T}$，$q_{s,C}$が、対立遺伝子を構成する４つの核酸（Ａ，Ｃ，Ｔ，Ｇ）の各々についての特定のＳＮＰ　ｓに対する遺伝物質の増幅された量を表すものとする。これらの量は、増幅で用いる技術に依存して、正確にゼロであり得ることを注記する。また、これらの量は、典型的には、特定のハイブリダイゼーションプローブからのシグナルの強度から測定されることも注記する。この強度測定を量の測定の代わりに用いることができるか、あるいは発明の性質を変化させることなく標準的な技術を用いて量の見積もりに変換することができる。q_sを特定のＳＮ

Ｐの全ての対立遺伝子から生じた全ての遺伝物質の合計とする：$q_s = q_{s,A} + q_{s,G} + q_{s,T} + q_{s,C}$。ＮをＳＮＰ　ｓを含有する細胞中のセグメントの数とする。Ｎは典型的には２であるが、０、１または３以上であってよい。議論したいずれの高または中程度スループットのゲノタイピング方法についても、遺伝物質の得られた量は$q_s = (A + A_{\theta,s})N + \theta_s$として表すことができ、ここで、Ａは事前に見積もられたか、または経験的に容易に測定される合計増幅であり、$A_{\theta,s}$はＳＮＰ　ｓについてのＡの見積もりにおける誤差であって、θ_sはそのＳＮＰについての増幅、ハイブリダイゼーションおよび他のプロセスで導入される相加的ノイズである。ノイズの項$A_{\theta,s}$およびθ_sは、典型的には、q_sがＮの信頼性がある測定ではないのに十分に大きい。しかしながら、これらのノイズの項の効果は、染色体上の多数のＳＮＰを測定することによって緩和することができる。Ｓを、染色体２１のような特定の染色体上で測定されるＳＮＰの数とする。以下のように、特定の染色体上の全てのＳＮＰにわたる遺伝物質の平均量を得ることが可能である：

【０２０５】

【数３８】

$$q = \frac{1}{S}\sum_{s=1}^{S} q_s = NA + \frac{1}{S}\sum_{s=1}^{S} A_{\theta,s}N + \theta_s \qquad (16)$$

$A_{\theta,s}$およびθ_sは正規分布したランダム変数であり、平均０、および偏差

【０２０６】

【数３９】

$$\sigma^2_{A_{\theta,S}} \text{および} \sigma^2_{\theta_s}$$

であると仮定し、ｑ＝ＮＡ＋φをモデル化することができ、ここで、φは正規分布したランダム変数であり、平均０および偏差

【０２０７】

【数４０】

$$\frac{1}{S}\left(N^2\sigma^2_{A_{\theta,S}} + \sigma^2_{\theta}\right)$$

である。その結果、もし十分な数のＳＮＰが、

【０２０８】

【数４１】

$$S >> \left(N^2\sigma^2_{A_{\theta,S}} + \sigma^2_{\theta}\right)$$

となるように染色体上で測定されるならば、Ｎ＝ｑ／Ａは正確に見積もることができる。

【０２０９】

もう１つの実施形態において、増幅は、１つのＳＮＰからのシグナルレベルがｓ＝ａ＋αであり、ここで、（ａ＋α）が図１４左側の図に似た分布を有するモデルに従うと仮定する。０におけるδ関数はおよそ３０％の対立遺伝子ドロップアウトの速度をモデル化し、平均はａであり、もし対立遺伝子ドロップアウトがなければ、増幅は０ないしa_0の均一な分布を有する。この分布の平均の項において、a_0は$a_0 = 2.86a$であることが判明する。さて、図１４右側の図面を用いてαの確率密度関数をモデル化する。s_cをｃ遺伝子座から生起するシグナルとし；ｎをセグメントの数とし；α_iを、遺伝子座ｉからのシグナルに寄与する図１４に従って分布したランダム変数とし；およびσを全ての$\{\alpha_i\}$についての標準偏差とする。$s_c = anc + \Sigma_{i=1..nc}\alpha_i$；平均$(s_c) = anc$；$std(s_c) = sqrt(nc)\sigma$。もしσを図１４左側における分布に従って計算すれば、それは$\sigma = 0.907a^2$であることが判明する。ｎ＝s_c／（ａｃ）からのセグメントの数を見出すことができ、＜５－シグマ統計学＞については、ｓｔｄ（ｎ）＜０．１、従って、ｓｔｄ（s_c）／（ａｃ）＝０．１＝＞０．９５ａ．ｓｑｒｔ（ｎｃ）／（ａｃ）＝０．１、従って、$c = 0.95^2 n / 0.1^2 = 181$を必要とする。

【０２１０】

要求における信頼性を見積もるためのもう１つのモデル、およびどのようにして多くの遺伝子座またはＳＮＰを測定して、所与の程度の信頼性を確実としなければならないかは、相加的ノイズ源、すなわち、ｓ＝ａ（１＋α）の代わりに増幅のマルチプライアーとしてのランダム変数を取り込む。ｌｏｇを取り、ｌｏｇ（ｓ）＝ｌｏｇ（ａ）＋ｌｏｇ（１＋α）となる。さて、新しいランダムな変数γ＝ｌｏｇ（１＋α）を作り出し、この変数は、正規分布していると仮定することができる～Ｎ（０，σ）。このモデルにおいて、増幅は、σに依存して非常に小さいないし非常に大きいを範囲とすることができるが、決して負ではない。従って、α＝ｅ$^{\gamma}$－１であり；およびｓ$_{c}$＝Σ$_{i=1...cn}$ａ（１＋α$_{i}$）である。表記方法については、平均（ｓ$_{c}$）および予測値Ｅ（ｓ$_{c}$）を相互交換的に用いる。

【０２１１】

【数４２】

$$E(s_c) = acn + aE\left(\sum_{i=1...cn} \alpha_i\right) = acn + aE\left(\sum_{i=1...cn} \alpha_i\right) = acn\left(1 + E(\alpha)\right)$$

Ｅ（α）を見出すためには、確率密度関数（ｐｄｆ）が、可能であるαについて見出されなければならない。というのは、αは公知のガウスｐｄｆを有するγの関数だからである。ｐ$_{\alpha}$（α）＝ｐ$_{\gamma}$（γ）（ｄγ／ｄα）である。従って、

【０２１２】

【数４３】

$$p_\gamma(\gamma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-\gamma^2}{2\sigma^2}} \quad \text{および} \quad \frac{d\gamma}{d\alpha} = \frac{d}{d\alpha}\left(\log(1+\alpha)\right) = \frac{1}{1+\alpha} = e^{-\gamma}$$

および：

$$p_\alpha(\alpha) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-\gamma^2}{2\sigma^2}} e^{-\gamma} = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(\log(1+\alpha))^2}{2\sigma^2}} \frac{1}{1+\alpha}$$

である。これは、σ＝１についての図１５で示された形態を有する。さて、Ｅ（α）は、このｐｄｆにわたって、多数の異なるσについて数値的に行うことができる。

【０２１３】

【数４４】

$$E(\alpha) = \int_{-\infty}^{\infty} \alpha p_\alpha(\alpha) d\alpha$$

を積分することによって見出すことができる。これは、Ｅ（ｓ$_{c}$）または平均（ｓ$_{c}$）をσの関数として与える。さて、このｐｄｆを用いて、ｖａｒ（ｓ$_{c}$）を見出すこともできる：

【０２１４】

【数４５】

$$\mathrm{var}(s_c) = \mathrm{E}\left(s_c - \mathrm{E}(s_c)\right)^2 = \mathrm{E}\left(\sum_{i=1..cn} a(1+\alpha_i) - acn - aE\left(\sum_{i=1...cn}\alpha_i\right)\right)^2$$
$$= \mathrm{E}\left(\sum_{i=1..cn} a\alpha_i - aE\left(\sum_{i=1...cn}\alpha_i\right)\right)^2$$
$$= a^2\mathrm{E}\left(\sum_{i=1..cn}\alpha_i - cnE(\alpha)\right)^2$$
$$= a^2\mathrm{E}\left(\left(\sum_{i=1..cn}\alpha_i\right)^2 - 2cnE(\alpha)\left(\sum_{i=1..cn}\alpha_i\right) + c^2n^2E(\alpha)^2\right)$$
$$= a^2\mathrm{E}\left(cn\alpha^2 + cn(cn-1)\alpha_i\alpha_j - 2cnE(\alpha)\left(\sum_{i=1..cn}\alpha_i\right) + c^2n^2E(\alpha)^2\right)$$
$$= a^2c^2n^2\left(\mathrm{E}(\alpha^2) + (cn-1)E(\alpha_i\alpha_j) - 2cnE(\alpha)^2 + cnE(\alpha)^2\right)$$
$$= a^2c^2n^2\left(\mathrm{E}(\alpha^2) + (cn-1)E(\alpha_i\alpha_j) - cnE(\alpha)^2\right)$$

ここで、これは多数の異なるσについてｐ$_{\alpha}$（α）を用いて数値的に解いて、σの関数としてｖａｒ（ｓ$_{c}$）を得ることもできる。次いで、公知の数の遺伝子座ｃおよび公知の数のセグメントｎを持つ試料から一連の測定を取ることができ、このデータからｓｔｄ（ｓ$_{c}$）／Ｅ（ｓ$_{c}$）を見出すことができる。それにより、σについての値を計算することが可能となる。ｎを見積もるためには、Ｅ（ｓ$_{c}$）＝ｎａｃ（１＋Ｅ（α））、従って、

【０２１５】

【数４６】

$$\hat{n} = \frac{s_c}{ac\left(1+E(\alpha)\right)}$$

を、

【０２１６】

【数４７】

$$std(\hat{n}) = \frac{std(s_c)}{ac\left(1+E(\alpha)\right)}\ \mathrm{std(n)}$$

となるように測定することができる。０ないし平均の十分に多数の独立したランダムな変数を合計すると、分布はガウス形態に近づき、かくして、ｓ$_{c}$（および

【０２１７】

【数４８】

$\hat{n}$

）は正規分布しているとして処理することができ、前記したように、５－σ統計学：

【０２１８】

【数４９】

$$std(\hat{n}) = \frac{std(s_c)}{ac\left(1+E(\alpha)\right)} < 0.1$$

を用いて、２ｎｏｒｍｃｄｆ（５，０，１）＝２．７ｅ－７の誤差確率を有するようにすることができる。これより、遺伝子座ｃの数について解くことができる。

【０２１９】

雌雄鑑別

システムの１つの実施形態において、遺伝子データを用いて、標的個体の性別を決定することができる。本明細書中に開示した該方法を用いて、親からのいずれの染色体のいずれのセグメントが標的の遺伝物質に貢献したかを決定した後、性染色体のいずれが父親から遺伝したかを見るためにチェックすることによって標的の性別を決定することができる：Ｘは女性を示し、およびＹは男性を示す。この方法をどのようにして用いて、標的の性

別を決定するかは当業者に明らかなはずである。

【０２２０】

仮説の確証

システムのいくつかの実施形態において、１つの決定は、最高の可能な信頼性でもって正しい遺伝子状態の予測を行うためには、各可能な状態について仮説を立てる必要があることである。しかしながら、遺伝子状態の可能な数が指数関数的に大きくなり、計算時間が制限されるにつれ、各仮説を検定するのは合理的でないであろう。これらの場合において、別のアプローチは、仮説確証の概念を用いることである。これは、ある値、値のセット、もしある仮説、または仮説のクラスが真実であるならば測定されたデータにおいて観察されることが期待される特性またはパターンに対する制限を見積もることを含む。次いで、測定された値を検定して、それらが予測された制限に入るか、および／またはある予測された特性またはパターンを検定することができるか、および予測が適合しないかを見ることができ、次いで、アルゴリズムはさらなる調査のための測定に警告を与えることができる。

【０２２１】

例えば、染色体の１つのアームの端部が標的ＤＮＡにおいて破壊されている場合、最もありそうな仮説は（例えば、「異数体」とは反対に）「正常である」と計算することができる。これは、遺伝物質の真の状態に対応し、すなわち、染色体の１つの端部が破壊された特定の仮説は、その状態の尤度が非常に低いので検定されていないからである。もし確証の概念を用いれば、アルゴリズムは、多数の値、染色体の破壊されたセクションに存在する対立遺伝子に対応するものは、測定の期待された限界の外にあることを注記するであろう。フラグが生起され、この場合についてのさらなる調査を促し、遺伝物質の真の状態が発見される尤度を増大させる。

【０２２２】

どのようにして、開示された方法を修飾して、確証技術を含ませるかは当業者に明らかなはずである。開示された方法を用いて検出するのは非常に困難であると予測される１つの異常は、バランスしたトランスロケーションであることを注記する。

【０２２３】

汚染されたＤＮＡでの方法の適用

システムの１つの実施形態において、外来性ＤＮＡで明確にまたは可能性として汚染された標的ＤＮＡからの遺伝子データもまた、開示された方法を用いて正常化することができる。先に概説した概念、仮説確証のそれを用いて、予測される限界の外になる遺伝子試料を同定することができ；汚染された試料の場合には、この確証は警告を生起させ、試料を汚染したものとして同定することができると予測される。

【０２２４】

標的ＤＮＡの大きなセグメントは親遺伝子データから知られておるので、かつ汚染の程度は十分に低く、十分なＳＮＰが測定されるものとすると、外来性遺伝物質による誤ったデータを同定しかねない。本明細書中に開示された方法は、依然として、より低い信頼性のレベルに拘わらず、標的ゲノムの再構築を可能とするはずである。汚染のレベルが十分に低いものとすれば、最もありそうであると計算される仮説は、依然として、標的ＤＮＡ試料中の遺伝物質の真の状態に対応すると予測される。

【０２２５】

どのようにして、外来性ＤＮＡにより誤ったシグナルで汚染された遺伝子データを清浄化する目的でこれらの方法を最適かするかは当業者に明らかなはずである。

【０２２６】

実施例

システムの１つの実施形態において、前記した方法は、関連ＳＮＰのリストにおける各ＳＮＰの最もありそうな同一性、ならびに各ＳＮＰ要求についての信頼性レベルを計算するアルゴリズムのセットを用いて実行することができる。本明細書中に記載するのは、この特許に開示した方法を実行するための１つの可能な方法である。図１６および図１７は

、開示された方法のこの実施を頓挫、入力の要件、および出力のフォーマットを実質的に表す。

【０２２７】

図１６は入力データ（１６０１）およびそのフォーマットおよび要件、ならびに出力データ（１６０５）およびそのフォーマットに焦点を当てる。アルゴリズムへの入力は、ユーザーによる入力を含めた測定されたデータ（１６０２）、および結果的には新しく収集されたデータによって更新されるデータベースに保存された現存データ（１６０３）よりなる。測定されたデータ（ＭＤ，１６０２）は胚、および父性および母性対立遺伝子についての所望のＳＮＰについて測定された遺伝子データ、ならびに対立遺伝子の各々が知られている精度または信頼性よりなる。現存データ（１６０３）は集団頻度データ（ＦＤ）、測定バイアスデータ（ＢＤ）および交差データ（ＣＤ）よりなる。

【０２２８】

集団頻度データ（ＦＤ）は利用可能なＳＮＰの各々について（値Ａ、Ｃ、Ｔ、Ｇの各々についての）対立遺伝子頻度を含有する。これらのデータは従前に知られているか、または測定することができ、本書類中の他の箇所に記載されたように新しく収集されたデータで更新することができる。

【０２２９】

測定バイアスデータ（ＢＤ）は、ある種の値に向けての測定プロセスのバイアスを捕獲する。例えば、対立遺伝子の真の値がＸ＝Ａであって、正しい測定の確率はp_xであると仮定し、測定された値ｘの分布は：

【０２３０】

【数５０】

	A	C	T	G
確率	p_X	p_C	p_T	p_G
バイアスの無い確率	p_X	$(1-p_X)/3$	$(1-p_X)/3$	$(1-p_X)/3$

であり、ここで、$p_X+p_C+p_T+p_G=1$である。もし値のいずれかに向けての測定のバイアスがなければ、$p_C=p_T=p_G=(1-p_X)/3$である。この情報は、測定プロセスのメカニズムおよび関連機器についての経験的および理論的知識から区別することができる。

【０２３１】

交差データ（ＣＤ）は、ＨＡＰＭＡＰデータから収集された、スニップの対の間の遺伝子距離および交差確率のデータベースよりなる。

【０２３２】

一緒にすると、（ＭＤ）、（ＦＤ）、（ＢＤ）、（ＣＤ）は、開示された方法（「親サポート」，１６０４という）アルゴリズムに対する必要な入力をなす。次いで、このアルゴリズム（１６０４）を入力データとして操作して、出力データ（１６０５）を生じさせ、これは測定値を仮定した標的の遺伝子データのもっともありそうな「真の」値、ならびに親対立遺伝子に関する各ＳＮＰの最もありそうな起源を記載する。

【０２３３】

図１７は（「親サポート」という）アルゴリズムそれ自体の構造、およびどのようにしてこれらの入力データの各々がアルゴリズムによって利用されるかに焦点を当てる。逆に作業し、最もありそうな仮説を見出すためには、全ての可能な仮説Ｈについての、測定を仮定した仮説の確率Ｐ（Ｈ｜Ｍ）１７０７を計算する必要がある。

先に記載したように：

【０２３４】

【数５１】

$$P(H \mid M) = \frac{P(M \mid H)}{P(M)} P(H),\ P(M) = \sum_{h \in S_H} P(M \mid h) P(h)$$

である。

Ｐ（Ｈ｜Ｍ）（１７１０）を見出すためには、全ての仮説Ｈについて、Ｐ（Ｍ｜Ｈ）（１７０７）およびＰ（Ｈ）（１７０８）を見出すことがまず必要である。これは、先に示した方程式によるＰ（Ｍ），１７０９の計算を可能とする。仮説の確率Ｐ（Ｈ）（１７０８）は、先に説明したようにどれくらい多くの交差が推定されるか、およびこれらの交差の各々の尤度（ＣＤ，１７０４）に依存する。

【０２３５】

Ｐ（Ｍ｜Ｈ）は、先に説明したように、以下の方程式

【０２３６】

【数５２】

$$P(M \mid H) = \sum_{t} P(M \mid H \& t) P(t)$$

を用いて計算することができる。

【０２３７】

Ｐ（Ｔ），１７０６は父性および母性対立遺伝子についての特定の値ｔの頻度であり、集団頻度データ（ＦＤ，１７０３）に由来する。Ｐ（Ｍ｜Ｈ＆ｔ），１７０５は、特定の「真の」値ｔを仮定し、胚、父親および母親の対立遺伝子を正しく測定する確率である。ユーザーによってエンターされた測定データおよび精度（ＭＤ，１７０１）、および測定バイアスデータベース（ＢＤ，１７０２）は、Ｐ（Ｍ｜Ｈ＆ｔ），１７０５を計算するのに必要な出力である。

【０２３８】

該方法のより詳細な記載を以下に掲げる。ｉｄ　S_1，・・・，S_kで同定される、ｋのＳＮＰについての、ＳＮＰ　Ｒ＝｛r_1，．．．，r_k｝、（ｋのＳＮＰのセット）、および親および胚の対応する測定された同一性Ｍ＝（e_1，e_2，p_1，p_2，m_1，m_2）で開始し、ここに：

e_1＝（e_{11}，e_{12}，．．．，e_{1k}）は、全てのＳＮＰについての、胚の染色体の１つでの測定であり（それらは、全てが、同一親染色体に由来する必要はない）、

e_2＝（e_{21}，e_{22}，．．．，e_{2k}）は胚の他の染色体での測定であり、

p_1＝（p_{11}，p_{12}，．．．，p_{1k}）は、（全て同一染色体に由来する）父親の第一の染色体での測定であり、

p_2＝（p_{21}，p_{22}，．．．，p_{2k}）は、（全て同一染色体に由来する）父親の第二の染色体での測定であり、

m_1＝（m_{11}，m_{12}，．．．，m_{1k}）は、（全て同一染色体に由来する）母親の第一の染色体での測定であり、

m_2＝（m_{21}，m_{22}，．．．，m_{2k}）は、（全て同一染色体に由来する）母親の第二の染色体での測定である。

【０２３９】

また、Ｍ＝（M_1，．．．，M_k）を書くことができ、ここで、M_1＝（e_{1i}，e_{2i}，p_{1i}，p_{2i}）である。

【０２４０】

該方法の目標は、「真の」胚値Ｔ＝（Ｅ１，Ｅ２）、すなわち、測定Ｍを仮定した最もありそうな場合を決定することであり、ここに：

E_1＝（E_{11}，E_{12}，．．．，E_{1k}）は、父性染色体に対応する胚の第一の染色体での測定、$E_{1i} \in \{p_{1i}, p_{2i}\}$であり、

E_2＝（E_{21}，E_{22}，．．．，E_{2k}）は、母性値に対応する胚の第二の染色体で

の測定、$E_{2i} \in \{m_{1i}, m_{2i}\}$である。

【０２４１】

また、$T = \{T_1, \ldots, T_k\}$を書くことができ、ここで、$T_i = (E_{1i}, E_{2i})$である。

【０２４２】

効果的には、親染色体値（p_1，p_2，m_1，m_2）は、（E_1，E_2）の測定された値をチェックし、確証し、および修正するためのサポートとして用いる、よって、用語「親サポートアルゴリズム」。

【０２４３】

この目標を達成するためには、胚値の起源についての全ての可能な仮説を開発し、測定Mを仮定して最もありそうなものを選択する。仮説空間は$S_H = \{H^1, \ldots, H^q\}$＝｛全ての仮説のセット｝であり、ここで、各仮説はフォーマット$H^j = (H^j_1, \ldots, H^j_k)$のものであり、ここで、$H^j_1$は、$p_i^* \in \{p_{1i}, p_{2i}\}$および$m_i^* \in \{m_{1i}, m_{2i}\}$であるフォーマット$H^j_1 = (p_i^*, m_1^*)$の、ＳＮＰ　ｉについての「ミニ」仮説である。４つの異なる「ミニ」仮説H^j_1、特に：

H^j_i1：$(e_{1i}, e_{2i}) = \{(p_{1i}, m_{1i})$または$(m_{1i}, p_{1i})\}$

H^j_i2：$(e_{1i}, e_{2i}) = \{(p_{1i}, m_{2i})$または$(m_{2i}, p_{1i})\}$

H^j_i3：$(e_{1i}, e_{2i}) = \{(p_{2i}, m_{1i})$または$(m_{1i}, p_{2i})\}$

H^j_i4：$(e_{1i}, e_{2i}) = \{(p_{2i}, m_{2i})$または$(m_{2i}, p_{2i})\}$

がある。

【０２４４】

理論において、S^Hは$q = 4^k$の異なるメンバーを有して、ピックアップすることができるが、後に、この空間は父性および母性染色体の最大数の交差で限定されるであろう。

【０２４５】

最もありそうな仮説H^*は：

【０２４６】

【数５３】

$$H^* = \arg\max_{H \in S_H} P(H \mid M)$$

であると選択される。特定のＨについては：

【０２４７】

【数５４】

$$P(H \mid M) = \frac{P(M \mid H)}{\sum_{h \in S_H} P(M \mid h) P(h)} P(H)$$

である。

各仮説についてのそのような由来：

（１）Ｐ（Ｍ／Ｈ）は、特定の仮説Ｈを仮定した測定Ｍの確率である。

（２）Ｐ（Ｈ）は特定の仮説Ｈの確率である。

（３）Ｐ（Ｍ）は測定Ｍの確率である。

全てのＨについてＰ（Ｈ｜Ｍ）を導いた後、最大の確率を持つものを選択する。

【０２４８】

Ｐ（Ｍ｜Ｈ）の誘導

各ＳＮＰについての測定は、全てのｋのＳＮＰでの、Ｍ＝（M_1，．．．，M_k）および特定の仮説Ｈ＝（H_1，．．．，H_k）について独立しているので：Ｐ（Ｍ｜Ｈ）＝Ｐ（M_1｜H_1）＊．．．＊Ｐ（M_k｜H_k）である。特定のＳＮＰ　ｒでは、Ｐ（M_r｜H_r）を誘導する。Ω＝｛Ａ，Ｃ，Ｔ，Ｇ｝Ｘ｛Ａ，Ｃ，Ｔ，Ｇ｝Ｘ＝｛Ａ，Ｃ，Ｔ，Ｇ｝Ｘ｛Ａ，Ｃ，Ｔ，Ｇ｝については、ベイズ式による「真の親値（P_{1r}，P_{2r}，M_{1r}，M_{2r}）についての全ての可能な空間は：

【０２４９】
【数５５】

$$P(M_r / H_r) = \sum_{t \in \Omega} P(M_r / H_r \& (P_{1r}, P_{2r}, M_{1r}, M_{2r}) = t) * P((P_{1r}, P_{2r}, M_{1r}, M_{2r}) = t)$$

である。
【０２５０】
Ｐ（Ｍ$_{r}$｜Ｈ$_{r}$＆（Ｐ$_{1r}$，Ｐ$_{2r}$，Ｍ$_{1r}$，Ｍ$_{2r}$）＝ｔ）の誘導
Ｍ$_{r}$＝（ｅ$_{1r}$，ｅ$_{2r}$，ｐ$_{1r}$，ｐ$_{2r}$，ｍ$_{1e}$，ｍ$_{2r}$）はこのＳＮＰでの所与の測定である。
【０２５１】
Ｔ＝（Ｅ$_{1r}$，Ｅ$_{2r}$，Ｐ$_{1r}$，Ｐ$_{2r}$，Ｍ$_{1r}$，Ｍ$_{2r}$）は、仮説によるＴから固定されたｔ＝（Ｐ$_{1r}$，Ｐ$_{2r}$，Ｍ$_{1r}$，Ｍ$_{2r}$）および（Ｅ$_{1r}$，Ｅ$_{2r}$）での推定された「真の」値である。（Ｅ$_{1r}$はＰ$_{1r}$、Ｐ$_{2r}$の一方であり、Ｅ$_{2r}$はＭ$_{1r}$、Ｍ$_{2r}$の一方である）。
Ｐ（Ｍ$_{r}$＝（ｅ$_{1r}$，ｅ$_{2r}$，ｐ$_{1r}$，ｐ$_{2r}$，ｍ$_{1r}$，ｍ$_{2r}$）／Ｔ＝（Ｅ$_{1r}$，Ｅ$_{2r}$，Ｐ$_{1r}$，Ｐ$_{2r}$，Ｍ$_{1r}$，Ｍ$_{2r}$））＝
Ｐ（ｅ$_{1r}$／Ｅ$_{1r}$）＊Ｐ（ｅ$_{2r}$／Ｅ$_{2r}$）＊Ｐ（ｐ$_{1r}$／Ｐ$_{1r}$）＊Ｐ（ｐ$_{2r}$／Ｐ$_{2r}$）＊Ｐ（ｍ$_{1r}$／Ｍ$_{1r}$）＊Ｐ（ｍ$_{2r}$／Ｍ$_{2r}$）
ｐ$_{eir}$＝ｐ（ＳＮＰ　ｒについての胚値ｉを正確に測定）
ｐ$_{pri}$＝Ｐ（ＳＮＰ　ｒについての父親値ｉを正確に測定）
ｐ$_{mri}$＝Ｐ（ＳＮＰ　ｒについての母親値ｉを正確に測定）とすれば、
【０２５２】
【数５６】

$$P(e_{1r} / E_{1r}) = \begin{cases} p_{er1} & e_{1r} = E_{1r} \\ (1 - p_{er1}) * p(e_{1r}, E_{1r}, r) & e_{1r} \neq E_{1r} \end{cases}$$
$$= I_{e_{1r}=E_{1r}} * p_{er1} + I_{e_{1r}\neq E_{1r}} * (1 - p_{pr1}) * p(e_{1r}, E_{1r}, r) = F(e_{1r}, E_{1r}, p_{er1}, r)$$

であり、ここで、測定バイアスがなければ、ｐ（ｅ$_{1r}$，Ｅ$_{1r}$，ｒ）＝１／３であり、そうでなければ、それはＨａｐｍａｐプロジェクトからのデータのような実験データから決定することができる。
【０２５３】
Ｐ（（Ｐ$_{1r}$，Ｐ$_{2r}$，Ｍ$_{1r}$，Ｍ$_{2r}$）＝ｔ）の誘導
ｔ＝（ｔ$_{1}$，ｔ$_{2}$，ｔ$_{3}$，ｔ$_{4}$）については：
Ｐ（（Ｐ$_{1r}$，Ｐ$_{2r}$，Ｍ$_{1r}$，Ｍ$_{2r}$）＝（ｔ$_{1}$，ｔ$_{2}$，ｔ$_{3}$，ｔ$_{4}$））＝Ｐ（Ｐ$_{1r}$＝ｔ$_{1}$）＊Ｐ（Ｐ$_{2r}$＝ｔ$_{2}$）＊Ｐ（Ｍ$_{1r}$＝ｔ$_{3}$）＊Ｐ（Ｍ$_{2r}$＝ｔ$_{4}$）である。（Ｐ$_{1}$，Ｐ$_{2}$，Ｍ$_{1}$，Ｍ$_{2}$）のｎの試料があると仮定し、全ての父性および母性値は独立しており、｛Ａ，Ｃ，Ｔ，Ｇ｝におけるｔ$_{i}$についてはｔ＝（ｔ$_{1}$，ｔ$_{2}$，ｔ$_{3}$，ｔ$_{4}$）であると推定される。
【０２５４】
ｔ$_{1}$＝Ａについて特定のｐ$_{1A}$＝Ｐ（Ｐ$_{1}$＝ｔ$_{1}$）を得るためには、いずれものデータの不存在下において、この確率は０および１の間の何かであり得ると推定され、従って、それはＵ（０，１）の値が割り当てられる。データの獲得に関しては、これは新しい値で更新され、このパラメーターの分布はベータ分布となる。Ｐ$_{1}$のｎの観察のうち、ｈの値Ｐ１＝Ａ、およびｗ＝（事象Ｐ$_{1}$＝Ａ）およびＤ＝（所与のデータ）がある。先のセクションにおいて、ｐ（ｗ｜データ）についてα＝ｈ＋１、β＝ｎ－ｈ＋１でのβ分布Ｂ（α，β）の形式が記載されている（方程式（８）参照）。予測された値およびＸ～Ｂ（α，β）分布の偏差は：
【０２５５】

