(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号
特開2003－52383
(P2003－52383A)
(43)公開日 平成15年2月25日(2003.2.25)

(51)Int.Cl.[7]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｃ１２Ｎ 15/09	ZNA		Ｃ１２Ｑ 1/68	Ａ	
Ｃ１２Ｑ 1/68			Ｇ０１Ｎ 33/53	Ｍ	
Ｇ０１Ｎ 33/53			33/566		
33/566			Ｃ１２Ｎ 15/00	ZNA Ａ	

審査請求　未請求　請求項の数 49ＯＬ（全140数）

(21)出願番号　特願2002－99196(P2002－99196)

(22)出願日　平成14年4月1日(2002.4.1)

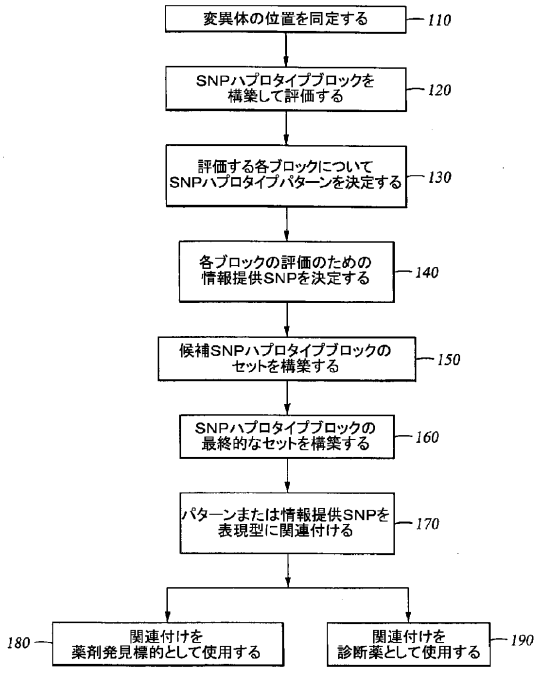
(31)優先権主張番号 60/280530
(32)優先日　平成13年3月30日(2001.3.30)
(33)優先権主張国　米国(US)
(31)優先権主張番号 60/313264
(32)優先日　平成13年8月17日(2001.8.17)
(33)優先権主張国　米国(US)
(31)優先権主張番号 60/327006
(32)優先日　平成13年10月5日(2001.10.5)
(33)優先権主張国　米国(US)

(71)出願人　502116106
パーレジェン　サイエンシーズ　インコーポレイテッド
アメリカ合衆国，カリフォルニア州，マウンテン　ヴュー，スティアリン　コート　2021
(72)発明者　ニラ　パティル
アメリカ合衆国，カリフォルニア州，マウンテン　ヴュー，スティアリン　コート　2021　パーレジェン　サイエンシーズ　インコーポレイテッド内
(74)代理人　100094318
弁理士　山田　行一　（外１名）

最終頁に続く

(54)【発明の名称】 ゲノム分析方法

(57)【要約】　（修正有）
【課題】ヒトゲノム分析方法の提供。
【解決手段】本発明は、ヒトゲノムにおいて生じる変異を同定するための方法、ならびにこれらの変異を疾患および薬物応答の遺伝的根拠に関連付ける方法に関する。特に本発明は、個々のＳＮＰを同定し、ＳＮＰハプロタイプブロックおよびパターンを決定し、そして更に、該ＳＮＰハプロタイプブロックおよびパターンを用いて疾患および薬物応答の遺伝的根拠を詳細に分析することに関する。本発明の方法は、全ゲノムの分析に有用である。



1

【特許請求の範囲】
【請求項１】　ＳＮＰハプロタイプパターンを選択するための方法であって、
複数の異なる起源から実質的に同じ核酸鎖を単離して分析する工程、
各核酸鎖の中の２以上のＳＮＰ位置を決定する工程、
ＳＮＰハプロタイプブロックを形成する、前記核酸鎖の中の連鎖したＳＮＰ位置を同定する工程、
孤立したＳＮＰハプロタイプブロックを同定する工程、
各ＳＮＰハプロタイプブロックおよび孤立したＳＮＰハプロタイプブロック中に生じるＳＮＰハプロタイプパターンを同定する工程、ならびに異なる起源源から得た前記実質的に同じ核酸鎖の少なくとも２つにおいて生じる各同定されたＳＮＰハプロタイプパターンを選択する工程、を含む、上記方法。
【請求項２】　前記第１の同定工程が、欲張りアルゴリズムまたは最短路アルゴリズムにより決定される、請求項１記載の方法。
【請求項３】　前記ＳＮＰハプロタイプブロックが重複しない、請求項１記載の方法。
【請求項４】　前記実質的に同じ核酸鎖が少なくとも約１０～約１００個の異なる起源に由来する、請求項１記載の方法。
【請求項５】　前記実質的に同じ核酸鎖が少なくとも約１６個の異なる起源に由来する、請求項４記載の方法。
【請求項６】　前記実質的に同じ核酸鎖が少なくとも約２５個の異なる起源に由来する、請求項５記載の方法。
【請求項７】　前記実質的に同じ核酸鎖が少なくとも約５０個の異なる起源に由来する、請求項６記載の方法。
【請求項８】　前記実質的に同じ核酸鎖がゲノムＤＮＡ鎖である、請求項１記載の方法。
【請求項９】　ある生物に由来するゲノムＤＮＡの少なくとも１０％が単離および分析される、請求項１記載の方法。
【請求項１０】　前記実質的に同じ核酸鎖から得た少なくとも1×10^{8}個の塩基が単離および分析される、請求項１記載の方法。
【請求項１１】　前記実質的に同じ核酸鎖から選択された反復領域は分析されない、請求項１記載の方法。
【請求項１２】　前記決定工程の後に、前記複数の同じ核酸鎖の中に１回しか生じないＳＮＰ位置を同定する工程、および前記１回しか生じないＳＮＰ位置を分析から除外する工程、をさらに含む、請求項１記載の方法。
【請求項１３】　前記実質的に同じ核酸鎖の中で最も頻繁に生じるＳＮＰハプロタイプパターンを選択する工程、
前記実質的に同じ核酸鎖の中で次に最も頻繁に生じるＳＮＰハプロタイプパターンを選択する工程、および前記選択されたＳＮＰハプロタイプパターンが前記実質的に同じ核酸鎖の一部分を同定するまで前記第２選択ステップを繰返す工程、をさらに含む、請求項１記載の方法。

2

【請求項１４】　前記一部分が前記実質的に同じ核酸鎖の約７０％～９９％である、請求項１３記載の方法。
【請求項１５】　前記一部分が前記実質的に同じ核酸鎖の少なくとも約８０％である、請求項１４記載の方法。
【請求項１６】　約３個以下のＳＮＰハプロタイプパターンが選択される、請求項１３記載の方法。
【請求項１７】　データ分析のためのＳＮＰハプロタイプブロックのデータセットを選択するための方法であって、
情報提供性についてＳＮＰハプロタイプブロックを比較する工程、
高い情報提供性を有する第１ＳＮＰハプロタイプブロックを選択する工程、
前記第１ＳＮＰハプロタイプブロックを前記データセットに追加する工程、
高い情報提供性を有する第２のＳＮＰハプロタイプブロックを選択する工程、
前記第２の選択されたＳＮＰハプロタイプブロックを前記データセットに追加する工程、および核酸鎖の目的の領域がカバーされるまで前記選択および追加工程を繰返す工程、を含む、上記選択方法。
【請求項１８】　前記選択されたＳＮＰハプロタイプブロックが重複しない、請求項１７記載の方法。
【請求項１９】　欲張りアルゴリズムを用いて前記選択工程を実行する、請求項１７記載の方法。
【請求項２０】　ＳＮＰハプロタイプパターンの中の情報提供ＳＮＰを決定するための方法であって、
ＳＮＰハプロタイプブロックのＳＮＰハプロタイプパターンを決定する工程、前記ＳＮＰハプロタイプブロックの中の目的の各ＳＮＰハプロタイプパターンを、前記ＳＮＰハプロタイプブロックの中の目的の他のＳＮＰハプロタイプパターンと比較する工程、および目的の第１ＳＮＰハプロタイプパターンの中の少なくとも１つのＳＮＰを選択する工程であって、前記ＳＮＰハプロタイプブロックの中の他の目的のＳＮＰハプロタイプパターンからこのような目的の第１ＳＮＰハプロタイプパターンを識別する工程、を含み、前記選択された少なくとも１つのＳＮＰが、前記ＳＮＰハプロタイプブロックの中の前記第１ＳＮＰハプロタイプパターンの情報提供ＳＮＰである、上記方法。
【請求項２１】　ＳＮＰハプロタイプブロックの中のＳＮＰハプロタイプパターンの一部分を識別するために十分な数の情報提供ＳＮＰが選択されるまで前記選択工程を繰返す工程を更に含む、請求項２０記載の方法。
【請求項２２】　ＳＮＰハプロタイプパターンの前記選択された一部分が、前記ＳＮＰハプロタイプブロックの中のＳＮＰハプロタイプパターンの約７０％～約９９％である、請求項２１記載の方法。
【請求項２３】　ＳＮＰハプロタイプパターンの前記選

択された一部分が目的の疾患の同定を可能とする、請求項２１記載の方法。

【請求項２４】　ＳＮＰハプロタイプブロックの情報提供性を決定するための方法であって、
前記ＳＮＰハプロタイプブロックの中のＳＮＰ位置の数を決定する工程、
前記ＳＮＰハプロタイプブロックの中の目的のＳＮＰハプロタイプパターンを識別するのに必要な情報提供ＳＮＰの数を決定する工程、および前記ＳＮＰ位置の数を前記情報提供ＳＮＰの数で割って商を生成する工程、を含み、前記商が前記ＳＮＰハプロタイプブロックの前記情報提供性である、上記方法。

【請求項２５】　ＳＮＰハプロタイプブロックの情報提供性を決定する方法であって、
前記ＳＮＰハプロタイプブロックの中のＳＮＰ位置の数を決定する工程、および前記ＳＮＰハプロタイプブロックの中の目的のＳＮＰハプロタイプパターンを互いに区別するのに必要な情報提供ＳＮＰの数を決定する工程、を含み、目的のＳＮＰハプロタイプパターンを区別するのに必要な前記情報提供ＳＮＰの数が、前記ＳＮＰハプロタイプブロックの前記情報提供性である、上記方法。

【請求項２６】　疾患関連遺伝子の配列または該遺伝子座の位置の事前知識もなく、該疾患関連遺伝子を決定するための方法であって、対照集団の少なくとも１６個体からＳＮＰハプロタイプパターンを決定する工程、
疾患に罹った集団の個体からＳＮＰハプロタイプパターンを決定する工程、および前記対照集団の前記ＳＮＰハプロタイプパターンの頻度を、前記疾患に罹った集団の前記ＳＮＰハプロタイプパターンの頻度と比較する工程を含み、前記頻度の差が疾患関連遺伝子座を示す、上記方法。

【請求項２７】　前記ＳＮＰハプロタイプパターンが、対照集団の少なくとも５０個体において決定される、請求項２６記載の方法。

【請求項２８】　前記集団からの前記ＳＮＰハプロタイプパターンが情報提供ＳＮＰを用いて決定される、請求項２６記載の方法。

【請求項２９】　複数の全ゲノムを用いてＳＮＰハプロタイプブロックのマップを作製する方法であって、
前記全ゲノムの少なくとも約１０％において見られるＳＮＰをＳＮＰハプロタイプブロックの中に並べる工程を含む、上記方法。

【請求項３０】　ＳＮＰハプロタイプパターンと目的の表現型特性とを関連付ける方法であって、
本発明の方法によってＳＮＰハプロタイプパターンのベースラインを作製する工程、
目的の共通表現型特性を有する集団から得た全ゲノムＤＮＡをプールする工程、および前記目的の表現型特性に関連付けられたＳＮＰハプロタイプパターンを同定する工程、を含む、上記方法。

【請求項３１】　前記作製工程および前記同定工程に情報提供ＳＮＰが使用される、請求項３０記載の方法。

【請求項３２】　請求項２０記載の情報提供ＳＮＰを同定する工程を含み、前記情報提供ＳＮＰが関連付けに基づく診断マーカーである、診断マーカーの同定方法。

【請求項３３】　薬剤発見標的を同定するための方法であって、
ＳＮＰハプロタイプパターンを疾患と関連付ける工程、
前記関連付けられたＳＮＰハプロタイプパターンの染色体位置を同定する工程、
前記染色体位置と前記疾患との前記関連付けの性質を決定する工程、および前記疾患に関連付けられた染色体位置またはその染色体位置の発現産物を選択する工程、を含み、前記疾患に関連付けられた前記選択された染色体位置またはその染色体位置の発現産物が薬剤発見標的である、上記方法。

【請求項３４】　高度保存領域中の位置と遺伝子間領域中の位置とを含む判断基準のセットに基づいた薬剤発見標的に対して、前記関連付けられた染色体位置が優先させられる、請求項３３記載の方法。

【請求項３５】　情報提供ＳＮＰが前記関連付け工程で使用される、請求項３３記載の方法。

【請求項３６】　個体のＳＮＰハプロタイプパターンを決定する方法であって、少なくとも１つの情報提供ＳＮＰを分析する工程を含む、上記方法。

【請求項３７】　ある種または種の部分集団のＳＮＰハプロタイプパターンを画定するための方法であって、
前記種の多数の生物のゲノム中に存在するＳＮＰを同定する工程、
曖昧位置の数が少ないＳＮＰハプロタイプパターンを反復的に選択することにより前記ＳＮＰをＳＮＰハプロタイプブロック中に並べる工程、を含む上記方法。

【請求項３８】　多数の生物のゲノムから得たＳＮＰハプロタイプブロックを含むデータベースであって、前記データベースが少なくとも１つの情報提供ＳＮＰを同定し、前記データベースがコンピュータが判読可能な媒体上にある、上記データベース。

【請求項３９】　１以上の特定の表現型特性に関連するものとして同定されるＳＮＰハプロタイプパターンを含むコンピュータが判読可能な媒体上のデータベース。

【請求項４０】　１以上の特定の表現型特性に関連するものとして同定される情報提供ＳＮＰを含むコンピュータが判読可能な媒体上のデータベース。

【請求項４１】　環境要因、他の遺伝的要因、関連する要因からなる群より選択される１以上の要因についての情報をさらに含み、前記要因は生化学的マーカー、行動および／または他の多型を含むがこれらに限定されず、前記多型は例えば低頻度ＳＮＰ、繰返し、挿入および欠失を含むがこれらに限定されない、請求項３８、３９または４０に記載のデータベース。

【請求項４２】　疾患、疾患に対する罹患性、または治療応答の診断キットであって、患者から得たゲノムＤＮＡのサンプル中のＳＮＰハプロタイプパターンまたは情報提供ＳＮＰの存在または不在を検出するための手段と、前記ＳＮＰハプロタイプパターンまたは情報提供ＳＮＰと１以上の特異的表現型特性との関連付けのデータセットと、をコンピュータが判読可能な媒体上に含む、上記診断キット。

【請求項４３】　少なくとも１つの情報提供ＳＮＰを含む単離核酸であって、前記情報提供ＳＮＰは本発明の方法に従って決定されたＳＮＰハプロタイプパターンを示し、前記情報提供ＳＮＰが表現型特性に関連付けられる、上記単離核酸。

【請求項４４】　複数の個体中の遺伝学的変異を同定する工程、
前記遺伝学的変異のうちの少なくともある他の変異と共に生じる個体中の前記遺伝学的変異の少なくとも幾つかを同定する工程、およびある表現型状態と相関する、前記遺伝学的変異のうちの少なくともある他の変異と共に生じる前記遺伝学的変異の全てではなく幾つかを用いる工程、を含む、方法。

【請求項４５】　ある生物の配列を決定する工程、
前記生物の他の個体を前記配列の変異体について走査する工程、
第１グループにおいて前記変異体のうちの他の変異体と共に生じる前記変異体の幾つかを同定する工程、
第２グループにおいて前記変異体のうちの他の変異体と共に生じる前記変異体の幾つかを同定する工程、および前記第１グループおよび第２グループにおける前記変異体の全てではなく幾つかを用いて、前記グループをある表現型状態に関係付ける工程、を含む、方法。

【請求項４６】　ゲノム分析において有用なＳＮＰハプロタイプブロックを選択するための方法であって、
少なくとも約５つの異なる起源から実質的に同じＤＮＡ鎖を分析用に単離する工程、
少なくとも約５つの異なる起源から得た前記実質的に同じＤＮＡ鎖の各々から少なくとも約1×10^{6}塩基を分析する工程、
各ＤＮＡ鎖の中の２以上のＳＮＰ位置を決定する工程、
前記ＤＮＡ鎖の中の連鎖したＳＮＰ位置を同定する工程であって、前記連鎖したＳＮＰ位置がＳＮＰハプロタイプブロックを形成する工程、
各ＳＮＰハプロタイプ中において生じるＳＮＰハプロタイプパターンを同定する工程、および異なる起源から得た前記実質的に同じＤＮＡのいずれかにおいて生じる各同定されたＳＮＰハプロタイプパターンを選択する工程、を含む、上記方法。

【請求項４７】　薬理遺伝学に関連する遺伝子座の配列または位置についての事前知識も無く、前記薬理遺伝学に関連する遺伝子座を決定するための方法であって、前記方法が、
対照集団の少なくとも１６個体からＳＮＰハプロタイプパターンを決定する工程、
ある物質の投与に対して変わった反応を示す個体からＳＮＰハプロタイプパターンを決定する工程、および前記対照集団の前記ＳＮＰハプロタイプパターンの頻度を、ある物質の投与に対して変わった反応を示す前記個体の前記ＳＮＰハプロタイプパターンの頻度と比較する工程を含み、前記頻度の差が薬理遺伝学に関連する遺伝子座の位置を示す、上記方法。

【請求項４８】　前記ＳＮＰハプロタイプパターンが、対照集団の少なくとも５０個体において決定される、請求項４７記載の方法。

【請求項４９】　前記集団から得た前記ＳＮＰハプロタイプパターンが情報提供ＳＮＰを用いて決定される、請求項４７記載の方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】(関連出願に対する相互参照)
本発明は、２００１年3月３０日に出願された米国仮特許出願番号第６０／２８０，５３０号、２００１年8月１７日に出願された米国仮特許出願番号第６０／３１３，２６４号、２００１年１０月5日に出願された米国仮特許出願番号第６０／３２７，００６号（全て「ヒトＳＮＰハプロタイプの同定、情報提供ＳＮＰおよびその使用（Ｉｄｅｎｔｉｆｙｉｎｇ Ｈｕｍａｎ ＳＮＰ Ｈａｐｌｏｔｙｐｅｓ， Ｉｎｆｏｒｍａｔｉｖｅ ＳＮＰｓ ａｎｄ Ｕｓｅｓ Ｔｈｅｒｅｏｆ）」と題する）、および２００１年１１月２６日に出願された米国仮特許出願番号第６０／３３２，５５０号（「ゲノム分析方法（Ｍｅｔｈｏｄｓ ｆｏｒ Ｇｅｎｏｍｉｃ Ａｎａｌｙｓｉｓ）」と題する）を基に優先権主張を行う。これら全ての開示内容は本明細書中に参考として特に組み込まれるものとする。

【０００２】

【従来の技術】ヒト染色体を構成するＤＮＡは、体内の全てのタンパク質の産生を指揮するインストラクションを提供する。これらのタンパク質は、生命に不可欠な機能を発揮する。タンパク質をコードするＤＮＡの配列が変化すると、これによりコードされるタンパク質に変化または突然変異が生じて、細胞の正常な機能に影響を及ぼす。環境は疾患においてしばしば大きな役割を担うが、個体のＤＮＡにおける変化や突然変異は、感染性疾患、癌および自己免疫異常を含むほぼ全てのヒト疾患に直接関係する。さらに、遺伝学（特にヒト遺伝学）の知識により、多くの疾患は、幾つかの遺伝子もしくは遺伝子産物の複雑な相互作用から、または１つの遺伝子内で起こる任意の数の突然変異から生じるということが分かった。例えば、Ｉ型糖尿病およびＩＩ型糖尿病は、多数の遺伝子（各遺伝子は独自の突然変異パターンを有す

る）に関連付けられた。これに対し、嚢胞性線維症は、１つの遺伝子内における３００を超える様々な突然変異のいずれかにより引き起こされ得る。

【０００３】さらに、薬物応答（薬理遺伝学の分野）に関しては、ヒト遺伝学の知識は、個体間の差異の理解の範囲を限られたものとした。半世紀以上前、有害薬物応答は、血漿コリンエステラーゼおよびグルコース－６－リン酸デヒドロゲナーゼという２つの薬物代謝酵素におけるアミノ酸の変化と関連付けられた。それ以来、入念な遺伝学的分析により、３５を超える薬物代謝酵素、２５の薬物標的および５つの薬物トランスポーター（ｄｒｕｇ　ｔｒａｎｓｐｏｒｔｅｒ）の中の配列多型（変異）を薬物の効力または安全性の折衷レベルに関連付けた（ＥｖａｎｓおよびＲｅｌｌｉｎｇ，　Ｓｃｉｅｎｃｅ　２９６：　４８７－９１（１９９９））。診療所では、このような情報は、薬物毒性を防ぐために使用されている。例えば、患者は、６－メルカプトプリンまたはアザチオプリンの代謝を低下させるチオプリンメチルトランスフェラーゼ遺伝子の中の遺伝的差異について慣習的にスクリーニングされる。しかし現在までのところ、観察された薬物毒性のうちの僅かしか、薬理遺伝学マーカーのセットによって十分に説明されていない。毒性の問題よりもより一般的なのは、ある個体に対して安全且つ／もしくは効力があることが示された薬物が、他の個体においては十分な治療的効能を持たない、または予期しがたい副作用を持つことが判明した事例である。

【０００４】ヒトの遺伝子構成における変異の影響を理解する重要性に加え、他の非ヒト生物（特に病原体）の遺伝子構成における変異の影響を理解することは、これらのヒトへの影響またはヒトとの相互作用を理解する上で重要である。例えば、病原性細菌またはウイルスによる毒性因子の発現は、このような生物に接触したヒトにおける感染率および感染程度に多大な影響を及ぼす。さらに、実験動物（すなわちマウス、ラットなど）の遺伝子構成の詳細な理解もまた大きな価値がある。例えば、治療の評価のためのモデル系として使用される動物の遺伝子構成の変異を理解することは、これらの系を用いて得たテスト結果およびこれらをヒトに使用した場合の予測値を理解するために重要である。

【０００５】任意の２人のヒトはその遺伝子構成において９９．９％類似しているので、彼等のゲノムＤＮＡ配列の大部分は同一である。しかし、個体間ではＤＮＡ配列に変異がある。例えば、ＤＮＡの多数塩基ストレッチの欠失、ＤＮＡストレッチの挿入、および非コード領域における反復ＤＮＡエレメントの数の変異、ならびに「一塩基多型（ＳＮＰ）」と呼ばれるゲノム内の１つの窒素含有塩基位置における変化がある。ヒトＤＮＡ配列変異は、個体間で観察された差異（疾患への感受性を含む）の大部分を説明する。

【０００６】大部分のＳＮＰは稀なものではあるが、ヒト間のＤＮＡ配列の違いの大部分を説明する一般的なＳＮＰ（各ＳＮＰの頻度は１０～５０％）は５３０万個あると推定されている。このようなＳＮＰはヒトゲノム中に６００塩基対毎に１回存在する（ＫｒｕｇｌｙａｋおよびＮｉｃｋｅｒｓｏｎ，　Ｎａｔｕｒｅ　Ｇｅｎｅｔ．　２７：２３５（２００２））。物理的に近くに存在するこのようなＳＮＰのブロックを構成する対立遺伝子（変異体）はしばしば相関関係を有し、その結果、遺伝的変異性（ｇｅｎｅｔｉｃ　ｖａｒｉａｂｉｌｉｔｙ）は低下し、限られた数の「ＳＮＰハプロタイプ」（各々は１つの古来先祖の染色体からの遺伝継承を反映する：Ｆｕｌｌｅｒｔｏｎら、Ａｍ．　Ｊ．　Ｈｕｍ．　Ｇｅｎｅｔ．６７：８８１（２０００））を定義する。

【０００７】ヒトゲノムにおける局所的ハプロタイプ構造の複雑性、および個々のハプロタイプが広がっている距離は、殆ど定義されていない。異なる集団におけるヒトゲノムの異なるセグメントを調査する経験主義的な調査により、局所的ハプロタイプ構造において大きな変異性があることが分かった。これらの調査は、突然変異、組換え、選択、個体群の歴史および確率的事象の、ハプロタイプ構造への相互的な寄与が、予測不可能な形で変化することを示しており、その結果、あるハプロタイプはわずか数千塩基（ｋｂ）の長さしかないものとなり、またあるハプロタイプは１００ｋｂを超える長さのものとなる（Ａ．Ｇ．　Ｃｌａｒｋら、Ａｍ．　Ｊ．　Ｈｕｍ．　Ｇｅｎｅｔ．　６３：５９５（１９９８））。

【０００８】これらの知見は、ヒトゲノムのハプロタイプ構造（一般的ＳＮＰにより定義される）を包括的に説明するには、ヒトゲノムの沢山の独立したコピーにおけるＳＮＰの稠密なセットの経験的な分析を必要とすることを示唆している。このような全ゲノム分析は、かなり精密な遺伝子マッピングを提供し、および具体的な連鎖領域の位置を正確に示すであろう。しかし本発明以前は、適度なサイズの個体群の各個体の３，０００，０００　を超えるＳＮＰの遺伝子タイピングを行わなければならないこと及びこれにかかるコストにより、この試みは実行不能であった。本発明は、様々な用途の中でも特に、ＳＮＰハプロタイプを用いた個体群の全ゲノム関連付け分析（ｗｈｏｌｅ－ｇｅｎｏｍｅ　ａｓｓｏｃｉａｔｉｏｎａｎａｌｙｓｉｓ）を可能とする。

【０００９】

【課題を解決するための手段】本発明は、ヒトゲノムにおいて生じる変異を同定するための方法、およびこれらの変異を、疾患への耐性、疾患への感受性または薬物応答などの表現型の遺伝的根拠（ｇｅｎｅｔｉｃ　ｂａｓｅ）に関連付けるための方法に関する。「疾患」とは、変化が望まれる、生物の任意の状態、体質および特性を含むが、これらに限定されない。例えば、状態とは、身体的、生理学的または心理的なものであってもよく、ま

た症状性のものであっても無症状性のものであってもよい。この方法は、変異体の同定、ＳＮＰの同定、ＳＮＰハプロタイプブロックの決定、ＳＮＰハプロタイプパターンの決定、およびさらに各パターンの情報提供ＳＮＰの同定（遺伝子データの圧縮を提供する）を可能とする。

【００１０】このように、本発明の１つの態様は、データ分析に有用なＳＮＰハプロタイプパターンの選択方法を提供する。このような選択は、複数の個体から実質的に同じ（相同な）核酸鎖を単離し、各核酸鎖の中のＳＮＰ位置を決定し、核酸鎖の中の連鎖したＳＮＰ位置を同定し（連鎖したＳＮＰ位置はＳＮＰハプロタイプブロックを形成する）、孤立した（ｉｓｏｌａｔｅ）ＳＮＰハプロタイプブロックを同定し、各ハプロタイプブロック中に生じるＳＮＰハプロタイプパターンを同定し、および実質的に同じ複数の核酸鎖のうちの少なくとも２つにおいて生じる同定されたＳＮＰハプロタイプパターンを選択することにより、行うことができる。１つの好適な実施形態において、少なくとも１０の異なる個体または起源から得た核酸鎖が用いられる。より好適な実施形態において、少なくとも１６の異なる起源から得た核酸鎖が用いられる。さらに好適な実施形態において、少なくとも２５の異なる起源から得た核酸鎖が用いられ、より更に好適な実施形態において、少なくとも５０の異なる起源から得た核酸鎖が用いられる。さらに、より好適な実施形態は、少なくとも約１００の異なる起源から得た核酸鎖の中のＳＮＰ位置を決定する。さらに、この方法は、実質的に同じ複数の核酸鎖中で最も頻繁に生じるＳＮＰハプロタイプを選択する工程、これらの実質的に同じ複数の核酸鎖中で次に最も頻繁に生じるＳＮＰハプロタイプを選択する工程、および選択されたＳＮＰハプロタイプパターンがこれらの実質的に同じ複数の核酸鎖の目的の一部分を同定するまでこの選択工程を繰返す工程、をさらに含む。好適な実施形態において、目的の一部分とは、該実質的に同じ核酸鎖の７０％～９９％であり、より好適な実施形態において、目的の一部分は該実質的に同じ核酸鎖の約８０％である。あるいは、ＳＮＰハプロタイプパターンの選択を、１つのＳＮＰハプロタイプブロックにつき約３個以下のＳＮＰハプロタイプパターンに限定したい場合もある。

【００１１】さらに本発明は、データ分析のためにＳＮＰハプロタイプブロックのデータセットを選択するための方法であって、情報提供性（ｉｎｆｏｒｍａｔｉｖｅｎｅｓｓ）についてＳＮＰハプロタイプブロックを比較する工程、高い情報提供性を有する第１のＳＮＰハプロタイプブロックを選択する工程、該第１ＳＮＰハプロタイプブロックを該データセットに追加する工程、高い情報提供性を有する第２のＳＮＰハプロタイプブロックを選択する工程、選択された該第２ＳＮＰハプロタイプブロックを該データセットに追加する工程、およびＤＮＡ鎖の目的の領域がカバーされるまでこの選択工程および追加工程を繰返す工程を含むことを特徴とする上記方法を提供する。好適な実施形態において、選択されるＳＮＰハプロタイプブロックは重複しない。

【００１２】本発明はさらに、ＳＮＰハプロタイプパターンにおける少なくとも１つの情報提供ＳＮＰを決定するための方法であって、まず、あるＳＮＰハプロタイプブロックのＳＮＰハプロタイプパターンを決定し、次にそのＳＮＰハプロタイプブロックの中の目的の各ＳＮＰハプロタイプパターンをそのＳＮＰハプロタイプブロックの中の目的の他のＳＮＰハプロタイプパターンと比較し、そして該ＳＮＰハプロタイプブロックの中のこの目的のＳＮＰハプロタイプパターンを他のＳＮＰハプロタイプパターンと区別する各ＳＮＰハプロタイプパターンの中の少なくとも１つのＳＮＰを選択する上記方法を提供する。選択された１以上のＳＮＰは、そのＳＮＰハプロタイプパターンの情報提供ＳＮＰである。

【００１３】また本発明は、ゲノム領域の高速スキャニングを可能とし、疾患に関連する遺伝子座または薬理遺伝学に関係する遺伝子座の配列または位置についての事前知識無しで、このような疾患に関連する遺伝子座または薬理遺伝学的に関係する遺伝子座を決定するための方法を提供する。これは、対照集団の中の個体からＳＮＰハプロタイプパターンを決定した後、実験集団の中の個体（例えば疾患に罹った集団の中の個体または薬物を投与したときに特定の反応を示す個体）からＳＮＰハプロタイプパターンを決定することによって行うことができる。対照集団のＳＮＰハプロタイプパターンの頻度を、実験集団のＳＮＰハプロタイプパターンの頻度と比較する。これらの頻度における差は、疾患に関連する遺伝子座または薬理遺伝学的に関連する遺伝子剤の位置を示す。

【００１４】本発明の他の態様は、ＳＮＰハプロタイプパターンと目的の表現型特徴とを関連付ける方法であって、本発明の方法によって対照個体のＳＮＰハプロタイプパターンのベースラインを作製し、目的の共通の表現型特徴を有する臨床集団から全ゲノムＤＮＡをプールし、および目的の表現型特徴に関係するＳＮＰハプロタイプパターンを同定する、上記方法を提供する。このように、本発明は、表現型に関係する複数のハプロタイプブロックを同定するためのゲノムスキャニングを可能とし、これは特に多遺伝子性の特徴（ｐｏｌｙｇｅｎｉｃ　ｔｒａｉｔ）を調査する際に有用である。

【００１５】また本発明は、薬剤発見標的（ｄｒｕｇ　ｄｉｓｃｏｖｅｒｙ　ｔａｒｇｅｔ）を同定するための方法であって、ＳＮＰハプロタイプパターンを疾患と関連付け、関連付けられたＳＮＰハプロタイプパターンの染色体位置を同定し、染色体位置と前記疾患との関係の性質を同定し、およびその染色体位置の遺伝子または遺伝子産物を薬剤発見標的として用いることを特徴とする

上記方法を提供する。

【００１６】添付の図面は本明細書の一部をなし、本発明のある態様をさらに説明するために含まれる。本発明は、本明細書中に提供される具体的な実施形態の詳細な説明と一緒に、これらの図面の１以上を参照することによってより理解を深めることができる。

【００１７】本発明は、ヒトゲノムにおいて生じる変異を同定し、これらの変異を疾患の遺伝的根拠および薬物応答に関連付けるための方法に関する。特に本発明は、個々のＳＮＰの同定、ＳＮＰハプロタイプブロックおよびＳＮＰハプロタイプパターンの決定、ならびにさらに、疾患の遺伝的根拠および薬物応答を詳細に分析するためのＳＮＰハプロタイプブロックおよびＳＮＰハプロタイプパターンの使用に関する。本発明の方法は、全ゲノムの分析に有用である。

【００１８】

【発明の実施の形態】当業者であれば、本発明の範囲および精神を逸脱することなく、本願に開示された発明に対して様々な実施形態および修正を実施することができることは自明であろう。本明細書中に記載された全ての公表文献は、本発明に関して使用され得る試薬、方法論および概念を説明および開示するために引用される。本明細書中に記載されるいかなる文献も、これらの参考文献が本明細書中に記載された発明の先行技術であるということを認めるものではない。

【００１９】本明細書において、特に数が指定されていない場合は１以上のものを指すものとする。特許請求の範囲において「含む」という用語に関連して記載された物質および事柄について特に数を明記していない限り、１以上の物質および事柄を含むことを意味する。本明細書中において「他の」とは少なくとも２つ目以降の物質または事柄を指す。

【００２０】本明細書において、「異なる起源」という用語が使用される場合、この用語は異なる生物から得たＤＮＡ鎖が異なる起源に由来する事実を指す。さらに、１つの生物のゲノム中の各ＤＮＡ鎖は異なる起源に由来する。二倍体生物において、個々の生物のゲノムは、実質的に同じＤＮＡ鎖の複数の対からなるセットで構成される。つまり、１つの個体は２つの異なる起源に由来する実質的に同じＤＮＡ鎖を有する（その対の一方のＤＮＡ鎖は母方起源に由来し、その対の他方のＤＮＡ鎖は父方起源に由来する）。２以上の核酸配列（例えば２以上のＤＮＡ鎖）は、これらがヌクレオチドレベルで少なくとも約７０％、好ましくは約７５％、より好ましくは約８０％、さらに好ましくは約８５％、もっと好ましくは約９０％、よりさらに好ましくは約９５％の配列同一性を示す場合、実質的に同じであると考えられる。またもっとさらに好ましくは、ヌクレオチド配列は、これらがヌクレオチドレベルで少なくとも約９８％の配列同一性を示す場合、実質的に同じであるとみなされる。２以上の核酸配列間に関する配列同一性の程度は、それらの核酸の宿主起源によって異なる。例えば、同じ種の比較を見るときは９５％を超える配列同一性が適当であるが、種間比較を行うときには７０％以下の配列同一性が適当である。もちろん、本明細書中においてＤＮＡについて言及する場合、このような言及はアンプリコン、ＲＮＡ転写体、核酸模倣体等のＤＮＡ誘導体を含み得る。

【００２１】本明細書で使用される「個体」とは、単一の動物、ヒト、昆虫、細菌等の特定の単一生物を指す。

【００２２】本明細書で使用されるＳＮＰハプロタイプブロックの「情報提供性（ｉｎｆｏｒｍａｔｉｖｅｎｅｓｓ）」とは、あるＳＮＰハプロタイプブロックが遺伝子領域についての情報を提供する程度として定義される。

【００２３】本明細書で使用される「情報提供ＳＮＰ（ｉｎｆｏｒｍａｔｉｖｅ　ＳＮＰ）」という用語は、ＳＮＰハプロタイプブロックの中の１つのＳＮＰハプロタイプパターンを他のＳＮＰハプロタイプパターンと区別する傾向を示すＳＮＰまたは（２以上の）ＳＮＰからなるサブセットなどの遺伝子変異体（ｇｅｎｅｔｉｃｖａｒｉａｎｔ）を指す。

【００２４】本明細書で使用される「孤立したＳＮＰブロック（ｉｓｏｌａｔｅ　ＳＮＰｂｌｏｃｋ）」という用語は、１つのＳＮＰからなるＳＮＰハプロタイプブロックを指す。

【００２５】本明細書で使用される「連鎖不均衡（ｌｉｎｋａｇｅ　ｄｉｓｅｑｕｉｌｉｂｒｉｕｍ）」、「連鎖した（ｌｉｎｋｅｄ）」または「ＬＤ」という用語は、世代から世代へと一緒に受け継がれる傾向がある遺伝子座、例えば非無作為に継承される遺伝子座等を指す。

【００２６】本明細書で使用される「シングルトンＳＮＰハプロタイプ（ｓｉｎｇｌｅｔｏｎ　ＳＮＰ　ｈａｐｌｏｔｙｐｅ）」または「シングルトンＳＮＰ」という用語は、その集団のある一定の割合未満で生じる特定のＳＮＰ対立遺伝子または変異体を指す。

【００２７】本明細書で使用される「ＳＮＰ」または「単一ヌクレオチド多型」という用語は、個体間の遺伝子変異（例えば生物のＤＮＡの中の変異性の（ｖａｒｉａｂｌｅ）単一の窒素含有塩基位置など）を指す。本明細書中で使用される「ＳＮＰｓ」とは複数のＳＮＰである。もちろん、本明細書中でＤＮＡについて言及する場合、このような言及はアンプリコンやＲＮＡ転写体等のＤＮＡ誘導体を含み得る。

【００２８】本明細書で使用される「ＳＮＰハプロタイプブロック」という用語は、別々に組換えが起こらないと思われる変異体もしくはＳＮＰ位置のグループであって、変異体もしくはＳＮＰのブロックの中に一緒にグループ化され得る上記グループを意味する。

【００２９】本明細書で使用される「ＳＮＰハプロタイ

プパターン」という用語は、単一のＤＮＡ鎖の中のＳＮＰハプロタイプブロック中のＳＮＰの遺伝子型のセットを指す。

【００３０】本明細書で使用される「ＳＮＰ位置」という用語は、ＤＮＡ配列中のＳＮＰが生じる部位である。

【００３１】本明細書で使用される「ＳＮＰハプロタイプ配列」とは、少なくとも１つのＳＮＰ位置を含むＤＮＡ鎖中のＤＮＡ配列である。

【００３２】分析用核酸の調製

当業者に公知である任意の手法を用いて、分析用に核酸分子を調製することができる。好ましくはこのような手法によって、その核酸分子中の１以上の位置における１以上の変異の存在または不在を決定するのに十分な純度の核酸分子を産生する。このような手法は、例えばＳａｍｂｒｏｏｋら、ＭｏｌｅｃｕｌａｒＣｌｏｎｉｎｇ：Ａ Ｌａｂｏｒａｔｏｒｙ Ｍａｎｕａｌ（Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ Ｌａｂｏｒａｔｏｒｙ, Ｎｅｗ Ｙｏｒｋ）（１９８９）およびＡｕｓｕｂｅｌら、Ｃｕｒｒｅｎｔ Ｐｒｏｔｏｃｏｌｓ ｉｎ Ｍｏｌｅｃｕｌａｒ Ｂｉｏｌｏｇｙ（Ｊｏｈｎ Ｗｉｌｅｙ ａｎｄ Ｓｏｎｓ, Ｎｅｗ Ｙｏｒｋ）（１９９７）（本明細書中に参考として組み込まれる）に記載されている。

【００３３】目的の核酸が細胞内に存在する場合、まずその細胞の抽出物を調製した後、更なる工程（すなわちディファレンシャル沈降（ｄｉｆｆｅｒｅｎｔｉａｌ ｐｒｅｃｉｐｉｔａｔｉｏｎ）、カラムクロマトグラフィー、有機溶媒を用いた抽出など）を行って十分な純度の核酸調製物を得ることが必要である。抽出物は、当分野における標準的な手法、例えば細胞の化学的もしくは機械的溶解によって調製することができる。次に抽出物は、例えば濾過および／または遠心分離により、および／またはカオトロピック塩（例えばグアニジニウムイソチオシアナートや尿素等）を用いて、または有機溶媒（例えばフェノールおよび／またはＨＣＣｌ$_3$等）を用いて処理を行い、任意の汚染タンパク質および邪魔になる可能性のあるタンパク質を変性させることができる。カオトロピック塩を用いる場合、その核酸含有サンプルからカオトロピック塩を除去することが望ましい。これは、当分野における標準的な手法、例えば沈殿法、濾過法、サイズ排除クロマトグラフィー等を用いて行うことができる。