【数５７】

$$EX=\frac{\alpha}{\alpha+\beta}$$

$$VX=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

である。従って、パラメーターの事後平均値はp_{1rA}＝Ｐ（P_{1r}＝Ａ｜Ｄａｔａ）＝（ｈ＋１）／（ｎ＋２）である。同様に、p_{1rB}＝（＃（p_{1r}＝Ｂ）＋１）／（ｎ＋２），．．．m_{2rG}＝（＃（m_{2r}＝Ｇ）＋１）／（ｎ＋２）などである。かくして、全ての値p_{1rA}，．．．，m_{2rG}が導かれ：

【０２５６】

【数５８】

$$P((P_{1r},P_{2r},M_{1r},M_{2r})=(t_1,t_2,t_3,t_4))=p_{1rt_1}*p_{2rt_2}*m_{1rt_3}*m_{2rt_4}$$

である。

【０２５７】

Ｐ（Ｈ）の誘導

H_i＝（p_i＊，m_1＊）での仮説Ｈ＝（H_1，．．．，H_k）の確率は、染色体交差の量に依存する。例えば、

Ｐ（交差）＝０であれば、もし｛（ｐ１１，ｐ２１，．．．ｐｓ１），（ｐ１２，ｐ２２，．．．，ｐｓ２）におけるｐ＊、｛（ｍ１１，ｍ２１，．．．，ｍｓ１），（ｍ１２，ｍ２２，．．．，ｍｓ２）｝におけるｍ＊であれば、Ｐ（Ｈ＝１／４であって、Ｈ＝（ｐ＊，ｍ＊）であり、そうでなければ０であり

Ｐ（交差）＞０であれば、各ＳＮＰの間の交差の確率を一体化させるのは重要である。

【０２５８】

仮説Ｈは、各ＳＮＰについての父性および母性染色体についての仮説、独立している、p_i＊∈｛p_{1i}，p_{2i}｝およびm_i＊∈｛m_{1i}，m_{2i}｝、すなわち、Ｈ＝（H_p＊，H_m）よりなり、ここで、H_p＝（p_1＊，．．．p_k＊）およびH_m＝（m_1＊，．．．m_k＊）である。

Ｐ（Ｈ＝Ｐ（H_p）＊Ｐ（H_m）。ＳＮＰはロケーションを増大させることによって秩序化され、

【０２５９】

【数５９】

$$P(H_p)=\frac{1}{4}\prod_{i=2}^{k}(PC_i*(1-I_i)+(1-PC_i)*I_i$$

であると仮定し、ここで、PC_i＝Ｐ（交差（r_{i-1}，r_i））、すなわち、ＳＮＰr_{i-1}，r_iの間のどれかの交差の確率であり、もしp_i＊，p_{i-1}＊が共にp_1またはp_2に由来するならば、I_i＝１であり、そうでなければそれは０である。

【０２６０】

Ｐ（交差（ａ，ｂ））の誘導

（塩基で与えた）塩基ロケーション1_a、1_bにおけるＳＮＰ　ａ，ｂを仮定すれば、交差の確率は：

Ｐ（l_a，l_b）＝０．５（１－ｅｘｐ（－２Ｇ（l_a，l_b））

として近似され、ここで、Ｇ（1_a，1_b）＝ロケーション1_a，1_bの間のモルガンで表した遺伝子距離。Ｇについての正確な閉じた形態の関数はないが、それはＧ（l_a，l_b）＝｜l_a－l_b｜＊$1e^{-8}$として緩く見積もられる。良好な近似は、全てのロケーションにわたってのｉスパンニングについての、塩基ロケーションs_iおよび距離Ｇ（s_i，s_{i+1}）のＨａｐＭａｐデータベースを利用することによって用いることができる。特に、

【０２６１】
【数６０】

$$G(l_a,l_b)=\sum_{l_a<s_i<l_b}G(s_i,s_{i+1})$$

であり、従って、それは交差確率で用いることができる。
【０２６２】
Ｐ（Ｍ）の誘導
一旦Ｐ（Ｍ｜Ｈ）が知られていれば、Ｐ（Ｈ）はS_Hにおける全ての異なるＨについて見出すことができる。
【０２６３】
【数６１】

$$P(M)=\sum_{H\in S_H}P(M\mid H)P(H)$$

。
【０２６４】
最大確率の仮説を導くためのより便宜な方法
コンピュータ時間の制限、および前記した方法の複雑性の指数関数スケーリングを仮定すれば、ＳＮＰの数が増加するにつれ、ある場合には、より便宜な方法を用いて、最大確率の仮説を決定し、かくして、関連するＳＮＰ要求をなすのが必要であろう。これを達成するためのより迅速な方法は以下の通りであり：
以前より：
Ｐ（Ｈ｜Ｍ）＝Ｐ（Ｍ｜Ｈ）＊Ｐ（Ｈ）／Ｐ（Ｍ）、$argmax_H$Ｐ（Ｈ｜Ｍ）＝$argmax_H$およびＰ（Ｍ｜Ｈ）＊Ｐ（Ｈ）＝$argmax_H$Ｆ（Ｍ，Ｈ）であり、目的はＦ（Ｍ，Ｈ）を最大化するＨを見出すことである。
【０２６５】
$M_{(s,k)}$＝スニップｓないしｋについての測定、$H_{(s,k)}$＝スニップｓないしｋについての仮説、および短いものについて、$M_{(k,k)}=M_k$を仮定すれば、$H_{(k、k)}=H_k$＝スニップｋについての測定および仮説である。先に示したように：
【０２６６】
【数６２】

$$P(M_{(1,k)}\mid H_{(1,k)})=\prod_{i=1}^{k}P(M_i\mid H_i)=P(M_k\mid H_k)*\prod_{i=1}^{k-1}P(M_i\mid H_i)=P(M_k\mid H_k)*P(M_{(1,k-1)}\mid H_{(1,k-1)})$$

であり、また、
【０２６７】
【数６３】

$$P(H_{(1,k)})=1/4*\prod_{i=2}^{k}PF(H_{i-1},H_i)=PF(H_{k-1},H_k)*1/4*\prod_{i=2}^{k-1}PF(H_{i-1},H_i)=PF(H_{k-1},H_k)*P(H_{(1,k-1)})$$

であり、ここで、
【０２６８】
【数６４】

$$PF(H_{i-1},H_i)=\begin{cases}1-PC(H_{i-1},H_i) & H_{i-1}=H_i\\ PC(H_{i-1},H_i) & H_{i-1}\neq H_i\end{cases}$$

および$PC(H_{i-1},H_i)=H_{i-1},H_i$の間の交差の確率。
【０２６９】
従って、最後には、ｎのスニップについては：
Ｆ（Ｍ，Ｈ）＝Ｐ（Ｍ｜Ｈ）＊Ｐ（Ｈ）＝Ｐ（$M_{(1,n)}$，$H_{(1,n)}$）＊Ｐ（Ｈ（

１，ｎ））
＝Ｐ（Ｍ（１，ｎ－１），Ｈ（１，ｎ－１））＊Ｐ（Ｈ（１，ｎ－１））＊Ｐ（Ｍｎ｜Ｈｎ）＊ＰＦ（Ｈｎ－１，Ｈｎ）
であり、従って：Ｆ（Ｍ，Ｈ）＝Ｆ（Ｍ（１，ｎ），Ｈ（１，ｎ））＝Ｆ（Ｍ（１，ｎ－１），Ｈ（１，ｎ－１））＊Ｐ（Ｍｎ｜Ｈｎ）＊ＰＦ（Ｈｎ－１，Ｈｎ）である。かくして、ｎのスニップについての計算をｎ－１スニップについての計算に代えることが可能である。
ｎについてのスニップのｎについてのＨ＝（Ｈ１，．．．Ｈｎ）仮説では：
【０２７０】
【数６５】

$$\max_{H} F(M,H) = \max_{(H_{(1,n1)},H_n)} F(M,(H_{(1,n-1)},H_n) = \max_{H_n} \max_{H_{(1,n-1)}} F(M,(H_{(1,n-1)},H_n) = \max_{H_n} G(M_{(1,n)},H_n)$$

であり、ここに
【０２７１】
【数６６】

$$G(M_{(1,n)},H_n) = \max_{H_{(1,n-1)}} F(M_{(1,n)},(H_{(1,n-1)},H_n) =$$
$$\max_{H_{(1,n-1)}} F(M_{(1,n-1)},H_{(1,n-1)}) * P(M_n \mid H_n) * PF(H_{n-1},H_n) =$$
$$P(M_n \mid H_n) * \max_{H_{(1,n-1)}} F(M_{(1,n-1)},H_{(1,n-1)}) * PF(H_{n-1},H_n) =$$
$$P(M_n \mid H_n) * \max_{H_{n-1}} \max_{H_{(1,n-2)}} F(M_{(1,n-1)},(H_{(1,n-2)},H_{n-1}) * PF(H_{n-1},H_n) =$$
$$P(M_n \mid H_n) * \max_{H_{n-1}} PF(H_{n-1},H_n) * G(M_{(1,n-1)},H_{n-1})$$

である。まとめると：
【０２７２】
【数６７】

$$\max_{H} F(M,H) = \max_{H_n} G(M_{(1,n)},H_n)$$

であり、ここで、Ｇは帰納的に見出すことができ：ｉ＝２，．．ｎについては、
【０２７３】
【数６８】

$$G(M_{(1,n)},H_n) = P(M_n \mid H_n) * \max_{H_{n-1}} \left[PF(H_{n-1},H_n) * G(M_{(1,n-1)},H_{n-1}) \right]$$

および $G(M_{(1,1)},H_1) = 0.25 * P(M_1 \mid H_1)$

である。
【０２７４】
最良の仮説は以下のアルゴリズムに従って見出すことができる：
工程１：Ｉ＝１では、Ｈ１についての４つの仮説を作り出し、これらの各々についてのＧ（Ｍ１｜Ｈ１）を作り、Ｇ１、Ｇ２、Ｇ３、Ｇ４を覚える。
工程２：Ｉ＝２では：Ｈ２についての４つの仮説を作り出し、前記式を用いてＧ（Ｍ（１，２）｜Ｈ２）を作成し：
【０２７５】
【数６９】

$$G(M_{(1,2)},H_2) = P(M_2 \mid H_2) * \max_{H_1} \left[PF(H_1,H_2) * G(M_1,H_1) \right]$$

これらの新しい４つのＧｎを覚える。
ｋ＝ｎまで、ｋｉ＝ｋｉ－１＋１にてＩ＝ｋにつき工程２を反復し：Ｈｋについて４つの

仮説を作り出し、Ｇ（Ｍ（１，ｋ）｜Ｈｋ）

【０２７６】

【数７０】

$$G(M_{(1,k)},H_k)=P(M_k\mid H_k)*\max_{H_{k-1}}\left[PF(H_{k-1},H_k)*G(M_{(1,k-1)},H_{k-1})\right]$$

を作成し、これらの４つのＧｎを覚える。

【０２７７】

いずれかの時点において覚えるべきただ４つの仮説、および一定数の操作があるので、アルゴリズムは線形である。

【０２７８】

Ｐ（Ｍ）：Ｐ（Ｈ｜Ｍ）＝Ｐ（Ｍ｜Ｈ）＊Ｐ（Ｈ）／Ｐ（Ｍ）＝Ｆ（Ｍ，Ｈ）／Ｐ（Ｍ））を見出すために、前記したように：

【０２７９】

【数７１】

$$P(M)=P(M_{(1,k)})=\sum_{H_{(1,k)}}P(M_{(1,k)}\mid H_{(1,k)})*P(H_{(1,k)})$$

$$=\sum_{H_k}P(M_K\mid H_k)\sum_{H_{(1,k-1)}}P(M_{(1,k-1)}\mid H_{(1,k-1)})*P(H_{(1,k-1)})*PF(H_{k-1},H_k)$$

$$=\sum_{H_k}P(M_K\mid H_k)*W(M_{(1,k-1)}\mid H_k)$$

であり、ここで、

【０２８０】

【数７２】

$$W(M_{(1,k-1)}\mid H_k)=\sum_{H_{(1,k-1)}}P(M_{(1,k-1)}\mid H_{(1,k-1)})*P(H_{(1,k-1)})*PF(H_{k-1},H_k)$$

である。

Ｗ（Ｍ，Ｈ）は帰納を用いることによって解くことができる：

【０２８１】

【数７３】

$$W(M_{(1,k-1)}\mid H_k)=\sum_{H_{(1,k-1)}}P(M_{(1,k-1)}\mid H_{(1,k-1)})*P(H_{(1,k-1)})*PF(H_{k-1},H_k)$$

$$=\sum_{H_{k-1}}P(M_{k-1}\mid H_{k-1})\sum_{H_{(1,k-2)}}P(M_{(1,k-2)}\mid H_{(1,k-2)})*P(H_{(1,k-2)})*PF(H_{k-2},H_{k-1})*PF(H_{k-1},H_k)$$

$$=\sum_{H_{k-1}}P(M_{k-1}\mid H_{k-1})*PF(H_{k-1},H_k)*W(M_{(1,k-2)}\mid H_{k-1})$$

従って：

【０２８２】

【数７４】

$$W(M_{(1,k-1)}\mid H_k)=\sum_{H_{k-1}}P(M_{k-1}\mid H_{k-1})*PF(H_{k-1},H_k)*W(M_{(1,k-2)}\mid H_{k-1})$$

および$W(M_{(1,1)}\mid H_2)=\sum_{H_1}P(M_1\mid H_1)*0.25*PF(H_1,H_2)$

である。

【０２８３】

アルゴリズムは前記の場合に同様であり、ここで、ｉ＝２：ｎであって、各工程において、Ｗ（ｉ）の新しいセットを、最終工程が最適化されたＷを生じるまで作り出される。

【０２８４】

ｄ$_1$、ｄ$_2$、ｈ、ｐｄ$_1$、ｐｄ$_2$、ｐｈからのｐ$_1$、ｐ$_2$、ｐｐ$_1$、ｐｐ$_2$値の誘導

説明の目的で、このセクションは父親のジプロイドおよびハプロイドデータに焦点を合わせるが、同一アルゴリズムを母親に適用することができることに注意するのは重要である。

－ｄ$_1$，ｄ$_2$－ジプロイド測定での対立遺伝子要求

－ｈ－ハプロイド測定についての対立遺伝子要求

－ｐ$_{d1}$，ｐ$_{d2}$－ジプロイド測定の各々についての正しい対立遺伝子要求の確率

－ｐ$_h$－ハプロイド測定についての正しい対立遺伝子要求の確率

これらのデータは開示されたアルゴリズムについての以下の入力パラメーターにマッピングすべきである：

－ｐ$_1$－ハプロイド細胞および１つのジプロイド細胞に対応する対立遺伝子

－ｐ$_2$－残りのジプロイド細胞に対応する対立遺伝子

－ｐ$_{p1}$，ｐ$_{p2}$－正しい対立遺伝子要求の確率

ｈはｄ$_1$に対応するので、ｐ$_1$の値を見出すためには、ｈおよびｄ$_1$を用いる必要がある。次いで、ｐ$_2$は自動的にｄ$_2$に対応する。同様に、もしｈがｄ$_2$に対応すれば、ｐ$_1$の値を見出すためには、ｈおよびｄ$_2$を用いる必要があり、次いで、ｐ$_2$はｄ$_1$に対応するであろう。

【０２８５】

用語「対応する」を用いる。というのは、それは、異なる測定結果および集団頻度に依存して、「等しい」または「より高い確率で由来する」を意味することができるからである。

【０２８６】

アルゴリズムの目標は、生の測定ｈ、ｄ$_1$、ｄ$_2$、ｐ$_h$、ｐ$_{d1}$、ｐ$_{d2}$および集団頻度の結果を超えて隠された「真の」対立遺伝子値の確率を計算することである。

【０２８７】

基本的なアルゴリズム工程は以下の通りである：

（ｉ）ｈ、ｄ$_1$、ｄ$_2$、ｐ$_h$、ｐ$_{d1}$、ｐ$_{d2}$値、および集団頻度データに基づいてｈがｄ$_1$またはｄ$_2$に対応するかを決定する、

（ｉｉ）対立遺伝子要求をｐ$_1$およびｐ$_2$に帰属させ；工程（１）に基づいて確率ｐ$_{p1}$およびｐ$_{p2}$を計算する。

【０２８８】

ｈのｄ$_1$またはｄ$_2$への帰属

２つの仮説：

Ｈ$_1$：ｈはｄ$_1$に対応する（ｈはｄ$_1$に由来する）

Ｈ$_2$：ｈはｄ$_2$に対応する（ｈはｄ$_2$に由来する）

を確率する。仕事は、測定Ｍ：を仮定してこれらの２つの仮説の確率を計算することである：

Ｐ（Ｈ$_1$／Ｍ（ｈ，ｄ$_1$，ｄ$_2$，ｐ$_h$，ｐ$_{d1}$，ｐ$_{d2}$））およびＰ（Ｈ$_2$／Ｍ（ｈ，ｄ$_1$，ｄ$_2$，ｐ$_h$，ｐ$_{d1}$，ｐ$_{d2}$））。

以下、（テキストを単純化するために、これらをＰ（Ｈ$_1$／Ｍ）およびＰ（Ｈ$_2$／Ｍ））という）。

【０２８９】

これらの確率を計算するために、ベイズ則：

【０２９０】

【数７５】

$$P(H_1 \mid M) = \frac{P(M \mid H_1) * P(H_1)}{P(M)};\ P(H_2 \mid M) = \frac{P(M \mid H_2) * P(H_2)}{P(M)}$$

を適用し、ここで、Ｐ（Ｍ）＝Ｐ（Ｍ／Ｈ$_1$）＊Ｐ（Ｈ$_1$）＋Ｐ（Ｍ／Ｈ$_2$）＊Ｐ（Ｈ$_2$）である。仮説Ｈ$_1$およびＨ$_2$は同等にありそうなので、Ｐ（Ｈ$_1$）＝Ｐ（Ｈ$_2$）＝０．

５であり、従って：

【０２９１】

【数７６】

$$P(H_1 \mid M) = \frac{P(M \mid H_1)}{P(M \mid H_1) + P(M \mid H_2)} \text{ および } P(H_2 \mid M) = \frac{P(M \mid H_2)}{P(M \mid H_1) + P(M \mid H_2)}$$

である。

【０２９２】

Ｐ（Ｍ／Ｈ$_1$）およびＰ（Ｍ／Ｈ$_2$）を計算するためには、ジプロイド結果ｄ$_1$およびｄ$_2$の全ての可能な値のセット、Ω＝｛ＡＡ，ＡＣ，．．．，ＧＧ｝、すなわち、Ａ、Ｃ、Ｔ、Ｇのいずれかの組合せ、いわゆる基礎となる状態を考慮しなければならない。仮説を基礎となる状態に適用する場合（すなわち、仮説Ｈ$_1$またはＨ$_2$に基づいてｈの推定値を値ｄ$_1$およびｄ$_2$に伴わせる）、ｈ、ｂ$_1$およびｄ$_2$についての「真の値」Ｈ、Ｄ$_1$およびＤ$_2$の全ての可能な組合せ（状態Ｓ＝｛ｓ$_1$，ｓ$_2$，．．．，ｓ１６｝）の以下の表を、各々、作成することができる。

【０２９３】

【数７７】

仮説 H_1: $h=d_1$		Ω={AA,AC,…,GG}	
状態	H	D_1	D_2
s_1	A	A	A
s_2	A	A	C
s_3	A	A	T
s_4	A	A	G
s_5	C	C	A
s_6	C	C	C
s_7	C	C	T
s_8	C	C	G
s_9	T	T	A
s_{10}	T	T	C
s_{11}	T	T	T
s_{12}	T	T	G
s_{13}	G	G	A
s_{14}	G	G	C
s_{15}	G	G	T
s_{16}	G	G	G

仮説 H_2: $h=d_2$		Ω={AA,AC,…,GG}	
状態	H	D_1	D_2
s_1	A	A	A
s_2	C	A	C
s_3	T	A	T
s_4	G	A	G
s_5	A	C	A
s_6	C	C	C
s_7	T	C	T
s_8	G	C	G
s_9	A	T	A
s_{10}	C	T	C
s_{11}	T	T	T
s_{12}	G	T	G
s_{13}	A	G	A
s_{14}	C	G	C
s_{15}	T	G	T
s_{16}	G	G	G

。

【０２９４】

「真の値」H、D_1およびD_2は知られておらず、かつ生の測定結果h、d_1、d_2、p_h、p_{d1}、p_{d2}のみが知られているので、全セットΩにわたるP（M／H_1）およびP（M／H_2）の計算は以下のように行わなければならない：

【０２９５】

【数７８】

$$P(M \mid H_1) = \sum_{(D_1,D_2)\in\Omega} P(M(h,d_1,d_2) \mid H_1 \,\&\, D_1,D_2) * P(D_1,D_2)$$

$$P(M \mid H_2) = \sum_{(D_1,D_2)\in\Omega} P(M(h,d_1,d_2) \mid H_2 \,\&\, D_1,D_2) * P(D_1,D_2)$$

もし、計算の目的で、d_1およびd_2、ならびにp_{d1}およびp_{d2}が独立した変数であると仮定すれば：

【０２９６】

【数７９】

$$P(M \mid H_1) = \sum_{\Omega} P(M(h,d_1,d_2) \mid H_1 \,\&\, D_1,D_2) * P(D_1,D_2) =$$

$$\sum_{S} P(M(h) \mid H) * P(M(d_1) \mid D_1) * P(M(d_2) \mid D_2) * P(D_1) * P(D_2)$$

を示すことができる。｛h，d_1，d_2｝におけるhについての、前記した最後の合計：P（M（ｘ）／X）下で３項を計算する。

【０２９７】

（「真の対立遺伝子値」をヒットさせる）正しい対立遺伝子要求の確率の計算は、対立遺伝子Xの真の値を仮定した結果ｘの測定に基づく。もし測定された値ｘおよび真の値Xが等しいならば、確率はp_xである（正しい測定の確率）。もしｘおよびXが異なるならば、その確率は（１－p_x）／３である。例えば、Ｘ＝Ｃ、および測定された値がｘ＝Ａである条件下で「真の値」Ｃが見出される確率を計算する。Ａを得る確率はp_xである。Ｃ、ＴまたＧを得る確率は（１－p_x）である。従って、Ｃがヒットする確率は（１－p_x）／３である。というのは、Ｃ、ＴおよびＧは等しくありそうと仮定することができるからである。

【０２９８】

もしインジケーター変数I_xが計算に含まれ、ここで、もしｘ＝ＸであればI_x＝１であり、もしｘ≠ＸならばI_x＝０であれば、確率は以下の通りである：

P（M（ｘ）／X）＝$I_{\{x=X\}} * p_x$＋（１－$I_{\{x=X\}}$）＊（１／３）＊（１－p_x）、｛h，d_1，d_2｝におけるｘ。

さて、P（M｜H_1）における最後の２つの項を考える。P（D_1）およびP（d_2）は、事前の知識から知ることができる、対立遺伝子Ａ、Ｃ、ＴおよびＧの集団頻度である。

【０２９９】

特定の測定Ｍ（ｈ＝Ａ，d_1＝Ｇ，d_2＝Ｃ）を仮定して、特定の状態s_2について先に示した表現を考慮する：

【０３００】

【数８０】

$$P(M(h)|H)*P(M(d_1)|D_1)*P(M(d_2)|D_2)*P(D_1)*P(D_2) =$$
$$= P(M(h)=A|H=A)*P(M(d_1)=G|D_1=A)*P(M(d_2)=C|D_2=C)*P(D_1=A)*P(D_2=C) =$$
$$= p_h*((1-p_{d1})/3)*p_{d2}*f(D_1=A)*f(D_2=C)$$

同様に、残りの１５の状態、およびセットΩにわたる合計について、特定の測定（この場合、Ｍ（ｈ＝Ａ，d_1＝Ｇ，d_2＝Ｃ））を仮定して（１）を計算する。

【０３０１】

さて、P（M／H_1）およびP（M／H_2）は計算された。最後に、前記したようにP（H_1／M）およびP（H_1／M）を計算する：

【０３０２】
【数８１】

$$P(H_1 \mid M) = \frac{P(M \mid H_1)}{P(M \mid H_1) + P(M \mid H_2)}$$

$$P(H_2 \mid M) = \frac{P(M \mid H_2)}{P(M \mid H_1) + P(M \mid H_2)}$$

。
【０３０３】
対立遺伝子要求の帰属および対応する確率
さて、４つの異なる仮説：
H_{p2A}：p_2の「真の値」はＡである、
H_{p2C}：p_2の「真の値」はＣである、
H_{p2T}：p_2の「真の値」はＴである、
H_{p2G}：p_2の「真の値」はＧである、
を確立し、Ｐ（H_{p2A}／Ｍ）、Ｐ（H_{p2C}／Ｍ）、Ｐ（H_{p2T}／Ｍ）、Ｐ（H_{p2G}／Ｍ）を計算する。最高の値は、特定の対立遺伝子の要求および対応確率を決定する。
【０３０４】
p_2の起源は未知である（それは、Ｐ（H_2／Ｍ）の確率でもってd_1から、および確率Ｐ（H_1／Ｍ）をもってd_2から由来する）ので、p_2対立遺伝子がd_1またはd_2に由来する双方の場合を考慮しなければならない。仮説H_Aについては、ベイズ則を適用し、
【０３０５】
【数８２】

$$P(H_{p2A} \mid M) = P(H_{p2A} \mid M, H_1) * P(H_1 \mid M) + P(H_{p2A} \mid M, H_2) * P(H_2 \mid M)$$

が得られる。
【０３０６】
Ｐ（H_1／Ｍ）およびＰ（H_2／Ｍ）は工程１においてすでに決定されている。ベイズ則によると、
【０３０７】
【数８３】

$$P(H_{p2A} \mid H_1, M) = \frac{P(M \mid H_1, H_{p2A}) * P(H_1, H_{p2A})}{P(H_1, M)}$$

である。H_1は、p_2がd_2に由来することを示唆するので、前記したように、
【０３０８】
【数８４】

$$P(M \mid H_1, H_{p2A}) = P(M(d_2) \mid D_2 = A) = I_{\{d_2=D_2\}} * p_{d2} + (1 - I_{\{d_2=D_2\}}) * (1 - p_{d2})/3$$

$$P(H_1, M/H_{p2A}) = P(M(d_2)/D_2 = A) = I_{\{d2=D2\}} * p_{d2} + (1 - I_{\{d2=D2\}}) * (1/3) * (1 - p_{d2})$$

である。
Ｐ（H_{p2A}）＝Ｐ（D_2＝Ａ）＝f_{d2}（Ａ）であり、ここで、f_{d2}（Ａ）は集団頻度データから得られる。
Ｐ（H_1，Ｍ）＝Ｐ（H_1，Ｍ／H_{p2A}）＊Ｐ（H_{p2A}）＋Ｐ（H_1，Ｍ／H_{p2C}）＊Ｐ（H_{p2C}）＋Ｐ（H_1，Ｍ／H_{p2T}）＊Ｐ（H_{p2T}）＋Ｐ（H_1，Ｍ／H_{p2G}）＊Ｐ（H_{p2G}）
同様に、Ｐ（H_{p2A}＆H_2／Ｍ）を計算する。

P（H_{p2A}／M）＝P（H_{p2A}＆H_1／M）＋P（H_{p2A}＆H_2／M）であり、したがって、p_2がAに等しい確率は計算された。C、TおよびGについての計算を反復する。最高の値は、p_2対立遺伝子要求および対応する確率の回答を与えるであろう。