【００３４】幾つかの例では、細胞からメッセンジャーＲＮＡを抽出および分離することが望ましい場合がある。このような目的ための手法および材料は当業者に公知であり、固相支持体（ビーズやプラスチック表面など）に固定されたオリゴｄＴの使用などが挙げられる。好適な条件および材料は当業者に公知であり、ＳａｍｂｒｏｏｋおよびＡｕｓｕｂｅｌの上記参考文献に記載されている。例えば逆転写酵素を用いてｍＲＮＡをｃＤＮＡへと逆転写することが望ましい場合もある。好適な酵素は、例えばＩｎｖｉｔｒｏｇｅｎ（ＣａｌｓｂａｄＣＡ）から市販されている。その後、場合により、ｍＲＮＡから調製したｃＤＮＡを増幅してもよい。

【００３５】ハプロタイプパターンおよびハプロタイプブロックの調査に特に適した１つの方法は、体細胞の遺伝学的特質を利用して染色体を二倍体状態から一倍体状態へと分離するものである。１つの実施形態において、二倍体であるヒトリンパ芽球細胞系を同じく二倍体であるハムスター線維芽細胞系に、該ヒト染色体が該ハムスター細胞中に導入されて細胞ハイブリッドを産生するように、融合することができる。得られた細胞ハイブリッドを調べてどのヒト染色体が導入されたか、および（あれば）導入されたヒト染色体のどれが一倍体状態であるかを決定する（例えばＰａｔｔｅｒｓｏｎら、Ａｎｎａｌ. Ｎ. Ｙ. Ａｃａｄ. Ｏｆ Ｓｃｉｅｎｃｅｓ, ３９６：６９－８１（１９８２））。

【００３６】この手法の概略図を図１０に示す。図１０は、チミジンキナーゼ遺伝子中に突然変異を含む二倍体ハムスター繊維芽細胞系に融合された、チミジンキナーゼ遺伝子について野生型である二倍体ヒトリンパ芽球細胞系を示す。得られた細胞の部分集団において、ヒト染色体はハイブリッド中に存在する。ヒトＤＮＡ含有ハイブリッド細胞の選択は、ＨＡＴ培地（選択培地）を利用して行われる。野生型ヒトチミジンキナーゼ遺伝子を有するヒトＤＮＡ鎖が安定に組み込まれたハイブリッド細胞のみが、ＨＡＴを含む細胞培養培地中で増殖する。得られたハイブリッドのうち、幾つかのハイブリッドは、幾つかのヒト染色体の両方のコピーを含むか、ヒト染色体の１つのコピーのみを含むか、または特定のヒト染色体のコピーを持たない。例えば、ＡまたはＢ対立遺伝子を持つ遺伝子座を有するヒト第２２番染色体の場合、得られるハイブリッド細胞には、一方のヒト第２２番染色体変異体（例えば「Ａ」変異体）またはその一部を含むもの、他方のヒト第２２番染色体変異体（「Ｂ」変異体）またはその一部を含むもの、これら両方のヒト第２２染色体変異体またはこれらの一部を含むもの、またはヒト第２２番染色体変異体のどの部分も含まないものがある。図１０において、得られるハイブリッド集団のうち２つのみを示す。適当なハイブリッドを選択したら、これらのハイブリッドから得た核酸を例えば上記に記載した手法によって単離し、その後にＳＮＰ発見、そして本発明のハプロタイプブロックおよびハプロタイプパターンの分析にかけることができる。

【００３７】増幅手法

核酸内の１以上の変異の存在もしくは不在を決定する前に、１以上の目的の核酸を増幅することが望ましい場合がある。核酸増幅は目的の核酸配列のコピー数を増やす。当業者に公知である任意の増幅手法（例えばポリメラーゼ連鎖反応（ＰＣＲ）手法があるがこれに限定され

ない）を本発明と一緒に用いることができる。ＰＣＲは当業者に公知の材料および方法を用いて実施することができる。

【００３８】ＰＣＲ増幅は一般に、鋳型として核酸配列の一方の鎖を用いてその配列に相補的な多数の相補体を産生する。この鋳型を、その鋳型配列の一部に相補的な配列を有するプライマーにハイブリダイズさせ、ｄＮＴＰおよびポリメラーゼ酵素を含む好適な反応混合物に接触させる。プライマーはポリメラーゼ酵素により伸長され、もとの鋳型に相補的な核酸が産生される。

【００３９】二本鎖核酸分子の両方の鎖を増幅するためには、２つのプライマー（各々はその核酸のそれぞれ一方の鎖の一部に相補的な配列を有する）が用いられる。ポリメラーゼ酵素によるプライマーの伸長により、二本鎖核酸分子（各鎖は鋳型鎖および新しく合成された相補的な鎖を含む）が産生される。プライマーの配列は典型的には、該プライマーの各々の伸長方向が、その核酸分子の中の他方のプライマーがハイブリダイズする部位に向かうように選択される。

【００４０】核酸分子の鎖を例えば加熱により変性し、プロセスを繰返す。このとき、前回の工程で新しく合成された鎖は後続工程において鋳型として使われる。ＰＣＲ増幅プロトコールは、変性、ハイブリダイゼーション、および伸長反応を数回以上繰返して、十分な量の所望の核酸を産生するものである。

【００４１】ＰＣＲ法は、典型的には熱を用いて鎖の変性を行い、後のプライマーのハイブリダイゼーションを可能とするが、核酸をプライマーにハイブリダイズできるようにする任意の手段を用いることができる。このような手法は、物理的、化学的、または酵素による手段、例えばヘリカーゼの封入（ｉｎｃｌｕｓｉｏｎ　ｏｆｈｅｌｉｃａｓｅ）（Ｒａｄｄｉｎｇ，　Ａｎｎ．　Ｒｅｖ．　Ｇｅｎｅｔｉｃｓ　１６：４０５－４３６（１９８２）を参照されたい）や電気化学手段（国際特許出願公開番号ＷＯ９２／０４４７０号およびＷＯ９５／２５１７７号を参照されたい）を含むが、これらに限定されない。

【００４２】ＰＣＲにおける鋳型依存的なプライマーの伸長は、適当な塩、金属陰イオン、およびｐＨ緩衝系を含む反応培地中において、少なくとも４つのデオキシリボヌクエオチド３リン酸（典型的にはｄＡＴＰ、ｄＧＴＰ、ｄＣＴＰ、ｄＵＴＰおよびｄＴＴＰから選択される）の存在下で、ポリメラーゼ酵素により触媒される。好適なポリメラーゼ酵素は当業者に公知であり、天然起源からクローニングまたは単離することができ、これは、その酵素の天然形態または突然変異形態であってもよい。その酵素がプライマーを伸長する能力を維持する限り、これらの酵素は本発明の増幅反応に使用することができる。

【００４３】本発明の方法において使用される核酸は、後の工程での検出を容易とするために標識することができる。１以上の標識したヌクレオチド３リン酸および／または１以上の標識したプライマーを増幅配列中に組み込むことによって、増幅反応中に標識を行うことができる。核酸は、増幅後に、例えば１以上の検出可能基を共有結合させることにより標識してもよい。当業者に公知である任意の検出可能基、例えば蛍光基、リガンドおよび／または放射能基を用いることができる。好適な標識手法の例としては、末端デオキシヌクレオチジルトランスフェラーゼ（ＴｄＴ）酵素を用いて目的の核酸中に標識を含むヌクレオチドを導入する手法である。例えば、標識を含むヌクレオチド（好ましくはジデオキシヌクレオチド）を、標識しようとする核酸および該ヌクレオチドを導入するのに十分な量のＴｄＴと一緒にインキュベートする。好適なヌクレオチドは、ビオチン標識を結合させた、ジデオキシヌクレオチドすなわちｄｄＡＴＰ、ｄｄＧＴＰ、ｄｄＣＴＰ、ｄｄＴＴＰ等である。

【００４４】長い配列の増幅を最適化する手法を用いることができる。このような手法はゲノム配列に対して非常に有効である。２００１年９月５日に出願された継続中の米国特許出願番号６０／３１７，３１１号、２００２年１月９日に出願された「Ａｌｇｏｒｉｔｈｍｓ　ｆｏｒ　Ｓｅｌｅｃｔｉｏｎ　ｏｆ　Ｐｒｉｍｅｒ　Ｐａｉｒｓ」と題した代理人整理番号１０１１Ｎ－１（出願番号は未定）、および２００２年１月９日に出願された「Ｍｅｔｈｏｄｓ　ｆｏｒ　Ａｍｐｌｉｆｉｃａｔｉｏｎ　ｏｆ　Ｎｕｃｌｅｉｃ　Ａｃｉｄｓ」と題した代理人整理番号１０１１Ｎ１Ｄ１（出願番号は決定済み）に開示された方法は、本発明の方法で使用されるゲノムＤＮＡの増幅に特に適している。

【００４５】増幅された配列は、標識する前または後に、他の増幅後処理にかけることができる。例えば、幾つかのケースにおいて、ヌクレオチドアレイにハイブリダイズさせる前に増幅配列を断片化することが望ましい。核酸の断片化は、当分野で公知である物理的方法、化学的方法または酵素による方法によって実施することができる。好適な手法としては、該核酸を含む液体サンプルを狭い開口部に通過させて増幅配列をせん断力にかけるもの、またはＰＣＲ産物をヌクレアーゼ酵素で消化するもの等が挙げられるが、これらに限定されない。好適なヌクレアーゼ酵素の１つの例は、ＤＮアーゼＩである。増幅後、適当なサイズの断片を産生するように設定された一定時間の間、ＰＣＲ産物をヌクレアーゼの存在下でインキュベートする。断片サイズは、例えばヌクレアーゼの量やインキュベーション時間を増やしてより小さな断片を産生させたり、またはヌクレアーゼの量やインキュベーション時間を減らしてより大きな断片を産生させたりして、好きなように変えることができる。所望のサイズの断片を産生するための消化条件の調節は、当業者の能力の範囲内である。このように産生された断片

を上記のように標識する。

【００４６】ＳＮＰの検出方法（ＳＮＰ発見）

核酸中の１以上の変異の存在または不在の決定は、当業者に公知である任意の手法を用いて行うことができる。変異の正確な決定を可能とする任意の手法を用いることができる。好適な手法は、最小限のサンプルの扱いのみで複数の変異の迅速且つ正確な決定を可能とする。好適な手法の幾つかの例を以下に挙げる。

【００４７】幾つかのＤＮＡ配列決定法は当分野において周知且つ一般に利用可能であり、ゲノム中のＳＮＰの位置を決定するために用いることができる。例えばＳａｍｂｒｏｏｋら、Ｍｏｌｅｃｕｌａｒ Ｃｌｏｎｉｎｇ：Ａ Ｌａｂｏｒａｔｏｒｙ Ｍａｎｕａｌ（Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ Ｌａｂｏｒａｔｏｒｙ，Ｎｅｗ Ｙｏｒｋ）（１９８９）およびＡｕｓｕｂｅｌら、Ｃｕｒｒｅｎｔ Ｐｒｏｔｏｃｏｌｓ ｉｎ Ｍｏｌｅｃｕｌａｒ Ｂｉｏｌｏｇｙ（Ｊｏｈｎ Ｗｉｌｅｙ ａｎｄ Ｓｏｎｓ，Ｎｅｗ Ｙｏｒｋ）（１９９７）を参照されたい（これらの文献は本明細書中に参考として組み込まれる）。このような方法は、異なるＤＮＡ鎖に由来する同じゲノム領域の配列を決定するために使用され得る。次にこれらの配列は比較され、その差（鎖間の変異）が調べられる。ＤＮＡ配列決定法は、ＤＮＡポリメラーゼＩのクレノウ断片、シークエナーゼ（ＵＳ Ｂｉｏｃｈｅｍｉｃａｌ Ｃｏｒｐ，Ｃｌｅｖｅｌａｎｄ，Ｏｈｉｏ．）、Ｔａｑポリメラーゼ（Ｐｅｒｋｉｎ Ｅｌｍｅｒ）、熱安定性ＴＰポリメラーゼ（Ａｍｅｒｓｈａｍ，Ｃｈｉｃａｇｏ，Ｉｌｌ．）またはポリメラーゼとプルーフリーディング（校正）エキソヌクレアーゼ（例えばＧｉｂｃｏ／ＢＲＬ（Ｇａｉｔｈｅｒｓｂｕｒｇ，Ｍｄ．）により市販されている伸長酵素増幅システム（Ｅｌｏｎｇａｓｅ Ａｍｐｌｉｆｉｃａｔｉｏｎ Ｓｙｓｔｅｍ）に見られるものなど）等の酵素を使用する。好ましくは、このプロセスはＨａｍｉｌｔｏｎ Ｍｉｃｒｏ Ｌａｂ ２２００（Ｈａｍｉｌｔｏｎ，Ｒｅｎｏ，Ｎｅｖ．）、Ｐｅｌｔｉｅｒ Ｔｈｅｒｍａｌ Ｃｙｃｌｅｒ（ＰＴＣ２００；ＭＪ Ｒｅｓｅａｒｃｈ，Ｗａｔｅｒｔｏｗｎ，Ｍａｓｓ．）およびＡＢＩ Ｃａｔａｌｙｓｔ、ならびに３７３および３７７ＤＮＡシーケンサ（Ｐｅｒｋｉｎ Ｅｌｍｅｒ，Ｗｅｌｌｅｓｌｅｙ，ＭＡ）等の機械により自動化される。

【００４８】さらに、市販されているキャピラリー電気泳動システムを用いて変異もしくはＳＮＰの分析を行うことができる。特にキャピラリー配列決定は、電気泳動による分離のための流動可能なポリマー、レーザーにより活性化される４つの異なる蛍光染料（各ヌクレオチド毎に1つ）、およびＣＣＤカメラによる発光波長の検出を用いることができる。出力／光強度は、適当なソフトウェア（例えばＧｅｎｏｔｙｐｅｒ ａｎｄ Ｓｅｑｕｅｎｃｅ Ｎａｖｉｇａｔｏｒ，Ｐｅｒｋｉｎ Ｅｌｍｅｒ，Ｗｅｌｌｅｓｌｅｙ，ＭＡ）を用いて電気信号に変換され、サンプルを入れてからコンピュータ分析および電子データ表示までの全プロセスはコンピュータにより制御することができる。またこの方法は、異なるＤＮＡ鎖に由来する同じゲノム領域の配列を決定するために使用することもできる。これらの配列は後に比較され、その差（鎖間の変異）が調べられる。

【００４９】選択肢として、１つの基準ＤＮＡ鎖に由来するゲノム配列を配列決定により決定したら、この基準鎖と他のＤＮＡ鎖との配列の変異を決定するためにハイブリダイゼーション手法を用いることが可能である。これらの変異はＳＮＰであり得る。好適なハイブリダイゼーション手法の１つの例は、例えばＡｆｆｙｍｅｔｒｉｘ，Ｉｎｃ．Ｓａｎｔａ Ｃｌａｒａ，ＣＡより入手可能なもの等のＤＮＡチップ（オリゴヌクレオチドアレイ）の使用を含む。例えばＳＮＰ検出のためのＤＮＡチップの使用についての詳細については、Ｌｉｐｓｈｕｌｔｚ等に付与された米国特許第６，３００，０６３号およびＣｈｅｅらに付与された米国特許第５，８３７，８３２号、ＨｕＳＮＰ Ｍａｐｐｉｎｇ Ａｓｓａｙの試薬キットおよびユーザマニュアル、Ａｆｆｙｍｅｔｒｉｘ Ｐａｒｔ Ｎｏ．９００９４（Ａｆｆｙｍｅｔｒｉｘ，Ｓａｎｔａ Ｃｌａｒａ，ＣＡ）を参照されたい（これらは全て本明細書中に参考として組み込まれる）。

【００５０】好適な実施形態において、基準配列および他のＤＮＡ鎖の１０，０００を超える塩基を、変異体について調べる。より好適な実施形態において、基準配列および他のＤＮＡ鎖の1×10^{6}を超える塩基を変異体について調べ、より好ましくは基準配列および他のＤＮＡ鎖の2×10^{6}を超える塩基を調べ、さらに好ましくは1×10^{7}を超える塩基を調べ、もっと好ましくは1×10^{8}を超える塩基を調べ、さらに好ましくは基準配列および他のＤＮＡ鎖の1×10^{9}を超える塩基を変異体について調べる。好適な実施形態において、少なくともエキソンを変異体について調べ、より好適な実施形態においては、イントロンおよびエキソンの両方を変異体について調べる。さらに好適な実施形態において、イントロン、エキソンおよび遺伝子間配列を変異体について調べる。好適な実施形態において、調べられる核酸はゲノムＤＮＡ（コード領域および非コード領域の両方を含む）である。最も好適な実施形態において、このようなＤＮＡは、ヒト等の哺乳動物生物に由来する。好適な実施形態において、その生物に由来するゲノムＤＮＡの１０％を超える部分について調べ、より好適な実施形態において、その生物のゲノムＤＮＡの２５％を超える部分について調べ、さらに好適な実施形態において、その生物に由来するゲノムＤＮＡの５０％を超える部分について調べ、最も好適な実施形態において、ゲノムＤＮＡの

７５％を超える部分について調べる。本発明の幾つかの実施形態において、ゲノムの既知の反復領域については調べず、調べるゲノムＤＮＡのパーセンテージには入れない。このような既知の反復領域には、短い散在性の反復配列（Ｓｉｎｇｌｅ Ｉｎｔｅｒｓｐｅｒｓｅｄ Ｎｕｃｌｅａｒ Ｅｌｅｍｅｎｔｓ，ＳＩＮＥ、例えばａｌｕおよびＭＩＲ配列等）、長い散在性の反復配列（Ｌｏｎｇ Ｉｎｔｅｒｓｐｅｒｓｅｄ Ｎｕｃｌｅａｒ Ｅｌｅｍｅｎｔｓ，ＬＩＮＥ、例えばＬＩＮＥ１およびＬＩＮＥ２配列等）、長い末端反復配列（Ｌｏｎｇ Ｔｅｒｍｉｎａｌ Ｒｅｐｅａｔｓ，ＬＴＲ、例えばＭａＬＲ、ＲｅｔｒｏｖおよびＭＥＲ４配列等）、トランスポゾン、ＭＥＲ１配列およびＭＥＲ２配列が含まれる。

【００５１】要約すると、１つの実施形態において、好適な溶液中の標識した核酸を、例えば９５℃に加熱して変性させ、変性した核酸を含む溶液をＤＮＡチップと共にインキュベートする。インキュベーション後、溶液を除去し、チップを好適な洗浄液で洗浄してハイブリダイズしなかった核酸を除去し、チップ上のハイブリダイズした核酸の存在を検出する。洗浄条件のストリンジェンシーは、安定なシグナル生成するのに必要なだけ調節することができる。ハイブリダイズした核酸の検出は直接行うことができる。例えば核酸が蛍光レポーター基を含む場合、蛍光は直接検出され得る。核酸上の標識が直接検出できないもの（例えばビオチンなど）である場合、検出前に、検出可能標識（例えばフィコエリトリンに結合したストレプトアビジン）を含む溶液を加えることができる。また、シグナルレベルを高めるよう設計された他の試薬を検出前に加えることもできる。例えばストレプトアビジンに特異的なビオチン化抗体を、該ビオチン・ストレプトアビジン－フィコエリトリン検出システムと一緒に用いることもできる。幾つかの実施形態において、本発明の方法で用いられるオリゴヌクレオチドアレイは、アレイ１つあたり少なくとも1×10^{6}プローブを含む。好適な実施形態において、本発明の方法で使用されるオリゴヌクレオチドアレイは、アレイ１つあたり少なくとも10×10^{6}プローブを含む。より好適な実施形態において、本発明の方法で使用されるオリゴヌクレオチドアレイは、アレイ１つあたり少なくとも50×10^{6}プローブを含む。

【００５２】例えば配列決定またはマイクロアレイ分析を用いて変異体の位置を決定したら（ＳＮＰ発見）、対照集団およびサンプル集団のＳＮＰを遺伝子タイピングすることが必要となる。上記ハイブリダイゼーション方法は、この目的に非常に役立ち、複数サンプルにおけるＳＮＰの検出および遺伝子タイピングのための正確且つ高速な技術を提供する。さらに、増幅を行わないゲノムＤＮＡ中のＳＮＰ検出に適した手法は、Ｔｈｉｒｄ Ｗａｖｅ Ｔｅｃｈｎｏｌｏｇｉｅｓ，Ｉｎｃ．，Ｍａｄｉｓｏｎ，ＷＩから入手可能なＩｎｖａｄｅｒテクノロジーである。ＳＮＰを検出するためのこのテクノロジーの使用については、例えばＨｅｓｓｎｅｒら、Ｃｌｉｎｉｃａｌ Ｃｈｅｍｉｓｔｒｙ ４６（８）：１０５１－５６（２０００）；Ｈａｌｌら、ＰＮＡＳ ９７（１５）：８２７２－７７（２０００）；Ａｇａｒｗａｌら、Ｄｉａｇ．Ｍｏｌｅｃ．Ｐａｔｈ．９（３）：１５８－６４（２０００）；およびＣｏｏｋｓｅｙら、Ａｎｔｉｍｉｃｒｏｂｉａｌ ａｎｄ Ｃｈｅｍｏｔｈｅｒａｐｙ ４４（５）：１２９６－１３０１（２０００）に記載されている。Ｉｎｖａｄｅｒプロセスでは、２つの短いＤＮＡプローブが標的核酸にハイブリダイズして、ヌクレアーゼ酵素により認識される構造を形成する。ＳＮＰ分析のために、２つの別々の反応、つまり各ＳＮＰ変異体毎に１つの反応が行われる。プローブのうちの一方がその配列に相補的である場合、ヌクレアーゼはこれを切断して、「フラップ（ｆｌａｐ）」と呼ばれる短いＤＮＡ断片を放出する。フラップは、蛍光標識したプローブに結合して、ヌクレアーゼ酵素により認識される他の構造を形成する。この酵素が標識プローブを切断すると、プローブは検出可能な蛍光シグナルを発することにより、どのＳＮＰ変異体が存在するかを示す。

【００５３】Ｉｎｖａｄｅｒテクノロジーの代替法であるローリングサークル型増幅は、環状ＤＮＡ鋳型に相補的なオリゴヌクレオチドを利用して増幅シグナルを生成する（例えばＬｉｚａｒｄｉら、Ｎａｔｕｒｅ Ｇｅｎｅｔｉｃｓ １９（３）：２２５－３２（１９９８）；およびＺｈｏｎｇら、ＰＮＡＳ ９８（７）：９９４０－４５（２００１）を参照されたい）。オリゴヌクレオチドの伸長により、長いコンカテマー中に該環状鋳型の多重コピーが生成される。典型的には、検出可能標識は、伸長反応中に伸長されたオリゴヌクレオチド中に導入される。伸長反応は検出可能な量の伸長産物が合成されるまで進められる。

【００５４】ローリングサークル型増幅を用いてＳＮＰを検出するためには、３つのプローブおよび２つの環状ＤＮＡ鋳型を用いることができる。第１プローブ（標的特異的プローブ）は、該プローブの５’側末端が、標的核酸中のＳＮＰ部位の５’側に隣接したヌクレオチドにハイブリダイズするように、該標的核酸分子に相補的となるよう構築される。このＳＮＰ部位は、第１プローブと塩基対を形成しない。

【００５５】他の２つのプローブ（ローリングサークル型プローブ）は、２つの３’側末端を持つように構築される。これは、様々な方法で（例えばこれらのプローブの中央部分に５’－５’結合を導入してそのポイントで該ヌクレオチド配列の極性を逆転させることにより）行うことができる。これらのプローブの各々の一方の端部は、異なる環状鋳型分子の一部分に相補的な配列を有す

るが、他方の端部は、標的核酸配列の一部分に相補的である。この標的配列相補的末端は、最も３’側のヌクレオチドがＳＮＰ部位のヌクレオチドと揃えられるように構築される。これらのプローブの一方は、標的配列中のＳＮＰ部位のヌクレオチドに相補的なヌクレオチドを含むが、他方は相補的でないヌクレオチドを含む。ＳＮＰの２以上の変異体がその集団の中に存在する場合、プローブは、検出される変異体に相補的な３’側ヌクレオチドを有するよう構築される。

【００５６】プローブ（標的特異的プローブおよびローリングサークル型プローブの両方）を標的配列にハイブリダイズさせ、リガーゼ酵素に接触させる。ローリングサークル型プローブの最も３’側のヌクレオチドがＳＮＰ部位のヌクレオチドと塩基対を形成すると、これら２つのプローブ（標的特異的プローブおよびローリングサークル型プローブ）は効率的に連結される。ローリングサークル型プローブの最も３’側のヌクレオチドが標的中のＳＮＰ部位にあるヌクレオチドと塩基対を形成できない場合、プローブは連結されない。連結されなかったプローブを洗い流し、サンプルを鋳型環状体、ポリメラーゼおよび標識したヌクレオチド３リン酸に接触させる。

【００５７】ＳＮＰの検出に適した他の手法は、プローブ分子を消化して蛍光標識されたヌクレオチドを放出することによりシグナルを生成するＤＮＡポリメラーゼの５’側エキソヌクレアーゼ活性を利用する。このアッセイはしばしばＴａｑｍａｎアッセイと呼ばれる（例えばＡｒｎｏｌｄら、ＢｉｏＴｅｃｈｎｉｑｕｅｓ　２５（１）：９８－１０６（１９９８）；およびＢｅｃｋｅｒら、Ｈｕｍ．　Ｇｅｎｅ　Ｔｈｅｒ．　１０：２５５９－６６（１９９９）を参照されたい）。ＳＮＰを含む標的ＤＮＡを、そのＳＮＰ部位にハイブリダイズするプローブ分子の存在下で増幅する。このプローブ分子は、５’側末端に蛍光レポーター標識ヌクレオチド、および３’側末端に消光剤標識ヌクレオチドを両方含む。このプローブ配列は、正しくマッチしたプローブとミスマッチプローブとの融解温度の差を最大とするために、標的ＤＮＡの中のＳＮＰ部位に揃えられるそのプローブの中のヌクレオチドができるだけそのプローブの中央に近くなるように選択される。ＰＣＲ反応が行われるとき、正しくマッチしたプローブは標的ＤＮＡ中のＳＮＰ部位にハイブリダイズし、ＰＣＲアッセイで用いられるＴａｑポリメラーゼによって消化される。この消化により、蛍光標識されたヌクレオチドは消光剤から物理的に分離され、それに伴い蛍光が増大する。ミスマッチプローブはＰＣＲ反応の伸長工程の間はハイブリダイズしたままではないので消化はされず、蛍光標識されたヌクレオチドは消光されたままとなる。

【００５８】ポリスチレン－ジビニルベンゼン逆相カラムおよびイオン・ペアリング移動相（ｉｏｎ－ｐａｉｒｉｎｇ　ｍｏｂｉｌｅ　ｐｈａｓｅ）を用いた変性ＨＰＬＣを用いて、ＳＮＰを同定することができる。ＳＮＰを含むＤＮＡセグメントをＰＣＲ増幅する。増幅後、ＰＣＲ産物を加熱して変性させ、ＳＮＰ位置に既知のヌクレオチドを有する第２の変性ＰＣＲ産物と混合する。ＰＣＲ産物をアニーリングし、高温にてＨＰＬＣにより分析する。温度は、ＳＮＰ位置にミスマッチを含む二本鎖分子（ｄｕｐｌｅｘ　ｍｏｌｅｃｕｌｅ）を変性させ且つ完全にマッチした二本鎖分子を変性させないよう選択される。このような条件下において、ヘテロ二本鎖分子は典型的にはホモ二本鎖分子の前に溶離する。このような手法の使用例については、Ｋｏｔａら、Ｇｅｎｏｍｅ　４４（４）：５２３－２８（２００１）を参照されたい。

【００５９】固相増幅および増幅産物のミクロ配列決定を用いてＳＮＰを検出することができる。プライマーを共有結合させたビーズを用いて、増幅反応を行う。プライマーは、ＩＩ型制限酵素の認識部位を含むように設計される。増幅（これによりＰＣＲ産物がビーズに結合する）後、産物を制限酵素で消化する。制限酵素による産物の切断により、ＳＮＰ部位および３’－ＯＨ（伸長されてその一本鎖部分を埋めることができる）を含む一本鎖部分が産生される。伸長反応にｄｄＮＴＰを含めることにより、産物の直接的な配列決定が可能となる。ＳＮＰを同定するためのこの手法の使用例については、Ｓｈａｐｅｒｏら、Ｇｅｎｏｍｅ　Ｒｅｓｅａｒｃｈ　１１（１１）：１９２６－３４（２００１）を参照されたい。

【００６０】データ分析

図１は、本発明の方法の１つの実施形態の工程を示す概略図である。例えば上記方法によってＳＮＰ（変異体）を位置付けまたは発見したら（図１のステップ１１０）、ＳＮＰハプロタイプブロック、各ＳＮＰハプロタイプブロックの中のＳＮＰハプロタイプパターン、およびそのＳＮＰハプロタイプパターンの情報提供ＳＮＰを決定することができる。位置付けされた全てのＳＮＰまたは変異体を使用しても良いし、あるいは位置付けされたＳＮＰの一部のみについて分析を行っても良い。例えば分析されるＳＮＰのセットは、Ｃｇ＜－＞ＴｇまたはｃＧ＜－＞ｃＡの転位ＳＮＰ（ｔｒａｎｓｉｔｉｏｎ　ＳＮＰ）を除外してもよい。さらに、本発明の１つの実施形態において、注目されるのは共通ＳＮＰである。共通ＳＮＰは、そのより一般的でない形態が所与の集団において最少頻度で存在するこれらのＳＮＰである。例えば、共通ＳＮＰは、その集団の少なくとも約２％～２５％で見られるこれらのＳＮＰである。好適な実施形態において、共通ＳＮＰは、その集団の少なくとも約５％～１５％で見られるこれらのＳＮＰである。より好適な実施形態において、共通ＳＮＰは、その集団の少なくとも約１０％において見られるＳＮＰである。共通ＳＮＰ

は、ヒトの進化過程において初期に生じた突然変異から生じたものであると思われる。共通ＳＮＰに焦点を絞ることにより、疾患または薬物応答に関連すると思われ且つ移住歴（ｍｉｇｒａｔｏｒｙｈｉｓｔｏｒｙ）または交配（ｍａｔｉｎｇ　ｐｒａｃｔｉｃｅｓ）によってのみ生じる、実験集団および対照集団における系統的な対立遺伝子または変異体の差を最少にする。すなわち、共通ＳＮＰに注目することによって、近年の集団の変則性（ｐｏｐｕｌａｔｉｏｎ　ａｎｏｍａｌｙ）から生じる偽陽性が減る。さらに共通ＳＮＰは、ヒト人口の大部分に関係があり、このことは、本発明を、疾患および薬物応答の調査により広く応用できるようにしている。これにならい、本発明の幾つかの実施形態（例えばシングルトンＳＮＰなど）において、変異体が１回だけ観察されるＳＮＰを分析から排除してもよい。しかし、特に移住歴等の影響を受けた特定の部分集団または集団を見る場合に、これらのシングルトンＳＮＰの一部または全てを含むある種の分析を行ってもよい。

【００６１】図１のステップ１２０では、目的の変異体またはＳＮＰをハプロタイプブロックに割り当てて評価する。全ゲノムまたは染色体から得た変異体またはＳＮＰを分析し、ＳＮＰハプロタイプブロックに割り当てることができる。あるいは、ある疾患または薬物応答メカニズムに特異的な特定のゲノム領域のみからの変異体をＳＮＰハプロタイプブロックに割り当てることができる。

【００６２】図２は、ゲノム内のハプロタイプブロック中において変異体（通常はＳＮＰ）がどのように生じるのか、および１以上のハプロタイプパターンが各ハプロタイプブロック内で起こり得ることを示す１つの例である。ＳＮＰハプロタイプパターンが完全に無作為である場合、Ｎ個のＳＮＰからなるＳＮＰハプロタイプブロックについて観察される可能なＳＮＰハプロタイプパターンの数は2^N個であると推測される。しかし、本発明の方法を実施した際に、ＳＮＰは連鎖しているため、各ＳＮＰハプロタイプブロックにおけるＳＮＰハプロタイプパターンの数は、2^N個より少ないことが分かった（変異体は最も一般的にはｂｉａｌｌｅｌｉｃである、すなわち４つのヌクレオチド塩基の可能性全てではなく２つの形態のうち一方の形態でのみ生じるので、4^Nではない）。あるＳＮＰハプロタイプパターンは、非連鎖ケースで予想される頻度よりはるかに高い頻度で観察された。したがって、ＳＮＰハプロタイプブロックは、ユニットとして継承される傾向がある染色体領域であり、比較的少ない数の共通パターンを有する。図２中の各行は、異なる個体の一倍体ゲノム配列の一部を表す。図２に示すように、個体Ｗは、２４１位に「Ａ」を、２４２位に「Ｇ」を、および２４３位に「Ａ」を有する。個体Ｘは、２４１位、２４２位および２４３位に同じ塩基を有する。これに対し、個体Ｙは、２４１位および２４３位にＴを有し２４２位にはＡを有する。個体Ｚは２４１位、２４２位および２４３位に個体Ｙと同じ塩基を有する。ブロック２６１における変異体は、一緒に生じる傾向がある。同様に、ブロック２６２における変異体は一緒に生じる傾向があり、ブロック２６３中における変異体も同様である。もちろん、ゲノム中の少数の塩基しか図２に表されていない。実際、多くの塩基は２４５位および２４８位の塩基と同様であり、個体間でばらつきはない。

【００６３】ＳＮＰハプロタイプブロックへのＳＮＰの割り当て（図１のステップ１２０）は、あるケースでは、目的のゲノム領域に沿ったＳＮＰ位置からのＳＮＰハプロタイプブロックの構築を含む反復プロセスである。１つの実施形態において、最初のＳＮＰハプロタイプブロックを構築したら、その構築されたＳＮＰハプロタイプブロックの中に存在するＳＮＰハプロタイプパターンを決定する（図１のステップ１３０）。幾つかの具体的な実施形態において、ステップ１３０において各ＳＮＰハプロタイプブロック毎に選択されたＳＮＰハプロタイプパターンの数は、約５未満である。他の具体的な実施形態において、各ＳＮＰハプロタイプブロック毎に選択されたＳＮＰハプロタイプパターンの数は、分析されるＤＮＡ鎖の５０％を超える中のＳＮＰハプロタイプパターンを同定するのに必要なＳＮＰハプロタイプパターンの数に等しい。言い換えると、十分なＳＮＰハプロタイプパターンが選択される。例えば、分析されるＤＮＡ鎖の少なくとも半分が、各ＳＮＰハプロタイプブロックで選択された４つのパターンのうちの１つにマッチするＳＮＰハプロタイプパターンを有するように、１ブロックあたり４つのパターンが選択される。好適な実施形態において、各ＳＮＰハプロタイプブロックにつき選択されたＳＮＰハプロタイプパターンの数は、分析されるＤＮＡ鎖の７０％を超える中のＳＮＰハプロタイプパターンを同定するのに必要なＳＮＰハプロタイプパターンの数に等しい。１つの好適な実施形態において、各ＳＮＰハプロタイプブロックにつき選択されたＳＮＰハプロタイプパターンの数は、分析されるＤＮＡ鎖の８０％を超える中のＳＮＰハプロタイプパターンを同定するのに必要なＳＮＰハプロタイプパターンの数に等しい。さらに、発明の幾つかの実施形態において、分析されるＤＮＡ鎖のある一定割合未満で生じるＳＮＰハプロタイプパターンは、分析から排除される。例えば、１つの実施形態において、１０本のＤＮＡ鎖を分析する場合、１０本のうち１本のサンプルにおいてのみ生じることが分かったＳＮＰハプロタイプパターンは、分析から排除される。

【００６４】目的のＳＮＰハプロタイプパターンを選択したら、これらのＳＮＰハプロタイプパターンのための情報提供ＳＮＰを決定する（図１のステップ１４０）。ブロックのこの最初のセットから、情報提供性について

の一定の基準を満たす候補ＳＮＰブロックのセットを作製する（図１のステップ１５０）。図４および図５は、ステップ１２０、１３０、１４０および１５０についてより詳細に記載している。

【００６５】図３において、ステップ３１０では、評価のためにＳＮＰの新しいブロックが選択される。１つの実施形態において、選択された第１ブロックは、ＳＮＰハプロタイプ配列の中の第１ＳＮＰのみを含む。したがってステップ３２０では、最初の１つのＳＮＰがブロックに追加される。ステップ３３０では、このブロックの情報提供性が決定される。

【００６６】１つの実施形態において、ＳＮＰハプロタイプブロックの「情報提供性」は、そのブロックが遺伝子領域についての情報を提供する程度として定義される。例えば、本発明の１つの実施形態において、情報提供性は、あるＳＮＰハプロタイプブロックの中のＳＮＰ位置の数を、考慮される各ＳＮＰハプロタイプパターンをそのブロックの中の考慮される他のＳＮＰハプロタイプパターンと区別するのに必要なＳＮＰの数（情報提供ＳＮＰの数）で割った比として計算することができる。情報提供性の他の測度は、そのブロックの中の情報提供ＳＮＰの数である。当業者であれば、情報提供性は、色々な方法で決定することができることが分かるであろう。

【００６７】再び図２を参照すると、ＳＮＰハプロタイプブロック２６１は３つのＳＮＰおよび２つのＳＮＰハプロタイプパターン（ＡＧＡおよびＴＡＴ）を含む。存在するこの３つのＳＮＰのどれを使用してもこれらのパターンを見分けることができるので、これらのＳＮＰのうち任意の１つを、このＳＮＰハプロタイプパラーンのための情報提供ＳＮＰとして選択することができる。例えば、あるサンプル核酸が第１位置にＴを含むと決定された場合、そのサンプルは第２位置にＡおよび第３位置にＴを含む。第２サンプルにおいて第２位置のＳＮＰがＧであると決定された場合、第１および第３のＳＮＰはＡである。このように、情報提供性の１つの測度として、この第１ブロックの情報提供性値は３である（全ＳＮＰの数３を、該パターンを互いから区別するのに必要な情報提供ＳＮＰの数１で割った）。同様に、ＳＮＰハプロタイプブロック２６２は、３つのＳＮＰ（２つの位置は変異体を持たない）および２つのハプロタイプパターン（ＴＣＧおよびＣＡＣ）を含む。前に分析したブロックを用いて、該３つのＳＮＰのうちのどれか１つを評価して１つのパターンを他のパターンと区別することができる。したがって、このブロックの情報提供性は３である（全ＳＮＰの数３を、パターンを区別するのに必要な情報提供ＳＮＰの数１で割った）。ＳＮＰハプロタイプブロック２６３は、５つのＳＮＰおよび２つのＳＮＰパターン（ＴＡＡＣＧおよびＡＴＣＡＣ）を含む。この場合も、５つのＳＮＰのうちのどれか１つを用いて１つのパターンを他のパターンと区別することができる。したがって、このブロックの情報提供性は、５である（全ＳＮＰの数５を、パターンを区別するのに必要な情報提供ＳＮＰの数１で割った）。

【００６８】図２は、遺伝子分析の単純な例を提供する。幾つかのＳＮＰハプロタイプパターンが１つのブロックの中に存在する場合、情報提供ＳＮＰとして２以上のＳＮＰを用いることが必要となる場合がある。例えば、１つのブロックが例えば６つのＳＮＰを含み且つ目的のパターンを区別するために２つのＳＮＰが必要である場合、そのブロックの情報提供性は３である（全ＳＮＰの数６を、パターンを区別するのに必要な情報提供ＳＮＰの数２で割った）。一般的に言うと、適切に選択されたＮ個のＳＮＰの遺伝子タイプを用いて、2^N個もの異なるＳＮＰハプロタイプパターンを区別することができる。したがって、そのＳＮＰハプロタイプブロックの中にたった２つのＳＮＰハプロタイプパターンしか存在しない場合、１つのＳＮＰはその２つを差別化できなければならない。３もしくは４つのパターンがある場合、少なくとも２つのＳＮＰが必要となるであろう。

【００６９】図３のステップ３４０において、ＳＮＰハプロタイプブロックの情報提供性が決定されたら、テストが行われる。このテストは本質的に、選択された基準（例えばあるブロックが情報提供性の閾値を満たすか否か）に基づいてＳＮＰハプロタイプブロックを評価し、テスト結果は、例えば他のＳＮＰが分析のためにそのブロックに追加されるか否か、またはその分析が異なるＳＮＰ位置で始まる新しいブロックを用いて進行するか否かを決定する。図４は、このプロセスの１つの実施形態を示す。