【０３０９】

対立遺伝子要求のp_1への帰属（ハプロイド細胞、および１つのジプロイド細胞に対応する対立遺伝子）

前記したように、４つの異なる仮説：

H_{p1A}：p_1の「真の値」はAである、

H_{p1C}：p_1の「真の値」はCである、

H_{p1T}：p_1の「真の値」はTである、

H_{p1G}：p_1の「真の値」はGである、

を確立し、P（H_{p1A}／M）、P（H_{p1C}／M）、P（H_{p1T}／M）、P（H_{p1G}／M）を計算する。

【０３１０】

これは、H_{p1A}の仕上げである。「真の場合」の場合においては、ハプロイドおよび対応するジプロイド細胞がAと等しい場合にのみp_1はAと等しい。したがって、p_1およびp_{p1}を計算するためには、ハプロイドおよび対応するジプロイド細胞が等しい状況を考慮しなければならない。したがって、仮説H_{p1A}：p_1の「真の値」はAであって、H_{hdA}となる：ハプロイド細胞および対応するジプロイド細胞の「真の値」はAである。

【０３１１】

hの起源は未知である（それは、P（H_1／M）の確率でもってd_1から、および確率P（H_2／M）でもってd_2から由来する）ので、h対立遺伝子がd_1またはd_2に由来する双方の場合を考慮し、p_1の決定におけるそれを実行しなければならない。それは、ベイズ則を用いると：

P（H_{hdA}｜M）＝P（H_{hdA}｜M，H_1）＊P（H_1｜M）＋P（H_{hdA}｜M，H_2）＊P（H_2｜M）

を意味する。

【０３１２】

前記したように、P（H_1／M）およびP（H_2／M）は先の計算から知られている。

【０３１３】

【数８５】

$$P(H_{hdA} \mid H_1, M) = \frac{P(H_1, M \mid H_{hdA}) * P(H_{hdA})}{P(H_1, M)}$$

P（H_1，M／H_{hdA}）＝P（M（h）／H＝A）＊P（M（d_1）／D_1＝A）＝
＝［$I_{\{h=H\}}$＊p_h＋（１－$I_{\{h=H\}}$）＊（１／３）＊（１－p_h）］＊［$I_{\{d1=D1\}}$＊p_{d1}＋（１－$I_{\{d1=D1\}}$）＊（１／３）＊（１－p_{d1}）］

である。というのは、H１は、p_1がd_1に由来することを示唆するからである。P（H_{hdA}）＝P（h＝A）＊P（D_1＝A）f_h（A）＊f_{d1}（A）であり、ここで、f_h（A）およびf_{d2}（A）は集団頻度データから得られる。P（H_1，M）＝P（H_1，M／H_{hdA}）＊P（H_{hdA}）＋P（H_1，M／H_{hdC}）＊P（H_{hdC}）＋P（H_1，M／H_{hdT}）＊P（H_{hdT}）＋P（H_1，M／H_{hdG}）＊P（H_{hdG}）。

【０３１４】

同様に、P（H_{hdA}＆H_2／M）を計算する。

P（H_{hdA}／M）＝P（H_{hdA}＆H_1／M）＋P（H_{hdA}＆H_2／M）であり、今や、我々はp_1がAと等しい確率を計算した。C、TおよびGについての計算を反復する。最高の値はp_1対立遺伝子要求および対応する確率の回答を与えるであろう。

【０３１５】

例としての入力

２つの入力の例を示す。最初の例は、共分離する低い傾向があるＳＮＰのセットのものであり、すなわち、ＳＮＰは染色体を通って拡大し、入力データを表３に示す。第二の例は、共分離する高い傾向があるＳＮＰのセットのものであり、すなわち、ＳＮＰは染色体上にクラスター形成し、インプットデータを表４に示す。双方のデータのセットは個体の測定されたＳＮＰデータ、個体の親のＳＮＰデータおよび対応する信頼性値を含む。このデータは現実の人々から測定された現実のデータであることを注記する。各列は、１つの特定のＳＮＰロケーションについての測定を表す。行は、行の見出しによって示されるデータを含む。行の見出し中の略語に対する鍵は以下の通りである：

ｆａｍｉｌｙ＿ｉｄ＝各人についてのユニークなｉｄ（事務的理由について含む）

ｓｎｐ＿ｉｄ＝ＳＮＰ同定番号

ｅ１，ｅ２＝胚についてのＳＮＰヌクレオチド値

ｐ１，ｐ２＝父親についてのＳＮＰヌクレオチド値

ｍ１，ｍ２＝母親についてのＳＮＰヌクレオチド値

ｐｅ１，ｐｅ２＝ｅ１，ｅ２についての測定精度ｐｐ１，ｐｐ２＝ｐ１，ｐ２についての測定精度

ｐｍ１，ｐｍ２＝ｍ１，ｍ２についての測定精度

例としての出力

出力データの２つの例を表５および表６に示し、これは各々、表３および表４に掲げたデータからの出力データに対応する。双方の表は、個体の測定されたＳＮＰデータ、個体の親のＳＮＰデータ、個体のＳＮＰデータの最もありそうな真の値、および対応する信頼性を示す。各列は、１つの特定のＳＮＰに対応するデータを表す。行は行の見出しによって示されるデータを含まれる。行の見出し中の略語に対する鍵は以下の通りである：

ｓｎｐ＿ｉｄ＝ＳＮＰ同定番号

ｔｒｕｅ＿ｖａｌｕｅ＝ｅ１，ｅ２についての提案されたヌクレオチド値

ｔｒｕｅ＿ｈｙｐ＝ｅ１，ｅ２の起源についての仮説

ｅｅ＝ｅ１，ｅ２についての測定されたＳＮＰヌクレオチド値

ｐｐ＝ｐ１，ｐ２についての測定されたＳＮＰヌクレオチド値

ｍｍ＝ｍ１，ｍ２についての測定されたＳＮＰヌクレオチド値

ＨｙｐＰｒｏｂ＝最終仮説の確率。出力についてはただ１つの数があるが、優れた行構造のため、この数字は全ての列中に複製される。

【０３１６】

このアルゴリズムは手動で、またはコンピュータによって実施することができることを注記する。表３および表４は、該方法のコンピュータで実施されたバージョンについての入力データの例を示す。表５は表３に示された入力データに対する出力データを示す。表６は、表４に示された入力データに対する出力データを示す。

【０３１７】

シミュレーションアルゴリズム

以下に、システムの一体性を確実とし、およびより広く種々の状況におけるアルゴリズムの現実の効率を評価するためになされた第二のシミュレーションを示す。これを行うために、１，０００のフルシステムシミュレーションを実行した。これは、親遺伝子データをランダムに作り出し、イン・シリコにて減数分裂を模倣して、胚データが得られ、胚データの不完全な測定をシミュレートし、次いで、本明細書中に開示された方法を実行して、シミュレートされた測定胚データを清浄化し、次いで、その「清浄化された」データを「現実の」データと比較することを含む。シミュレーションのより詳細な説明を以下に掲げ、事象のフローの目に見える表示を図１８に掲げる。理論の２つの異なる実施を検定した。より十分な説明を以下に掲げる。

【０３１８】

ＤＨおよびＰＳについてのシミュレーションアルゴリズムおよび結果

双方のアルゴリズムについて、初期入力変数は：

（ｉ）検定すべきＳＮＰのリスト、

（ｉｉ）母性（ｐｏｐｆｒｅｑｌｉｓｔＭＭ）および父性（ｐｏｐｆｒｅｑｌｉｓｔＰＰ）染色体の集団頻度、
（ｉｉｉ）ハプロイド測定（ｐｈ，ｐｅ）についての、および秩序立っていないジプロイド測定（ｐｄ）についての正しい対立遺伝子要求の確率、
である。

【０３１９】

これらの値は、関連するＳＮＰについての経験的なデータ（集団頻度）からの、および測定機器性能（ｐｈ，ｐｄ，ｐｅ）からの結果に基づいて固定すべきである。シミュレーションは、最もありそうな（通知された）、均一な（通知されていない）および非常にありそうにない（極端な場合）のようないくつかのシナリオについて実行した。

【０３２０】

一旦、前記した静的なパラメーターが固定されれば、特定のＳＮＰを仮定した交差確率はすべてのシミュレーションについて同一であり、スニッププロケーション（ＳＮＩＰＬＯＣ＿ＮＡＭＥ＿ＭＡＴ）および遺伝子距離（ＨＡＰＬＯＣ＿ＮＡＭＥ＿ＭＡＴ）についてのデータベースを仮定して該時点に先立って誘導されるであろう。

［ｃｒｏｓｓｐｒｏｂ，ｓｎｉｐｓ］＝
ＧｅｔＣｒｏｓｓＰｒｏｂ（スニップ，ＳＮＩＰＬＯＣ＿ＮＡＭＥ＿ＭＡＴ，パラメーター，ＨＡＰＬＯＣ＿ＮＡＭＥ＿ＭＡＴ）

予備的シミュレーションループ

予備的シミュレーションループは、十分なシミュレーションで用いられるであろう遺伝子データが現実的であることを示すものである。工程１ないし５を１０，０００回反復した。このシミュレーションが、いずれかのまたは双方の親について実行することができ；該工程は同一であることを注記する。この場合、シミュレーションは説明目的のために父性ケースで実行され、図１８への言及はカッコに入れた図１８中の対応する母性エントリーも含む。

【０３２１】

工程１：オリジナルの親ジプロイド細胞（Ｐ１，Ｐ２）の創製

［Ｐ１，Ｐ２］＝オリジナルの染色体の創製（ｓｎｉｐｓ，ｐｏｐｆｒｅｑｌｉｓｔＰＰ）；１８０１（１８０２）

父親細胞についての各ＳＮＰに対する集団頻度に依存して、オリジナルの父性細胞を創製する。

【０３２２】

工程２：ＤＨＡｌｇｏについてのハプロイドおよび秩序立っていないジプロイドデータの創製

親染色体１８０３の交差をシミュレートして、染色体、交差の２つのセット：Ｐ１Ｃ１、Ｐ２Ｃ１およびＰ１Ｃ２、Ｐ２Ｃ２；１８０４（１８０５）を得る。ハプロイド対立遺伝子ＨＰ　１８０７（１８０８）、この場合は、Ｐ１（というのは、いずれについても差はないからである）についての、（第一のセットからの）交差１８０６後に父親対立遺伝子のうちの１つをピックアップし、ジプロイド対立遺伝子中の順序を混合して、（Ｄ１Ｐ，Ｄ２Ｐ）１８０７（１８０８）を得る。

ＨＰ＝ＰｉｃｋＯｎｅ（Ｐ１Ｃ１，Ｐ２Ｃ１）；
［Ｄ１Ｐ，Ｄ２Ｐ］＝Ｊｕｍｂｌｅ（Ｐ１，Ｐ２）。

【０３２３】

工程３：オリジナルなデータセットへエラーを導入して、測定をシミュレートする。

【０３２４】

正しい測定（ｐｈ－ハプロイド、ｐｄ－ジプロイド測定）の所与の確率に基づき、エラーを測定に導入して、シミュレートされた測定親データ１８１１（１８１２）を得る。

ｈｐ＝ＭａｋｅＥｒｒｏｒ（ＨＰ，ｐｈ）；
ｄ１ｐ＝ＭａｋｅＥｒｒｏｒ（Ｄ１Ｐ，ｐｄ）；
ｄ２ｐ＝ＭａｋｅＥｒｒｏｒ（Ｄ２Ｐ，ｐｄ）。

【０３２５】
工程４：ＤＨＡｌｇｏを適用して、（ｐ１，ｐ２）、（ｐｐ１，ｐｐ２）を得る。
【０３２６】
ＤＨＡｌｇｏは、ハプロイド細胞からの対立遺伝子、およびジプロイド細胞からの秩序立っていない対立遺伝子を取り、これらを生起したもっともありそうな秩序立ったジプロイド対立遺伝子を戻す。ＤＨＡｌｇｏは（Ｐ１，Ｐ２）を再形成するよう試み、また、父親についての見積もり誤差（ｐｐ１，ｐｐ２）を戻す。比較のために、単純な対立遺伝子マッチングを行う経験的アルゴリズムを用いる。目標は、単純な経験的アルゴリズムと比較して、どれくらい開示されたアルゴリズムが良好であるかを比較することである。
［ｐ１，ｐ２，ｐｐ１，ｐｐ２］＝ＤＨＡｌｇｏ（ｈｐ，ｄ１ｐ，ｄ２ｐ，ｐｈ，ｐｄ，ｓｎｉｐｓ，ｐｏｐｆｒｅｑｌｉｓｔＰＰ，‘ＤＨ’）；
［ｐ１ｓ，ｐ２ｓ，ｐｐ１ｓ，ｐｐ２ｓ］＝ＤＨＡｌｇｏ（ｈｐ，ｄ１ｐ，ｄ２ｐ，ｐｈ，ｐｄ，ｓｎｉｐｓ，ｐｏｐｆｒｅｑｌｉｓｔＰＰ，‘ＳＴ’）；
。
【０３２７】
工程５：実行のための統計学の収集
（Ｐ１，Ｐ２）を誘導された（ｐ１，ｐ２）と比較する。
［Ｐ１ｃｍｐ（　：，ｉ），　Ｐ２ｃｍｐ（　：，ｉ），Ｐ１ｐｒｏｂ（　：，ｉ），Ｐ２ｐｒｏｂ（　：，ｉ），Ｐ１ｍｎ（ｉ），Ｐ２ｍｎ（ｉ）］＝ＤＨＳｉｍＶａｌｉｄａｔｅ（Ｐ１，Ｐ２，ｐ１，ｐ２，ｐｐ１，ｐｐ２）；
注意：（$P1S_i$，$P2S_i$，$P1P_i$，$P2P_i$，$P1A_i$，$P2A_i$）＝（$I_{\{P1=p1\}}$，$I_{\{P2=p2\}}$，p_{p1}，p_{p2}，$p1_{acc}$，$p2_{acc}$）であり、ここで、$I_{\{P1=p1\}}$は全てのＳＮＰについての、同様に、$I_{\{P2=p2\}}$についての、ＤＨアルゴリズム精度の見積もり用のバイナリインジケーターアレイである。P_{p1}，P_{p2}は該アルゴリズムに由来する正しい対立遺伝子要求および$p1_{acc}$＝平均（$I_{\{P1=p1\}}$）、すなわち、$p2_{acc}$についてと同様な、ｐ１についてのこの実行に対する平均精度の確率である。
【０３２８】
予備的シミュレーションの結果
１０，０００のシミュレーションを用いて、Ｐ１，Ｐ２からのＤＨアルゴリズムの総じての精度を示す、アルゴリズム精度ＤＨＡｃｃｕｒａｃｙ．Ｐ１＝平均（$P1A_i$）、ＤＨＡｃｃｕｒａｃｙ．Ｐ２＝平均（$P2A_i$）を見積もった。個々のＳＮＰに基づき、各ＳＮＰ　ＳＮＰＡｃｃ．Ｐ１＝平均（$P1S_i$）についての平均精度は、ＳＮＰ，ＳＮＰＰｒｏｂ．Ｐ１＝平均（$P2P_i$）であると正しく測定する見積もられた確率の平均に合致すべきであり、すなわち、もしアルゴリズムが正しく作動すれば、ＳＮＰＡｃｃ．Ｐ１に対する値はＳＮＰＰｒｏ．Ｐ１に密接に対応すべきである。これらの２つの間の関係はそれらの相関によって反映される。
【０３２９】
シミュレーションの１００００ループは異なる設定シナリオで実行した：
（１）基礎となる集団頻度は、より現実的である現存のゲノタイピングデータ、およびＡ、Ｃ、Ｔ、Ｇが各ＳＮＰについて同一の確率を有する均一な集団頻度によって与えられた。
（２）ハプロイドおよび秩序立っていないジプロイド測定（ＰＨ，ＰＤ）についての測定精度に対するいくつかの組合せ。種々の仮定を行った；測定は共に非常に精度があり（０．９５，０．９５）、精度が低く（０．７５，０．７５）、および精度なしまたはランダムであり（０．２５，０．２５）、ならびに（０．９，０．５）、（０．５，０．９）のバランスが取れていない組合せである。現実に最も近いであろうものは、ほぼ０．６ないし０．８の精度であろう。
（３）シミュレーションを、ＤＨＡｌｇｏｒｉｔｈｍおよび単純なマッチングＳＴＡｌｇｏｒｉｔｈｍ双方についてのすべてのこれらの場合に実行して、開示されたアルゴリズム

の性能を評価した。
これらの全ての実行の結果を表７にまとめる。
【０３３０】
開示されたアルゴリズムは、これらのシミュレーションにおいて、特に、不均一な集団頻度、および正しい測定のアンバランスな、または低下した確率の現実的場合について、現存の経験的アルゴリズムよりも良好に実行される。また、個々のＳＮＰについてのアルゴリズム精度の１つの見積もりはこれらの場合において非常に良好であることが確認された。というのは、正しい対立遺伝子要求の見積もられた精度、およびシミュレーション平均精度の間の相関は９９％程度であり、平均比率は１だからである。
【０３３１】
最も現実的場合において、データ集団頻度および（ｐｈ，ｐｄ）＝（０．６，０．８）については、（Ｐ１，Ｐ２）についての正しく検索されたＳＮＰの平均パーセントは実行１において（０．８５２，０．８１６）であって、実行２において（０．６０１，０．６７３）である。
【０３３２】
表７および表８については、「データ」使用集団頻度データで始まる列は経験的結果から取られたものであり、他方、「均一」で始まる列は均一な集団を仮定することに注意されたし。
【０３３３】
表７および表８においては、精度は、正しいＳＮＰ要求がなされ、正しい元の染色体が同定されたＳＮＰの平均パーセントとして定義されることに注意するのは重要である。また、これらのシミュレーションはアルゴリズムの２つの可能な実行を反映するのに注意するのも重要である。良好な結果を与えることができるアルゴリズムを実行する他の方法があり得る。このシミュレーションは、該方法が実施できることを示すつもりだけである。
【０３３４】
十分なシミュレーションループ
工程１ないし８を１００００回反復した。これは、関連する個体、この場合は、親から測定された遺伝子データを用いて標的固体についての測定された遺伝子データを清浄化する十分に開示された方法を検定するためのシミュレーションである。
【０３３５】
工程１：オリジナルの親ジプロイド細胞（Ｐ１，Ｐ２），（Ｍ１，Ｍ２）の創製
［Ｐ１，Ｐ２］＝オリジナルな染色体の創製（ｓｎｉｐｓ，ｐｏｐｆｒｅｑｌｉｓｔＰＰ）；（１８０１）
［Ｍ１，Ｍ２］＝オリジナルな染色体の創製（ｓｎｉｐｓ，ｐｏｐｆｒｅｑｌｉｓｔＭＭ）；（１８０２）
母親および父親細胞についての各ＳＮＰに対する集団頻度に依存して、オリジナルな親細胞を創製する。
【０３３６】
工程２：交差親細胞（Ｐ１Ｃ，Ｐ２Ｃ），（Ｍ１Ｃ，Ｍ２Ｃ）（１８０３）
交差を持つ父性細胞の２つのセットを創製して：第一に、ＤＨＡｌｇｏで用いる（Ｐ１Ｃ１，Ｐ２Ｃ１）が得られ、第二に、ＰＳＡｌｇｏで用いる（Ｐ１Ｃ２，Ｐ２Ｃ２）を得る。（１８０４）
交差を持つ母性細胞の２つのセットを創製して：第一に、ＤＨＡｌｇｏで用いる（Ｍ１Ｃ１，Ｍ２Ｃ１）、およびＰＳＡｌｇｏで用いる（Ｍ１Ｃ２，Ｍ２Ｃ２）を得る。（１８０５）
［Ｐ１Ｃ１，Ｐ２Ｃ１］＝（Ｐ１，Ｐ２，ｓｎｉｐｓ，ｆｕｌｌｐｒｏｂ）を交差させる；
［Ｐ１Ｃ２，Ｐ２Ｃ２］＝（Ｐ１，Ｐ２，ｓｎｉｐｓ，ｆｕｌｌｐｒｏｂ）を交差させる；
［Ｍ１Ｃ１，Ｍ２Ｃ１］＝（Ｍ１，Ｍ２，ｓｎｉｐｓ，ｆｕｌｌｐｒｏｂ）を交差させる

；
［Ｍ１Ｃ２，Ｍ２Ｃ２］＝（Ｍ１，Ｍ２，ｓｎｉｐｓ，ｆｕｌｌｐｒｏｂ）を交差させる
；
。
【０３３７】
工程３　ＤＨＡｌｇｏについてのハプロイド細胞および無秩序ジプロイド細胞を作成する（１８０６）。
【０３３８】
ハプロイド細胞ＨＰについての父性細胞のセットのうち１つ（１８０４，第一のセット）をピックアップし、ジプロイド細胞中の順序を混合して、（Ｄ１Ｐ，Ｄ２Ｐ）（１８０７）を得る。母性細胞（１８０５，第一のセット）についても同様にして、ＭＨ，（Ｄ１Ｍ，Ｄ２Ｍ）を得る。（１８０８）
ＨＰ＝１つの（Ｐ１Ｃ１，Ｐ２Ｃ１）をピックアップする；
ＨＭ＝１つの（Ｍ１Ｃ１，Ｍ２Ｃ１）をピックアップする；
［Ｄ１Ｐ，Ｄ２Ｐ］＝（Ｐ１，Ｐ２）を乱雑とする；
［Ｄ１Ｍ，Ｄ２Ｍ］＝（Ｍ１，Ｍ２）を乱雑とする；
。
【０３３９】
工程４：ジプロイド胚細胞の作成（１８０９）
胚細胞について父性細胞の１つ（１８０４，第二のセット）および母性細胞の１つ（１８０５，第二のセット）をピックアップする。測定目的で順序を混合する。
Ｅ１＝１つの（Ｐ１Ｃ２，Ｐ２Ｃ２）をピックアップする；
Ｅ２＝１つの（Ｍ１Ｃ２，Ｍ２Ｃ２）をピックアップする；
［Ｅ１Ｊ，Ｅ２Ｊ］＝（Ｅ１，Ｅ２）を乱雑とする；（１８１０）。
【０３４０】
工程５：測定（１８１１，１８１２，１８１３）に誤差を導入する
所与の測定誤差（ＨＰ－ハプロイド細胞，ＰＤ－無秩序ジプロイド細胞，ｐｅ－胚細胞）に基づいて、測定に誤差を導入する。
ｈｐ＝誤差（ＨＰ，ｐｈ）を作りだす；（１８１１）
ｄ１ｐ＝誤差（Ｄ１Ｐ，ｐｄ）を作りだす；（１８１１）
ｄ２ｐ＝誤差（Ｄ２Ｐ，ｐｄ）を作りだす；（１８１１）
ｈｍ＝誤差（ＨＭ，ｐｈ）を作りだす；（１８１２）
ｄ１ｍ＝誤差（Ｄ１Ｍ，ｐｄ）を作りだす；（１８１２）
ｄ２ｍ＝誤差（Ｄ２Ｍ，ｐｄ）を作りだす；（１８１２）
ｅ１＝誤差（Ｅ１Ｊ，ｐｅ１）を作りだす；（１８１３）
ｅ２＝誤差（Ｅ２Ｊ，ｐｅ２）を作りだす；（１８１３）。
【０３４１】
工程６：ＤＨＡｌｇｏを適用して、（ｐ１，ｐ２）、（ｍ１，ｍ２）、（ｐｐ１，ｐｐ２）、（ｐｍ１，ｐｍ２）を得る。
【０３４２】
ＤＨＡｌｇｏはハプロイド細胞および無秩序ジプロイド細胞を取り、これらを生起させた最もありそうな秩序立ったジプロイド細胞を戻す。ＤＨＡｌｇｏは父親染色体について（Ｐ１Ｃ１，Ｐ２Ｃ１）、および母親染色体について（Ｍ１Ｃ１，Ｍ２Ｃ１）を再形成するよう試み、また父親（ｐｐ１，ｐｐ２）および母親（ｐｍ１，ｐｍ２）細胞についての見積もり誤差を戻す。
［ｐ１，ｐ２，ｐｐ１，ｐｐ２］＝ＤＨＡｌｇｏ（ｈｐ，ｄ１ｐ，ｄ２ｐ，ｓｎｉｐｓ，ｐｏｐｆｒｅｑｌｉｓｔＰＰ）；（１８１４）
［ｍ１，ｍ２，ｐｍ１，ｐｍ２］＝ＤＨＡｌｇｏ（ｈｍ，ｄ１ｍ，ｄ２ｍ，ｓｎｉｐｓ，ｐｏｐｆｒｅｑｌｉｓｔＭＭ）；（１８１５）。
【０３４３】

工程７：ＰＳＡｌｇｏを適用して、（ＤＥ１，ＤＥ２）（１８１６）を得る。

【０３４４】

ＰＳＡｌｏｇは再形成された親細胞（ｐ１，ｐ２，ｍ１，ｍ２）および無秩序な測定胚細胞（ｅ１，ｅ２）を取って、最もありそうな秩序立った真の胚細胞（ＤＥ１，ＤＥ２）を戻す。ＰＳＡｌｇｏは、（Ｅ１，Ｅ２）を再形成するよう試みる。

［ＤＥ１，ＤＥ２，ａｌｌｄａｔａ］＝ＰＳＡｌｇｏ（ｓｎｉｐｓ，ｅ１，ｅ２，ｐ１，ｐ２，ｍ１，ｍ２，ｐｅ，ｐｐ１，ｐｐ２，ｐｍ１，ｐｍ２，ｐａｒａｍｅｔｅｒｓ，ｃｒｏｓｓｐｒｏｂ，ｐｏｐｆｒｅｑｌｉｓｔＰＰ，ｐｏｐｆｒｅｑｌｉｓｔＭＭ）；

。

【０３４５】

工程８：このシミュレーション実行からの望まれる統計学の収集

実行についての統計学を得る：

ｓｉｍｄａｔａ＝ＳｉｍＶａｌｉｄａｔｅ（ａｌｌｄａｔａ，ＤＥ１，ＤＥ２，Ｐ１，Ｐ２，Ｍ１，Ｍ２，Ｅ１，Ｅ２，ｐ１，ｐ２，ｍ１，ｍ２，ｅ１，ｅ２，ｐｅ，ｐｅ，ｐｐ１，ｐｐ２，ｐｍ１，ｐｍ２）；

。

【０３４６】

シミュレーションの結果

１００００のシミュレーションを実行し、Ｅ１，Ｅ２からのＰＳアルゴリズムの全精度を我々に告げる、アルゴリズム精度についての最終見積もりＰＳＡｃｃｕｒａｃｙ．Ｅ１＝平均（$E1A_i$）、ＰＳＡｃｃｕｒａｃｙ．Ｅ２＝平均（$E2A_i$）を計算した。個々のＳＮＰに基づき、各ＳＮＰ　ＳＮＰＡｃｃ．Ｅ１＝平均（$E1S_i$）についての平均精度は、ＳＮＰ，ＳＮＰｒｏｂ．Ｅ１＝平均（$E2P_i$）であると正しく測定する見積もられた確率の平均に合致するはずであり、すなわち、もしアルゴリズムが正しく書かれれば、ＳＮＰＡｃｃ．Ｅ１は、ＳＮＰｒｏｂ．Ｅ１に相関するように観察されるはずである。これらの２つの間の関係はそれらの相関によって反映される。

【０３４７】

シミュレーションの１００００ループを異なる設定シナリオについて実行した：

（１）より現実的である現存のゲノタイピングデータ、およびＡ、Ｃ、Ｔ、Ｇが各ＳＮＰにおいて同一の確率を有する均一な集団頻度によって与えられる基礎となる集団頻度

（２）ハプロイド、無秩序ジプロイドおよび胚測定（ｐｈ，ｐｄ，ｐｅ）についての測定精度のいくつかの組合せ。種々の精度をシミュレートした：非常に精度がある（０．９５，０．９５，０．９５）、精度が低い（０．７５，０．７５，０．７５）、および精度なしまたはランダム（０．２５，０．２５，０．２５）、ならびに（０．９，０．５，０．５）、（０．５，０．９，０．９）のアンバランスな組合せ。現実に最も近いであろうものは、ほぼ（０．６，０．８，０．８）である。

（３）すべてのこれらの場合において、我々のＰＳＡｌｇｏｒｉｔｈｍおよび単純なマッチングＳＴＰＳＡｌｇｏｒｉｔｈｍ双方についてシミュレーションを行って、開示されたアルゴリズムの性能を評価した。

これらの実行の結果を表８にまとめる。

【０３４８】

開示されたアルゴリズムは、これらのシミュレーションにおいて、特に、不均一な集団頻度および正しい測定のアンバランスな、または低下した確率の現実的場合について、現存の経験的アルゴリズムよりも良好に実行される。また、ここのＳＮＰについてのアルゴリズム精度の見積もりはこれらの場合において非常に良好であることが示された。というのは、正しい対立遺伝子要求の見積もられた精度、およびシミュレーション平均精度の間の相関は９９％程度であり、平均率は１だからである。

【０３４９】

最も現実的な場合において、データ集団頻度および（ｐｈ，ｐｄ，ｐｅ）＝（０．６，０．８，０．８）については、（Ｅ１，Ｅ２）について正しく検索されたＳＮＰの平均パ

ーセントは実施１において（０．７７７，０．７８８）および実施２において（０．８３５，０．８２８）である。前記したように、アルゴリズムの平均精度を示す数は正しいＳＮＰの要求のみならず、ＳＮＰの正しい親起源の同定もいう。効果的であるためには、アルゴリズムは、それが測定されるにつれデータを単純に許容するアルゴリズムよりも良好な結果を戻さなければならない。ある場合には、アルゴリズムの精度には測定のリストされた精度よりも低いのを見て驚くであろう。このシミュレーションの目的では、もしそれが共に正しく要求され、また、その親および元の染色体が正しく同定された場合のみ、ＳＮＰの要求は正確であると考えられる。偶然にこれを正しくするチャンスは測定精度よりもかなり低い。

【０３５０】

出生前および胚遺伝物質を得るのに必要な実験室的技術

ゲノタイピングのための細胞およびＤＮＡ断片の単離を可能とする多くの利用できる技術がある。本明細書中に記載されたシステムおよび方法をこれらの技術、特に、母性血液からの胎児細胞またはＤＮＡの単離、またはＩＶＦの関係で胚からの胚盤胞の単離を含むものの中でいずれにも適応することができる。それはイン・シリコにてゲノムデータに同等に適応することができ、すなわち、遺伝物質から直接的に測定できない。

【０３５１】

システムの１つの実施形態において、このデータは以下に記載するように獲得することができる。

【０３５２】

細胞の単離

成人ジプロイド細胞はバルク組織または血液試料から得ることができる。成人ジプロイド単一細胞は、ＦＡＣＳ、または蛍光活性化細胞ソーティングを用い、全血液試料から得ることができる。成人はプロイド単一精子細胞もまた、ＦＡＣＳを用いて精子試料から単離することができる。成人ハプロイド単一卵細胞は、ＩＶＦ手法の間に卵収穫に関して単離することができる。

【０３５３】

ヒト胚からの標的単一胚盤胞の単離は、体外受精クリニックにおいて普通の技術に従って行うことができる。母性血液中の標的胎児細胞の単離は、モノクローナル抗体、あるいはＦＡＣＳまたは密度勾配遠心のような他の技術を用いて達成することができる。

【０３５４】

ＤＮＡ抽出は本出願についての標準的でない方法も含むであろう。ＤＮＡ抽出についての種々の方法を比較する文献の報告は、いくつかの場合において、Ｎ－ラウロイルサルコシンの添加の使用のような新規なプロトコルは、より効果的であることが判明し、最も少ない偽陽性を生じることを見出している。

【０３５５】

ゲノムＤＮＡの増幅

ゲノムの増幅は、連結－媒介ＰＣＲ（ＬＭ－ＰＣＲ）、縮重オリゴヌクレオチドプライマーＰＣＲ（ＤＯＰ－ＰＣＲ）、および多重置換増幅（ＭＤＡ）を含めた多数の方法によって達成することができる。これらの方法のうち、ＤＯＰ－ＰＣＲは、染色体の単一コピーを含めた、少量のＤＮＡから多量のＤＮＡを信頼性よく生じさせ；この方法は、データ忠実度が臨界的である親ジプロイドデータをゲノタイピングするために最も適しているであろう。ＭＤＡは最速な方法であり、数時間以内にＤＮＡの１００倍増幅を生じる；この方法は、胚細胞をゲノタイピングするのに、あるいは時間が必須である他の状況において最も適切であろう。

【０３５６】

バックグラウンド増幅はこれらの方法の各々で問題である。というのは、各方法は、潜在的に、汚染ＤＮＡを増幅するだろうからである。非常に少量の汚染はアッセイを不可逆的に毒し、偽データを与えかねない。従って、増幅前および後ワークフローが完全に物理的に分離されたクリーンな実験室条件を用いるのが非常に重要である。ＤＮＡ増幅のため

のクリーンな汚染なしのワークフローは、今日、産業的分子生物学においてルーチン的であって、単に詳細に対して思慮深い注意を必要とする。

【０３５７】

ゲノタイピングアッセイおよびハイブリダイゼーション

増幅されたＤＮＡのゲノタイピングは、Ａｆｆｙｍｅｔｒｉｘ’ｓ Ｇｅｎｆｌｅｘ Ｔａｇ Ａｒｒａｙのような分子逆転プローブ（ＭＩＰ）、Ａｆｆｙｍｅｔｒｉｘ’ｓ ５００Ｋアレイまたは Ｉｌｌｕｍｉｎａ Ｂｅａｄ Ａｒｒａｙｓのようなマイクロアレイ、またはＡｐｐｌｉｅｄＢｉｏｓｃｉｅｎｃｅ‘ｓ ＴａｑｍａｎアッセイのようなＳＮＰゲノタイピングアッセイを含めた多くの方法によって行うことができる。Ａｆｆｙｍｅｔｒｉｘ’ｓ ５００Ｋアレイ、ＭＩＰｓ／ＧｅｎＦｌｅｘ、ＴａｑＭａｎおよびＩｌｌｕｍｉｎａアッセイは、全て、マイクログラム量のＤＮＡを必要とし、従って、いずれかのワークフローでの単一細胞のゲノタイピングはいくつかの種類の増幅を必要とするであろう。これらの技術の各々は、とりわけ、コスト、データの質、定量的ｖｓ定性的データ、慣用化性、アッセイを完了するための時間、および測定可能なＳＮＰの数の点で種々の釣り合いを有する。５００ＫおよびＩｌｌｕｍｉｎａアッセイの利点は、１０，０００のＳＮＰのオーダーで検出できるＭＩＰ、およびより少数さえを検出できるＴａｑｍａｎアッセイとは反対に、それがデータを集めることができるＳＮＰの大きな数、およそ２５０，０００である。５００Ｋアレイよりも優れたＭＩＰ、ＴａｑｍａｎおよびＩｌｌｕｍｉｎａアッセイの利点は、それらが固有に慣用化可能であり、ユーザーがＳＮＰを選択するのを可能とすることであり、他方、５００Ｋアレイはそのような慣用化を可能としない。

【０３５８】

ＩＶＦの間における着床前の診断の関係では、固有の時間の制限は重要であり；この場合、応答時間に換えてデータの質を犠牲にするのは有利であろう。それは他の明瞭な利点を有するが、標準ＭＩＰアッセイプロトコルは、典型的には、完了するのに２．５ないし３日かかる比較的時間を消費するプロセスである。ＭＩＰにおいて、ＤＮＡを標的とするためのプローブのアニーリング、および増幅後ハイブリダイゼーションは特に時間を消費し、これらの時間からのいずれの偏差もデータ質の劣化をもたらす。プローブはＤＮＡ試料に一晩アニーリングさせる（１２ないし１６時間）。増幅後ハイブリダイゼーションはアレイに一晩アニーリングさせる（１２ないし１６時間）。アニーリングおよび増幅双方の前および後の多数の他の工程は、プロトコルの合計標準タイムラインを２．５日とする。スピードについてのＭＩＰアッセイの最適化は、潜在的に、プロセスを３６時間未満に低下させることができよう。５００ＫアレイおよびＩｌｌｕｍｉｎａアッセイは共により速い応答時間：ほぼ１．５ないし２日を有して、標準的プロトコルにおいて高度に信頼性があるデータを生じる。これらの方法の双方は最適化可能であり、５００Ｋアレイについてのゲノタイピングアッセイおよび／またはＩｌｌｕｍｉｎａアッセイのための応答時間は２４時間未満まで低下させることができようと見積もられる。なおより速いのはＴａｑｍａｎアッセイであり、これは３時間以内に実行することができる。これらの方法の全てについて、アッセイ時間の低下の結果、データの質の低下をもたらすが、それは、正確には、開示された発明が何を取り組むように設計されているかである。より速いいくつかの利用可能な技術は、特に高－スループットではなく、従って、この時点において高度に平衡な出生前遺伝子診断で使用できない。

【０３５９】

当然に、ＩＶＦの間における胚盤胞のゲノタイピングのような、時間が臨界的である状況においては、より速いアッセイはより遅いアッセイよりも明瞭な利点を有し、他方、ＩＶＦの前に出生前ＤＮＡをゲノタイピングすることが開始されている場合のような、そのような時間圧力を有しない場合には、他の因子が適当な方法を選択するのに支配的であろう。例えば、もう１つの技術に対する１つの技術から出てくるもう１つの釣り合いは、価格ｖｓデータ質のものである。より重要な測定のための高い質のデータを与えるより効果な技術および忠実度が臨界的でない測定用のより低い質のデータを与える安価な技術を用

いるのは理にかなっているであろう。十分に迅速な高－スループットゲノタイピングの点まで開発されたいずれの技術を用いて、この方法で用いる遺伝物質をゲノタイピングすることもできよう。

【０３６０】

該方法の関連数例

どのようにして、開示された方法を、ＩＶＦ手法の時間拘束内に全ての生きた胚の十分なゲノタイピングを可能とするであろう。ＩＶＦ実験室の関係で用いることができるかの例をここに記載する。卵受精から胚着床までの、ＩＶＦ実験室で必要な応答時間は３日下である。これは、関連する実験室的作業、データの清浄化および表現型予測がその時間内に完了させなければならないことを意味する。このシステムの模式的ダイヤグラムを図１９に示し、本明細書中に記載する。このシステムは、ゲノタイピングシステムを用いてＩＶＦ　ｌａｂ１９０４で分析されるＩＶＦユーザー（母親）１９０２およびＩＶＦユーザー（父親）１９０３からの親遺伝子試料１９０１よりなることができる。それは、母親１９０２から収穫され、父親１９０３からの精子で受精させて、多数の受精した胚１９０５を作り出す多数の卵を含むことができる。それは、各胚について胚盤胞を抽出し、各胚盤胞のＤＮＡを増幅し、高スループットゲノタイピングシステム１９０６を用いてそれらを分析する実験室技術者を含むことができる。それは、親からの、および胚盤胞からの遺伝子データをデータ保護プロセッシングシステム１９０７に送ることを含む、該システムは胚遺伝子データを確証し、清浄化する。それは、フェノタイピングアルゴリズム１９０９によって操作されて、各胚の表現型感受性を予測する清浄化胚データ１９０８を含むことができる。それは、ＩＶＦユーザー１９０２および１９０３が母親１９０１における着床について胚を選択するのを助ける医師１９１０に送られる関連信頼性レベルと共にこれらの予測を含むことができる。

【０３６１】

遺伝子データの清浄化に関連する雑多な注意

本明細書中に記載される方法は遺伝子データの清浄化に関することに注意するのは有用であり、すべての生き物は遺伝子データを含有するので、該方法は親から染色体を受け継ぐいずれのヒト、動物または植物にも等しく適用することができる。動物および植物のリストは、限定されるものではないが、ゴリラ、チンパンジー、ピグミーチンパンジー、ネコ、イヌ、パンダ、ウマ、ウシ、ヒツジ、ヤギ、ブタ、チーター、トラ、ライオン、サケ、サメ、クジラ、ラクダ、バイソン、マナティー、ウナギ、メカジキ、イルカ、アルマジロ、カリバチ、ゴキブリ、虫、コンドル、ワシ、スズメ、チョウ、セコイア、トウモロコシ、小麦、米、ペチュニア、カウズベッチ、ヒマワリ、ブタクサ、カシノキ、栗の木およびアタマジラミを含む。

【０３６２】

遺伝子データの測定は、特に、遺伝物質の試料が少量である場合に完全なプロセスではない。測定は、しばしば、正しくない測定、不明瞭な測定、誤った測定、および失われた測定を含む。本明細書中に記載された方法の目的は、これらの誤差のいくつかまたはすべてを検出し、修正することにある。この方法を用い、遺伝子データがかなり知られる信頼性を改良することができる。例えば、現行の技術を用い、単一細胞から増幅されたＤＮＡからの不明瞭な測定遺伝子データは、２０％および５０％の間の未測定領域、または対立遺伝子ドロップアウトを含み得る。いくつかの場合において、遺伝子データは２０％および９９％の間の未測定領域、または対立遺伝子ドロップアウトを含み得るであろう。加えて、所与の測定ＳＮＰの信頼性は同様に誤差に従う。

【０３６３】

未清浄化データがほぼ５０％の対立遺伝子ドロップアウト率を有する場合において、本明細書中に開示された方法を適応した後に、清浄化されたデータは少なくとも９０％の場合において正しい対立遺伝子要求を有し、理想的な状況下では、これは９９％またはそれを超えるまで上昇し得ると予測される。未清浄化データがほぼ８０％の対立遺伝子ドロップアウト率を有する場合において、本明細書中に開示された方法を適用した後に、清浄化

データは少なくとも９５％の場合において、正しい対立遺伝子要求を有し、理想的な状況下ではこれは９９％またはそれを超えるまで上昇し得ると予測される。未清浄化データがほぼ９０％の対立遺伝子ドロップアウト率を有する場合、本明細書中に開示された方法を適用した後に、清浄化データは少なくとも９９％の場合において正しい対立遺伝子要求を有し、理想的な状況下ではこれは９９％以上まで上昇し得ると予測される。特定のＳＮＰ測定が９０％近くの信頼性率でもってなされる場合、清浄化データは９５％を超える、および理想的な場合には、９９％を超える、またはそれを超える信頼性率でもってＳＮＰ要求を有すると予測される。特定のＳＮＰ測定が９９％近くの信頼性率でもってなされる場合において、清浄化データは、９９．９％を超えるおよび理想的な場合には９９．９９％を超える、またはそれよりも高い信頼性率でもってＳＮＰ要求を有すると予測される。

【０３６４】

また、１つの胚盤胞からの増幅されたＤＮＡを測定することによって創製することができる胚遺伝子データは、多数の目的で使用することができるのみ注意するのも重要である。例えば、それは、異数体、片親二染色体を検出し、個体の性別を鑑定し、ならびに複数の表現型予測を行うのに用いることができる。現在、ＩＶＦ実験室においては、用いる技術のため、しばしば、それは、胚盤胞が異数性のような１つの障害、または特定の単一遺伝子病についてテストするのに十分な遺伝物質を供することができるに過ぎない場合である。本明細書中に開示された方法は、なされる予測のタイプに拘わらず、胚盤胞からＳＮＰの大きなセットを測定する通常の最初の工程を有するので、医師または親は、スクリーニングすべき限定された数の障害を選択することを強制されない。その代わり、医療的知識の状態が許容する程度に多くの遺伝子および／または表現型についてスクリーニングするオプションが存在する。開示された方法では、胚盤胞のゲノタイピングに先立ってスクリーニングするための特定の条件を同定する唯一の利点は、もしあるＰＳＮＰが特に関連すると決定されたならば、注目するＰＳＮＰとより共分離するようなＮＳＮＰのより適切なセットを選択することができ、かくして、注目する対立遺伝子の要求の信頼性を増大させることである。ＳＮＰが先立って個人化されない場合においてさえ、信頼性は、本明細書中に記載された種々の目的で適切なものを超えると予測されることを注記する。

【０３６５】

表現型および臨床的予測

遺伝子型および臨床的情報から表現型データを予測するのに利用できる多くの方法がある。異なるモデルは、利用できるデータの量およびタイプに基づいて、異なる状況においてより適切である。表現型予測のための最も適切な方法を選択するためには、テストデータのセットについて多数の方法をテストし、テストデータの測定された結果と比較する場合に、予測の最良の精度をいずれの方法が提供するかを決定するのがしばしば最良である。本明細書中に記載されたある実施形態は、組合せて採用され、かつテストデータでの性能に基づいて選択された場合に、正確な表現型予測を行う高い尤度を供する方法のセットを含む。まず、（ｉｉ）偶発事象表を用いるシナリオでの遺伝子型－表現型モデリングのための技術を記載する。次に、（ｉｉｉ）凸最適化によって形成された回帰モデルを用いるシナリオにおける遺伝子型－表現型モデリングのための技術を記載する。次いで、予測すべき特定の表現型、特定の患者のデータ、およびモデルを訓練し、テストするためのデータの特定のセットを仮定して最良のモデルを選択するための技術を記載する。

【０３６６】

今日のデータ：偶発事象表に基づく表現型結果のモデリング

公知の遺伝子的欠陥、および病気表現型の確率を増加させる対立遺伝子がある場合、およびプレディクターの数が十分に少数である場合、表現型確率は偶発事象表でモデル化することができる。もしただ１つの関連遺伝子対立遺伝子があれば、特定の対立遺伝子の存在／不存在はＡ＋／Ａ－として記載することができ、病気表現型の存在／不存在はＤ＋／Ｄ－として記載することができる。（f_1，N_1，f_2，N_2）を含有する偶発事象表は：

【０３６７】

【数８６】

	D+	D-	#
G+	f_1	1-f_1	N_1
G-	f_2	1-f_2	N_2

$$S^2 = \frac{N_1 N_2 (N_1 + N_2)(p_1(1-p_2) - (1-p_1)p_2)^2}{(p_1 N_1 + p_2 N_2)((1-p_1)N_1 + (1-p_2)N_2)}$$

である。ここで、f_1およびf_2は測定された頻度または異なる結果の確率を表し、対象の合計数は$N = N_1 + N_2$である。この表から、独立変数（ＩＶ）Ｇ＋またはＧ－を有する２つの場合において病気状態Ｄ＋を有する確率についてのオッズ比は、９５％信頼区間を持つ$OR = f_1(1-f_2)/f_2(1-f_1)$：Ｓが標準偏差である$OR^{1 \pm 1.96/S}$として報告することができる。例えば、１０，０００の個体における乳癌の実験を用い、ここで、Ｎ＋はＢＲＣＡ１またはＢＲＣＡ２対立遺伝子の存在を表す。

【０３６８】

【数８７】

	D+	D-	#
M+	.563	.437	1720
M-	.468	.532	8280

このデータの結果、信頼区間［１．３１；１．６２］でのオッズ比率ＯＲ＝１．４６３がもたらされ、これを用いて、所与の突然変異を持つ乳癌の出現の増大した確立を予測することができる。２×２よりも大きな偶発事象表を用いて、より独立した変数または結果変数を収容することができることを注記する。例えば、乳癌の場合には、偶発事象Ｍ＋およびＭ－は４つの偶発事象：ＢＲＣＡ１およびＢＲＣＡ２、ＢＲＡＣＡ１およびＢＲＣＡ２ではない、およびＢＲＣＡ１ではなくおよびＢＲＣＡ２、および最後にＢＲＣＡ１でなくＢＲＣＡ２でない；で置き換えることができよう。どのようにして２×２を超える偶発事象表についての信頼区間を決定するのは当業者によってよく理解されるこの技術は、独立変数の異なる偶発事象によって定義される異なる群における患者をカウントすることによって低い標準偏差を持つモデルを形成するのに十分に少数のＩＶおよび十分なデータがある場合に用いられる。このアプローチは、回帰モデルを構築する場合に必要なように異なるＩＶをモデル化すべき結果に関連させる数学モデルを設計する困難性を回避する。

【０３６９】

特定のＳＮＰからの遺伝子データは、特に、ＨａｐＭａｐプロジェクトで認識されるＳＮＰの異なるパターンのような独立した変数の他の空間へ投影することもできることを注記する。ＨａｐＭａｐ投影は個体をビンにクラスター化し、各ビンはＳＮＰの特定のパターンによって特徴づけられるであろう。例えば、１つのビン（Ｂ１）はＢＲＣＡ１およびＢＲＣＡ２を含有するＳＮＰパターンを有し、もう１つのビン（Ｂ２）はＢＲＣＡ１を含有するが、ＢＡＣＡ２を含有しないＳＮＰパターンを有し、および第三のビンは、突然変異のすべての他の組合せに関連するＳＮＰパターン（Ｂ３）を含有すると考える。これらのＳＮＰのすべての異なる組合せを表す偶発事象表を作成するよりはむしろ、偶発事象Ｂ１、Ｂ２およびＢ３を表す偶発事象表を作成することができる。

【０３７０】

ＨａｐＭａｐ投影によって記載されるように、あるＳＮＰが一緒におこる傾向を用いて、プレディクターとして多数のＳＮＰを用いるモデルを作成することができ、次いで、データは患者の別々の群よりなり、ここで、各群はただ１つの測定されたＳＮＰを有することにさらに注意されたし。この問題は、ＯＭＩＭから入手可能なもののような公に入手可能な研究論文からモデルを作成する場合に普通に遭遇し、多数のＳＮＰは表現型を予測するものではあるが、各論文は唯一の測定された関連ＳＮＰを有するコホートについてのデ

ータを含有する。今日利用可能なデータを用いて予測モデルを形成するのに有用なこの態様を説明するために、ＩＶ：アルツハイマー病の家族履歴、性別、人種、年齢、３つの遺伝子、すなわち、ＡＰＯＥ、ＮＯＳ３、およびＡＣＥの種々の対立遺伝子に基づいて予測モデルを形成することができるアルツハイマー病に特に言及する。この病気との関係では、アルツハイマー病を超える多くの病気に適用される普及した論点を議論し：多くの遺伝子は特定の表現型についての特性の決定に関与するが、履歴研究のほとんど大部分は特定の遺伝子の対立遺伝子をサンプリングしたのに過ぎなかった。アルツハイマー病の場合においては、ほとんど全ての研究コホートは唯一の遺伝子をサンプリングしたに過ぎなかった；すなわち、ＡＰＯＥ、ＮＯＳ３、またはＡＣＥ。それにも拘わらず、利用可能なデータの大部分が、唯一の遺伝子を調べる研究から由来する場合でさえ、多数の遺伝子対立遺伝子を入力するモデルを形成するのが重要である。この問題は、２つの表現型状態の単純化された場合、および各々が丁度２つの状態を持つ、２つの関連遺伝子を表す唯２つの独立した変数を考慮することによって説明される１つの態様において取り組まれる。病気表現型を記載するランダム変数Ｄ∈［Ｄ＋，Ｄ－］、および遺伝子を記載する２つのランダムな変数Ａ∈［Ａ＋，Ａ－］およびＢ∈［Ｂ＋，Ｂ］を仮定すれば、目標はＰ（Ｄ／Ａ，Ｂ）の最良の可能な見積もりを見出すことである。これは、Ｐ（Ｄ／Ａ，Ｂ）＝Ｐ（Ａ，Ｂ／Ｄ）Ｐ（Ｄ）／Ｐ（Ａ，Ｂ）を用いてベイズ則を適用することによって見出すことができる。Ｐ（Ｄ）およびＰ（Ａ，Ｂ）は公のデータから入手可能である。特に、Ｐ（Ｄ）とは、集団における病気の全罹患率をいい、これは公に入手可能な統計学から見出すことができる。加えて、Ｐ（Ａ，Ｂ）とは、個体において一緒に起こる遺伝子ＡおよびＢの特定の状態の罹患率をいい、これは、異なる人種群における多数の個体での測定された多くの異なるＳＮＰを有するＨａｐＭａｐ　Ｐｒｏｊｅｃｔのような公のデータベースから見出すことができる。好ましい実施形態においては、これらの確率の全ては、全ヒト集団についてよりはむしろ、確率バイアスがある、特定の人種群および特定の性別について計算することができることを注記する。一旦、これらの確率が決定されたならば、挑戦は正確にＰ（Ａ，Ｂ／Ｄ）を見積もることから由来する。というのは、コホートデータの大部分はＰ（Ａ／Ｄ）およびＰ（Ｂ／Ｄ）の見積もりを供するからである。関連情報は、異なる遺伝子対立遺伝子の間の統計学的関連についての、すなわち、Ｐ（Ａ／Ｂ）についての、ＨａｐＭａｐ　Ｐｒｏｊｅｃｔのような種々の公のデータベースで見出すことができる。しかしながら、Ｐ（Ａ／Ｂ）、Ｐ（Ａ／Ｄ）、Ｐ（Ｂ／Ｄ）のみを仮定し、依然として、Ｐ（Ａ，Ｂ／Ｄ）については何も言うことができない。というのは、拘束されない自由度があるからである。それにも拘わらず、もしなんらかの情報が、（Ａ－，Ｂ－）のような丁度単一偶発事象についてさえ、遺伝子ＡおよびＢ双方をそれにつきサンプリングしたコホートからのＰ（Ａ，Ｂ／Ｄ）について知られていれば、Ｐ（Ａ／Ｄ）、Ｐ（Ｂ／Ｄ）、Ｐ（Ａ／Ｂ）についての情報の価値を利用して、Ｐ（Ａ，Ｂ／Ｄ）の見積もりを改良することができる。この概念は、偶発事象の表を用いて説明されるであろう。

【０３７１】

遺伝子状態Ａ＋およびＡ－に従う結果Ｄ＋およびＤ－の確率を表す以下の２つの偶発事象表を考える。この実験はＡと言及される。Ａについての測定された頻度はＦと言及され、見積もりを求める現実の確率はｐを伴って言及される。

【０３７２】

【数８８】

A	D+	D-
A+	f_1	f_2
A-	f_3	f_4

A	D+	D-
A+	p_1	p_2
A-	p_3	p_4

ここで、$f_3 = 1 - f_1$、$f_4 = 1 - f_2$および$p_3 = 1 - p_1$、$p_4 = 1 - p_2$である。K_1が、Ａについての場合の群における対象の数、すなわち、結果Ｄ＋を有する対象の数を表すものとする。K_2が、Ａについての対照群における数、すなわち、結果Ｄ－を有する対象の数であるとする。

【０３７３】

同様に、遺伝子状態Ｂ＋およびＢ－に従う結果Ｄ＋およびＤ－の確率を表す以下の２つの偶発事象の表を考える。この実験はＢと言及される。測定された頻度はｆを伴って言及され、見積もりを求める現実の確率はｐを伴って言及される。

【０３７４】

【数８９】

B	D+	D-
B+	f_5	F_6
B-	f_7	F_8

B	D+	D-
B+	P_5	p_6
B-	P_7	p_8

ここで、$f_7 = 1 - f_5$、$f_8 = 1 - f_6$および$p_7 = 1 - p_5$、$p_8 = 1 - p_6$である。K_3がＢについての場合の群における数を表すものとし、K_4がＢについての対照群における数であるとする。前記偶発事象の表は、遺伝子状態ＡおよびＢが別々に測定される試験を表す。しかしながら、理想的に求められる偶発事象表は、組み合わされたＡおよびＢの異なる状態を含む。偶発事象表は、ＡＢという仮定実験について以下に示し、ここで、ｆは測定された確率を表し、およびｐは現実の確率を表す。

【０３７５】

【数９０】

AB	D+	D-
A+B+	f_9	f_{10}
A+B-	f_{11}	f_{12}
A-B+	f_{13}	f_{14}
A-B-	f_{15}	f_{16}

AB	D+	D-
A+B+	p_9	p_{10}
A+B-	p_{11}	p_{12}
A-B+	p_{13}	p_{14}
A-B-	p_{15}	p_{16}

ここで、$f_{15} = 1 - f_9 - f_{11} - f_{13}$、$f_{16} = 1 - f_{10} - f_{12} - f_{14}$および$p_{15} = 1 - p_9 - p_{11} - p_{13}$、$p_{16} = 1 - p_{10} - p_{12} - p_{14}$である。$K_5$がＡＢについての場合の群における数とし、$K_6$がＡＢについての対照群における数とする。

【０３７６】

表記方法目的では、$K_7 = K_9 = K_5$および$K_8 = K_{10} = K_6$であることを注記する。従って、事実、群のサイズは：

【０３７７】

【数９１】

#	D+	D-
A	K_1	K_2
B	K_3	K_4
AB	K_5	K_6

である。

【０３７８】

統計学の基本則を用いて、仮定偶発事象表ＡＢの細胞の間の依存性を強制することができる。この例においては、Ｄ＋に対応する細胞について、以下の関係を強制することができる：

Ｐ（Ａ＋Ｂ－／Ｄ＋）＝Ｐ（Ａ＋／Ｄ＋）－Ｐ（Ａ＋Ｂ＋／Ｄ＋）、

Ｐ（Ａ－Ｂ＋／Ｄ＋）＝Ｐ（Ｂ＋／Ｄ＋）－Ｐ（Ａ＋Ｂ＋／Ｄ＋）

Ｐ（Ａ－Ｂ－／Ｄ＋）＝１－Ｐ（Ａ＋／Ｄ＋）－Ｐ（Ｂ＋／Ｄ＋）＋Ｐ（Ａ＋Ｂ＋／Ｄ＋）

同様に、Ｄ－に対応する細胞については：

Ｐ（Ａ＋Ｂ－／Ｄ－）＝Ｐ（Ａ＋／Ｄ－）－Ｐ（Ａ＋Ｂ＋／Ｄ－）

Ｐ（Ａ－Ｂ＋／Ｄ－）＝Ｐ（Ｂ＋／Ｄ－）－Ｐ（Ａ＋Ｂ＋／Ｄ－）

Ｐ（Ａ－Ｂ－／Ｄ－）＝１－Ｐ（Ａ＋／Ｄ－）－Ｐ（Ｂ＋／Ｄ－）＋Ｐ（Ａ＋Ｂ＋／Ｄ－）

である。

前記偶発事象表中の表記方法を用い、および余分な最後の関係を残し、これらの関係は：

$p_{11}=p_1-p_9$

$p_{13}=p_5-p_9$

$p_{12}=p_2-p_{10}$

$p_{14}=p_6-p_{10}$

に移され、あるいは同等に、

$p_1=p_9+p_{11}$

$p_2=p_{10}+p_{12}$

$p_5=p_9+p_{13}$

$p_6=p_{10}+p_{14}$

に移される。

全ての関係をまとめると、$p_9,\cdots,p_{16}$に対する$p_1,\cdots,p_{16}$の全ての依存性の表を以下に掲げる。値の間の依存性を得るために、列内の確率は、値＝１を有する行内の確率の合計であり、例えば、第一列は$p_1=p_9+p_{11}$を与える。