【００７０】図４では、６つのＳＮＰ位置を有するＤＮＡ配列があると想定する。上記のＳＮＰハプロタイプブロック分析は、以下のように行うことができる：ＳＮＰ位置１にあるＳＮＰのみを含むＳＮＰハプロタイプブロックＡが選択される（図３のステップ３１０および３２０）。このブロックの情報提供性を計算し（ステップ３３０）、このブロックの情報提供性が情報提供性の閾値を満たすか否かを決定する（ステップ３４０）。この場合、これは「合格」し、２つの事が起こる。まず第１に、１つのＳＮＰ（ＳＮＰ位置１）のこのブロックを候補ＳＮＰハプロタイプブロックのセットに追加する（ステップ３５０）。第２に、他のＳＮＰ（ここではＳＮＰ位置２）をこのブロックに追加し（ステップ３２０）、ＳＮＰ位置１および２を含む新しいブロックＢを作製し、次にこれを分析する。この例では、ブロックＢも情報提供性の閾値を満たすので（ステップ３４０）、候補ＳＮＰハプロタイプブロックのセットに追加され（ステップ３５０）、他のＳＮＰ（この場合はＳＮＰ位置３）がこのブロックに追加されて（ステップ３２０）、ＳＮＰ位置１、２および３を含む新しいブロックＣが作製さ

れ、次にこれを分析する。この例では、Ｃも情報提供性の閾値を満たすので、候補ＳＮＰハプロタイプブロックのセットに追加され（ステップ３５０）、他のＳＮＰ（この場合はＳＮＰ位置４）がこのブロックに追加されて（ステップ３２０）、ＳＮＰ位置１、２、３および４を含む新しいブロックＤが作製され、次にこれを分析する。図４の例では、ＳＮＰブロックＤは情報提供性の閾値を満たさない。ＳＮＰブロックＤは候補ＳＮＰハプロタイプブロックのセットに追加されず（ステップ３５０）、また他のＳＮＰが分析のためにブロックＤに追加されることもない。その代わりに、新しいＳＮＰ位置が選択され、これについてＳＮＰブロック評価が行われる。

【００７１】図４では、ブロックＤが情報提供性の閾値を満たさないと分かった後、位置２にのみＳＮＰを含む新しいブロックＥが選択される。ブロックＥの情報提供性について評価し、閾値を満たすことが分かり、候補ＳＮＰハプロタイプブロックのセットに追加され（ステップ３５０）、および他のＳＮＰ（この場合ＳＮＰ位置３）がこのブロックに追加されて（ステップ３２０）、ＳＮＰ位置２および３を含む新しいブロックＦを作製し、次にこれが分析され、上記工程がまた繰返される。ここで、ブロックＨは情報提供性の閾値を満たさず、候補ＳＮＰハプロタイプブロックのセットに追加されず（ステップ３５０）、また他のＳＮＰが分析のためにブロックＨに追加されることもない。その代わり、位置３のＳＮＰのみを含む新しいブロックＩが選択されて、上記工程が繰返される。

【００７２】候補ＳＮＰブロックのセットが作製されたら（図３のステップ３５０）、このセットに対して分析が行われ、ＳＮＰブロックの最終セットが選択される（図１のステップ１６０）。ＳＮＰブロックの最終セットの選択は、様々な方法で行うことができる。例えば、再び図４を参照すると、閾値テストを合格するＳＮＰ位置１を含む最も大きなブロック（ＳＮＰ１、２および３を含むブロックＣ）を選択し、同じＳＮＰを含むより小さなブロック（ブロックＡおよびＢ）を破棄することができる。そして、次に選択されるブロックは、情報提供性についての閾値テストに合格する最も大きなブロックであるＳＮＰ位置４で始まる次のブロック（ブロックＧ）であり、同じＳＮＰを含むより小さなブロック（ブロックＥおよびＦ）は破棄される。このような方法により、目的のゲノム領域にまたがり、目的のＳＮＰを含み、且つ高レベルの情報提供性を有する、最終的な非重複ＳＮＰハプロタイプブロックのセットが得られる。このように、全ての候補ＳＮＰハプロタイプブロックが評価されたら、結果は、好適な実施形態において、元のセットの中の全てのＳＮＰを包含する非重複ＳＮＰハプロタイプブロックのセットである。孤立体（ｉｓｏｌａｔｅ）と呼ばれる幾つかのグループはたった１つのＳＮＰからなり、定義により情報提供性は１である。他のグループは、１００以上のＳＮＰからなり、情報提供性は３０を超え得る。

【００７３】ＳＮＰハプロタイプブロックの最終セットを選択するための他の方法を図５Ａおよび図５Ｂに示す。まず図５Ａのステップ５１０を見ると、候補ＳＮＰハプロタイプブロックセット（例えば本願の図３および図４に記載された方法により作製されたもの）の情報提供性について分析する。ステップ５２０において、その候補セット全体の中で最も高い情報提供性を有する候補ＳＮＰハプロタイプブロックを選択して、最終ＳＮＰハプロタイプブロックセットに追加する（ステップ５３０）。この候補ＳＮＰハプロタイプブロックを最終ＳＮＰハプロタイプブロックセットのメンバーとして選択したら、このブロックを候補ブロックセットから削除し（ステップ５４０）、この選択されたブロックと重複する全ての他の候補ＳＮＰハプロタイプブロックをこの候補ＳＮＰハプロタイプブロックセットから削除する（ステップ５５０）。次に、該候補セットの中に残っている候補ＳＮＰハプロタイプブロックの情報提供性について分析し（ステップ５１０）、最も高い情報提供性を有する候補ＳＮＰハプロタイプブロックセットを選択して最終ハプロタイプブロックセットに追加する（ステップ５２０および５３０）。先に記載したように、このＳＮＰハプロタイプブロックを最終ＳＮＰハプロタイプブロックセットのメンバーとして選択したら、このブロックを候補ブロックセットから削除し（ステップ５４０）、この選択されたブロックと重複する全ての他の候補ＳＮＰハプロタイプブロックをこの候補ＳＮＰハプロタイプブロックセットから削除する（ステップ５５０）。このプロセスは、元のセットの中の全てのＳＮＰを包含する非重複ＳＮＰハプロタイプブロックの最終セットが作製されるまで続けられる。

【００７４】図５Ｂは、図５Ａに記載されたＳＮＰハプロタイプブロックの最終セットの選択方法の単純な使用例を示す。図５Ｂにおいて、本発明の方法に従い、配列５’側～３’側を、ＳＮＰ、ＳＮＰハプロタイプパターンおよび候補ＳＮＰハプロタイプブロックについて分析する。この配列中に含まれる候補ＳＮＰハプロタイプブロックは、その配列の下のこれらの配置により示され、文字によって指定される。さらに、文字の後に各ブロックの情報提供性が示される。例えば、候補ＳＮＰハプロタイプブロックＡは、その配列の最も５’側端部に位置し、情報提供性は１である。候補ＳＮＰハプロタイプブロックＲは、その配列の最も３’側末端に位置し、情報提供性は２である。

【００７５】図５Ａに従い、第１ステップ５１０において、候補ＳＮＰハプロタイプブロックの情報提供性について分析し、ステップ５２０で、最も高い情報提供性を有するＳＮＰハプロタイプブロックを選択して、最終Ｓ

ＮＰハプロタイプブロックセットに追加する（ステップ５２０および５３０）。図５Ｂの場合、情報提供性が６である候補ＳＮＰハプロタイプブロックＭが、最終ＳＮＰハプロタイプブロックセットに追加するために選択された最初の候補ＳＮＰハプロタイプブロックである。ＳＮＰハプロタイプブロックＭを選択したら、このブロックはＳＮＰハプロタイプブロックの候補セットから削除もしくは除外され（ステップ５４０）、ＳＮＰハプロタイプブロックＭと重複する全ての他の候補ＳＮＰハプロタイプブロック（ブロックＪ、Ｎ、Ｋ、Ｌ、ＯおよびＰ）は、候補ＳＮＰハプロタイプブロックセットから削除される（ステップ５５０）。次に、候補ＳＮＰハプロタイプブロックセットの残りのブロック、つまりＳＮＰハプロタイプブロックＡ、Ｂ、Ｃ、Ｄ、Ｅ、Ｆ、Ｇ、Ｈ、Ｉ、ＱおよびＲの情報提供性について分析し、およびステップ５２０において、最も高い情報提供性（情報提供性５）を有する残りのＳＮＰハプロタイプブロックＩを選択して最終ＳＮＰハプロタイプブロックセットに追加し（５３０）、ＳＮＰハプロタイプブロックの候補セットから削除または除外する（ステップ５４０）。次に、ステップ５５０でＳＮＰハプロタイプブロックＩと重複する全ての他の候補ＳＮＰハプロタイプブロック（この場合はブロックＨ）が、候補ＳＮＰハプロタイプブロックセットから削除される。この場合も、候補ＳＮＰハプロタイプブロックセットの残りのブロック、つまりＳＮＰハプロタイプブロックＡ、Ｂ、Ｃ、Ｄ、Ｅ、Ｆ、Ｇ、ＱおよびＲの情報提供性について分析する。ステップ５２０では、最も高い情報提供性（情報提供性４）を有する残りのＳＮＰハプロタイプブロック（ブロックＦ）を選択して最終ＳＮＰハプロタイプブロックセットに追加し（５３０）、ＳＮＰハプロタイプブロックの候補セットから削除または除外する（ステップ５４０）。次に、ＳＮＰハプロタイプブロックＦと重複する全ての他の候補ＳＮＰハプロタイプブロック（この場合はブロックＥ、Ｇ、ＣおよびＤ）が、候補ＳＮＰハプロタイプブロックセットから削除され、候補ＳＮＰハプロタイプブロックセットの残りのブロック、つまりＳＮＰハプロタイプブロックＡ、Ｂ、ＱおよびＲの情報提供性について分析する。このように上記工程が繰返される。

【００７６】他の方法を用いて、候補ＳＮＰハプロタイプブロックのセットから分析用のＳＮＰハプロタイプブロックの最終セットを選択することができる（図１のステップ１６０）。例えば、当分野で公知のアルゴリズムをこの目的で応用することができる。例えば、最短路アルゴリズムを用いることができる（一般にＣｏｒｍｅｎ，　ＬｅｉｓｅｒｓｏｎおよびＲｉｖｅｓｔ，　Ｉｎｔｒｏｄｕｃｔｉｏｎｔｏ　Ａｌｇｏｒｉｔｈｍｓ（ＭＩＴ　Ｐｒｅｓｓ）　ｐｐ．５１４－７８（１９９４）を参照されたい）。最短路問題において、辺を実数値の重みにマッピングする重み関数ｗ：Ｅ→Ｒが重みとして与えられた有向グラフＧ＝（Ｖ，Ｅ）が与えられる。路ｐ＝（v_0，v_1，．．．v_k）の重みはその構成辺の重みの総和である。

【００７７】

【数１】

$$w(p) = \sum_{i=1}^{k} w(v_{i-1}, v_i).$$

【００７８】μからνまでの路がある場合、ｕからｖまでの最短路の重みはδ（ｕ，ｖ）＝ｍｉｎ　ｗ（ｐ）ｕ→ｖであり、そうでない場合はδ（ｕｓ，ｖ）＝∞である、として定義される。次に、頂点ｕから頂点ｖまでの最短路は、重みｗ（ｐ）＝δ（ｕ，ｖ）が与えられた任意の路ｐとして定義される。辺の重みは、様々な測定基準、例えば距離、時間、費用、ペナルティ、損失、または最小にしたい路に沿って線形に蓄積する任意の他の数量として解釈することができる。本発明の用途で使用される最短路アルゴリズムの実施形態において、各ＳＮＰハプロタイプブロックは、そのブロックの各境界毎に画定された「辺」を有する「頂点」であると考えられる。各ＳＮＰハプロタイプブロックは、他の各ＳＮＰハプロタイプブロックとの関係を有し、各エッジ毎に「費用」を有する。費用は、一般に好まれるパラメータ、例えば頂点の重複（もしくは重複の程度）または頂点間のギャップ等により決定される。

【００７９】単一出発点最短路問題は、所与の出発点である頂点ｓεＶから全ての頂点ｖεＶまでの最短路を決定する所与のグラフＧ＝（Ｖ，Ｅ）に焦点を当てる。更に、単一出発点アルゴリズムの異形を応用することもできる。例えば、全ての頂点ｖから所与の到達点である頂点ｔまでの最短路を見つける単一到達点最短路解決法を応用してもよい。グラフの中の各辺の向きを反対にすると、この問題が単一出発点問題に還元される。あるいは、所与の頂点ｕおよびｖのためのｕからｖまでの最短路を見つける単一ペア最短路問題を応用してもよい。出発点が頂点ｕである単一出発点問題を解決すれば、単一出発点最短路問題も解決する。また、全ペア最短路法を使用してもよい。この場合、頂点ｕおよび頂点ｖの全てのペアのためのｕからｖまでの最短路が見つかる、つまり、単一出発点アルゴリズムは各頂点から実行される。

【００８０】本発明の方法で用いることができる１つの単一出発点最短路アルゴリズムは、ダイクストラ法である。ダイクストラのアルゴリズムは、全ての辺の重みが非負である場合、重みが与えられた有向グラフＧ＝（Ｖ，Ｅ）に対して単一出発点最短路問題を解決する。ダイクストラのアルゴリズムは、出発点ｓから最終的な最短路の重みが既に決定された頂点のセットＳを維持する。つまり、全ての頂点ｖはＳ，ｄ［ｖ］＝δ（ｓ，ｖ）の要素である。このアルゴリズムは最小最短路推定値を有するＶ－Ｓの要素として頂点ｕを繰返し選択し、

ｕをＳに挿入し、ｕから出ている全ての辺を緩和する。１つの処理系において、Ｖ－Ｓの中の全ての頂点を含む（それらのｄ値によりキーイングされた）優先順位付き待ち行列Ｑが維持される。この処理系は、グラフＧが隣接リストにより表されると想定する。

【００８１】

【数２】

```
Dijkstra (G, w, s)
1    INITIALIZE-SINGLE SOURCE (G,s)
2    S ← Ø
3    Q ← V[G]
4    while Q ≠ Ø
5    do  u ← EXTRACT-MIN (Q)
6    S ← S U {u}
7    for each vertex v ϵ Adj[u]
8    do RELAX (u,v,w)
```

【００８２】このように、この場合のＧは分析されるゲノム配列の線状網羅度（ｌｉｎｅａｒ　ｃｏｖｅｒａｇｅ）のグラフであり、Ｓは選択された頂点のセットである。ゲノム配列の特定の領域を網羅する１つの頂点が選択されたら、この配列と重複する他の頂点を破棄することができる。

【００８３】ＳＮＰハプロタイプブロックを選択するために使用することができる他のアルゴリズムとしては、欲張り法アルゴリズムが挙げられる（再びＣｏｒｍｅｎ，ＬｅｉｓｅｒｓｏｎおよびＲｉｖｅｓｔ，　Ｉｎｔｒｏｄｕｃｔｉｏｎ　ｔｏＡｌｇｏｒｉｔｈｍｓ　（ＭＩＴ　Ｐｒｅｓｓ）　ｐｐ．３２９－５５（１９９４）を参照されたい）。欲張りアルゴリズムは、選択肢のシーケンスを作製することにより、ある問題に対する最適解を得る。このアルゴリズムにおける各決定ポイント毎に、その時点で最良であると思われる選択肢が選択される。このヒューリスティック法は、いつも最適解をもたらすわけではない。ダイナミックプログラミングにおいて、選択は各ステップで行われるがその選択は部分問題に対する解に依存するという意味において、欲張りアルゴリズムはダイナミックプログラミングとは異なる。欲張りアルゴリズムにおいて、その時点でどの選択肢が最良に見えたとしても、その選択が行われた後に生じる部分問題は解決される。このように、欲張りアルゴリズムによってなされた選択は、それまでになされた選択に依存するが、将来の選択または部分問題に対する解に依存することはできない。欲張りアルゴリズムの１つの変形は、ハフマン記号である。ハフマンの欲張りアルゴリズムは、最適な語頭条件を満足する符号を作成し、このアルゴリズムは、その最適符号に対応するボトムアップ式の木Ｔを構築する。このアルゴリズムは｜Ｃ｜葉のセットから始まって｜Ｃ｜－１「マージング」演算のシーケンスを実行し、最終的な木を作成する。例えば、Ｃがｎ個の文字からなるセットであり、各文字ｃεＣは決められた頻度（ｆｒｅｑｕｅｎｃｙ）ｆ［ｃ］を有するオブジェクトであると仮定すると、ｆについてキーイングされた優先順位付き待ち行列Ｑを用いて２つの最少頻度オブジェクトを同定し、これらをマージングする。２つのオブジェクトがマージングされたもの（マージャー、ｍｅｒｇｅｒ）の結果は、マージングされたこれら２つのオブジェクトの頻度の合計である頻度を有する新しいオブジェクトである。例えば：

【００８４】

【数３】

```
1.  n←|C|
2.  Q←C
3.  for i←1 to n-1
4.  do z←ALLOCATE-NODE()
5.  x←left[z] ←EXTRACT-MIN(Q)
6.  y←right[z] ←EXTRACT-MIN(Q)
7.  f[z] ←f[x] +f[y]
8.  INSERT (Q,z)
9.  return EXTRACT-MIN(Q)
```

【００８５】２行目は、Ｃの中の文字を用いて優先順位付き待ち行列Ｑを開始する。３～８行目のｆｏｒループは、この待ち行列から最低頻度の２つの節点ｘおよびｙを繰返し抽出し、この待ち行列の中のこれらの節点を、これらのマージャーを表す新しい節点ｚ点に置き換える。ｚの頻度は７行目でｘおよびｙの頻度の総和として算出される。節点ｚはその左側の子としてｘをおよびその右側の子としてｙを有する。ｎ－１個のマージャーの後、９行目において、その待ち行列の中に残った１つの節点（その符号木の根）が戻される。

【００８６】また、これらの方法によっても、特定のゲノム領域の中で評価された全てのＳＮＰを包含する最終的な非重複ＳＮＰハプロタイプブロックのセットが得られる。本発明の方法に従ってＳＮＰ、ＳＮＰハプロタイプブロックおよびＳＮＰハプロタイプパターンを選択して得られる結果として重要なのは、幾つかの実施形態においてＳＮＰハプロタイプブロックの情報提供性の計算中に、各ＳＮＰハプロタイプブロックおよびパターン毎の情報提供ＳＮＰが決定されることである。情報提供ＳＮＰは、データ圧縮を可能とする。本発明の１つの実施形態において、ｐ個のパターンを含む各グループから少なくとも$\log_2 p$個のＳＮＰを選択する（最も近い整数に切り上げる）ことにより、遺伝子型／表現型の関連を予測するのに非常に有力な情報提供ＳＮＰからなる１つのセットが提供される。当業者であれば、他の分析において、このような部分集合を決定するために空間的に連続的な基を用いる必要はないことが分かる。例えば、本発明の幾つかの実施形態において、ＳＮＰハプロタイプブロックのそれと同じように統計学的に進められる非隣接ＳＮＰのセットはそのＤＮＡ鎖上で空間的に連続していないが、それらを同定することが望ましい。

【００８７】関連付け調査で使用されるＳＮＰハプロタイプブロックを正確に決定する（ＳＮＰ、ＳＮＰハプロタイプブロックおよびＳＮＰハプロタイプパターンの正確なベースラインを作製する）ためには、２～３以上の

個体ＤＮＡ鎖を調べる必要がある。図６は、ＳＮＰハプロタイプブロックを決定するためおよび情報提供ＳＮＰの選択のために少なくとも約５つの異なるＤＮＡ鎖を調べることの重要性を示す。図６の一番上の部分は、ＤＮＡの仮定的なストレッチの配列を示すとともに、変異体の位置を示し、変異体ブロックの境界が引かれている。しかし、ＳＮＰハプロタイプブロックの境界は最初から分かっているわけではない。配列決定の結果６１０は、３つの個体の一倍体ＤＮＡを配列決定した結果を表す。図示したように、一般に、比較的少数の個体を配列決定した後に、共通ＳＮＰの大きなフラクションを同定することが可能である。図６の場合、図６の一番上の部分にチェックマークで示した各位置のＳＮＰを同定した。