【０３７９】

【数９２】

	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}
p_1	1		1					
p_2		1		1				
p_3					1		1	
p_4						1		1
p_5	1				1			
p_6		1				1		
p_7			1				1	
p_8				1				1
p_9	1							
p_{10}		1						
p_{11}			1					
p_{12}				1				
p_{13}					1			
p_{14}						1		
p_{15}							1	
p_{16}								1

頻度および確率の間の関係から、ｎ＝９．．．１６についての測定方程式$f_i = p_i + n_i$を作成でき、ここで、n_iは、出現f_iの頻度に基づいた確率p_iの不完全な測定を表すノイズ項である。前記した関係にこれを適用し、かつ偶発事象ＡＢの細胞の全ては測定されていると仮定し（これは、丁度説明目的のためであり、以下に議論する）、これらの１０の観察を表すことができる。

これらの測定方程式は：

$F = XP + N$

として行列表記方法で表すことができる。ここで、$F = [F_1, \ldots, F_{16}]^T$、$P = [p_9, \ldots, p_{16}]^T$および$N = [n_9, \ldots, n_{16}]^T$であり、Ｘは前記表中に表した行列である。この行列方程式を用いて、８つの未知の係数、$p_9 \ldots p_{16}$、を解くことができる。この特別な場合において、我々は、全てのパラメーター$p_9 \ldots p_{16}$について解く。もし我々が組み合わされたＡ，Ｂ遺伝子について全ての測定を有しないならば、我々は、Ｄ＋についての少なくとも１つの測定、およびＤ－についての１つの測定を必要とする。前記関係を仮定すれば、次いで、我々は表の残りを満たすことができる。言い換えれば、仮定実験ＡＢについての偶発事象表を埋めることができるためには、望ましくは、ＡおよびＢの特定の状態が、Ｄ＋およびＤ－の結果を有する対象について同時に測定される少なくとも１つの例がある。これは、なされた測定を表す行列Ｘについて十分なランクを達成することを可能とし、従って、値$p_9 \ldots p_{16}$を解き、偶発事象表ＡＢに満たす。もしより多くの実験データが存在すれば、さらなる列を、前記で示したのと同様な構造を持つ行列Ｘの底部に加えることができる。

【０３８０】

正確な回帰を行うためには、群試料のサイズによって決定される各観察ｆｉについての重みを持つ重み付け回帰が望ましく、従って、さらに多くの観察を持つ実験および細胞は

より重み付けされる。測定方程式$f_i = p_i + n_i$では、n_iは全て同一の偏差を有さず、回帰は等分散性でない。具体的には、$f_i = 1/K_i * Binomial(p_i, K_i) \sim N(p_i, p_i(1-p_i)/K_i)$であり、ここで、$Binomial(p_i, K_i)$は、各テストがケース結果$p_i$の確率を有し、および$K_i$テストを行う二項分布を表す。この二項分布は$N(p_i, p_i(1-p_i)/K_i)$によって近似することができ、これは平均$p_i$および偏差$p_i(1-p_i)/K_i$を持つ二項分布である。この結果、ノイズは、理論的偏差$V_i = p_i * (1-p_i)/K_i$を有する正規変数$n_i \sim N(0, p_i(1-p_i)/K_i)$としてモデル化することができる。この偏差は、試料頻度$v_i = f_i * (1-f_i)/K_i$で近似することができる。

【０３８１】

偏差v_iに逆比例する各観察ｉについての重みを持つ重み付け回帰を行った。Ｖが直交要素$[v_9, \ldots, v_{16}]$を持つ行列であって、全ての他の要素は０である$\sim N(0, V)$としてのノイズ行列Ｎの分布は、今や、記載することができる。これは$V = diag([v_9, \ldots, v_{16}])$として示される。同様に、$W = diag([1/v_9, \ldots, 1/v_{16}])$とする。さて、重み付け回帰：

$P = (X'WX)^{-1} X'WY$

を用いてＰについて解くことが可能である。

Ｐの偏差は

$Var(P) = (X'WX)^{-1}$

であることは直接的に示され、これを用いて、Ｐの決定における信頼性を示すことができる。

【０３８２】

まとめると、我々は、ＡおよびＢの組合せからのデータ（ＡＢ：$f_9, \ldots, f_{16}$）と共に、個々の遺伝子からのデータ（Ａ：$f_1, \ldots, f_4$，Ｂ：$f_5, \ldots, f_8$）を用いて、ＡおよびＢの組合せについての確率（$p_9, \ldots, p_{16}$）およびそれらの偏差（$v_9, \ldots, v_{16}$）を見積もるのを助けた。最後に、我々の研究においては、我々は、確率ではなくｌｏｇオッズ比をほとんど取り扱い、従って、我々は、これらの確率をＬＯＲに移す必要がある。一般に、事象Ｈについて確率および偏差を以下のように仮定する。

【０３８３】

【数９３】

	D+	D-
H+	p1	p2
H-	1-p1	1-p2
V	v1	v2

ＬＯＲについての式は（デルタ方法によって）偏差を伴って、$LOR = [\log(p1) - \log(1-p1)] - [\log(p2) - \log(1-p2)]$である。$V = [(p1)^{-1} + (1-p1)^{-1}]^{-2} * V(p1) + [(p2)^{-1} + (1-p2)^{-1}]^{2} * V(p2)$。以下の表は、Ａ，Ｂの組合せについての確率、対応するＬＯＲおよび偏差を示す。

【０３８４】

【数９４】

	D+	D-	LOR	Var
A+B+	P_9	p_{10}	lor_1	$V_1 = [1/p_9 + 1/(1-p_9)]^2 v_9 + [1/p_{10} + 1/(1-p_{10})]^2 v_{10}$
A+B-	P_{11}	p_{12}	lor_2	$V_2 = [1/p_{11} + 1/(1-p_1)]^2 v_1 + [1/p_{12} + 1/(1-p_{12})]^2 v_{12}$
A-B+	P_{13}	p_{14}	lor_3	$V_3 = [1/p_{13} + 1/(1-p_{13})]^2 v_3 + [1/p_{14} + 1/(1-p_{14})]^2 v_{14}$
A-B-	P_{15}	p_{16}	lor_4	$V_4 = [1/p_{15} + 1/(1-p_{15})]^2 v_{15} + [1/p_{16} + 1/(1-p_{16})]^2 v_{16}$

これは、ｌｏｇオッズ比および各偏差の見積もりを提供する。

【０３８５】

この方法の説明として、該技術を使用して、Ｐ（Ａ，Ｂ／Ｄ）の改良された見積もりが得られ、ここで、Ｄはアルツハイマー病を有する状態を表し、およびここで、ＡおよびＢは、各々、ＡＰＯＥおよびＡＣＥ遺伝子の２つの異なる状態を表す。表９は、唯一の遺伝子Ａがサンプリングされた１９９９年にＡｌｖａｒｅｚによって；唯一の遺伝子Ｂがサンプリングされた１９９８年にＬａｂｅｒｔによって；および遺伝子ＡおよびＢがサンプリングされた２００５年Ｆａｒｒｅｒによって行われた３つの異なる実験を表す。結果の２つのセットはこれらの実験から作成されたものであり、表１０に示す。最初のセット（表１０、行２、３、４および５参照）は、全てのコホートを分析し、本明細書中に開示された方法を用いてＰ（Ａ／Ｄ）Ｐ（Ｂ／Ｄ）を仮定してＰ（Ａ，Ｂ／Ｄ）の見積もりを改良する。第二のセット（表１０、行６、７、８および９参照）は、Ｐ（Ａ，Ｂ／Ｄ）についてのＦａｒｒｅｒ（２００５）の近代コホートから生じた結果のみを用い、そこでは、双方の遺伝子がサンプリングされた。前者の場合における予測の信頼限界は低下したと考えられる。これらの予測は、公の源からのＰ（Ａ／Ｂ）を記載するデータを用いてさらに改良することができ－これらの測定は前記したようにＸ行列に加えることができることを注記する。また、本明細書中に記載された技術を用いて、前記したｐ１＝ｐ５＋ｐ７のような関係を用い、Ｐ（Ａ＋／Ｄ＋）、Ｐ（Ａ＋／Ｄ－）、Ｐ（Ｂ＋／Ｄ＋）およびＰ（Ｂ－／Ｄ－）のような別々のＡ，Ｂ確率についての見積もりを改良することができることも注記する。

【０３８６】

この方法はただ２つの変数ＡおよびＢについて説明してきたが、偶発事象の表は、アルツハイマー予測の関係で前記したもの：アルツハイマー病の家族履歴、性別、人種、年齢、および３つの遺伝子、すなわち、ＡＰＯＥ、ＮＯＳ３、およびＡＣＥの種々の対立遺伝子のような多くの異なるＩＶを含むことができることに注意すべきである。年齢のような連続的変数は、値のビンにカテゴリー化することによってカテゴリーを作成して、偶発事象表の処方に適当とすることができる。好ましい実施形態において、最大数を用いて、結果の確率をモデル化し、確率の標準偏差は、典型的には、いくつかの特定の閾値未満である。還元すれば、可能な最も特別な偶発事象は、その偶発事象についての十分な関連訓練データを維持して、関連する確率の見積もりを意味のあるものとしつつ、特定の患者に利用可能なＩＶを仮定して創製することができる。

【０３８７】

また、本開示を読んだ後に、病気－遺伝子関連、遺伝子－遺伝子関連、および／または集団における遺伝子頻度についてのデータを用いるために同様な技術をどのようにして適用して、多変数線形および非線形回帰および論理回帰モデルの精度を改良することができることは当業者に明らかであろうことを注記する。さらに、本開示を読んだ後に、病気－遺伝子関連、遺伝子－遺伝子関連、および／または集団における遺伝子頻度についてのデ

ータを用いるための同様な技術を適用して、どのように適用して、結果データの利用を可能として、モデルに関連するその全ての独立変数がその結果データにつき測定されるものではないモデルを訓練することによって、多変数線形および非線形回帰および論理回帰モデルの精度を改良することができることは当業者に明らかであろう。さらに、本開示を読んだ後には、病気－遺伝子関連、遺伝子－遺伝子関連、および／または集団における遺伝子頻度についてのデータを用いるための同様な技術をどのようにして適用して、当該分野で良く理解される期待値最大化（ＥＭ）アルゴリズムのような他の技術を用いて形成された偶発事象表モデルの精度を改良することができるかは当業者に明らかであろう。これらの技術は、ＨａｐＭａｐ　Ｐｒｏｊｅｃｔからの活用データ、およびＮａｔｉｏｎａｌ　Ｃｅｎｔｅｒ　ｆｏｒ　Ｂｉｏｔｅｃｈｎｏｌｏｇｙ　Ｉｎｆｏｒｍａｔｉｏｎ（ＮＣＢＩ）Ｏｎｌｉｎｅ　Ｍｅｎｄｅｌｉａｎ　Ｉｎｈｅｒｉｔａｎｃｅ　ｉｎ　Ｍａｎ（ＯＭＩＭ）およびｄｂＳＮＰデータベースのような公のデータベースに含まれる他のデータに特に関連する。

【０３８８】

また、当該特許を通じて、我々が個体または対象に関連するデータに言及する場合、これは、該対象に感染したかもしれないいずれの病原体または該対象に感染しつつあるいずれの癌の該データは言及できるとも仮定する。該個体または対象データは、ヒト胚、ヒト胚盤胞、ヒト胎児、いくつかの他の細胞または細胞のセットについてのデータ、あるいはいずれかの種類の動物または植物にも言及することができる。

【０３８９】

明日のデータ：回帰モデルでの多因子表現型のモデル化

より多くのデータが多因子表現型での遺伝子型に関連して蓄積されるにつれ、支配的なシナリオは前記した（ｉｉｉ）となり、すなわち、表現型を正確に予測するためには遺伝子マーカーの複雑な組合せを考慮するのが望ましく、多次元線形または非線形回帰モデルが導かれる。典型的には、このシナリオについてのモデルを訓練するにおいて、潜在的プレディクターの数は、測定された結果の数と比較して大きいであろう。本明細書中に記載されたシステムおよび方法の例は、未決定の、または悪い条件の遺伝子型－表現型データセットについての疎なパラメーターモデルを創製する新規な技術を含む。該技術は、それについて多くのモデリング業績が比較のために利用でき、およびそれについてデータが多くの潜在的遺伝子プレディクターに関連して入手可能な抗－レトロウイルス療法（ＡＲＴ）に対するＨＩＶ／ＡＩＤＳの応答のモデル化に焦点を当てることによって説明される。現実の実験室測定で交差－確証によってテストする場合、これらのモデルは、文献中で以前に議論されたモデル、および本明細書中に記載された他のカノニカル技術よりも正確に薬物応答表現型を予測する。

【０３９０】

２つの回帰技術を、遺伝子配列データからの抗－レトロウイルス療法に対する応答においてウイルス表現型を予測する関係で記載し、説明する。双方の技術は、モデルパラメーターの粗なセットの連続的サブセット選択のために凸最適化を使用する。最初の技術は、最小絶対収縮および選択オペレーター（ＬＡＳＳＯ）を用い、これはl^1ノルム喪失関数を適用して、疎な線形モデルを作り出し；第二の技術は径ベースの核関数と共にサポートベクトルマシーン（ＳＶＭ）を用い、これは、ε－非感受性喪失関数を適用して、疎な非線形モデルを創製する。該技術は、１０の逆転写酵素阻害剤（ＲＴＩ）および７つのプロテアーゼ阻害剤薬物（ＰＩ）に対するＨＩＶ－１ウイルスの応答の予測に適用される。遺伝子データは、逆転写酵素およびプロテアーゼ酵素についてのＨＩＶコーディング配列に由来する。この性能を可能とするこれらのモデルの鍵となる特徴は、喪失関数が、パラメーターの多くがゼロである単純モデルを創製する傾向があり、およびコスト関数の凸性が、モデルパラメーターを見出して、特定の訓練データセットについてのコスト関数を全体的に最小化することができることを確実とすることである。

【０３９１】

ＬＡＳＳＯおよびL^1選択関数

プレディクターＭの数が訓練試料の数Ｎを超える場合、モデル化の問題は過剰決定系、または不適切である。というのは、Ｎのプレディクターのいずれかの任意のサブセット、Ｘ行列における関連行が直線的に独立している限り、訓練データについてのゼロ誤差を持つ線形モデルを生じるのに十分だからである。その結果、線形回帰方法によって戻されたＮ－プレディクターモデルに信頼を置く気がしない。しかしながら、Ｎよりもかなり少数の変数が低い訓練誤差を有するモデルを仮定する。モデルがより疎であれば、低い訓練誤差は偶然人工物である確率は低く；よって、プレディクターが独立した変数に因果的に関連するのがよりありそうである。これは、ＲＴＩデータの場合のように、過剰決定系の問題における疎な解の重要性の基礎となる。同様な議論を、ＰＩデータに当てはまるように、行列$X^{T}X$での大きな条件数によって特徴付けられる悪条件の問題に適用することができる。この場合、見積もられたパラメーター

【０３９２】

【数９５】

$$\hat{b}$$

はモデル誤差に対して、ならびに測定ノイズに対して高度に感受性であり、結果として、正確に一般化されないようである。過剰決定系および悪条件の問題は、可能なプレディクター－遺伝子、蛋白質、または我々の場合には、突然変異部位の数が、測定された結果の数に対して大きな遺伝子データに典型的である。

【０３９３】

そのような場合に対する１つのカノニカルアプローチはサブセットの選択である。例えば、段階的選択にて、各工程において、その変数が予測誤差に相関する優位性のレベルを示す最高Ｆ－検定統計学を有することに基づいて、単一プレディクターをモデルに加える。各変数を加えた後、残りの変数を全てチェックして、モデルのプレディクター誤差とのそれらの関連性において統計学的有意性の閾値未満までそれらのいずれも降下しないことを確実とする。この技術は、薬物応答予測の問題に成功して適用されてきた。しかしながら、選択プロセスの区別される性質のため、データの小さな変化はプレディクターの選択されたセットをかなり改変することができる。１つの変数の存在または不存在は、もう１つの変数と関連する統計学的有意性、およびその変数がモデルに含まれ、またはそこから拒絶されるかに影響し得る。これは、特に、悪条件の問題について一般化での精度に影響する。

【０３９４】

もう１つのアプローチは、収縮関数によって拘束されるべき見積もられたパラメーター

【０３９５】

【数９６】

$$\hat{b}$$

の値についてである。カノリカル収縮関数は該パラメーターの平方の合計であり、これは：

【０３９６】

【数９７】

$$\hat{b} = \arg\min_b \|y - Xb\|^2 + \lambda \|b\|^2 \qquad (17)$$

［式中、λは、典型的には、交差－確証によって決定されるチューニングパラメーターである］

に従ってパラメーターを見出すｒｉｂｇｅ回帰において適用される。この方法は非疎であって、パラメーターを０に設定しない。これは、一般化における精度を低める傾向があり、解を解釈するのを困難とする。

【０３９７】

これらの問題はＬＡＳＳＯ技術によって取り組まれる。サブセット選択とは対照的に、

ＬＡＳＳＯはプレディクター変数の離散的許容または拒絶を行わず；むしろ、それは連続的サブセット最適化を介して、一緒になって最も効果的なプレディクターとなる変数のセットを一斉に選択することを可能とする。それは１^{1}ノルム収縮関数：

【０３９８】

【数９８】

$$\hat{b} = \arg\min_b \left\| y - Xb \right\|^2 + \lambda \sum_{i=1...M} |b_i| \qquad (18)$$

［式中、λは典型的には交差－確証によって設定される］

を用いる。ＬＡＳＳＯはパラメーターの多くを０に設定する傾向がある。図２０は、選択性と言及されるＬＡＳＳＯのこの特徴に対する洞察を供する。丁度２つの突然変異に基づくモデルは訓練データＸ＝［１　０；０１］^{T}，ｙ＝［２　１］^{T}で創製され、ｘ－軸およびｙ－軸は、各々、２つのパラメーターｂ$_{1}$およびｂ$_{2}$を表す。１^{1}および１^{2}収縮関数の使用を比較し、ここで、双方の場合において、｜｜ｙ－Ｘｂ｜｜^{2}＝２となるように訓練データを同等によくフィットさせる解が見出される。大きな円（２００１）、小さな円（２００２）、および（２００３）は、各々、コスト関数｜｜ｙ－Ｘｂ｜｜^{2}、１^{2}ノルム｜｜ｂ｜｜^{2}、および１^{1}ノルム｜ｂ$_{1}$｜＋｜ｂ$_{2}$｜についてのレベル曲線を表す。ｒｉｄｇｅ回帰に対する解（１^{2}）が見出され、ここで、２つの円が交差し（２００４）；ＬＡＳＳＯについての解（１^{1}）が見出され、ここで、四角および大きな円が交わる（２００５）。１^{1}ノルムについてのレベル曲線の「尖性」のため、軸ｂ$_{1}$上にある解が見出され、これは、従って、疎である。より高次元へ拡大されたこの議論は、疎な解を生じるＬＡＳＳＯの傾向を説明し、なぜ達成された結果が文献に報告されたものよりも測定可能に良好であるかを示唆する。

【０３９９】

１^{1}ノルムは、凸でありつつ、最も選択的収縮関数として見ることができる。凸性は、所与のデータセットに対して１つの全体的解を見出すことができることを保証する。最小角回帰と言及されるかなり有効な最近のアルゴリズムは、Ｍ工程においてＬＡＳＳＯの全体的解に収束することが保証されている。

【０４００】

本開示を読んだ後には、１^{1}ノルムをどのようにして論理回帰の関係で用いて、カテゴリー変数の各状態の確率をモデル化することもできることは当業者に明らかであろうことを注記する。論理回帰において、測定のセットの事後確率の逆数に対応する凸コスト関数を形成することができる。事後確率は、各結果の尤度のモデル見積もりを仮定する観測された訓練データの確率である。１^{1}ノルムを凸コスト関数に加えることによって、得られた凸コスト関数を最小化して、特定の結果の確率をモデル化するための疎パラメーターモデルを見出すことができる。論理回帰についての１^{1}ノルムの使用は、測定された結果の数がプレディクターの数に対して小さい場合に、特に関連し得る。

【０４０１】

サポートベクトルマシーンおよびＬ１－ノルム

ＳＶＭは、特に、モデルが独立変数の間の複雑な相互作用を含む場合に、薬物応答および他の表現型の良好なモデル化を達成するように構成することができる。ＳＶＭについての訓練アルゴリズムは、１^{1}ノルム選択関数の使用を黙示的とする。ＳＶＭは、現実の価値の関数近似を行うことができ、かつ見積もり問題がＨａｄａｍａｒｄの意味において不適切である場合でさえ、試料データの正確な一般化を達成することができる学習アルゴリズムである。正確に一般化されるＳＶＭの能力は、ＳＶＭモデルおよび訓練アルゴリズムにおける２つの選択可能な特徴によって典型的には影響される。第一ものはコスト関数、または訓練において最小化されるべき関数の選択である。第二のものは、ＳＶＭの核、または線形回帰パラメーターの比較的小さなセットを用いて、ＳＶＭが、独立変数の間の相互作用を含む複雑な非線形関数をマッピングするのを可能とする関数の選択である。これらの特徴は以下に議論する。

【０４０２】
線形関数近似：
【０４０３】
【数９９】

$$\hat{y}_i = f(x_i, b) = b^T x_i$$

を持つ対象ｉ　ｙ$_i$についての表現型をモデル化することを考える。まず、いくらかε＞０未満の誤差にペナルティを与えない「ε－非感受性喪失」関数と共に、パラメーターでの１^{2}収縮関数よりなるコスト関数を最小化することによってｂを見積もる。ＳＶ回帰を拘束：
【０４０４】
【数１００】

$$y_i - b^T x_i \le \varepsilon + \xi_i^+, i = 1...N \quad (20)$$

$$b^T x_i - y_i \le \varepsilon + \xi_i^-, i = 1...N \quad (21)$$

$$\xi_i^+ \ge 0, \xi_i^- \ge 0, i = 1...N \quad (22)$$

を条件として、以下の最適化：
【０４０５】
【数１０１】

$$\hat{b} = \arg\min_{b,\xi^-,\xi^+} \frac{\|b\|^2}{2} + C\sum_{i=1}^{N}\left(\xi_i^- + \xi_i^+\right) \quad (19)$$

として公式化することができる。
【０４０６】
コスト関数の第二項は、「非感受性」閾値εを超えてモデル化誤差の絶対値を最小化する。パラメーターＣは、誤差ｖｓ重みに対する収縮の相対的重要性を見積もることを可能とする。この拘束された最適化を、ラグランジュの鞍点を見出す標準的技術を用いて解いて、Ｋｕｈｎ－Ｔｕｃｋｅｒ拘束を満足させることができる。前記したコストおよび拘束を適合させるラグランジュは：
【０４０７】
【数１０２】

$$L(b,\xi^+,\xi^-,\alpha^+,\alpha^-,\lambda^+,\lambda^-) = \frac{\|b\|^2}{2} + C\sum_{i=1}^{N}\left(\xi_i^- + \xi_i^+\right) - \sum_{i=1}^{N}\alpha_i^-\left(y_i - b^T x + \varepsilon + \xi_i^-\right) - \sum_{i=1}^{N}\alpha_i^-\left(y_i - b^T x + \varepsilon + \xi_i^-\right) - \sum_{i=1}^{N}\left(\lambda^-{}_i\xi_i^- + \lambda^+{}_i\xi_i^+\right) \quad (23)$$

である。パラメーターｂ、ζ$^-$、ζ$^+$のベクトルに関して最小化し、ラグランジュ乗数α$^-$、α$^+$、λ$^-$、λ$^+$のベクトルに関して最小化する。ラグランジュ乗数はＫｕｈｎ－Ｔｕｃｋｅｒ拘束に従って望ましくは正であることを注記する。よって、パラメーターの最適なセットは、
【０４０８】
【数１０３】

$$\alpha_i^+, \alpha_i^-, \lambda_i^+, \lambda_i^- \ge 0, i = 1...N \quad (25)$$

を条件として、

【０４０９】
【数１０４】

$$\left(b_*,\xi_*^+,\xi_*^-\right)=\arg\min_{b,\xi^+,\xi^-}\max_{\alpha^+,\alpha^-,\beta^+,\beta^-}L\left(b,\xi^-,\xi^+,\alpha^+,\alpha^-,\lambda^+,\lambda^-\right) \quad (24)$$

に従って見出すことができる。最小化／最大化の順序は相互交換できるので、これらの変数に関するＬの部分的導関数を０に設定することによって、変数ｂ、ζ$_i^-$、ζ$_i^+$に関してまず最小化する。得られた方程式から、重みベクトルを
【０４１０】
【数１０５】

$$b=\sum_{i=1}^{N}\left(\alpha_i^+-\alpha_i^-\right)x_i \quad (26)$$

の項で表すことができることが判明する。また、得られた方程式から、
【０４１１】
【数１０６】

$$\sum_{i=1}^{N}\alpha_i^+=\sum_{i=1}^{N}\alpha_i^- \quad (28)$$

$$0\le\alpha_i^+\le C,i=1...N \quad (29)$$

$$0\le\alpha_i^-\le C,i=1...N \quad (30)$$

を条件として、二次形式：
【０４１２】
【数１０７】

$$W\left(\alpha^+,\alpha^-\right)=-\sum_{i=1}^{N}\varepsilon\left(\alpha_i^-+\alpha_i^+\right)+\sum_{i=1}^{N}y_i\left(\alpha_i^+-\alpha_i^-\right)-\frac{1}{2}\sum_{i,j=1}^{N}\left(\alpha_i^--\alpha_i^+\right)\left(\alpha_j^--\alpha_j^+\right)x_i^Tx_j \quad (27)$$

を最大化することによって、係数α$_i^+$、α$_i^-$、ｉ＝１．．．Ｎを見出すことができるように、ラグランジュから変数を排除する。
【０４１３】
これは、ベクトルｂが計算されるのを可能とし、ε－非感受性喪失関数に対するＳＶＭモデルを十分に定義する。方程式（１１）から、モデルは、
【０４１４】
【数１０８】

$$f(x)=\sum_{i=1}^{M}\beta_i\left(x^Tx_i\right)+b_0 \quad (31)$$

［式中、β$_i$＝α$_i^+$－α$_i^-$］
として特徴付けることができることを注記する。得られたモデルは、セット｛β$_i$，ｉ＝１．．．Ｍ｝中のパラメーターの多くが０となる点で、疎となる傾向があろう。非ゼロの値β$_i$に対応するベクトルｘ$_i$はモデルのサポートベクトルとして知られている。サポートベクトルの数は、チューナブルパラメーターＣの値、訓練データ、およびモデルの適当性に依存する。以下の説明においては、今や、どのようにしてモデルを増加させて、核関数の使用でもって複雑な非線形関数を適合させることができるかを示す。次に、ε－非感受性喪失関数は１^{1}ノルム収縮関数に関し、それは同じこと、すなわち、１^{1}ノルムによる疎なパラメーターセットの一斉選択を実質的に達成することが示されるであろう。
【０４１５】

変数の間で結合が可能な複雑な関数をモデル化するためには、方程式（１７）の単純な内積を、ベクトルの間のより複雑な相互作用を計算する核関数で置き換える。核関数を挿入し、（１７）中の我々の関数の近似は形態：

【０４１６】

【数１０９】

$$f(x)=\sum_{i=1}^{N}\beta_i K(x,x_i)+\beta_0=\sum_{i=0}^{N}\beta_i K(x,x_i) \tag{32}$$

［式中、定義によるとＫ（ｘ，ｘ$_0$）＝１である］

を採る。これらのパラメーターを見出すためには、前記したのと正確に同一の最適化方法を用い、全ての項ｘ^{T}ｘ$_i$をＫ（ｘ，ｘ$_i$）で置き換える。前記したように、前記したのと同一の拘束に従い、

【０４１７】

【数１１０】

$$W(\alpha^+,\alpha^-)=-\sum_{i=1}^{N}\varepsilon(\alpha_i^-+\alpha_i^+)+\sum_{i=1}^{N}y_i(\alpha_i^+-\alpha_i^-)-\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i^--\alpha_i^+)(\alpha_j^--\alpha_j^+)K(x_i^T x_j) \tag{33}$$

を最大化する独立変数を見出すことによって、β$_i$＝α$_i^+$－α$_i^-$に従ってパラメーターセットを計算する。前記したＳＶＭ結果では、径基礎核関数を選択した。

【０４１８】

さて、１^{1}ノルムの黙示的使用を説明するために：方程式（１７）を最適化することを試みる代わりに、最適化：

【０４１９】

【数１１１】

$$\beta^*=\arg\min_\beta\int_{-\infty}^{\infty}\left(f(x)-\sum_{i=0}^{N}\beta_i K(x_i,x)\right)^2 dx+\varepsilon\sum_{i=0}^{N}|\beta_i| \tag{34}$$

で開始し、ここで、１^{1}収縮を明示的に用いて、βの値を拘束させてあり、訓練データの離散的試料に対して定義される代わりに、データフィッティング誤差を、モデル化すべき仮定関数のドメインに対して定義する。さて、変数置換：β$_i$＝α$_i^+$－α$_i^-$；α$_i^+$，α$_i^-$≧０，α$_i^+$α$_i^-$≧０，ｉ＝１．．．Ｎを行う。次いで、拘束

【０４２０】

【数１１２】

$$\alpha_i^+,\alpha_i^-\geq 0 \tag{36}$$

$$\alpha_i^+\alpha_i^-=0 \tag{37}$$

に従い、

【０４２１】

【数１１３】

$$W(\alpha^+,\alpha^-)=-\sum_{i=1}^{N}\varepsilon(\alpha_i^-+\alpha_i^+)+\sum_{i=1}^{N}y_i(\alpha_i^+-\alpha_i^-)-\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i^--\alpha_i^+)(\alpha_j^--\alpha_j^+)K(x_i,x_j) \tag{35}$$

として、最適化を書き直すことができる。異なる拘束を有するこの解は、それにも拘わらず、もしＳＶ方法についての値Ｃが、拘束０≦α$_i^+$，α$_i^-$≦Ｃが単純に拘束（２１）および（２２）となるのに十分に大きく選択されるならば、ε－非感受性喪失関数に一致し、また、基礎関数の１つは、我々の場合についての方程式（１７）におけるように定数

である。この場合、ＳＶ方法によって用いられるさらなる拘束

【０４２２】

【数１１４】

$$\sum_{i=1}^{N}\alpha_i^{+}=\sum_{i=1}^{N}\alpha_i^{-}$$

を必要としない。拘束（２５）は既に方程式（１５）において黙示的であることを注記する。というのは、拘束（８）および（９）は同時に活性となることはできず、従って、ラグランジェ乗数α_i^{+}またはα_i^{-}のうちの１つはスラックであるか、または０であるべきであるからである。

【０４２３】

これらの条件下で、1^{1}収縮関数のアプローチを黙示的に用いて、ε－非感受性喪失関数が疎な関数近似を達成することを見ることができる。

【０４２４】

多因子表現型予測：ＨＩＶ－１薬物応答のモデリングの例

サルベージＡＲＴの表現型結果の予測に対する現在のアプローチは、主として、薬物養生法および遺伝子突然変異の多くの異なる順列と組み合わせた、主として、統計学的に有意な結果データの欠如のため、良好な予測パワーを示さない。この分野は、多数の不均一データセットの統合、および薬物応答予測の増強の双方についての逼迫した必要性を有する。

【０４２５】

本明細書中で示されたモデルは、訓練およびテスト目的のためのＳｔａｎｆｏｒｄ ＨＩＶｄｂ ＲＴおよびプロテアーゼ薬物耐性データベースからのデータを用いた。このデータは、逆転写酵素（ＲＴ）またはプロテアーゼコーディングセグメントが配列決定されているＨＩＶ－１ウイルスの６６４４イン・ビトロ表現型テストよりなる。テストは、１０の逆転写酵素阻害剤（ＲＴＩ）および７つのプロテアーゼ阻害剤（ＰＩ）について行われた。ＲＴＩはラミブジン（３ＴＣ）、アバカビール（ＡＢＣ）、ジドブジン（ＡＺＴ）、スタビジン（Ｄ４Ｔ）、ザルシタビン（ＤＤＣ）、ジダノシン（ＤＤＩ）、デラビラジン（ＤＬＶ）、エバビレンズ（ＥＦＶ）、ネビラピン（ＮＶＰ）およびテノフォビール（ＴＤＦ）を含む。ＰＩはアムプラナビール（ＡＰＶ）、アナザナビール（ＡＴＶ）、ネルフィナビール（ＮＦＶ）、リトナビール（ＲＴＶ）、サキナビール（ＳＱＶ）、ロピナビール（ＬＰＶ）およびインジナビール（ＩＤＶ）を含む。

【０４２６】

各薬物については、データは形式（x_i，y_i），ｉ＝１．．．Ｎの対に構造化されており、ここで、Ｎは訓練データを構成する試料の数であり、y_iは測定された薬物の倍耐性（または表現型）であって、x_iは突然変異のベクトル＋定数、$x_i=[1\ x_{i1}, x_{i2}\ldots x_{iM}]^{T}$であり、ここで、Ｍは関連酵素についての可能な突然変異の数である。もしｍ番目の突然変異がｉ番目の試料に存在すれば要素$x_{im}=1$であり、その他の場合$x_{im}=0$に設定する。各突然変異はコドン遺伝子座および置換されたアミノ酸の双方によって特徴付けられる。アミノ酸配列に影響しない突然変異は無視する。各薬物についての試料に１％を超えて存在する突然変異のみがモデルについての可能なプレディクターのセットに含まれることを注記する。というのは、耐性に関連する突然変異はあまり頻繁でなく起きるのはありそうもないからである。測定y_iは野生型と比較した突然変異ウイルスについての薬物の倍耐性を表す。具体的には、y_iは、野生型ウイルスのIC_{50}と比較した、突然変異したウイルスのIC_{50}（複製を５０％だけ遅らせるのに必要な薬物の濃度）の比率のｌｏｇである。目標は、x_iからy_iを正確に予測する各薬物についてのモデルを開発することである。データに対してバッチ最適化を行うためには、Ｎ×Ｍ＋１行列、$X=[x_1, x_2\ldots x_N]^{T}$に独立変数をスタックし、ベクトル$y=[y_1, y_2\ldots y_N]^{T}$に全ての観察をスタックする。

【０４２７】

各アルゴリズムの性能は交差－確証を用いて測定する。各薬物については、一次相関係数Ｒを、モデルの予測された表現型応答、およびテストデータの現実の測定されたイン・ビトロ表現型応答の間で計算する。

【０４２８】

【数１１５】

$$R=\frac{\left(\hat{y}-\bar{\hat{y}}\vec{1}\right)^{T}\left(y-\bar{y}\vec{1}\right)}{\left\|\hat{y}-\bar{\hat{y}}\vec{1}\right\|_{2}\left\|y-\bar{y}\vec{1}\right\|_{2}} \qquad (38)$$

ベクトル

【０４２９】

【数１１６】

$\hat{y}$

が表現型ｙの予測である場合、

【０４３０】

【数１１７】

$\bar{y}$

はベクトルｙにおける要素の平均を示し、

【０４３１】

【数１１８】

$\vec{1}$

は全てのもののベクトルを示す。各薬物および各方法については、各々、訓練およびテストのためにデータを比率９：１にランダムに細分化する。１つの例において、１０の異なる細分化を行って、訓練およびテストデータのいずれの重複もなくしてベクトル

【０４３２】

【数１１９】

$\hat{y}$

およびＲを得る。次いで、この全プロセスを１０回反復して、Ｒの１０の異なる値を得ることができる。Ｒの１０の異なる値を平均して、報告されたＲを得る。また、１０の異なる実験にわたって測定されたモデルの各々についてＲの標準偏差を決定して、モデルが統計学的に有意な方法で比較されることを確実とする。

【０４３３】

表１１はＰＩ薬物についての前記したモデルの結果を示し；表１２は１０のＲＴＩ薬物についての結果を示す。結果は、訓練およびテストデータの１０細分化にわたって平均した、修正係数Ｒの形式で示す。試料偏差から計算したＲの平均値の見積もった標準偏差も示す。各薬物についての利用可能な試料の数を最後の列に示す。平均性能を増加させるためにテストした方法は：ｉ）ＲＲ－Ｒｉｄｇｅ回帰、ｉｉ）ＤＴ－検出ツリー、ｉｉｉ）ＮＮ－神経ネットワーク、ｉｖ）ＰＣＡ－主成分分析、ｖ）ＳＳ－段階的選択、ｖｉ）直線核でのＳＶＭ＿Ｌ－サポートベクトルマシーン、ｖｉｉ）ＬＡＳＳＯ－最小絶対収縮および選択オペレーター、およびｖｉｉｉ）径基礎核でのＳＶＭ－サポートベクトルマシーンである。表１１および１２の最後の行中の情報を図２１に示す。図２１中の円は、各ＰＩについての１０の異なる実験にわたって平均し、かつ７の異なるＰＩにわたって平均した相関係数Ｒを示す。図２１中の菱形は、各ＲＴＩについての１０の異なる実験にわたって平均し、かつ１０の異なるＲＴＩにわたって平均した相関係数Ｒを呈する。１標準偏差誤差棒も示す。

【０４３４】

モデル化技術がチューニングパラメーターを含む場合は常に、これらは、グリッドサーチアプローチを用い、交差－確証によって測定されたように技術の最適性能のために調整されている。全ての場合において、グリッド量子化は、グリッドからの最良の実行パラメーターが所与のデータについての最適パラメーターから現実的には識別可能であるのに十分良好であった。というのは、グリッド量子化による予測の差は実験ノイズを低めるからである。

【０４３５】

データには強い傾向があるが、試料の数の差のため、基礎となる遺伝子プレディクター、および薬物の間で変化するデータ中の他の特異性、各アルゴリズムによって達成されるＲの相互作用は薬物間で変化し得ることは注意すべきである。この変動は、表１１（３ないし９行）および表１２（３ないし１２行）の個々の薬物行を調べることによって見ることができる。

【０４３６】

全ての方法のうち、ＳＶＭは良好に実行され、ＬＡＳＳＯを僅かに凌ぐ（ＲＴＩについてＰ＜０．００１；ＰＩについてＰ＝０．１８）。ε－非感受性喪失関数で訓練したＳＶＭの性能は、サポートベクトルマシーンに基づいた従前に報告された方法のそれよりもかなり良好である。非線形核関数を用いるＳＶＭは、線形核関数を用い、およびε－非感受性喪失関数を用いても訓練されるＳＶＭ　Ｌを凌ぐ（ＲＴＩについてＰ＝０．００３；ＰＩについてＰ＜０．００１）。ＳＶＭは、神経ネットワークを用い、かつ凸コスト関数および連続的サブセット選択を創製しない他の非線形技術をかなり凌ぐ（ＲＴＩおよびＰＩ双方についてＰ＜０．００１）。凸コスト関数を用いて線形回帰モデルを訓練し、ＬＯＳＳＯ技術は、ＳＳ技術をかなり凌ぐ（ＰＩおよびＲＴＩ双方についてＰ＜０．００１）。トップの５つの方法、すなわち、ＳＳ、ＰＣＡ、ＳＶＭ＿Ｌ、ＬＡＳＳＯ、ＳＶＭ＿Ｒは、全て、疎であるモデルを創製する傾向があるか、または限定された数の非ゼロパラメーターを有する。

【０４３７】

プレディクターとして選択された突然変異のサブセットを説明するために、本明細書中に開示されたある実施形態は第二の最良の実行モデル、すなわち、ＳＶＭとは異なり、プレディクターの間の非線形または論理的結合を模倣することを試みない線形回帰モデルを創製するＬＡＳＳＯに焦点を当てる。結果として、どのようにして多くのプレディクターを選択するかを示すのは直接的である。表１３は、各モデルを訓練するにおいて用いられる、突然変異の数（表１３、３列）、および試料の合計数（表１３、２列）と共に、各ＰＩ薬物についてのプレディクターとしてのＬＡＳＳＯによって選択された突然変異の数（表１３、４列）を示す。同一の表が、ＲＩＴについて示される（表１４、同一列は同一事項に対応する）。

【０４３８】

選択された突然変異もまた薬物耐性の原因の理解を高めることができる。図２２、２３および２４は、各々、ＰＩ、ヌクレオチドＲＴＩ（ＮＲＴＩ）、および非ヌクレオチドＲＴＩ（ＮＮＲＴＩ）に対する応答を予測するためにＬＡＳＳＯによって選択されたパラメーターの値を示す。図面中の各列は薬物を表し；各行は突然変異を表す。関連突然変異はＰＩ薬物についてはプロテアーゼ酵素に対する、およびＲＮＴＩおよびＮＮＲＴＩ薬物についてはＲＴ酵素に対するものである。各四角の陰影は、その薬物についてのその突然変異に関連するパラメーターの値を示す。右側の色付き棒線（各々、２２０１、２３０１および２４０１）によって示されるように、陰影を付したダーカーであるプレディクターは増大した耐性に関連し；陰影を付したライターであるパラメーターは増大した感受性に関連する。突然変異は、関連パラメーターの平均の大きさを減少させる順序で左側から右側の順序とする。関連パラメーターをクラスにおいて全ての列または薬物にわたって平均する。４０の最大のパラメーターの大きさに関連する突然変異を示す。特定の突然変異、または行については、パラメーターの値は、列、または同一クラスにおける異なる薬物にわたってかなり変化する。

【０４３９】
アルゴリズムＲＲ、ＤＴ、ＮＮ、およびＳＳについては、モデルは、全ての遺伝子突然変異についてではなく、むしろＤｅｐａｒｔｍｅｎｔ ｏｆ Ｈｅａｌｔｈ ａｎｄ Ｈｕｍａｎ Ｓｅｒｖｉｃｅｓ（ＤＨＨＳ）によって耐性に影響すると考えられる部位で起こる突然変異のサブセットについて訓練した。独立変数の数の低下は、これらのアルゴリズムの性能を改良することが判明した。ＳＶＭ＿Ｌアルゴリズムの場合には、全ての突然変異についてモデルを訓練することによってＰＩに対する最良の性能を達成しつつ、ＤＨＨＳ突然変異サブセットのみを用いてＲＴＩに対する最良の性能を達成した。全ての他のアルゴリズムについては、最良の全性能は、全ての突然変異についてモデルを訓練することによって達成された。

【０４４０】
プレディクターとしてのＬＡＳＳＯによって選択されたが、現在、耐性に影響するとＤＨＨＨＳによって判断された遺伝子座と関連付けられていない図２２、２３、および２４に示された突然変異のセットは：ＰＩについては－１９Ｐ、９１Ｓ、６７Ｆ、４Ｓ、３７Ｃ、１１Ｉ、１４Ｚ；ＮＲＴＩについては－６８Ｇ、２０３Ｄ、２４５Ｔ、２０８Ｙ、２１８Ｅ、２０８Ｈ、３５Ｉ、１１Ｋ、４０Ｆ、２８１Ｋ；およびＮＮＲＴＩについては－１３９Ｒ、３１７Ａ、３５Ｍ、１０２Ｒ、２４１Ｌ、３２２Ｔ、３７９Ｇ、２９２Ｉ、２９４Ｔ、２１１Ｔ、１４２Ｖである。ＬＡＳＳＯおよびＳＶＭのようないくつかの場合においては、ＬＰＶのような特定の薬物についての性能は、ＤＨＨＳよって耐性に影響すると認識された遺伝子座のみが含まれた場合（Ｒ＝８１．７２，Ｓｔｄ．ｄｅｖ．＝０．１８）と比較して、全ての突然変異がモデルに含まれた場合（Ｒ＝８６．７８，Ｓｔｄ．ｄｅｖ＝０．１７）、有意に改良された（Ｐ＜０．００１）ことを注記する。これは、ＤＨＨＳによって認識されたものを超えた他の突然変異が薬物耐性において役割を演じることができることを説明する。

【０４４１】
凸最適化技術の使用は、本明細書中において、疎なパラメーターセットの連続的サブセット選択を達成して、正確に一般化される表現型予測モデルを訓練することが示された。ＬＡＳＳＯは、1^1ノルム収縮関数に適用して線形回帰パラメーターの疎なセットを生じる径基礎核関数でのかつε－非感受性喪失関数で訓練したＳＶＮは疎な非線形モデルを創製する。これらの技術の優れた性能は、それらのコスト関数の凸性、および疎なモデルを生じるそれらの傾向の点で、説明することができる。凸性は多くの潜在的プレディクターがある場合に、特定の訓練データセットについて全体的に最適なパラメーターを見出すことができるのを確実とする。疎なモデルは、遺伝子データに典型的なように、特に劣決定または悪条件データの関連でよく一般化される傾向がある。1^1ノルムは、最も選択的な凸関数として見ることができる。選択的収縮関数を用いる疎なパラメーターセットの選択は、Ｏｃｃａｍ’ｓ Ｒａｚｏｒと同様な原理で：多くの可能な理論が観察されたデータを説明できる場合、最も単純なものは最も正しいようである：を発揮する。ε－非感受性喪失関数と共に1^2収縮関数を用いるＳＶＭは、サポートベクトルと関連するパラメーターに適用された収縮関数として1^1ノルムの明示的な使用と同様な効果を生じる傾向がある。

【０４４２】
1^1収縮関数を用いる技術は、しばしば、ＩＶの数が大きくて、データが未決定または悪条件である場合、正確に一般化することができる。結果として、独立変数の非線形または論理的組合せをモデルに加え、良好なプレディクターである組合せを訓練で選択されると予測することが可能である。ＳＭＶは、線形核関数よりも有意に良好に実行される、径基礎関数のような非線形核関数の使用と独立変数との相互作用をモデル化することが可能である。結果的に、本明細書中に開示した基本的な概念を変えることなく、独立変数の論理的組合せをモデルに加えることによって、ＬＡＳＳＯの性能を高めることができる。論理的項は、決定ツリーによって生じたものから、専門家則によって記載された論理的相互作用から、論理的回帰の技術から、または論理的項のランダム順列のセットさえから由来

することができる。ＬＡＳＳＯの利点は、パラメーターが、サポートベクトルよりはむしろ独立変数、または独立変数を含む表現を直接的に組み合わせるので得られるモデルが解釈するのが容易であることである。モデルにおける多数の独立変数に対するＬＡＳＳＯの頑強性は、1^{1}ノルムの選択的性質およびその凸性双方によるものである。

【０４４３】

1^{1}ノルムよりも収縮関数をより選択的に使用する他の技術が存在する。例えば、ｌｏｇ－収縮回帰は、モデルパラメータセットに存在する情報の量を測定する暗号理論に由来する収縮関数を用いる。この技術は1^{1}－ノルムの代わりに収縮関数としてｌｏｇ関数を用い、その結果、非凸である。パラメーターの疎なセットを求めるための理論的に興味があるアプローチを供しつつ、ペナルティ関数の非凸性は、対応する回帰を解くことが、ＬＡＳＳＯよりも依然として計算の、扱いやすくなく、プレディクターの大きなセットについては所与のデータについての全体的最小よりはむしろ局所的な最小のみを生じさせることができることを意味する。

【０４４４】

本明細書中に記載された技術は、広い範囲の表現型予測問題についての線形および非線形回帰モデルの創製に適用することができる。それらは潜在的遺伝子プレディクターの数が測定された結果の数と比較して大きい場合に特に関連する。

【０４４５】

遺伝子独立変数を異なる空間へマッピングすることによる回帰モデルの単純化

前記したように、遺伝子マーカーの複雑な組合せを考える場合、ＳＮＰ変数をもう１つの変数空間に投影して、分析を単純化することが可能であることを注記する。この変数空間は、ＨａｐＭａｐ　Ｐｒｏｊｅｃｔによって記載されたクラスターまたはビンのような、突然変異の公知のパターンを表すことができる。言い換えれば、前記した特定のＳＮＰ突然変異を表すベクトルx_iよりはむしろ、それは、個体が特定のＨａｐＭａｐクラスターまたはビンに入るか否かを表すことができる。例えば、前記した表記方法に従い、Ｂが関連ＨａｐＭａｐビンの数であるベクトル$x_i = [x_{i1}, x_{i2} \cdots x_{iB}]^{T}$があると想像する。もし個体のＳＮＰＳパターンがｂ番目のビンに入るならば、要素x_{ib}＝１を設定し、そうでなければ０を設定することができる。別法として、もし個体ＳＮＰおよび特定のビンの間の重複が不完全であって、カテゴリー「他の」において単純に個体を置き換えるのが望ましくないのであれば、各x_{ib}を、ＳＮＰのパターンおよびビンｂのそれの間の重複の割合と等しく設定することができる。本明細書中に開示された概念を変えることなく多くの他の技術は回帰問題を公式化することが可能である。

【０４４６】

結果予測についての交差確証によるモデルの選択

この議論を進めた中で、専門家則、偶発事象表、線形および非線形回帰を含む異なる表現型予測技術を記載した。さて、訓練データの使用に基づき、特定の対象についての特定のカテゴリーまたは非カテゴリー結果をモデル化するのが最良であるモデル化技術のセットから選択する一般的アプローチを記載する。図２５は、システムについての説明的フローダイアグラムを供する。図２５に記載されたプロセスは、特定の患者、モデル化すべき表現型、およびデータをテストし訓練する所与のセットで利用できるデータを仮定して最良のモデルを選択する一般的アプローチであり、該プロセスは特定のモデル化技術から独立している。好ましい実施形態において、用いることができるモデル化技術のセットは、専門家則、偶発事象表、ＬＡＳＳＯで、またはデータが劣決定されていない場合は単純な最小二乗で訓練された線形回帰モデル、およびサポートベクトルマシーンを用いる非線形回帰モデルを含む。

【０４４７】

該プロセスは、モデル化されるであろう、あるいはもしそれがカテゴリー変数であれば、それについて確率をモデル化することができる、特定の対象および特定の従属変数（ＤＶ）を選択で開始する２５０１。次いで、該システムは、対象の記録に関連し、かつＤＶの結果のモデル化に関連し得る独立変数（ＩＶ）のセットを決定する２５０２。システム

のヒトユーザーは、ユーザーがモデルに関連して可能と考えるＩＶのそのサブセットを選択することもできる。次いで、システムはチェックして２５０３ａ、モデルが既に訓練され、独立変数の所与の組合せ、およびモデル化すべき所与の従属変数について選択されているか否かをみる。もしこれが当てはまり、かつ出来合いのモデルを訓練し、テストするのに用いるデータが旧式でなければ、システムは、そのモデルを用いる予測の創製に直接的に向かう２５１９。そうでなければ、システムは、注目する特定のＤＶを有し、かつ注目する特定の対象と同一のＩＶのセットを有しても有しなくてもよい。全ての他の記録をデータベースから抽出するであろう。そうすることにおいて、システムは、データがモデルを訓練しテストするのに利用できるか否かを決定する２５０３ｂ。もし答えが否であれば、システムは、いずれかの利用可能な専門家則があるかをみるためにチェックして２５１５、対象で利用可能なＩＶのサブセットに基づいて結果を予測する。もし専門家則が利用できなければ、システムは出て２５０４、それが有効な予測をできないと示す。もし１以上の専門家則ができれば、システムは、特定の対象のデータに最良に適する専門家則のサブセットを選択する２５０５。好ましい実施形態において、対象にいずれの専門家則を適用するかの選択は、その専門家則見積もりにおける信頼性のレベルに基づくであろう。もしそのような信頼性見積もりが利用できなければ、それらの特異性のレベルに基づいて、すなわち、注目する対象で利用できるどれくらい多くのＩＶを専門家則が予測で用いるかに基づいてランク付けすることができる。次いで、専門家則の選択されたサブセットを用いて予測を生じさせる２５０６。

【０４４８】

もしデータが利用できると判断されたならば２５０３ｂ、システムはチェックして２５１６、テストおよび訓練データで失われたいずれかのデータがあるか否かを決定する。言い換えれば、関連ＤＶを含む全ての記録について、システムはチェックして、全ての記録が、注目する患者について利用できるのと正確に同一のＩＶのセットを有するか、およびいずれがモデルにおいて潜在的予測であり得るかをチェックする。典型的には、答えは「否」であろう。というのは、異なる情報が異なる患者で利用可能だからである。もし失われたデータがあればシステムは四方を進んで対象にとって最良の可能な予測をなすのに用いるべきＩＶのセットを見出す。この手法は時間を消費するものである。というのは、それは多数ラウンドのモデル訓練および交差－確証を含むからである。その結果、この手法における最初の工程は、考えられるＩＶのセットを、利用可能な計算時間に基づいて管理可能なサイズに低下させることである２５０７。好ましい実施形態においてＩＶのセットは、やはり利用可能なＤＶを有する対象のあるパーセンテージについてのそのＩＶに関するデータがあることに基づいて低下させる。単純な線形回帰モデルを仮定し、それらはモデル化誤差に関連する程度に基づいてＩＶを選択する段階的選択のような当該分野で知られた他の技術を用いて、ＩＶのセットをさらに低下させることができる。次いで、システムはループに入り、そこでは、残りのＩＶの各組合せが調べられる。好ましい実施形態において、各ＩＶおよびＤＶについても以下の状態を考慮する：各ＩＶはモデルに含めることができるか、または含めることができず、全ての対象について陽性であるＩＶまたはＤＶについての数値データでは、該データはその対数を取ることによって進行させても、させなくてもよい。ＩＶの包含／排除および前処理の各特定の組合せについてモデル化技術のセットを適用する２５１０。

【０４４９】

ほとんどのモデル化技術は、テストデータでの交差－確証を用いるグリッド－サーチアプローチに基づいて最適化し、またはチューニングすることができるいくつかのチューニングパラメーターを有するであろう。例えば、先に議論したＬＡＳＳＯ技術については、多くの値が変数パラメーターλについて調べられる。λの各値について、回帰パラメーターを訓練することができ、モデルの予測をテストデータの測定された値と比較することができる。同様に、先に議論したサポートベクトルマシーンアプローチでは、グリッド－サーチアプローチを用いて最適化すべきチューニングパラメーターはＣ、ε、および、おそらくは核関数の特徴を記載するパラメーターを含む。偶発事象表に基づいた技術ではチュ

ーナブルパラメーターは、先に議論したように、偶発事象を所与の対象についてできるだけ特異的としつつ、偶発事象表モデルから許容できる最高の標準偏差と比較することができる。

【０４５０】

多くの異なる行列を用いて、モデル予測をテストデータと比較して、チューナブルパラメーターを最適化し、モデルを選択することができる。好ましい実施形態において、誤差の標準偏差を用いる。他の実施形態において、予測されたおよび測定された結果の間の相関係数Ｒを用いることができる。論理的回帰または偶発事象表の関係で、事後確率、すなわち、各テスト結果の尤度のモデルの予測を仮定するテストデータの所与のセットの確率を用いることもできる。いずれの測定基準を用いようとも、もし予測誤差の標準偏差をテスト測定基準として用いるならば、予測誤差の標準偏差の最小化のような、測定基準の値を最適化するチューニングパラメーターのその値を選択する。モデル訓練および交差－確証はゆっくりとしたプロセスであるので、この段階２５１０において、異なるチューニングパラメーターが調べられるように規定するグリッドは、最良のモデルおよび最良のチューニングパラメーターの粗いアイデアのみを得ることができるように、利用可能な時間の量に基づいておおまかセットされる。

【０４５１】

一旦、全ての異なるＩＶ／ＤＶ組合せがこのようにして調べられたならば２５１１、システムが、テスト測定基準の最良の値を達成した、ＩＶ／ＤＶの組合せ、モデルおよびチューニングパラメーターを選択する。もし失われたデータがなければ、システムはＩＶ／ＤＶのすべての組合せをチェックする工程をとばすことを注記する。代わりに、システムは、異なるモデル化技術およびチューニングパラメーターを調べ２５０８、テスト測定基準を最大化するモデル化方法およびチューニングパラメーターのセットを選択する。次いで、より細かく間隔を設けられたグリッドを用い、システムは最良の回帰モデルの洗練されたチューニングを行い、チューニングパラメーター値の各セットについて、テストデータとの相関を決定する。テスト測定基準の最良な値を生じるチューニングパラメーターのセットを選択する。次いで、システムは、予測誤差の標準偏差のようなテスト測定基準が、予測が有効と考えられるように、選択された閾値未満であるか否かを決定する２５１８。例えば、１つの実施形態において、Ｒ＞０．５の相関係数は予測が有効とみなされるのに望ましい。もし得られたテスト測定基準が閾値を満足しないならば、予測を行うことができない２５１７。もしテスト測定基準が必要な閾値を満足するならば、予測で用いたＩＶおよびモデルがテストデータで達成した相関係数の組合せと共に、表現型予測を生じさせることができる。

【０４５２】

失われたデータでの癌コホートにおける交差確証によるモデル選択の説明

この態様を示すためには、Ｎａｔｉｏｎａｌ Ｉｎｓｔｉｔｕｔｅ ｏｆ Ｈｅａｌｔｈ’ｓ Ｐｈａｒｍａｃｏｇｅｎｏｍｉｃ Ｒｅｓｅａｒｃｈ Ｎｅｔｗｏｒｋの一部であって、どのようにして個々の遺伝子変異が異なる薬物応答に寄与するかを発見する使命を有するＰｈａｒｍＧＫＢで見出すことができる結腸癌に関連する遺伝子および表現型データの利用に焦点を当てた。このデータベースについては、鍵となる挑戦は失われた情報であった。理想的には、前記した回帰技術を適用して、特定の患者に利用できるすべてのＩＶからのモデルについてのＩＶサブセットを自動的に選択したいであろう。しかしながら、これは、モデルを訓練し、テストするために他の患者から入手できるデータの量を制限する。その結果、あまり十分でないＩＶを含有するデータベースについては、独立変数の全ての可能なサブセットを通じてサーチすることが可能である。各々について、前記したように、必要な結果が測定され、および独立変数の関連セットが利用できる患者のセットを抽出することができる。前記したように、可能な方法の空間をサーチして、陽性数的独立変数のｌｏｇを取ることのような含まれた独立変数を前処理することもできる。含まれた独立変数の各組合せ、および独立変数前処理技術については、テストデータでの交差－確証によってモデルを訓練し、テストする。テストデータでの最良の交差－確証を有す