【００８８】しかし、更なる個体を評価しない場合、ブロックの境界は、この段階では正しく同定されない。例えば、この段階でブロック６２０とブロック６３０との間にブロック境界線を引くことができるが（それぞれ最初のＣからＧへの変異は最初のＧからＡへの変異を予言し、最初のＣからＴへの変異は第２のＣからＴへの変異を予言している）、この段階ではブロック６３０とブロック６４０とを区別することはできない。この段階では、最初のＣからＴへの変異は第１および第２のＴからＡへの変異を予言するようである。従って、このブロック境界線を引くためには、より統計的に有意なサンプルセットが必要となる。例えば、本発明の方法において、ＳＮＰハプロタイプブロック、ＳＮＰハプロタイプパターンおよび／または情報提供ＳＮＰを決定するために分析されるＤＮＡ鎖の数は、複数（例えば少なくとも約５個または少なくとも約１０個）である。好適な実施形態において、ＤＮＡ鎖の数は少なくとも１６である。より好適な実施形態において、ＳＮＰハプロタイプブロック、ＳＮＰハプロタイプパターンおよび／または情報提供ＳＮＰを決定するために分析されるＤＮＡ鎖の数は、少なくとも２５個である。しかし、関連するＳＮＰを同定してしまえば（即ちＳＮＰ発見を行ってしまえば）、ゲノムＤＮＡの全ストレッチを配列決定しなくても、残りのサンプル中の変異体位置のみを遺伝子タイピングしてブロック境界を同定するプロセスを完了することが可能となる。このような方法の例については、２００２年１月６日に出願された米国特許出願第１０／０４２，８１９号（代理人整理番号１０１６Ｎ－１、発明の名称「全ゲノムの走査」）を参照されたい。

【００８９】他の仮定的なゲノムサンプルの中のＳＮＰのみについて遺伝子タイピングプロセスを行った結果を、図６の６５０に示す。図示されるように、この追加的な遺伝子タイピングステップを実行することによって、ブロック６３０とブロック６４０とを区別することが可能となったことが分かる。特に、第１のＣからＴへの変異が第１および第２のＴからＡへの変異を伴わない（ｎｏｔ　ｔｒａｃｋ　ｗｉｔｈ）が、その代わりに、最初のＣからＴへの変異は第２のＣからＴへの変異のみを予言するのに用いることができ（またその逆も可能である）、および最初のＴからＡへの変異は、第２のＴからＡへの変異を予言するためだけに使用することができる（またその逆も可能である）。

【００９０】本発明の上記態様に加え、本発明の特定の実施形態は、データ解析のために曖昧なＳＮＰハプロタイプ配列を解決するために使用することができることである。例えば、ゲル配列決定操作またはアレイハイブリダイゼーション実験から得たデータでは、はっきりした結果が得られないので、ＳＮＰは曖昧である。この場合の「解決する」とは、ＳＮＰハプロタイプ配列を、そのＳＮＰハプロタイプ配列が最も密接に関係するＳＮＰハプロタイプパターンにマッチングすることにより、ＳＮＰハプロタイプ配列中の曖昧なＳＮＰ位置を解決することを意味する。さらに、「解決する」とは、データ分析から曖昧なＳＮＰハプロタイプ配列を削除することを意味する。

【００９１】曖昧なＳＮＰハプロタイプ配列を解決する１つの実施形態において、ＳＮＰハプロタイプ配列は、パターンセットに追加することが可能となるようデータセット中に配置される。このデータセットは、ＳＮＰハプロタイプパターンへの可能な割当てについて評価しようとする全てのＳＮＰハプロタイプ配列を含む。ここで図７Ａを参照すると、ステップ７１０において、データセット中のＳＮＰハプロタイプ配列は、そのパターンセット中のパターン配列と１つずつ比較される。最初はそのパターンセットの中にパターンが１つもないこともあるし、幾つかまたは全てのパターン配列が予め分かっている場合もある。ステップ７２０において、「データセットからのＳＮＰハプロタイプ配列はそのパターンセットの中のパターン配列と一致するか？」という問合せが行われる。答えが「いいえ」である場合、ステップ７３０で、評価されているＳＮＰハプロタイプ配列はそのパターンセットに追加される。答えが「はい」である場合、他の問合せ「データセットからのＳＮＰハプロタイプ配列はそのパターンセットの中の２以上のパターン配列と一致するか？」が行われる（７４０）。

【００９２】答えが「ＹＥＳ」である場合、データセットからのＳＮＰ配列は破棄されるか、または幾つかの実施形態において、更なるもしくは異なる分析のために保持される（ステップ７５０）。２番目の問合せに対する答えが「ＮＯ」である場合、ステップ７６０において、そのデータセットからのそのＳＮＰ配列は、パターンセットからのそれが一致するパターン配列と比較される。これらの２つの配列から、曖昧性の数が最も少ないＳＮＰが選択され、そのパターンセットの中に配置される（７７０）。より多い曖昧性を含むＳＮＰ配列は破棄してもいし、または幾つかの実施形態において、更なるもしくは異なるタイプの分析のために保持してもよい。

【００９３】解決プロセスは、図７Ａおよび図７Ｂを参照することにより更に理解することができる。図７Ｂにおいて、最初のＳＮＰ配列ＴＴＣＧＡが、そのパターンセットの中に含まれる配列と比較される（ステップ７１０）。この時点で、そのパターンセットの中にパターン配列が１つも含まれていない場合、ＴＴＣＧＡはそのパターンセットの中のどのパターン配列とも一致しないことになる。従って、このＳＮＰ配列ＴＴＣＧＡの存在はそのデータセットから削除され（または他の分析用に保持され）、そのパターンセットに追加される（７３０）。これで、そのパターンセットは１つのパターン配列ＴＴＣＧＡを有することになる。

【００９４】再び図７Ｂを見ると、そのデータセットの中の第２のＳＮＰ配列「Ｔ？Ｃ？？」が、そのパターンセットの中に含まれる配列と比較される（ステップ７１０）。この時、そのパターンセットの中には１つのパターン配列「ＴＴＣＧＡ」があり、そして「Ｔ？Ｃ？？」はこの配列に一致する（ステップ７２０）。このときそのパターンセットの中には１つのパターン配列「ＴＴＣＧＡ」しかないので、２番目の問合せ「ＳＮＰ配列「Ｔ？Ｃ？？」はそのパターンセットの中の２以上のパターン配列と一致するか？」（７４０）に対する答えは「いいえ」である。ステップ７６０で、「Ｔ？Ｃ？？」は「ＴＴＣＧＡ」と比較され、どちらの配列がより多い曖昧性を有するかを決定する。これは明らかに「Ｔ？Ｃ？？」なので、「ＴＴＣＧＡ」はそのパターンセットの中に保持され（７７０）、「Ｔ？Ｃ？？」は破棄されるかまたは更なる分析のために保持される。

【００９５】図７Ｂのデータセットの３番目の配列は「Ｃ？？？？」である。「Ｃ？？？？」はまずＴＴＣＧＡと比較され（ステップ７１０）、「ＴＴＣＧＡ」とは一致しないことが分かったので（７２０）、そのパターンセットに追加される（７３０）。図７Ｂの中の４番目の配列は「ＣＴＡＣＡ」である。「ＣＴＡＣＡ」は「ＴＴＣＧＡ」および「Ｃ？？？？」（そのパターンセットの中のパターン配列）と比較され（ステップ７１０）、そして「Ｃ？？？？」と一致することが分かる（７２０）。ここで２番目の問合せ「「ＣＴＡＣＡ」は「Ｃ？？？？」および「ＴＴＣＧＡ」の両方に一致するか？」が行われる（７４０）。答えは「いいえ」なので、次に「Ｃ？？？？」および「ＣＴＡＣＡ」が比較され（７６０）、曖昧性が最も少ない配列（この場合「ＣＴＡＣＡ」）がそのパターンセットの中に保持され、「Ｃ？？？？」は破棄（分析から削除）されるかまたは更なる分析用に保持される（７７０）。

【００９６】図７Ｂのデータセット中の５番目のＳＮＰ配列は「？Ｔ？？Ａ」である。このＳＮＰ配列はパターン配列「ＴＴＣＧＡ」および「ＣＴＡＣＡ」と比較され（７１０）、「ＴＴＣＧＡ」および「ＣＴＡＣＡ」の両方と一致することが分かる。従って、問合せ「「？Ｔ？？Ａ」はそのパターンセットの中の２以上のパターン配列と一致するか？」（７４０）に対する答えは「ＹＥＳ」である。ステップ７５０において、ＳＮＰ配列「？Ｔ？？Ａ」は、更なる分析のために保持されるか、または破棄（分析から削除）される。他の解決方法は、例えばもし１つのパターン配列が「ＣＣＡＴＴ？」でありデータセットからのＳＮＰ配列が「Ｃ？ＡＴＴＧ」である場合、これらの配列を曖昧性を解決するために「組み合わせ」（ＣＣＡＴＴＧ）、この「組み合わせ」た配列をそのパターンセットに追加することを可能とする。更なるアレイハイブリダイゼーション、配列決定法、または当分野において公知である他の手法を用いて、曖昧なＳＮＰヌクレオチド位置を分析することができる。

【００９７】ＳＮＰハプロタイプブロックおよびパターンと表現型との関連付け

同定されたＳＮＰハプロタイプブロック、ＳＮＰハプロタイプパターンおよび／または情報提供ＳＮＰは、様々な遺伝子分析に用いることができる。例えば、情報提供ＳＮＰを同定してしまえば、これらを関連付け調査のための様々なアッセイにおいて用いることができる。例えば、これらの情報提供ＳＮＰを尋問するマイクロアレイ用のプローブを設計することができる。他の例示的アッセイとしては、例えばＴａｑｍａｎアッセイおよびＩｎｖａｄｅｒアッセイ（上記に記載）ならびに従来のＰＣＲおよび／または配列決定手法が挙げられる。

【００９８】幾つかの実施形態において、図１のステップ１７０に記載したように、同定されたハプロタイプパターンを上記アッセイで用いて、関連付け調査を行うことができる。これは、目的の表現型を有する個体（例えば特定の疾患を示す個体または薬物の投与に対して特定の反応を示す個体）においてハプロタイプパターンを決定し、これらの個体におけるそのハプロタイプパターンの頻度を、対照個体群における該ハプロタイプパターンの頻度と比較することにより、達成することができる。好ましくは、このようなＳＮＰハプロタイプパターンの決定は、ゲノム全体で行われるが、ゲノムの特定の領域のみに関心がある場合は、これらの特定の領域のＳＮＰハプロタイプパターンが用いられる。本発明の方法の本明細書中に開示される他の実施形態に加え、これらの方法はさらに、表現型の「詳細な分析（ｄｉｓｓｅｃｔｉｏｎ）」も可能とする。つまり、特定の表現型は２以上の異なる遺伝的根拠から生じる場合がある。例えば、ある個体における肥満は、Ｘ遺伝子中のある欠陥により生じたものかもしれないが、他の個体における肥満表現型は、Ｙ遺伝子およびＺ遺伝子における突然変異により生じたものである場合もある。このように、本発明のゲノム走査能力（ｇｅｎｏｍｅ　ｓｃａｎｎｉｎｇ　ｃａｐａｂｉｌｉｔｙ）により、類似した表現型についての様々な遺伝的根拠の吟味を可能とする。ゲノムの特定の領域が特定の表現型と関連性があることが同定されたな

ら、これらの領域は薬剤発見標的として（図１のステップ１８０）、または診断マーカーとして（図１のステップ１９０）使用することができる。

【００９９】先の段落に記載したように、関連付け調査を行う１つの方法は、目的の表現型を示す個体におけるＳＮＰハプロタイプパターンの頻度を、対照個体群におけるＳＮＰハプロタイプパターンの頻度と比較することである。好適な実施形態において、情報提供ＳＮＰは、ＳＮＰハプロタイプパターン比較を行うために用いられる。情報提供ＳＮＰを用いるアプローチは、現在までに当分野で公知である他の全ゲノム走査もしくは遺伝子タイピング法よりも大きな利点を有する。というのは、各個体のゲノムの３０億個の塩基全てを読み取る（または見つけられる３００～４００万個の共通ＳＮＰを読む）代わりに、サンプル集団からの情報提供ＳＮＰのみを決定すれば良いからである。これらの特定の情報提供ＳＮＰを読み取れば、上記のような特定の実験集団からの統計学的に正確な関連付けデータの抽出を可能するのに十分な情報が得られる。

【０１００】図８は、本発明の方法を用いて遺伝学的関連性を決定する１つの方法のある実施形態を示す。ステップ８００では、対照集団のゲノムについて情報提供ＳＮＰの頻度を決定する。ステップ８１０では、臨床集団のゲノムについて情報提供ＳＮＰの頻度を決定する。ステップ８００および８１０は、個体の集団において情報提供ＳＮＰを分析するための上記ＳＮＰアッセイを用いることにより行うことができる。ステップ８２０において、ステップ８００および８１０から得た情報提供限定ＳＮＰの頻度が比較される。頻度の比較は、例えば各集団における各情報提供性ＳＮＰ位置におけるマイナーな対立遺伝子の頻度（特定のマイナー対立遺伝子を有する個体数を全個体数で割った）を決定し、これらのマイナー対立遺伝子頻度を比較することにより行うことができる。ステップ８３０において、対照集団および臨床集団における発生頻度の差を示す情報提供ＳＮＰを選択して分析する。情報提供ＳＮＰを選択したら、この情報提供ＳＮＰを含むＳＮＰハプロタイプブロックが同定され、これを用いて目的のゲノム領域を同定する（ステップ８４０）。これらのゲノム領域を当分野で公知の遺伝学的方法または生物学的方法により分析し（ステップ８５０）、そしてこれらの領域が薬剤発見標的として（ステップ８６０）または診断マーカーとして（ステップ８７０）使用することができるかどうかについて分析する（以下に詳細に説明する）。

【０１０１】同定したゲノム配列の使用

ゲノム中の１以上の遺伝子座を特定の表現型特徴、例えば疾患罹患性遺伝子座等に関連付けたら、その特徴に関係のあるこれらの遺伝子または調節エレメントを同定することができる。次にこれらの遺伝子または調節エレメントは、疾患の治療のための治療標的として用いることができる（図１のステップ１８０または図８のステップ８６０に記載）。本発明の方法により同定されたゲノム配列は、遺伝子配列（ｇｅｎｉｃ　ｓｅｑｕｅｎｃｅ）または非遺伝子配列（ｎｏｎｇｅｎｉｃ　ｓｅｑｕｅｎｃｅ）であっても良い。「遺伝子」という用語は、特定のポリペプチドをコードするオープンリーディングフレーム（ＯＲＦ）、イントロン領域、ならびに該コード領域の範囲を超えて最長で約１０ｋｂにもなる遺伝子の発現の調節に関与する隣接する５’および３’側の非コードヌクレオチド配列（ただし、いずれかの方向にもっと長く延びる場合もある）を意味するものとする。同定された遺伝子のＯＲＦは、タンパク質構造に影響を及ぼすことにより疾患の状態に影響を及ぼし得る。あるいは、同定された遺伝子の非コード配列または非遺伝子配列は、タンパク質の発現レベルまたは発現特異性に影響することにより、疾患の状態に影響を及ぼし得る。一般に、ゲノム配列は、同定された遺伝子を、該遺伝子配列を含まない他の核酸配列を実質的に含まないように単離することによって、調査される。該ＤＮＡ配列は、様々に利用することができる。例えば、該ＤＮＡを用いて生物学的試料における該遺伝子の発現を検出または定量することができる。特定のヌクレオチド配列の存在について細胞を走査する方法は、文献に十分に記載されており、ここで詳細に述べる必要がないが、例えばＳａｍｂｒｏｏｋらのＭｏｌｅｃｕｌａｒ　Ｃｌｏｎｉｎｇ：Ａ　Ｌａｂｏｒａｔｏｒｙ　Ｍａｎｕａｌ（Ｃｏｌｄ　Ｓｐｒｉｎｇ　Ｈａｒｂｏｒ　Ｌａｂｏｒａｔｏｒｙ，　Ｎｅｗ　Ｙｏｒｋ）（１９８９）を参照されたい。

【０１０２】さらに、該遺伝子の配列（フランキングプロモーター領域およびコード領域を含む）を当分野で公知である様々な方法で突然変異させて、コードされるタンパク質の発現レベルまたは配列等に、意図する変化をもたらすことができる。この配列変化は、置換、挿入、トランスロケーションまたは欠失であってもよい。欠失は、あるドメインまたはエキソンの全体的欠失などの大きな変化を含み得る。クローニングされた遺伝子のインビトロ突然変異誘発法は公知である。部位特異的突然変異誘発のためのプロトコールの例は、Ｇｕｓｔｉｎら、Ｂｉｏｔｅｃｈｎｉｑｕｅｓ　１４：２２（１９９３）；　Ｂａｒａｎｙ，　Ｇｅｎｅ　３７：１１１－２３（１９８５）；　Ｃｏｌｉｃｅｌｌｉら、Ｍｏｌ．Ｇｅｎ．　Ｇｅｎｅｔ．　１９９：　５３７－９（１９８５）；　Ｐｒｅｎｔｋｉら、Ｇｅｎｅ　２９：３０３－１３（１９８４）；　Ｓａｍｂｒｏｏｋら、Ｍｏｌｅｃｕｌａｒ　Ｃｌｏｎｉｎｇ：Ａ　Ｌａｂｏｒａｔｏｒｙ　Ｍａｎｕａｌ（Ｃｏｌｄ　Ｓｐｒｉｎｇ　Ｈａｒｂｏｒ　Ｐｒｅｓｓ）ｐｐ．　１５．３－１５．１０８（１９８９）；　Ｗｅｉｎｅｒら、Ｇｅｎｅ　１２６：３５－４１（１９９３）；　Ｓａｙｅｒｓら、Ｂｉ

ｏｔｅｃｈｎｉｑｕｅｓ　１３：５９２：６（１９９２）；　ＪｏｎｅｓおよびＷｉｎｉｓｔｏｒｆｅｒ，Ｂｉｏｔｅｃｈｎｉｑｕｅｓ　１２：５２８－３０（１９９２）；およびＢａｒｔｏｎら、Ｎｕｃｌｅｉｃ　Ａｃｉｄｓ　Ｒｅｓ．　１８：７３４９－５５（１９９０）に記載されている。このような突然変異遺伝子を用いて、そのタンパク質産物の構造／機能の関係を調査したり、またはその機能または調節に影響を及ぼす該タンパク質の特性を変更したりすることができる。

【０１０３】同定された遺伝子を用いて、得られるポリペプチドの全てまたは一部を産生することができる。タンパク質産物を発現するために、同定された遺伝子を組み込む発現カセットを用いても良い。発現カセットまたはベクターは、一般には転写および翻訳開始領域（誘導可能または構成的なもの）を提供し、この場合、コード領域は、転写開始領域ならびに転写および翻訳停止領域の転写制御下において機能上連結されている。これらの制御領域は、同定された遺伝子にもともと備わっているものであっても良いし、また外来起源に由来するものであってもよい。

【０１０４】該ペプチドは、従来法に従って、発現の目的に応じて原核生物または真核生物内で発現させることができる。該タンパク質の大規模生産を行う場合、大腸菌、Ｂ．　ｓｕｂｔｉｌｉｓ、Ｓ．　ｃｅｒｅｖｉｓｉａｅ等の単細胞生物、バキュロウイルスベクターと一緒に用いた昆虫細胞、または脊椎動物（特に哺乳動物）等のより高等な生物の細胞（例えばＣＯＳ７細胞）を発現宿主細胞として用いることができる。多くの状況において、真核生物中で該遺伝子を発現させることが望ましい。なぜなら、真核生物では、該遺伝子は天然の折畳み構造および翻訳後修飾による恩恵を得るであろうからである。また、小さなペプチドは実験室で合成することもできる。タンパク質またはその断片が入手できれば、従来方法に従って、そのタンパク質を単離および精製することができる。発現宿主のライゼートを調製し、ＨＰＬＣ、排除クロマトグラフィー、ゲル電気泳動法、アフィニティークロマトグラフィーまたは他の精製手法を用いて、該タンパク質もしくはその断片を精製することができる。

【０１０５】発現されたタンパク質を抗体産生に用いることができる。この場合、短い断片は特定のポリペプチドに対して特異的な抗体（モノクローナル抗体）の発現を誘導し、より大きな断片または全タンパク質は、該ポリペプチドの全長に対する抗体（ポリクローナル抗体）の産生を可能とする。抗体は従来法に従って調製され、この場合、発現されたポリペプチドまたはタンパク質は、そのまま、もしくは既知の免疫原性キャリア（例えばＫＬＨ、ｐｒｅ－Ｓ　ＨＢｓＡｇ、他のウイルスタンパク質または真核生物タンパク質等）とコンジュゲートさせて、免疫原として用いられる。様々なアジュバントを適当に、一連の注射を注射と共に用いることができる。モノクローナル抗体の場合、１以上の追加抗原刺激注射後に、脾臓を単離し、リンパ球を細胞融合により不死化させ、親和性の高い抗体結合についてスクリーニングする。次に、所望の抗体を産生する不死化細胞（すなわちハイブリドーマ）を増殖する。さらに詳しいことについてはＭｏｎｏｃｌｏｎａｌ　Ａｎｔｉｂｏｄｉｅｓ：　Ａ　Ｌａｂｏｒａｔｏｒｙ　Ｍａｎｕａｌ，　ＨａｒｌｏｗおよびＬａｎｅ編（Ｃｏｌｄ　Ｓｐｒｉｎｇ　Ｈａｒｂｏｒ　Ｌａｂｏｒａｔｏｒｉｅｓ，　Ｃｏｌｄ　Ｓｐｒｉｎｇ　Ｈａｒｂｏｒ，　Ｎ．Ｙ．）（１９８８）を参照されたい。所望により、重鎖および軽鎖をコードするｍＲＮＡを単離し、大腸菌内でのクローニングにより突然変異させて、これらの重鎖および軽鎖を混合して該抗体の親和性を更に強化することができる。インビボでの免疫化に代わる抗体の作製方法としては、ファージ「提示」ライブラリーへの結合（通常はインビトロ・アフィニティー熟成を伴う）が挙げられる。

【０１０６】同定された遺伝子、遺伝子断片またはこれにコードされるタンパク質もしくはタンパク質断片は、変性疾患および他の疾患を治療するための遺伝子療法において有用であり得る。例えば、発現ベクターを用いて同定された遺伝子を細胞内に導入することができる。このようなベクターは一般に、プロモーター配列近くに位置する、受容者（即ち宿主）のゲノム中への核酸配列の挿入をもたらすに便利な制限部位を有する。転写開始領域、標的遺伝子もしくはその断片、および転写終結領域を含む転写カセットを調製することができる。転写カセットは、例えばプラスミド、レトロウイルス（レンチウイルス等）、アデノウイルス等の様々なベクターの中に導入することができ、該ベクターは、細胞内で一時的または安定に維持されることができる。該遺伝子またはタンパク質産物は、様々な経路（例えばウイルス感染、マイクロインジェクションまたは小胞の融合（ｆｕｓｉｏｎ　ｏｆ　ｖｅｓｉｃｌｅ）等）により組織または宿主細胞中に直接導入することができる。また筋肉内投与のためにジェット注射を用いることもできる（Ｆｕｒｔｈら、Ａｎａｌ．　Ｂｉｏｃｈｅｍ，　２０５：３６５－６８（１９９２）に記載）。あるいは、ＤＮＡを金微粒子にコーティングして、粒子衝撃デバイス（ｐａｒｔｉｃｌｅ　ｂｏｍｂａｒｄｍｅｎｔ　ｄｅｖｉｃｅ）または「遺伝子銃」（文献記載）により皮内に送達することができる（例えばＴａｎｇら、Ｎａｔｕｒｅ，　３５６：１５２－５４（１９９２）を参照されたい）。

【０１０７】アンチセンス分子を用いて、細胞内における同定遺伝子の発現をダウンレギュレートすることができる。アンチセンス試薬は、アンチセンスオリゴヌクレオチド、特に化学修飾を施した合成アンチセンスオリゴヌクレオチド、またはこのようなアンチセンス分子を発現する核酸構築物（ＲＮＡなど）であってもよい。アン

チセンス分子の組合せ（この組合せは多数の異なる配列を含み得る）を投与することもできる。

【０１０８】アンチセンスインヒビターの代わりに、触媒的核酸化合物（例えばリボザイムやアンチセンスコンジュゲートなど）を用いて遺伝子発現を阻害することができる。リボザイムをインビトロで合成して患者に投与しても良いし、または発現ベクター上にコードさせて、そこからリボザイムを標的細胞内で合成してもよい（例えば国際特許出願公開番号WO９５／２３２２５号）およびＢｅｉｇｅｌｍａｎら、Ｎｕｃｌ．　Ａｃｉｄｓ　Ｒｅｓ．　２３：４４３４－４２（１９９５）を参照されたい）。触媒活性を有するオリゴヌクレオチドの例は、国際特許出願公開番号第WO９５／０６７６４号に記載されている。金属錯体とアンチセンスオリゴヌクレオチドとのコンジュゲート（例えばmRNAの加水分解を媒介することができるｔ－ピリジルＣｕ（ＩＩ）（ｔｅｒｐｙｒｉｄｙｌ　Ｃｕ（ＩＩ））等）は、Ｂａｓｈｋｉｎら、Ａｐｐｌ．　Ｂｉｏｃｈｅｍ．　Ｂｉｏｔｅｃｈｎｏｌ．　５４：４３－５６（１９９５）に記載されている。同定された遺伝子配列は遺伝子療法のために用いる以外に、同定された核酸を用いて遺伝子修飾された非ヒト動物を作製して疾患動物モデルを作製したり、またはタンパク質の機能および調節を調査するために細胞系において部位特異的遺伝子修飾を行うことができる。「トランスジェニック」という用語は、宿主細胞中に安定に送達された外来遺伝子を有する遺伝子修飾された動物を含むものとする。この場合、宿主細胞中において、例えば、修飾タンパク質を産生するために該遺伝子の配列を変更してもよいし、または該遺伝子は外来プロモーターに機能上連結されたレポーター遺伝子であってもよい。内因性性遺伝子座を変化、置き換えまたは破壊する相同性組換えにより、トランスゲニック動物を作製しても良い。あるいは、核酸構築物を無作為にゲノム中に組み込んでも良い。安定な組込みのためのベクターとしては、プラスミド、レトロウイルスおよび他の動物ウイルス、ＹＡＣ等が挙げられる。関心が持たれるものとしては、トランスジェニック哺乳動物（例えばウシ、ブタ、ヤギ、ウマ等、および特にラットやマウス等のげっ歯類）である。

【０１０９】また遺伝子機能の調査は、非哺乳動物モデル、特に生物学的および遺伝学的に十分に特徴付けられている非哺乳類生物（例えばＣ．　ｅｌｅｇａｎｓ、Ｄ．ｍｅｌａｎｏｇａｓｔｅｒおよびＳ．　ｃｅｒｅｖｉｓｉａｅ等）を用いることもできる。タンパク質の機能に関与する生理学的および生化学的経路を決定するために、対象となる遺伝子配列を用いて、対応する遺伝子機能をノックアウトしたり、または明らかにされた遺伝子病変を補うことができる。例えば変性疾患の進行を調査したり、治療法をテストしたり、または薬剤発見のために、補足もしくはノックアウト調査と共に、薬物スクリーニングを行ってもよい。

【０１１０】更に、改変された細胞または動物は、タンパク質の機能および調節の調査において有用である。例えば、同定された遺伝子の中で一連の小さな欠失および／または置換を行い、酵素活性、細胞輸送または局在化などにおける異なるドメインの役割を決定することができる。目的の具体的な構築物としては、遺伝子発現、ドミナント－ネガティブ遺伝子突然変異の発現、および同定遺伝子の過剰発現を遮断するアンチセンス構築物が挙げられるがこれらに限定されない。また、ある細胞もしくは組織内において通常は発現されないまたは異常な発生時期において発現される同定遺伝子またはその変異体を、該細胞もしくは組織内で発現させることもできる。さらに、ある細胞の中でその細胞中では通常は産生されないタンパク質を発現させることにより、そのタンパク質の正常な機能に関する情報を提供する細胞内挙動の変化を誘導することができる。

【０１１１】構造／機能パラメータを調査するためにタンパク質分子を分析することができる。例えば、ある同定遺伝子のタンパク質産物を大量に産生することにより、そのタンパク質産物に結合するまたは該タンパク質産物の機能をモジュレートもしくは模倣する基質またはリガンドを同定することができる。薬物スクリーニングにより、例えば影響を受けた細胞内におけるタンパク質機能に取って代わるもしくは増強する薬剤、またはタンパク質機能をモジュレートもしくは取り消す薬剤が同定される。本明細書中で用いられる「薬剤」という用語は、同定遺伝子または遺伝子産物の生理学的機能を直接または間接的に変更、模倣またはマスキングすることができるあらゆる分子（例えばタンパク質や小分子等）を指す。一般に、様々な濃度に応じて異なる反応を得るために、複数のアッセイ混合物を異なる濃度の薬剤と平行に泳動させる。典型的には、これらの濃度の１つが陰性対照（即ちゼロ濃度または検出レベル未満）として使用される。

【０１１２】この目的のために、標識インビトロタンパク質／タンパク質結合アッセイ、タンパク質－ＤＮＡ結合アッセイ、電気泳動移動度シフトアッセイ、タンパク質結合のイムノアッセイ等を含む様々なアッセイを使用することができる。また、精製されたタンパク質の全てまたはその断片を用いて、三次元結晶構造を決定することもできる。この三次元結晶構造は、タンパク質またはその一部の生物学的機能を決定するため、分子間の相互作用をモデリングするため、または膜融合等のために使用することができる。

【０１１３】候補となる薬剤は、多くの化学クラスを包含するが、典型的には、このような薬剤は有機分子や複合体（好ましくは分子量が５０ダルトンを超え且つ約２，５００ダルトン未満である小さな有機化合物）である。候補薬剤は、タンパク質との構造的相互作用に必要

な官能基、特に水素結合を含み、典型的には、少なくとも１つのアミン、カルボニル、ヒドロキシル、またはカルボニル基、およびしばしばこれらの官能性化学基のうちの少なくと２つを含む。候補薬剤はしばしば、炭素環または複素環式構造および／または芳香族もしくは多環芳香族構造が、１以上の上記官能基で置換されている。また候補薬剤は、生体分子（ペプチド、糖類、脂肪酸、ステロイド、プリン、ピリミジン、およびこれらの誘導体、構造的類似体または組合せ等を含むがこれらに限定されない）の中にも見られる。

【０１１４】候補薬剤は、合成もしくは天然化合物のライブラリーを含む様々な起源から得られる。例えば、種々の有機化合物および生体分子を無作為的および特異的に合成するためには、無作為なオリゴヌクレオチドおよびオリゴペプチドの発現を含む様々な手段が利用可能である。あるいは、細菌、真菌、植物および動物のエキスといった形態の天然化合物のライブラリーが入手可能であり、また簡単に生成することができる。更に、天然のまたは合成したライブラリーおよび化合物を、従来の化学的、物理的および生化学的手段により簡単に修飾し、これを用いてコンビナトリアルライブラリーを作製することができる。既知の薬物を、特異的もしくは無作為な化学修飾（例えばアシル化、アルキル化、エステル化、アミド化等）にかけて構造的類似体を作製することができる。

【０１１５】スクリーニングアッセイが結合アッセイである場合、該分子の１以上を、検出可能シグナルを直接もしくは間接的に発することができる標識に結合させることができる。様々な標識として、放射性同位体、蛍光体、化学発光体、酵素、特異的結合分子、粒子（例えば磁性粒子）等が挙げられる。特異的結合分子としては、ビオチンとストレプトアビジン、ジゴキシンと抗ジゴキシン抗体等のペアが挙げられる。特異的結合メンバーの場合、通常は相補的メンバーを、既知の手法によって検出され得る分子で標識する。種々の他の試薬をスクリーニングアッセイに含めることができる。これらには、最適なタンパク質－タンパク質結合を容易とするため、および／または非特異的もしくはバックグラウンド相互作用を低減するために使用される、塩、中性タンパク質（例えばアルブミン等）、洗浄剤等の試薬が含まれる。アッセイの効率を上げる試薬（例えばプロテアーゼインヒビター、ヌクレアーゼインヒビター、抗菌剤等）を使用してもよい。

【０１１６】薬剤は、薬学的に許容される担体（あらゆる全ての溶剤、分散媒、コーティング、抗酸化剤、等張性剤、吸収遅延剤等と組み合わせても良い。薬学的に活性な物質のためのこのような媒体および物質の使用は、当分野において周知である。該活性物質に対して従来の媒体または薬剤がいずれも適合しない場合でない限り、本明細書中に記載される治療的組成物および治療方法において該活性物質が使用できると考えられる。また、組成物中に補助的な活性成分を組み込むこともできる。

【０１１７】様々な投与法で使用するための製剤を調製することができる。製剤は、経口投与、吸入、または注射（例えば静脈内、腫瘍内、皮下内、腹膜内、筋肉内等）により投与することができる。治療製剤の投薬量は、疾患の性質、投与頻度、投与方法、宿主からの薬剤のクリアランス等に応じて広く異なる。最初の投与量を多くして、後により少ない維持量を投与してもよい。有効な投薬レベルを維持するために、用量を１週間または２週間に１回といった少ない頻度で投与しても良いし、あるいは用量を小分けにして、毎日または週２回等で投与してもよい。幾つかのケースにおいて、経口投与は静脈内投与の場合とは異なる用量を必要とする。本発明の同定薬剤は、様々な治療投与用製剤に組み込むことができる。より具体的には、適切な薬学的に許容される担体もしくは希釈剤と組合せた複合体を医薬組成物として調製し、および固体、半固体、液体または気体状の調製物（例えば錠剤、カプセル、粉末、顆粒、軟膏、溶液、坐剤、注射液、吸入剤、ゲル、微小球、エアロゾル等）として製剤化することができる。このように、該薬剤の投与は様々な方法で行うことができる。薬剤は、投与後に全身をめぐるものであってもよいし、また内植した部位に活性用量を維持する働きをするインプラントを用いて局所化させてもよい。

【０１１８】以下の方法および賦形剤は単に例示的なものであり、限定的なものではない。経口用調製物の場合、薬剤を単独または適当な添加物（例えばラクトース、マンニトール、コーンスターチまたはポテトスターチ等の従来添加物；結晶セルロース、セルロース誘導体、アカシアゴム、コーンスターチまたはゼラチン等の結合剤；コーンスターチ、ポテトスターチまたはカルボキシルメチルセルロースナトリウム等の崩壊剤；タルクまたはステアリン酸マグネシウム等の潤滑剤；ならびに、場合により、希釈剤、緩衝剤、湿潤剤、保存剤および着香料等）と一緒に使用して、錠剤、粉末、顆粒またはカプセルを作製することができる。

【０１１９】さらに、薬剤は、水系もしくは非水系溶剤（例えば植物油や他の類似油、合成脂肪族系酸グリセリド、高級脂肪族系酸のエステルまたはプロピレングリコール等）中にこれらを溶解、懸濁または乳化し、場合により従来添加物（例えば可溶化剤、等張剤、懸濁剤、乳化剤、安定化剤および保存剤）を添加することによって、注射用調製物として製剤化してもよい。さらに薬剤は、吸入により投与されるエアロゾル製剤に入れて使用してもよい。本発明により同定された物質は、圧縮された許容可能な推進剤（例えばジクロロジフルオロメタン、プロパン、窒素等）中に製剤化することができる。あるいは、薬剤は様々な基剤（例えば乳化基剤や水溶性基剤等）と混合して坐剤として調製してもよい。さら

に、本発明の同定物質は、坐剤として直腸内に投与することができる。坐剤は、ココアバター、カーボワックスおよびポリエチレングリコールなどのビヒクル（体温で溶け且つ室温では固化する）を含み得る。

【０１２０】持続放出性製剤用のインプラントは当分野において周知である。インプラントは、生分解性もしくは非生分解性高分子を用いて微小球、スラブ等として製剤化される。例えば、乳酸および／グリコール酸のポリマーは、宿主による寛容性が高い侵食性ポリマーを形成する。本発明の同定薬剤を含むインプラントは、薬剤の局所濃度が身体の他の部分に比べて高くなるように、作用部位の近くに設置することができる。シロップ、エリキシルおよび懸濁液等の、経口もしくは直腸内投与用の単位用量剤形を提供することができる。この場合、各投与ユニット（例えばティースプーン１杯分、テーブルスプーン１杯分、ゲルカプセル、錠剤または坐剤等）は所定量の本発明の組成物を含む。同様に、注射または静脈内投与用の単位用量剤形は、滅菌水、食塩水、または他の製薬上許容可能な担体中の溶液としての組成物中に本発明の組成物を含み得る。本発明の新規な単位用量剤形のための仕様は、使用される具体的な化合物、達成したい効果、ならびに宿主中における各活性薬剤に関係する薬理学的作用に応じて異なる。

【０１２１】薬学的に許容される賦形剤（例えばビヒクル、アジュバント、担体または希釈剤等）は一般に簡単に入手可能である。さらに、薬学的に許容される補助剤（例えばｐＨ調節および緩衝剤等、張度調節剤、安定化剤、浸潤剤等）も、一般に簡単に入手可能である。

【０１２２】同定薬剤の治療的用量を、疾患または障害を患う宿主に投与する。投与は、具体的な疾患に応じて、表面塗布または局所もしくは全身投与であってもよい。化合物は、適当な期間の間その疾患の進行が実質的に抑えられるような有効用量で投与される。インビボで使用する場合は、組成物は医師の指導にもとに獲得および使用されるものとする。用量は使用される特定の薬剤および製剤、疾患のタイプ、患者の状態等に応じて、副作用を最小限に抑えつつその疾患もしくは症状を緩和するのに十分な量として選択される。治療は短期的なもの（例えば外傷後の治療等）であっても長期的なもの（例えば精神分裂病の予防または治療等）であってもよい。

【０１２３】本発明により同定されるＳＮＰを用いて、関連する遺伝子の発現パターンおよびその生物の表現型特徴（例えば疾患罹患性または薬物応答性）に関係のある発現パターンを分析することができる。様々な組織における発現パターンを決定し、遍在的な発現パターン、組織特異的発現パターン、一時的発現パターン、および様々な外的刺激（例えば化学物質または電磁放射等）により誘導される発現パターンを同定するために使用することができる。このような決定は、その遺伝子および／またはそのタンパク質産物の機能に関する情報を提供する。

【０１２４】また新しく同定された配列は、診断マーカーとして（すなわち疾患疾患罹患性または薬物応答性などの表現型特徴を予測するために）使用することもできる。さらに、本発明の方法は、臨床実験のために集団を等級別に分類するために使用することができる。したがって、該遺伝子またはその断片は、テストされている生物のゲノム中に同じ核酸配列が存在するか否かを決定するためのプローブとして用いることができる。さらに該プローブは、その生物の特定の表現型特徴に相関性を有し得るマーカーの発現レベルを決定するために、テストされる生物の体内またはその一部（例えば特定の組織や器官等）におけるＲＮＡもしくはｍＲＮＡレベルをモニターするために用いることができる。同様に、該マーカーは、免疫学的方法（ウェスタンブロットや放射線免疫沈降法等）等の慣習的手法、または遺伝子産物に関係する活性を測定するための活性ベースのアッセイを用いて、タンパク質レベルで分析することができる。さらに、異なる遺伝的根拠を有する類似疾患の間で表現型が明確に区別できない場合、本発明の方法を用いてその疾患を正確に同定することができる。

【０１２５】また、本発明の方法をヒト以外の生物に対して用いることができることも明白であろう。例えば、その生物が動物である場合、本発明の方法は、例えば疾患に対する耐性／もしくは疾患罹患性、環境耐性（ｅｎｖｉｒｏｎｍｅｎｔａｌ　ｔｏｌｅｒａｎｃｅ）、薬物応答などに関係する遺伝子座を同定するために使用することができる。またその生物が植物である場合、本発明の方法は、疾患に対する耐性／もしくは疾患罹患性、環境耐性、および／または除草剤耐性などに関係する遺伝子座を同定するために使用することができる。

【０１２６】本発明は、記載された特定の方法論、プロトコール、細胞系、動物の種や属、ならびに試薬に限定されず、様々である、ということを理解されたい。また、本明細書中で使用される用語は、単に特定の実施形態を説明するためのものであって、本発明の範囲を限定することを意図するものではなく、本発明の範囲は特許請求の範囲によってのみ限定されるものとする。