るモデルを選択する。一旦、ＩＶについての所与のセットのためにモデルを創製したならば、網羅的なモデルサーチを必要とすることなくＩＶの同一セットが供給された新しい患者データにそのモデルを適用する。

【０４５３】

この技術は、結直腸癌薬物イリノテカンについての臨床的副作用を予測するのに用いられてきた。ひどい毒性がイリノテカンを受ける癌患者で共通して観察される。イリノテカン薬物動態学および副作用等、イリノテカン代謝酵素および推定関連性のトランスポーターをコードする遺伝子の対立遺伝子変種との間の関係を記載するデータが含まれた。患者を、ＭＤＲ１　Ｐ－糖蛋白質（ＡＢＣＢ１）、多薬物耐性－関連蛋白質ＭＲＰ－１（ＡＢＣＣ１）およびＭＲＰ－２（ＡＢＣＣ２）、乳癌耐性蛋白質（ＡＢＣＧ２）、チトクロームＰ４５０イソ酵素（ＣＹＰ３Ａ４，ＣＹＰ３Ａ５）、カルボキシルエステラーゼ（ＣＥＳ１，ＣＥＳ２）、ＵＤＰグルクロノシル－トランスフェラーゼ（ＵＧＴ１Ａ１，ＵＧＴ１Ａ９）、および肝臓転写因子ＴＣＦ１をコードする遺伝子における変異について遺伝子タイプ分けした。この研究のための遺伝子配列データに関連する表現型データを表１５に記載する。

【０４５４】

図２６は、ファルマコゲノミック移動エンジンを用いて供給された利用可能なＰｈａｒｍＧＫＢデータが所与の、イリノテカンでの結腸癌治療のための予測結果のモデルを説明する。図２６において、モデルは、関連遺伝子座（２６０１）、用いるインジケーター、この場合、０ないし２４時間からのＣＰＴ－１１の濃度曲線下面積（ＡＵＣ）のｌｏｇ（２６０２）、および１２日ないし１４日に絶対好中球カウントのＮａｄｉｒのｌｏｇ（２６０４）を予測するための０ないし２４時間のＳＮ－３８　ＡＵＣのｌｏｇ（２６０３）を示す。テストデータでモデルを交差－確証し、Ｒ＝６４％の相関係数が達成された（２６０５）。モデル予測の経験的標準偏差はモデルを訓練するのに用いられた（２６０７）結果のヒストグラムに重ねて示す（２６０６）。これらの統計学を用いて、イリノテカン治療を完全に差し控えるような通知した治療決定を行い、あるいは顆粒球コロニー刺激因子のような第２の薬物を投与して低いＡＮＣおよび得られた感染を妨げることができる。

【０４５５】

高められた診断報告

病気治療の関係では創製された遺伝子データは、データを用いて治療用療法を選択するのを助けることができる臨床家にとって最も用いられるものである。１つの態様において、表現型予測を状況に当てはめ、臨床家または患者に対する報告に組織化する。もう１つの態様において、本明細書中に開示されたシステムおよび方法は、診断ｌａｂ２７０３がｌａｂテスト２７０１および医療報告２７０２からのデータを確証し、それをデータセンター２７０４に送り、そこで、それは開示された方法を用いて分析された標準的腫瘍学に一体化されるより大きなシステム（図２７参照）の一部として用いることができ、高められた診断報告２７０５が創製され、医師２７０６に送られる。

【０４５６】

報告を創製することができる１つの可能な状況は、イリノテカンで治療される結腸癌患者についての予測臨床結果に関するであろう。それは、治療のための禁忌の概念、投与スケジュール、副作用プロフィールをコードすることができる。そのような副作用の例は、２つとも普通である骨髄抑制および後期－開始下痢、緊急の医療的看護を必要とするイリノテカン治療の用量－律速副作用を含む。加えて、ひどい好中球減少症およびひどい下痢は、各々、患者の２８％および２１％に影響する。あるＵＧＴ１Ａ１対立遺伝子、肝臓帰納テスト、ギルバート症候群の過去の医療的履歴、および抗－痙攣薬およびいくつかの抗－催吐薬のようなチトクロームｐ４５０を誘導する患者投薬の同定は、イリノテカン用量調整を警告するインジケーターである。

【０４５７】

図２８は、表現型予測を用いるイリノテカンでの結直腸癌治療についての高められた報告のモック－アップである。治療に先立ち、報告は患者の癌の段階、過去の医療履歴、現

在の投薬および薬物用量を推奨するためのＵＧＴ１Ａ１遺伝子型を考慮する。最初の薬物投与からほぼ１日後に報告は、ＵＧＴ１Ａ１遺伝子中の突然変異、および患者の血液から測定された代謝産物（例えば、ＳＮ－３８、ＣＰＴ－１１）に基づいた、ほぼ２週間の時間における患者の絶対好中球カウントの予測されたＮａｄｉｒの予測を含む。この予測に基づき、医師は、患者にコロニー刺激因子薬物を与え、またはイリノテカン用量を変更するか否かを決定することができる。また、患者を血液カウント、下痢のグレードについてモニターする。データ源および推奨の正当性を供する。

【０４５８】

態様の組合せ

先に述べたように、本開示の利点を仮定すれば、他の態様、特徴および実施形態は本明細書中に開示された方法およびシステムの１以上を実行することができる。以下に、開示された発明の種々の態様を複数の方法で組み合わせることができる状況を説明する例の短いリストを掲げる。このリストは包括的であることを意図せず、本発明の態様、特徴および実施形態の多くの他の組合せが可能であることに注意するのは重要である。

【０４５９】

１つの例は、各々の値を最適化する方法での種々のゲノタイピング測定技術を利用することができる。例えば、ｌａｂは、Ａｐｐｌｉｅｄ　Ｂｉｏｓｃｉｅｎｃｅ　Ｔａｑｍａｎアッセイのような低シグナルの場合において、高価であるが、高い品質のデータを与えることができる技術を用いて、標的ＤＮＡを測定し、およびＡｆｆｙｍｅｔｒｉｘ’ｓ５００Ｋ　Ｇｅｎｅｃｈｉｐ、またはＭＩＰＳのような高価であるが、多量の遺伝物質を必要とする技術を用いて、良好な質のデータを与え、親ＤＮＡを測定することができる。

【０４６０】

もうひとつの例は、ＩＶＦ治療を受けているカップルが婦人から収穫された卵を有し、男性からの精子で受精させ、８つの生きた胚を生じる状況であろう。胚盤胞を各胚から収穫し、胚盤胞からのゲノムデータを、Ｔａｑｍａｎゲノタイピングアッセイを用いて測定する。他方、Ｍｏｌｅｃｕｌａｒ　Ｉｎｖｅｒｓｉｏｎ　Ｐｒｏｂｅｓ（分子逆転プローブ）を用い、双方の親から採った組織からジプロイドデータを測定する。男性の精子の１つからの、および婦人の卵の１つからのハプロイドデータもＭＩＰを用いて測定する。親の遺伝子データを用いて８つの胚盤胞のＳＮＰデータを清澄化する。次いで、清澄化された遺伝子データを用いて胚の潜在的表現型に関して予測を行う。最も有望なプロフィールを有する２つの胚を選択し、婦人の子宮に着床させる。

【０４６１】

もう１つの例は、その夫がテイ・サックス病の家族履歴を有する妊娠した婦人が、彼女が担う胎児が遺伝子的に罹患性であるかを知りたがっているが、羊水穿刺は流産のかなりの危険性があるのでそれを受けることを望まない状況であろう。彼女は血液を吸い取り、幾らかの胎児ＤＮＡを彼女の血液から単離し、ＭＩＰを用いてそのＤＮＡを分析する。彼女および彼女の夫は、従前に分析された彼らの十分なゲノムデータを既に有しており、それはイン・シリコで利用可能である。医師は、親ゲノムのイン・シリコ知識および本明細書中に開示した方法を用いて胎児ＤＮＡデータを清澄化し、テイ・サックス病の原因である臨界的遺伝子が胎児のゲノムに存在するかをチェックすることもできる。

【０４６２】

もう１つの例は、４４歳の妊娠した婦人が、彼女が担う胎児がダウン症候群を有し得るかに関心がある状況であろう。彼女は、流産の個人的履歴を仮定すれば、出生前診断で用いる煩わしい技術を有することを警戒しており、従って、彼女は自分の血液を分析することを選択する。健康ケア実践者は、母体血液試料中の胎児細胞を見出すことができ、婦人自身も遺伝子データの知識とともに本明細書中に開示した方法を用い、異数性について診断することができる。

【０４６３】

もう１つの例は、カップルがＩＶＦ治療を受けており；彼らは婦人から収穫した卵を有し、男性からの精子で受精させ、９つの生きた胚を生じる状況であろう。胚盤胞が各胚か

ら収穫され、胚盤胞からのゲノムデータをＩｌｌｕｍｉｍａビーズアッセイを用いて測定する。他方、分子逆転プローブを用いて双方の親から採取された組織からジプロイドデータを測定する。同一方法を用い、父親の精子からのハプロイドデータを測定する。母親から入手できる過剰な卵はなく、従って、バルクジプロイド組織試料は彼女自身の父親および母親から採取され、精子試料は彼女の父親から採取される。それらはすべてＭＩＰを用い分析され、本明細書中で開示された方法を用いて、母親のゲノムについての遺伝子分析を供する。次いで、父親のジプロイドおよびハプロイドデータとともにそのデータを用いて胚盤胞の各々の遺伝子データの高度に正確な分析を行う。表現型予測に基づき、カップルは３つの胚を着床させることを選択する。

【０４６４】

もう１つの例は、競走馬飼育者が、彼の優勝競走馬によって種付けされた子馬がそれ自体が優勝馬となる尤度を増加させることを望む状況である。彼は所望の雌馬がＩＶＦによって妊娠されるように手配し、雄馬および雌馬からの遺伝子データを用いて、生きた胚から測定された遺伝子データを清澄化する。清澄化された胚遺伝子データは、育種者が関連遺伝子型－表現型相関を見出し、望ましい競走馬を最も生産するような着床用の胚を選択することを可能とする。

【０４６５】

もう１つの例は、妊娠した婦人が彼女が担う胎児がいずれかの深刻な病気に対する素因があるか否かを知りたい状況であろう。父親は既に亡くなっており、従って父親の兄弟および父親の父親から創製されたハプロイドおよびジプロイドデータを用いて、胎児血液サンプリングの間に集められた胎児細胞から測定された胎児の遺伝子データを清澄化することを助ける。健康ケア実践者によって契約された会社は清澄化された遺伝子データを用いて各予測の信頼性とともに、胎児が呈するような表現型のリストを提供する。

【０４６６】

もう１つの例は、乏しい研究室技術のため、汚染された胎児遺伝子データと場合によっては闘わなければならない羊水穿刺であろう。開示された方法を用いて、母性および父性遺伝子データを用いて汚染された胎児遺伝子データを清澄化することができる。開示された方法が汚染ＤＮＡの増大した速度を補うことができることを知って、稔性手法を緩和させることによって研究室がコストを切り詰めることができる状況を想像することができる。

【０４６７】

もう１つの例は、４０代の婦人が妊娠を得ようとしてＩＶＦを受けている状況であろう。彼女は、胚をスクリーニングして、遺伝病を最も有しないようであり、最も着床し、妊娠まで持っていけそうなものを選択することを望む。彼女が用いているＩＶＦクリニックは生きた胚の各々から胚盤胞を収穫し、標準的な手法を用いてＤＮＡを増幅し、鍵となるＳＮＰを測定する。次いで、技術者は本明細書中に開示された方法を用いて、染色体アンバランスについてスクリーニングし、また、胚の遺伝子データを見出し、それを清澄化して、各胚の表現型素因について予測を行う。

【０４６８】

もう１つの例は、妊娠した婦人が羊水穿刺を有し、本明細書中に開示された方法とともに、血液試料中の胎児細胞における遺伝物質を用いて異数性および他の染色体異常についてスクリーニングする。

【０４６９】

１つの例は、径基礎核関数およびノルム喪失関数とともにサポートベクトルマシーンを用いる非線形モデルがヒト成人の遺伝子型および表現型データを利用して、早期開始アルツハイマー病の尤度を予測し、該病気の開始を遅らせることができる可能なライフスタイルの変化および運動養生法を提案する。

【０４７０】

もう１つの例は、ＬＡＳＳＯ技術を用いる線形モデルが、癌の遺伝子データとともに、肺癌に罹った成人婦人の遺伝子型および表現型データを利用して、いずれの医薬が該病気

の進行を遅らせるのに最も効果的であるかを予測する婦人の医師についての医師用の報告を作成する。

【０４７１】

もう１つの例は、複数のモデルを、クローン病患者の遺伝子、表現型および臨床データより成る集合データについてテストし、次いで、最も正確であることが判明する非線形回帰モデルが成人男性の表現型および臨床データを利用して、彼のクローン病の徴候を緩和するようであるある種の栄養サプリメントを提案する報告を作成する状況であろう。

【０４７２】

もう１つの例は、Ｈａｐｍａｐ　Ｐｒｏｊｅｃｔを通じて獲得されたデータから形成された偶発事象表を利用し、かつ胚からの胚盤胞から集めた遺伝子情報を利用するモデルを用いて、もし胚が着床すれば、結果をもたらす子供のありそうな表現型に関して予測を行う状況であろう。

【０４７３】

もう１つの例は、新生児に感染するＨＩＶの株の遺伝子情報を利用する線形回帰モデルを用いて、いずれの抗ウイルス薬物が、もし投与されたならば、成人に達する最大のチャンスを彼女に与えるかを示唆する赤ん坊の医師用の報告を作成する状況であろう。

【０４７４】

もう１つの例は、新しい研究が公表され、中年婦人における心筋梗塞の罹患率、およびある遺伝子および表現型マーカーの間のある相関を示唆する状況であろう。次いで、これは非線形回帰モデルの使用を促進して、中年データの集合データならびにそのデータがシステムに知られている個体の遺伝子および表現型データを再度調べ、次いで、該モデルは、心筋梗塞の危険性が最もある婦人を同定し、予測される危険性を彼らに通知する各医師に送られる報告書を作成する。

【０４７５】

もう１つの例は、複数のモデルを試みられた種々の薬物介入を含めた、結腸癌に罹った人々の集合データについてテストされる状況であろう。最良の予測を可能とすることが判明するモデルを用いて、実験的新しい医薬から最も利益を受けるであろう患者、および該新しい医薬に対する権利を所有する会社によってそれらの結果が用いられ、臨床試験に行うにおいて彼らを助ける。

【０４７６】

定義

ＳＮＰ（単一ヌクレオチド多形）：個体間変異を示す傾向がある染色体上の特別な遺伝子座。

ＳＮＰを要求すること：直接的および間接的証拠を考慮し、特定の塩基対の同一性を質問すること。

対立遺伝子を要求すること：ＳＮＰを要求すること遺伝子データを清澄化すること：関連する個人の遺伝子データ、および本明細書中に記載された方法を用いて不完全な遺伝子データを取り、誤差のいくつかまたは全てを修正すること。

不完全な遺伝子データ：以下の：対立遺伝子ドロップアウト、不明瞭な塩基対測定、正しくない塩基対測定、偽シグナル、または失われた測定のいずれかを持つ遺伝子データ

信頼性：要求されたＳＮＰ、対立遺伝子、または対立遺伝子のセットが個体の現実の遺伝子状態を表す統計学的尤度

多重遺伝子：複数の遺伝子または対立遺伝子によって影響される

ノイジーな遺伝子データ：不完全な遺伝子データとも呼ばれる不完全遺伝子データ；

未清浄化遺伝子データ：測定された遺伝子データ、すなわち、生の遺伝子データにおいてノイズの存在について修正するのにいずれの方法も用いられたことがない；また、粗遺伝子データとも呼ばれる

直接的関係：母親、父親、息子、または娘

染色体領域：染色体のセグメント、または全染色体

親サポート：遺伝子データを清浄化する開示された方法で時々用いられる名称

染色体のセクション：１塩基対ないし全染色体のサイズの範囲とすることができる染色体のセクション。

【０４７７】

（表）

【０４７８】

【表１】

病気－連鎖遺伝子のOMIMまとめ

	常染色体	X－連鎖	Y－連鎖	ミトコンドリア	合計
*公知の配列を持つ遺伝子	9916	457	18	37	10428
公知の配列および表現型を持つ遺伝子	359	30	0	0	389
#表現型の記載、分子基礎既知	1552	147	2	26	1727
?メンデル表現型または遺伝子座、分子基礎未知	1362	135	1	0	1498
疑われるメンデル基礎を持つ表現型	2092	147	2	0	2241
合計	15381	916	56	63	16416

表1

【０４７９】

【表２】

異数性についてのスクリーニングの異なる技術の特徴

技術の記載	計算に適用された式 N=	アッセイ名称	E(m1)	E(m0)	Std(m1)	Std(m0)	コピー数の比較	モザイク現象	N	SNP当たりの価格	さらなるコスト	因子シグマ	1-Conf	セグメント	ウェル当たりの細胞	総細胞 w/o PCR	コスト
平均の比較 参照試料なし	2*2*5^2*s^2/(m1-m0)^2	Taqman	32.26	30.75	0.597	0.597	2 vs 1	100.0%	16	$0.20	30	5	2.87E-07	5	50	800	$46.00
平均の比較 参照試料あり	2*2*5^2*(2^s^2)/(m1-m0)^2	Taqman	32.26	30.75	0.597	0.597	2 vs 1	100.0%	32	$0.20	30	5	2.87E-07	5	50	1600	$62.00
平均の比較 参照試料なし	4*(2^2)*(0.95^s^2+0.05^s^2)/(0.05^2*(m1-m0)^2)	Taqman	32.26	30.75	0.597	0.597	2 vs 1	5.0%	1001	$0.20	30	2	2.30E-02	5	50	50050	$1,031.00
平均の比較 参照試料あり	4*(2^2)*(1.95^s^2+0.05^s^2)/(0.05^2*(m1-m0)^2)	Taqman	32.26	30.75	0.597	0.597	2 vs 1	5.0%	2002	$0.20	30	2	2.30E-02	5	50	100100	$2,032.00
平均の比較 参照試料なし	4*(2^2)*(0.95^s^2+0.05^s^2)/(0.05^2*(m1-m0)^2)	MPs	32.26	30.75	0.597	0.597	2 vs 1	5.0%	1001	$0.01	200	2	2.30E-02	5	50	50050	$250.05
平均の比較 参照試料あり	4*(2^2)*(1.95^s^2+0.05^s^2)/(0.05^2*(m1-m0)^2)	MPs	32.26	30.75	0.597	0.597	2 vs 1	5.0%	2002	$0.01	200	2	2.30E-02	5	50	100100	$300.10
平均の比較 参照試料なし	4*(2^2)*((1-ff)^s^2+ff^s^2)/(ff^2*(m1-m0)^2); ff = 0.061	Taqman	32.26	30.75	0.597	0.597	2 vs 1	6.1%	384	$0.20	30	2	2.30E-02	5	50	19200	$414.00
平均の比較 参照試料なし	4*(1^2)*((1-ff)^s^2+ff^s^2)/(ff^2*(m1-m0)^2); ff = 0.058	Taqman	32.26	30.75	0.597	0.597	2 vs 1	4.1%	384	$0.20	30	1	1.59E-01	5	50	19200	$414.00
平均の比較 参照試料なし	4*(2^2)*((1-ff)^s^2+ff^s^2)/(ff^2*(m1-m0)^2); ff = 0.19	Taqman	32.26	30.75	0.597	0.597	2 vs 1	19.0%	70	$0.20	30	2	2.30E-02	5	50	3500	$100.00
平均の比較 参照試料なし	2*2*5^2*s^2/(m1-m0)^2	qPCR	27.65	26.64	0.53	0.53	2 vs 1	100.0%	27	$0.15	30	5	2.87E-07	5	50	1350	$50.25
平均の比較 参照試料あり	2*2*5^2*(2^s^2)/(m1-m0)^2	qPCR	27.65	26.64	0.53	0.53	2 vs 1	100.0%	55	$0.15	30	5	2.87E-07	5	50	2750	$71.25
平均の比較 参照試料なし	4*(2^2)*(0.95^s^2+0.05^s^2)/(0.05^2*(m1-m0)^2)	qPCR	27.65	26.64	0.53	0.53	2 vs 1	5.0%	1754	$0.15	30	2	2.30E-02	5	50	87700	$1,345.50
平均の比較 参照試料あり	4*(2^2)*(1.95^s^2+0.05^s^2)/(0.05^2*(m1-m0)^2)	qPCR	27.65	26.64	0.53	0.53	2 vs 1	5.0%	3508	$0.15	30	2	2.30E-02	5	50	175400	$2,661.00
平均の比較 参照試料なし	4*(2^2)*((1-ff)^s^2+ff^s^2)/(ff^2*(m1-m0)^2); ff = 0.107	qPCR	27.65	26.64	0.53	0.53	2 vs 1	10.7%	384	$0.15	30	2	2.30E-02	5	50	19200	$318.00
平均の比較 参照試料なし	4*(1^2)*((1-ff)^s^2+ff^s^2)/(ff^2*(m1-m0)^2); ff = 0.064	qPCR	27.65	26.64	0.53	0.53	2 vs 1	5.4%	384	$0.15	30	1	1.59E-01	5	50	19200	$318.00
平均の比較 参照試料なし	4*(2^2)*((1-ff)^s^2+ff^s^2)/(ff^2*(m1-m0)^2); ff = 0.19	qPCR	27.65	26.64	0.53	0.53	2 vs 1	19.0%	121	$0.15	30	2	2.30E-02	5	50	6050	$120.75

表2

【０４８０】

【表３】

snp_id	e1	e2	p1	p2	m1	m2	pe1	pe2	pp1	pp2	pm1	pm2
101100940	C	T	T	C	C	T	0.9538	0.8902	0.8626	0.8580	0.8654	0.9101
101164838	T	C	T	C	T	C	0.9359	0.9521	0.9406	0.9253	0.9957	0.8770
rs1463589	C	C	T	C	C	C	0.9428	0.9928	0.9841	0.9266	0.8661	0.9798
101028396	C	G	C	G	C	C	0.9252	0.8792	0.9246	0.9856	0.9819	0.8631
101204217	A	G	G	A	G	G	0.9799	0.9843	0.9194	0.9478	0.9438	0.9709
101214313	A	G	G	A	G	A	0.8513	0.9863	0.9521	0.9707	0.8570	0.9639
101231593	G	A	G	G	A	A	0.9857	0.9653	0.8908	0.9036	0.9431	0.9832
rs1426442	G	C	C	G	C	G	0.9338	0.9278	0.9469	0.9514	0.8766	0.9017
rs7486852	C	C	C	T	T	C	0.9566	0.9616	0.9390	0.8673	0.8785	0.8889
101266729	A	G	A	G	A	G	0.9238	0.9500	0.9026	0.9855	0.8760	0.9381

表3

【０４８１】

【表４】

snp_id	e1	e2	p1	p2	m1	m2	pe1	pe2	pp1	pp2	pm1	pm2
101019515	G	G	G	G	G	G	0.9134	0.8768	0.8666	0.9690	0.8679	0.8599
101100940	C	T	T	C	C	T	0.9538	0.8902	0.8626	0.8580	0.8654	0.9101
101160854	A	A	A	A	A	A	0.8705	0.9769	0.8763	0.8870	0.9311	0.9553
rs4980809	A	G	G	A	A	G	0.9638	0.9951	0.9582	0.9621	0.9197	0.9199
101058479	G	A	G	A	G	A	0.9003	0.9885	0.8906	0.9235	0.9787	0.8792
101236938	G	G	G	G	G	A	0.8528	0.9710	0.8810	0.9249	0.9274	0.9891
rs7137405	T	T	T	T	T	A	0.9360	0.9918	0.9148	0.9558	0.9135	0.9388
101251161	G	G	G	G	G	G	0.9802	0.8620	0.9372	0.8501	0.9891	0.8679
101270051	G	G	G	G	G	A	0.9004	0.9643	0.9778	0.9060	0.9943	0.8962
rs215227	G	G	G	G	G	A	0.9244	0.9236	0.9629	0.8575	0.9019	0.9362
101245075	G	G	G	G	G	G	0.9958	0.8593	0.9129	0.8504	0.8534	0.9866
101158538	A	G	A	G	G	G	0.9471	0.8909	0.8710	0.9581	0.8961	0.9046
rs2535386	A	A	A	A	A	A	0.9273	0.9479	0.9867	0.8918	0.9264	0.9750
rs6489653	T	T	T	T	T	T	0.9453	0.9776	0.9051	0.8547	0.9636	0.9532
101137205	C	G	C	C	G	G	0.8619	0.9503	0.9029	0.9426	0.8845	0.9282
101089311	T	C	C	C	C	T	0.8844	0.9381	0.9719	0.8636	0.9186	0.9652
101205712	A	A	A	A	A	A	0.8513	0.9226	0.8755	0.8999	0.9193	0.8535
101124605	G	G	G	G	G	G	0.8981	0.9093	0.9075	0.8676	0.8931	0.9258
101025989	G	T	T	G	G	T	0.9695	0.9016	0.8722	0.8821	0.9787	0.9273
rs4766370	T	A	A	T	T	A	0.8886	0.9166	0.8762	0.8767	0.9890	0.8536

表4

【０４８２】

【表５】

snp_id	true_value	true_hyp	ee	pp	mm	SnipProb	HypProb
101100940	CT	p2 m2	CT	TC	CT	0.8416	0.5206
101164838	CT	p2 m1	TC	TC	TC	0.9061	0.5206
rs1463589	CC	p2 m1	CC	TC	CC	0.9946	0.5206
101028396	GC	p2 m1	CG	CG	CC	0.9791	0.5206
101204217	AG	p2 m2	AG	GA	GG	0.9577	0.5206
101214313	GA	p1 m2	AG	GA	GA	0.9308	0.5206
101231593	GA	p1 m2	GA	GG	AA	1.0000	0.5206
rs1426442	CG	p1 m2	GC	CG	CG	0.9198	0.5206
rs7486852	CC	p1 m2	CC	CT	TC	0.9138	0.5206
101266729	AG	p1 m2	AG	AG	AG	0.9296	0.5206

表5

【０４８３】

【表６】

snp_id	true_value	true_hyp	ee	pp	mm	SnipProb	HypProb
101019515	GG	p1 m1	GG	GG	GG	1.0000	0.9890
101100940	TC	p1 m1	CT	TC	CT	0.9946	0.9890
101160854	AA	p1 m1	AA	AA	AA	1.0000	0.9890
rs4980809	GA	p1 m1	AG	GA	AG	0.9961	0.9890
101058479	GG	p1 m1	GA	GA	GA	0.9957	0.9890
101236938	GG	p1 m1	GG	GG	GA	1.0000	0.9890
rs7137405	TT	p1 m1	TT	TT	TA	1.0000	0.9890
101251161	GG	p1 m1	GG	GG	GG	1.0000	0.9890
101270051	GG	p1 m1	GG	GG	GA	1.0000	0.9890
rs215227	GG	p1 m1	GG	GG	GA	1.0000	0.9890
101245075	GG	p1 m1	GG	GG	GG	1.0000	0.9890
101158538	AG	p1 m1	AG	AG	GG	0.9977	0.9890
rs2535386	AA	p1 m1	AA	AA	AA	1.0000	0.9890
rs6489653	TT	p1 m1	TT	TT	TT	1.0000	0.9890
101137205	CG	p1 m1	CG	CC	GG	1.0000	0.9890
101089311	CC	p1 m1	TC	CC	CT	0.9940	0.9890
101205712	AA	p1 m1	AA	AA	AA	1.0000	0.9890
101124605	GG	p1 m1	GG	GG	GG	1.0000	0.9890
101025989	TG	p1 m1	GT	TG	GT	0.9973	0.9890
rs4766370	AT	p1 m1	TA	AT	TA	0.9973	0.9890

表6

【０４８４】

【表７】

			DHAlgorithm1		DHAlgorithm2	
Pop.Freq	ph	pd	P1精度	P2精度	P1精度	P2精度
データ	0.95	0.95	0.982	0.951	0.95	0.906
データ	0.75	0.75	0.891	0.811	0.749	0.618
データ	0.25	0.25	0.71	0.71	0.253	0.25
データ	0.5	0.9	0.849	0.838	0.499	0.768
データ	0.9	0.5	0.942	0.734	0.898	0.347
データ	0.6	0.8	0.852	0.816	0.601	0.673
均一	0.95	0.95	0.95	0.906	0.949	0.905
均一	0.75	0.75	0.749	0.612	0.749	0.612
均一	0.25	0.25	0.25	0.248	0.25	0.25
均一	0.5	0.9	0.69	0.669	0.501	0.671
均一	0.9	0.5	0.901	0.412	0.901	0.413
均一	0.6	0.8	0.678	0.618	0.6	0.618

表7

【０４８５】

【表８】

				PSAlgorithm1		PSAlgorithm2	
Pop.Freq	ph	pd	pe	P1精度	P2精度	P1精度	P2精度
データ	0.95	0.95	0.95	0.834	0.815	0.928	0.931
データ	0.75	0.75	0.75	0.797	0.769	0.819	0.819
データ	0.25	0.25	0.25	0.711	0.682	0.703	0.687
データ	0.5	0.9	0.9	0.849	0.838	0.866	0.864
データ	0.9	0.5	0.5	0.792	0.809	0.756	0.752
データ	0.6	0.8	0.8	0.777	0.788	0.835	0.828
均一	0.95	0.95	0.95	0.673	0.631	0.898	0.901
均一	0.75	0.75	0.75	0.549	0.497	0.635	0.65
均一	0.25	0.25	0.25	0.239	0.249	0.252	0.25
均一	0.5	0.9	0.9	0.601	0.611	0.814	0.818
均一	0.9	0.5	0.5	0.459	0.391	0.449	0.468
均一	0.6	0.8	0.8	0.544	0.511	0.672	0.679

表8

【０４８６】

【表９】

Farrer (2005)		
	D+	D-
A+B+	56/151	25/206
A+B-	26/151	59/206
A-B+	43/151	20/206
A-B-	26/151	20/206

Alvarez (1999)	D+	D-
A+	161/350	176/400
A-	189/350	224/400
Labert (1998)	D+	D-
B+	334/573	104/509
B-	239/573	405/509

表9

【０４８７】

【表１０】

prob.								
P(A+B+/D+)	0.28	0.024	0.24	0.32	0.37	0.039	0.29	0.45
P(A+B+/D-)	0.18	0.016	0.15	0.21	0.17	0.026	0.12	0.22
P(A+B-/D+)	0.22	0.027	0.16	0.27	0.28	0.037	0.21	0.35
P(A+B-/D-)	0.18	0.018	0.15	0.22	0.12	0.023	0.08	0.16
P(A-B+/D+)	0.31	0.026	0.26	0.36	0.39	0.038	0.26	0.4
P(A-B+/D-)	0.05	0.017	0.02	0.09	0.1	0.021	0.06	0.14
P(A-B-/D+)	0.2	0.032	0.13	0.28	0.02	0.011	0	0.04
P(A-B-/D-)	0.58	0.034	0.52	0.65	0.61	0.034	0.54	0.68

表10

【０４８８】

【表１１】

方法	平均	APV	ATV	NFV	RTV	SQV	LPV	IDV	平均
RR	平均	79.30	69.44	78.37	87.96	82.93	75.35	82.81	79.45
	標準偏差	0.26	0.90	1.18	0.49	0.60	1.36	0.65	0.66
NN	平均	83.16	68.45	87.59	90.88	89.10	73.67	89.07	83.13
	標準偏差	0.52	1.22	0.19	0.36	0.15	0.82	0.10	0.61
PCA	平均	86.49	73.52	87.65	91.98	88.72	84.79	87.86	85.86
	標準偏差	0.19	0.79	0.14	0.12	0.12	0.26	0.08	0.33
DT	平均	77.32	65.12	83.61	85.57	83.63	69.05	82.54	78.11
	標準偏差	0.33	0.93	0.15	0.21	0.31	0.87	0.34	0.54
SS	平均	85.25	73.27	89.18	91.65	89.81	77.15	87.79	84.87
	標準偏差	0.03	1.02	0.14	0.10	0.08	0.39	0.08	0.42
SVM L	平均	87.48	75.24	88.95	92.24	89.53	87.13	88.12	86.96
	標準偏差	0.15	1.05	0.09	0.09	0.10	0.23	0.10	0.42
LASSO	平均	89.17	76.60	89.88	93.47	91.00	86.78	88.61	87.93
	標準偏差	0.08	0.62	0.07	0.16	0.07	0.17	0.07	0.26
SVM R	平均	88.22	82.27	89.92	92.29	89.48	86.19	90.28	88.38
	標準偏差	0.03	0.46	0.06	0.07	0.08	0.32	0.08	0.22
試料		577	142	617	510	598	253	579	

表11

【０４８９】

【表１２】

方法	平均	3TC	ABC	AZT	D4T	DDC	DDI	DLV	EFV	NVP	TDF	平均
RR	平均	87.29	74.56	78.3	79.78	70.96	89.65	80.86	78.12	73.47	71.85	76.48
	標準偏差	0.72	1.15	0.94	0.61	1.06	0.66	0.17	0.29	0.63	0.56	0.69
NN	平均	94.33	74.96	71.12	80.29	73.42	85.79	84.16	80.5	84.05	68.71	77.73
	標準偏差	0.25	0.50	2.45	0.82	0.89	1.99	0.43	0.60	0.60	1.23	1.18
PCA	平均	92.23	83.3	84.14	86.83	85.28	85.19	80.88	77.49	77.34	68.36	82.30
	標準偏差	0.11	0.28	0.16	0.17	0.12	0.29	0.21	0.25	0.32	0.88	0.34
DT	平均	94.86	81.93	81.32	80.87	73.85	78.35	81.92	78.78	85.96	66.91	79.48
	標準偏差	0.22	0.31	0.58	0.59	0.78	0.61	0.30	0.82	0.36	2.05	0.84
SS	平均	94.91	94.43	81.86	86.88	86.99	82.17	83.97	80.15	76.8	67.27	82.52
	標準偏差	0.07	0.22	0.42	0.32	0.17	0.67	0.21	0.28	0.68	1.39	0.56
SVM L	平均	94.31	85.27	84.04	89.09	85.6	83.58	87.07	83.78	78.77	67.9	83.94
	標準偏差	0.12	0.21	0.21	0.15	0.17	0.32	0.16	0.24	0.48	1.67	0.69
LASSO	平均	95.48	86.32	85.54	91.37	87.54	85.74	85.16	82.31	82.19	70.68	85.22
	標準偏差	0.07	0.18	0.11	0.10	0.14	0.24	0.12	0.20	0.20	0.41	0.20
SVM R	平均	95.94	86.84	89.39	89.30	85.29	85.40	86.33	84.61	85.38	78.75	86.82
	標準偏差	0.11	0.19	0.13	0.18	0.26	0.22	0.20	0.28	0.27	0.40	0.25
試料		356	354	363	356	319	354	395	384	401	96	

表12

【０４９０】

【表１３】

薬物	APV	ATV	NFV	RTV	SQV	LPV	IDV
#試料	577	142	617	510	698	253	579
#突然変異	320	320	320	320	320	320	320
#プレディクター	120	50	90	120	120	100	110

表13

【０４９１】

【表１４】

薬物	3TC	ABC	AZT	D4T	DDC	DDI	DLV	EFV	NVP	TDF
#試料	366	364	353	356	319	364	335	334	401	96
#突然変異	791	791	791	791	791	791	791	791	791	791
#プレディクター	50	70	100	110	80	100	170	120	120	60

表14

【０４９２】

【表１５】

対象ID	Pharm_GKBにおける対象についてのPharm_GKBアクセションID
性別	男性、女性、または未知
人種民族性	カッコに入れた自己報告情報を伴うNIH_管理および予算オフィスによって規定
用量 (mg/m2)	イリノテカンの調整された容量
BSA(m2)	平方メートルで表した身体表面
用量 (mg)	ミリグラムで表したイリノテカンの用量
CPT-11 AUC24 SN38 AUC24 SN38 G AUC24	蛍光検出法と共に逆相HPLC法において測定した、1サイクルの1日目の、0ないし24時間の濃度曲線下のイリノテカン、SN－38およびSN－38Gの面積(hr*ng/ml)
APC AUC24	0ないし24時間の濃度曲線下の7－エチル－10－[4－N－(5－アミノペンタン酸)1－ピペリジノ]－カルボニルオキシカンプトテシン(APC)面積
ANC Nadir	最初のイリノテカンに用量(12ないし14日)後の絶対抗中球カウント(ANC)のNadir値
下痢グレード	処理のサイクルの間に観察された通常の毒性基準バージョン2．0下痢を用いて評価した

表15

。

【図１】

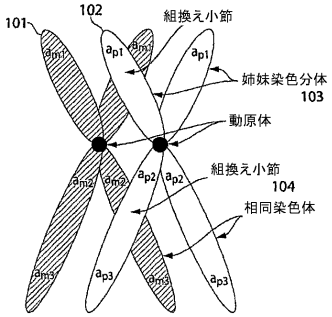

Fig. 1

【図２】

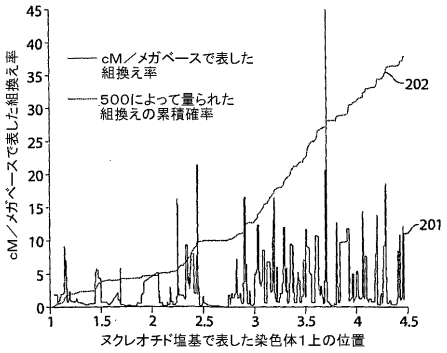

Fig. 2

【図３】

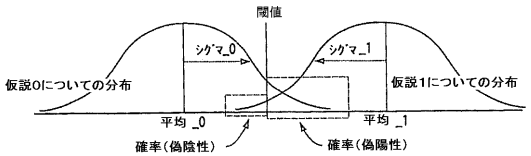

Fig. 3

【図４】

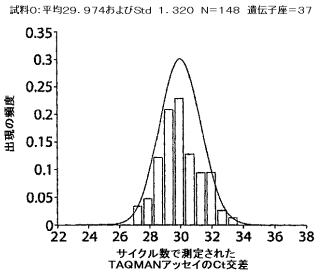

Fig. 4

【図５】

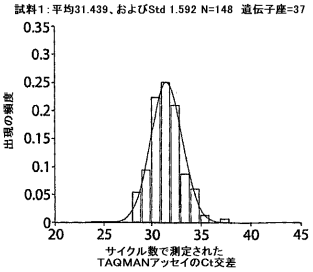

Fig. 5

【図６】

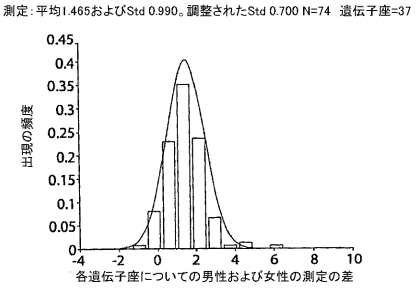

Fig. 6

【図７】

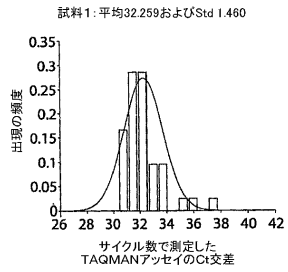

Fig. 7

【図８】

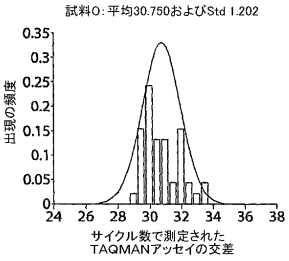

Fig. 8

【図９】

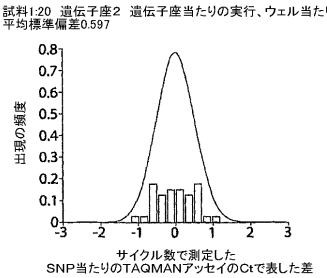

Fig. 9

【図１０】

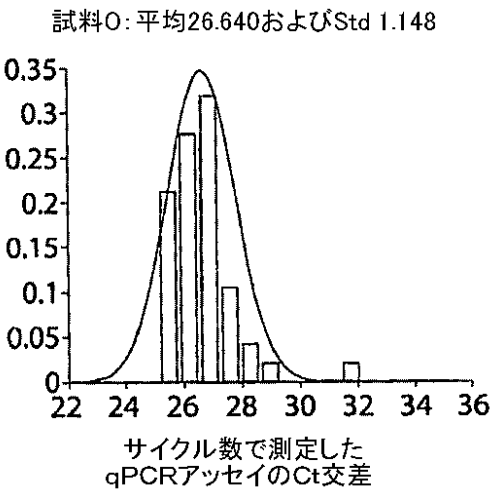

Fig. 10

【図１１】

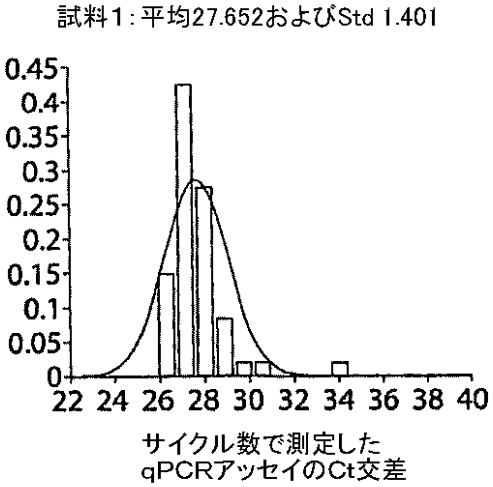

Fig. 11

【図１２】

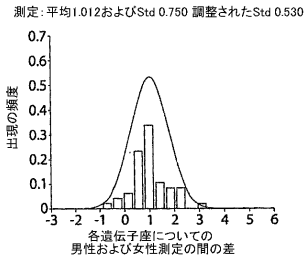

Fig. 12

【図１３】

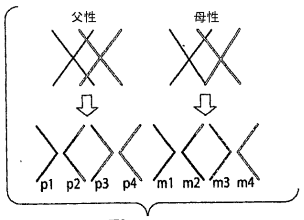

Fig. 13

【図１４】

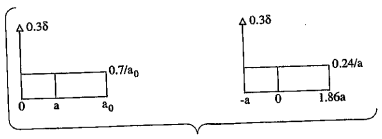

Fig. 14

【図１５】

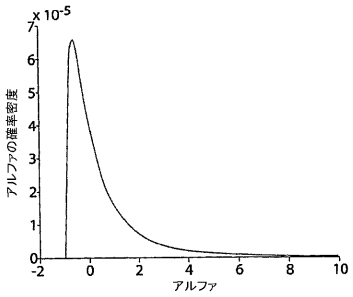

Fig. 15

【図１６】

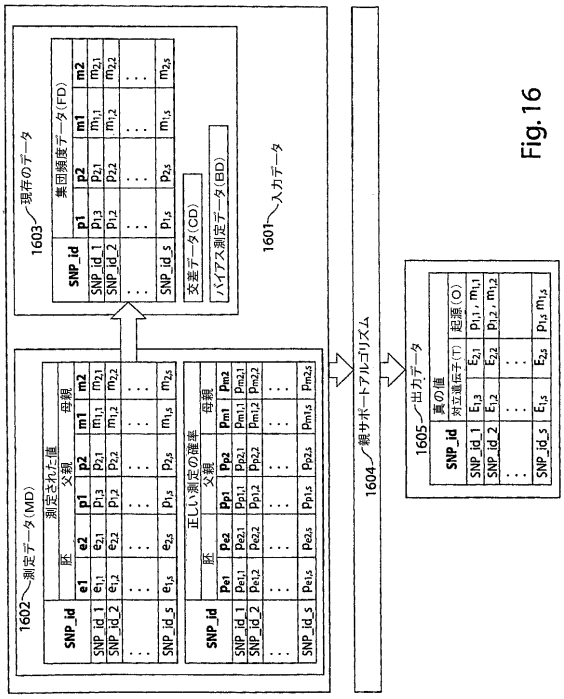

Fig. 16

【図１７】

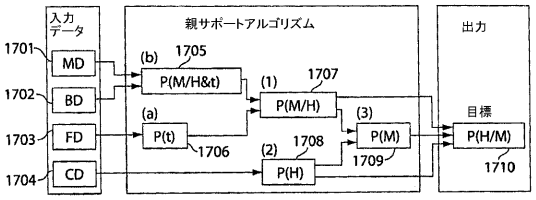

Fig. 17

【図１８】

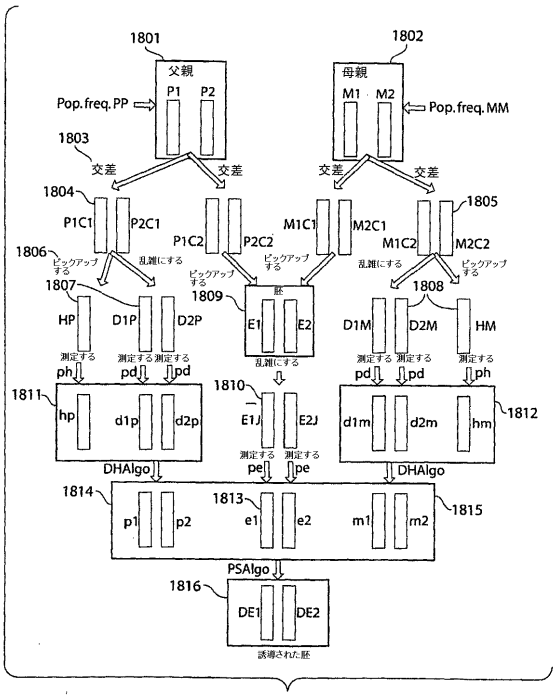

Fig. 18

【図１９】

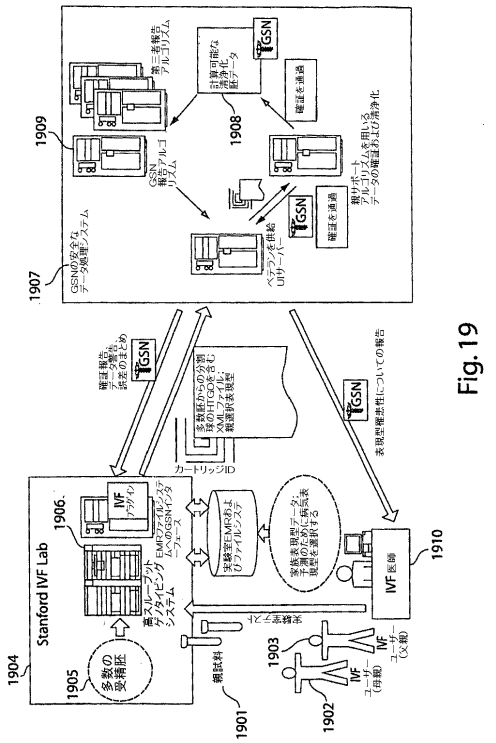

Fig. 19

【図２０】

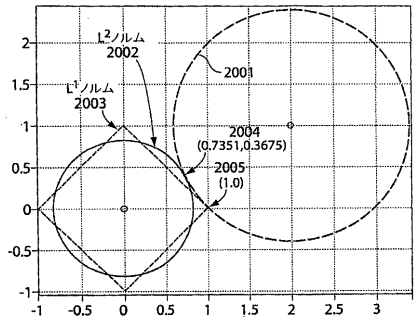

Fig. 20

【図２１】

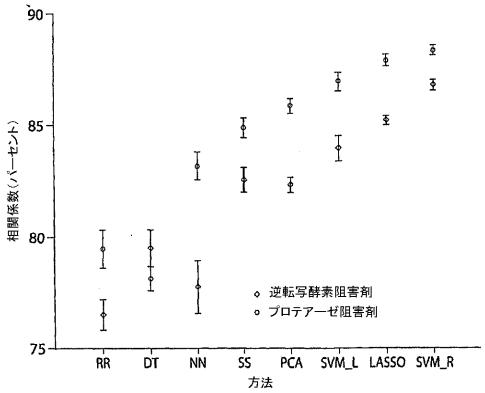

Fig. 21

【図２２】

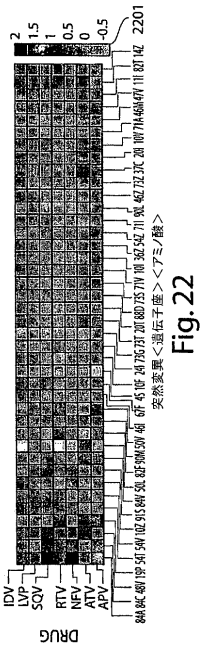

Fig. 22

【図２３】

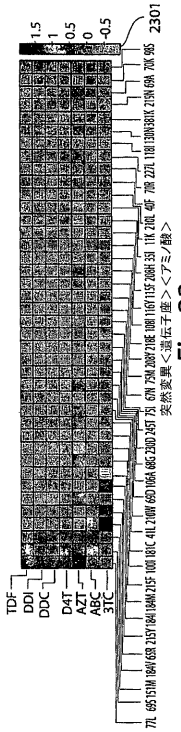

Fig. 23

【図２４】

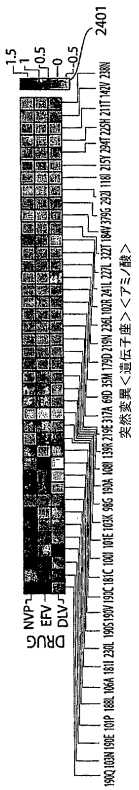

Fig. 24

【図２５】

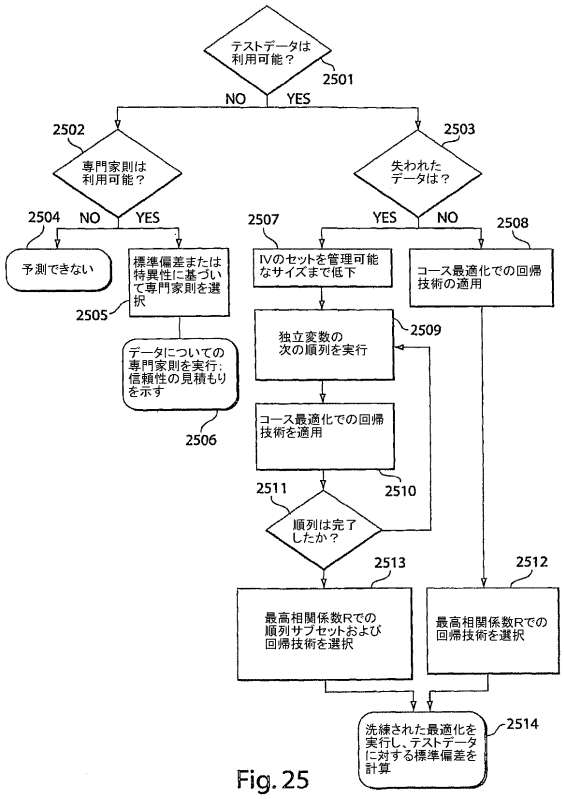

Fig. 25

【図２６】

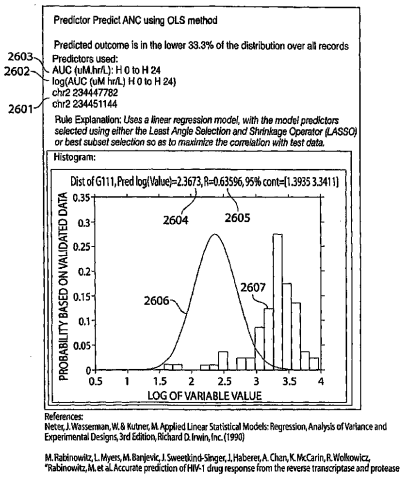

Fig. 26

【図２７】

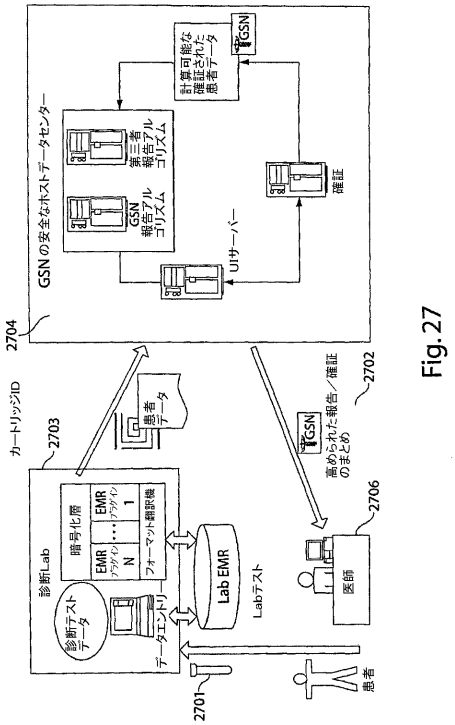

Fig. 27

【図２８】

Enhanced Colorectal Cancer Treatment Report

GSN GENESECURITY NETWORK

NAME: R.E. Mission	Report Date	March 26, 2006
MRN: 1122334455	Clinical Indication:	
DOB: 08/08/1958	ICD-9 153.9	Malignant neoplasm, colon
	Staging of Colorectal	
Sex: Male	Carcinoma	Dukes C; TNM Stage IIIA
Therapy	Irinotecan	Q3 Week Dosing Regimen
Date of therapy	March 25, 2006	
Other pertinent past medical history:		
Abnormal bilirubin glucuronidation (Gilbert's)		
Significant drug-drug interactions with patient current medications	Effect on irinotecan levels	
Phenytoin	↓	
Paroxetine	↑	
Ketoconazole	CONTRAINDICATED	
Laboratory and clinical EMR data		
Data Sources:		UGT1A1 Assay - 3/26/2006
Quest Labs	3/20/2006	Heterozygous for UGT1A1 *28
EMR Data	3/26/2006	AUC_{CPT-11} (ng • h/mL) 5582.0
		AUC_{SN-38} (ng • h/mL) 126.6
White blood cell count	3.0 K/uL	AUC_{SN-38G} (ng • h/mL) 1,251.3
Absolute neutrophil count	3000/mm³	ANC Prediction:
Neutropenia grade	None	
Hemoglobin	12 g/dL	
Platelet count	150 K/uL	
Hyperbilirubinemia	1.2 mg/dL	
Diarrhea grade	grade 2 (4-6 stools/day)	Grade 2 Neutropenia

SUMMARY RECOMMENDATIONS	RATIONALE
Pre-Treatment	
Consider reducing starting dose to 300 mg/m2	Gilbert's Syndrome
Discontinue ketoconazole 1 week prior to therapy, resume 2 days after therapy if indicated	Ketoconazole <--> Irinotecan interaction
Consider alternative regimens not including irinotecan	*28/*28 genotype
Day 1 Post Treatment Predictions	
Decrease dose by 50 mg/m2	Grade2 neutropenia (1000-1499/mm3)
Consider addition of colony stimulating factor	Grade2 neutropenia (1000-1499/mm3)
Omit dose until resolved to baseline, then reduce to 250 mg/m2	Grade2 diarrhea (4-6 stools/day)

Fig. 28

フロントページの続き

(31)優先権主張番号 60/754,396
(32)優先日 平成17年12月29日(2005.12.29)
(33)優先権主張国 米国(US)
(31)優先権主張番号 60/774,976
(32)優先日 平成18年2月21日(2006.2.21)
(33)優先権主張国 米国(US)
(31)優先権主張番号 60/789,506
(32)優先日 平成18年4月4日(2006.4.4)
(33)優先権主張国 米国(US)
(31)優先権主張番号 60/817,741
(32)優先日 平成18年6月30日(2006.6.30)
(33)優先権主張国 米国(US)
(31)優先権主張番号 11/496,982
(32)優先日 平成18年7月31日(2006.7.31)
(33)優先権主張国 米国(US)
(31)優先権主張番号 60/846,610
(32)優先日 平成18年9月22日(2006.9.22)
(33)優先権主張国 米国(US)

(74)代理人 100117019
弁理士 渡辺 陽一
(74)代理人 100150810
弁理士 武居 良太郎
(74)代理人 100141977
弁理士 中島 勝
(72)発明者 マシュー ラビノビッツ
アメリカ合衆国 カリフォルニア ９４０２８, ポルトラ バレー, ヘイフィールズ ロード ８０
(72)発明者 ミレナ バンジェビック
アメリカ合衆国 ニューヨーク １０００９, ニューヨーク, イースト １４ティーエイチ ストリート ４４５, アパートメント ２エフ
(72)発明者 ザカリー ポール デムコ
アメリカ合衆国 マサチューセッツ ０２１４４, サマービル, セント ジェームス アベニュー ３１ビー
(72)発明者 デイビッド スコット ジョンソン
アメリカ合衆国 カリフォルニア ９４０２８, ポルトラ バレー, ゴールデン オーク ドライブ ４２５

審査官 松原 寛子

(56)参考文献 特表２００４－５０２４６６（ＪＰ，Ａ）
特表２００４－５３３２４３（ＪＰ，Ａ）
厚生省精神・神経疾患研究委託費 筋ジストロフィーの臨床・疫学及び遺伝相談に関する研究 平成６・７年度
日本不妊学会雑誌，日本，２００１年，第46巻、第1号，p.43-46
遺伝学的検査に関するガイドライン，日本，遺伝医学関連学会（日本遺伝カウンセリング学会、日本，２００３年
Nature Reviews Genetics，２００４年，Vol.5，p.251-261

Ecology Letters, ２００４年, Vol.7, p.509-520

(58)調査した分野(Int.Cl., ＤＢ名)
Ｃ１２Ｎ　１５／１１
Ｃ１２Ｑ　　１／６８３７
Ｃ１２Ｑ　　１／６８７４
Ｇ０６Ｆ　１９／２２
ＪＳＴＰｌｕｓ／ＪＭＥＤＰｌｕｓ／ＪＳＴ７５８０（ＪＤｒｅａｍＩＩＩ）
ＣＡｐｌｕｓ／ＭＥＤＬＩＮＥ／ＥＭＢＡＳＥ／ＢＩＯＳＩＳ（ＳＴＮ）