【０１２７】データベース

本発明は、変異に関する情報、例えばＳＮＰ、ＳＮＰハプロタイプブロック、ＳＮＰハプロタイプパターンおよび情報提供ＳＮＰに関する情報を含むデータベースを含む。幾つかの実施形態において、本発明のデータベースは、１以上の表現型特性に関係のある１以上のハプロタイプパターンについての情報を含み得る。またデータベースは、所与の変異に関する情報、例えば変異が生じる全体的な（ｇｅｎｅｒａｌ）ゲノム領域についての記述的情報（例えば変異が既知の遺伝子の中に位置するか否か、近くに既知の遺伝子、遺伝子相同体または調節領域があるか否か等）等も含み得る。

【０１２８】本発明のデータベースに含まれ得る他の情報には、ＳＮＰ配列情報、ＳＮＰハプロタイプパターンについて分析される組織サンプルの臨床状態に関する記述的情報、またはそのサンプルが由来する患者の臨床状態が含まれるが、これらに限定されない。データベースは、異なるパーツ、例えば変異データベース、ＳＮＰデータベース、ＳＮＰハプロタイプブロックもしくはＳＮＰハプロタイプパターンデータベース、および情報提供ＳＮＰデータベース等を含むように設計することができる。データベースの構成および構築法は広く入手可能であり、例えばＡｋｅｒｂｌｏｍら（１９９９）、米国特許第５，９５３，７２７号（本明細書中に参考として全て組み込まれる）を参照されたい。

【０１２９】本発明のデータベースは、外側または外部のデータベースにリンクされていてもよい。図９は、該データベースに適しおよび本発明のソフトウェアを実行するコンピュータネットワークの例を示す。コンピュータワークステーション９０２は、イーサネット（登録商標）９０５等のローカルエリアネットワーク（ＬＡＮ）を介してアプリケーション／データサーバ９０６に接続される。プリンタ９０４はワークステーションに直接またはイーサネット９０５に接続されていてもよい。ＬＡＮは、ゲートウェイサーバ９０７（WAN９０８とＬＡＮ９０５との間でファイアウォールとしても機能し得る）を介してインターネット９０８などの広域ネットワーク（WAN）に接続され得る。好適な実施形態において、ワークステーションは、インターネット９０８を介してＴｈｅ　ＳＮＰ　Ｃｏｎｓｏｒｔｉｕｍ　（ＴＳＣ）またはｔｈｅ　Ｎａｔｉｏｎａｌ　Ｃｅｎｔｅｒ　ｆｏｒ　Ｂｉｏｔｅｃｈｎｏｌｏｇｙ　Ｉｎｆｏｒｍａｔｉｏｎ　９０９等の外部のデータソースと通信れていてもよい。

【０１３０】任意の適当なコンピュータプラットフォームを用いて、ＳＮＰハプロタイプブロックもしくはパターン、関連する表現型、そのデータベースの中の他の情報または入力された情報の間で必要な比較を行うことができる。例えば、様々な製造業者から多数のコンピュータワークステーションが入手可能であり、例えばＳｉｌｉｃｏｎ　Ｇｒａｐｈｉｃｓから入手可能なものがある。また、クライアント・サーバ環境、データベースサーバおよびネットワークも広く入手可能であり、本発明のデータベースに適したプラットフォームである。

【０１３１】また、本発明のデータベースは、個体におけるＳＮＰハプロタイプパターンを同定する情報を提供するために用いることができ、このように提供された情報は、その個体の１以上の表現型特性を予測するために用いることができる。このような方法を用いて、個体の疾患に対する罹患性／抵抗性および／または薬物応答を予測することができる。さらに、本発明のデータベースは、本発明の変異に関係がある１以上の遺伝子の発現レベルに関する情報を含み得る。

【０１３２】以下の実施例は、本発明の具体的な実施形態を説明するものであり、材料および方法は本発明を例示するものであり、本発明の範囲を限定するものではない。

【０１３３】（実施例１）　体細胞ハイブリッドの調製体細胞遺伝学における標準的な方法を用いて、ヒトＤＮＡ鎖（染色体）を分離して二倍体の状態から一倍体の状態とした。この場合、チミジンキナーゼ遺伝子について野生型である二倍体ヒトリンパ芽球細胞系を、チミジンキナーゼ遺伝子中に突然変異を含む二倍体ハムスター線維芽細胞系に融合した。得られた細胞の部分集団は、ヒト染色体を含むハイブリッド細胞であった。１０％ウシ胎児血清（ＦＢＳ）＋１×Ｐｅｎ／Ｓｔｒｅｐ＋１０％グルタミンを加えた１０ｍｌ　ＤＭＥＭを含む遠心分離管の中に、ハムスター細胞系Ａ２３細胞をピペッティングし、５分間１５００ｒｐｍにて遠心分離し、５ｍｌのＲＰＭＩ中に再懸濁し、１５ｍｌのＲＰＭＩ培地を含む組織培養フラスコ中にピペッティングした。リンパ芽球細胞を３７℃にてコンフルエントになるまで増殖させた。同時に、１５％ＦＢＣＳ＋１×Ｐｅｎ／Ｓｔｒｅｐ＋１０％グルタミンを加えた１０ｍｌ　ＲＰＭＩを含む遠心分離管の中に、ヒトリンパ芽球細胞をピペッティングし、５分間１５００ｒｐｍにて遠心分離し、５ｍｌのＲＰＭＩ中に再懸濁し、１５ｍｌのＲＰＭＩを含む組織培養フラスコ中にピペッティングした。リンパ芽球細胞を３７℃にてコンフルエントになるまで増殖させた。

【０１３４】Ａ２３ハムスター細胞を調製するために、増殖培地を吸引し、細胞を１０ｍｌのＰＢＳで濯いだ。次に細胞を2ｍｌのトリプシンで処理し、新しい培地（ＨＡＴを含まないDMEM）を含む３～５個のプレートに分け、３７℃にてインキュベートした。培地を遠心分離管に移し、５分間１５００ｒｐｍにて遠心分離し、増殖培地を吸引し、細胞を５ｍｌのＲＰＭＩ中に再懸濁し、細胞１～3ｍｌを、２０ｍｌのＲＰＭＩを含む2つのフラスコにピペッティングすることにより、リンパ芽球細胞を調製した。

【０１３５】細胞融合を行うために、約８～１０×10^6個のリンパ芽球細胞を１５００ｒｐｍにて５分間遠心分離した。次に、細胞をDMEM中に再懸濁し、再びこれらを遠心分離してからDMEMを吸引することにより、細胞ペレットをDMEMで濯いだ。次にリンパ芽球細胞を５ｍｌの新しいDMEM中に再懸濁した。受容者であるＡ２３ハムスター細胞をコンフルエントになるまで増殖させ、融合する３～４日前に分裂させ、この時点で５０～８０％コンフルエントであった。古い培地を除去し、細胞をDMEMで３回濯ぎ、トリプシン処理し、最後に５ｍｌのDMEMに懸濁した。リンパ芽球細胞を受容者であるＡ２３細胞上にゆっくりピペッティングし、合わせた培養物をゆっくりかき混ぜてから３７℃に

て１時間インキュベートした。インキュベーション後、Ａ２３細胞から培地を静かに吸引し、片手でプレートを回しながらプレートのエッジをピペットに付けてＰＥＧをプレートにゆっくり加えることにより、２ｍｌのＰＥＧ１５００（室温）を加えた。プレートを１回転させる間に全てのＰＥＧを加えるのに約１分かかった。次に、プレートをゆっくり回しながら８ｍｌのＤＭＥＭをプレートのエッジに加えた。細胞からＰＥＧ／ＤＭＥＭ混合物を静かに吸引した後、８ｍｌのＤＭＥＭを用いて細胞を濯いだ。このＤＭＥＭを除去し、１０ｍｌの新鮮なＤＭＥＭを加えて、細胞を３７℃にて３０分間インキュベートした。細胞から再びＤＭＥＭを吸引し、１０％　ＦＢＣＳおよび１×Ｐｅｎ／Ｓｔｒｅｐを加えた１０ｍｌのＤＭＥＭを細胞に加えた後、一晩インキュベートさせた。

【０１３６】インキュベーション後、培地を吸引し、細胞をＰＢＳで濯いだ。次に細胞をトリプシン処理し、各プレートに約１００，０００細胞が入るように、選択培地（１０％　ＦＢＳ＋１×Ｐｅｎ／Ｓｔｒｅｐ＋１×ＨＡＴを加えたＤＭＥＭ）を含む複数のプレートに分けた。プレートに入れた後３日目に培地を取り替えた。コロニーが肉眼で見えるようになったら（９～１４日目）コロニーを取り出して２４ウェルプレートに入れた。取り出したコロニーが５日以内にコンフルエントになったら、そのコロニーは健全であるとみなされ、細胞をトリプシン処理して６ウェルプレートに移した。

【０１３７】６ウェルプレート培養物から得た細胞から、ＤＮＡおよびストックハイブリッド細胞培養物を調製した。細胞をトリプシン処理し、１０ｍｌの選択培地を含む容量１００mmのプレートとエッペンドルフ試験管とに分けた。試験管内の細胞をペレット化し、２００μｌのＰＢＸ中に再懸濁し、Ｑｉａｇｅｎ　ＤＮＡミニキットを用いてスピンカラム１つあたり細胞５００万未満の濃度でＤＮＡを単離した。該１００mmプレートをコンフルエントになるまで増殖させ、細胞を培養し続けるかまたは冷凍した。

【０１３８】（実施例２）　一倍体ハイブリッドの選択単一チップのハイブリダイゼーションにおいて１４９４個のマーカーを評点することができる、Ａｆｆｙｍｅｔｒｉｘ，　ＨｕＳＮＰ遺伝子チップ（Ａｆｆｙｍｅｔｒｉｘ，　Ｉｎｃ．，　Ｓａｎｔａ　Ｃｌａｒａ，　ＣＡ，　ＨｕＳＮＰマッピングアッセイ、試薬キットおよびユーザマニュアル，　Ａｆｆｙｍｅｔｒｉｘ　Ｐａｒｔ　Ｎｏ．　９００１９４）を用いて、各ハイブリッド中のヒト染色体の存在、不在および二倍体／一倍体状態の評点を行った。ＨｕＳＮＰチップハイブリダイゼーションアッセイを用いて、対照としてハムスターおよびヒト二倍体リンパ芽球細胞系をスクリーニングした。親リンパ芽球二倍体細胞系中でヘテロ接合性であるＳＮＰを、各融合細胞系中における一倍体状態（ｈａｐｌｏｉｄｙ）について評点した。「Ａ」および「Ｂ」は各ＳＮＰ位置における二者択一の変異体であると仮定する。親二倍体細胞系において「ＡＢ」ヘテロ接合性として存在するマーカーを、ハイブリッド中において「Ａ」または「Ｂ」（半接合性）として存在する同じマーカーと比較することにより、各ハイブリッド系において一倍体状態であるヒトＤＮＡ鎖を決定した。

【０１３９】図１１は、２種類のヒト／ハムスター細胞ハイブリッド（ハイブリッド１およびハイブリッド２）を、ヒト第２１番染色体上の選択されたマーカーについてテストした結果を示す。一列目は、ＨｕＳＮＰチップマーカーの名称を示す。２列目は、ハムスター細胞核酸（融合なし）をＨｕＳＮＰチップとのハイブリダイゼーションに用いた場合にシグナルが得られたか否かを示す。予想通り、ハムスター細胞サンプル中のどのマーカーにもシグナルは見られなかった。三列目は、二倍体親ヒトリンパ芽球細胞系ＣＰＤ１７において各マーカー毎にどの変異体が検出されたか（「Ａ」、「Ｂ」または「ＡＢ」）を示す。幾つかのケースでは、Ａ変異体のみが存在し、幾つかのケースではＢ変異体のみが存在し、また幾つかのケースではＣＰＤ１７細胞がこれらの変異体についてヘテロ接合性（「ＡＢ」）であった。最後の二列は、２種類のヒト／ハムスターハイブリッド（ハイブリッド１およびハイブリッド２）から得た核酸サンプルをＨｕＳＮＰチップとハイブリダイズさせた結果を示す。親ＣＰＤ１７細胞系中にＡ変異体のみが存在する場合において、Ａ変異体のみが融合体に伝えられたことに留意されたい。親ＣＰＤ１７細胞系においてＢ変体のみが存在する場合は、Ｂ変異体のみが融合体に伝えられた。ＣＰＤ１７細胞系がヘテロ接合性である場合は、ある融合クローンにはＡ変異体は伝えられ、他の融合クローンにはＢ変異体が伝えられた。ただし、しばしば、この融合プロセスから得られるハイブリッド細胞系の中において染色体の一部しか存在しないこと、幾つかのハイブリッドは幾つかのヒト染色体もしくはその一部について二倍体であること、幾つかのハイブリッドは他のヒト染色体もしくはその一部について一倍体であること、ならびに幾つかのハイブリッドは幾つかの染色体のいずれの変異体も持たない場合があること、を理解されたい。特定のヒト染色体（例えば第２１番染色体）の１つの変異体のみを含むハイブリッドを選択して分析した。さらに好ましくは、（染色体の一部のみではなく）ある染色体全体を含むハイブリッドを選択して分析した。

【０１４０】（実施例３）　長領域ＰＣＲ（ｌｏｎｇ－ｒａｎｇｅ　ＰＣＲ）

ハムスター／ヒト細胞ハイブリッドから得たＤＮＡを用いて長領域ＰＣＲアッセイを行った。長領域ＰＣＲアッセイは当分野において一般公知であり、例えばベーリンガー・マンハイム社のＥｘｐａｎｄ　Ｌｏｎｇ　Ｒａｎｇｅ　ＰＣＲキットの標準的な長領域ＰＣＲプロトコー

ル（参考としてまたは全ての目的のために本明細書中に組み込まれる）に記載されている。

【０１４１】増幅反応に使用されるプライマーを、以下の通り設計した：所与の配列、例えば第２１番染色体上の２３×１０^{6}塩基のコンティグを、ゲノム中で繰返される配列（例えばＡｌｕおよびＬｉｎｅエレメントなど）を認識する当分野で公知のソフトウェアプログラム（本明細書中において「リピート・マスカー（ｒｅｐｅａｔ　ｍａｓｋｅｒ）」と呼ぶ）に入力した（Ａ．Ｆ．Ａ．　ＳｍｉｔおよびＰ．Ｇｒｅｅｎ，　ｗｗｗ．　ｇｅｎｏｍｅ．　ｗａｓｈｉｎｇｔｏｎ．　ｅｄｕ／ｕｗｇｃ／ａｎａｌｙｓｉｓｔｏｏｌｓ／ｒｅｐｅａｔｍａｓｋを参照されたい。これは本明細書中に参考として組み込まれる）。このプログラムにより、反復配列の各特定のヌクレオチド（Ａ、Ｔ、ＧまたはＣ）を「Ｎ」に置換することにより、反復配列を「マスキング」した。次に、この反復マスキング置換を行った後の配列の出力結果を、市販されているプライマー設計プログラム（Ｏｌｉｇｏ　６．２３）に供給し、長さが３０ヌクレオチドを超え且つ融解温度が６５℃を超えるプライマーを選択した。次に、Ｏｌｉｇｏ　６．２３から出力された設計されプライマーを、ゲノムの所与の領域をＰＣＲ増幅し且つ隣接するＰＣ産物と重複する部分が最も少ないプライマー対を「選択する」プログラムに供給した。市販されているプロトコールおよびこのプライマー設計を用いた長領域ＰＣＲの成功率は少なくとも８０％であり、またヒト染色体のある部分については９５％を超える成功率が得られた。

【０１４２】長領域ＰＣＲの１つの例示的なプロトコールは、ベーリンガー・マンハイム社のＥｘｐａｎｄ　Ｌｏｎｇ　Ｔｅｍｐｌａｔｅ　ＰＣＲ　Ｓｙｓｔｅｍ、カタログＮｏ．１６８１　８３４、１６８１　８４２または１７５９　０６０を用いる。この手法では、各５０μＬのＰＣＲ反応に２つのマスターミックス（ｍａｓｔｅｒ　ｍｉｘ）を要する。ある具体的な実施例では、各反応につき、氷上の１．５ｍｌの微量遠心機試験管の中でＭａｓｔｅｒ　Ｍｉｘ　１を調製した。Ｍａｓｔｅｒ　Ｍｉｘ　１は、最終体積で１９μＬのＭｏｌｅｃｕｌａｒ　Ｂｉｏｌｏｇｙ　Ｇｒａｄｅ　Ｗａｔｅｒ（Ｂｉｏ　Ｗｈｉｔｔａｋｅｒ、カタログＮｏ．１６－００１Ｙ）、２．５μＬの１０ｍＭ　ｄＮＴＰセット（ｄＡＴＰ、ｄＣＴＰ、ｄＧＴＰおよびｄＴＴＰを含む、各１０ｍＭ）（Ｌｉｆｅ　ＴｅｃｈｎｏｌｏｇｉｅｓカタログＮｏ．１０２９７－０１８）（各ｄＮＴＰの最終濃度は４００μＭ）、および５０ｎｇのＤＮＡ鋳型を含む。

【０１４３】全ての反応のためにＭａｓｔｅｒ　Ｍｉｘ　２を調製し、氷上に維持した。各ＰＣＲ反応につき、Ｍａｓｔｅｒ　Ｍｉｘ　２は最終体積で２５μＬのＭｏｌｅｃｕｌａｒ　Ｂｉｏｌｏｇｙ　Ｇｒａｄｅ　Ｗａｔｅｒ（Ｂｉｏ　Ｗｈｉｔｔａｋｅｒ）、２２．５０ｍＭ　$MgCl_2$を含む５μＬの１０×ＰＣＲ緩衝液３（Ｓｉｇｍａ、カタログＮｏ．　Ｍ　１０２８９）、２．５μＬの１０ｍＭ　$MgCl_2$（最終$MgCl_2$濃度は２．７５ｍＭ）、および０．７５μＬの酵素ミックス（最後に加える）を含む。

【０１４４】予め混合した６μＬのプライマー（２．５μＬのＭａｓｔｅｒ　Ｍｉｘ　１を含む）を適当な試験管に加えた後、２５μＬのＭａｓｔｅｒ　Ｍｉｘ　２を各試験管に加えた。試験管にキャップをし、混合し、短時間遠心分離をしてから氷に戻した。この時点で、以下のプログラムに従ってＰＣＲサイクルを開始した：ステップ１＝９４℃で３分間鋳型を変性させる；ステップ２＝９４℃で３０分間；ステップ３＝使用したプライマーに適した温度で３０秒間アニーリングする；ステップ４＝６８℃にて生成物１ｋｂあたり１分間伸長を行う；ステップ５＝ステップ２～４までを３８回繰返し、全部で３９サイクル行う；ステップ６＝９４℃にて３０秒間；ステップ７＝３０秒間アニーリングする；ステップ８＝６８℃にて生成物１ｋｂあたり１分間＋さらに５分間伸長を行う；およびステップ９＝４℃に保つ。あるいは、２段階ＰＣＲを行ってもよい：ステップ１＝９４℃にて３分間鋳型を変性させる；ステップ２＝９４℃にて３０秒間；ステップ３＝６８℃にて生成物１ｋｂあたり１分間アニーリングおよび伸長を行う；ステップ４＝ステップ２～３を３８回繰返し、全部で３９サイクル行う；ステップ５＝９４℃にて３０秒間；ステップ６＝６８℃にて生成物１ｋｂあたり１分間＋さらに５分間アニーリングおよび伸長を行う；およびステップ７＝４℃に保つ。

【０１４５】ヒト第１４番および２２番染色体上の様々な領域の長領域ＰＣＲ増幅反応の結果を、臭化エチジウムで染色したアガロースゲル上で可視化した（図１２）。本発明の長領域ＰＣＲ増幅法は、平均サイズが約８ｋｂである増幅断片を常套的に生成し、ゲノム領域の増幅に失敗したのは極めて稀なケースのようであった（第２２番染色体ゲル上のＧ１１を参照されたい）。

【０１４６】（実施例４）　ウェハの設計、製造、ハイブリダイゼーションおよび走査

オリゴヌクレオチドアレイ（チップまたはウェハ）に入れるオリゴヌクレオチドプローブのセットは、問い合わせするヒトＤＮＡ鎖配列に基づいて決定した。該オリゴヌクレオチド配列は、一般的に利用可能なデータベースに報告されているコンセンサス配列に基づいて決定した。プローブ配列を決定したら、コンピュータアルゴリズムを用いて、プローブ含有アレイを製造するために使用される写真平版マスク（ｐｈｏｔｏｌｉｔｈｏｇｒａｐｈｉｃ　ｍａｓｋ）を設計した。アレイは、固相化学合成を写真平版製造手法と組み合わせた光誘導化学合成プロセスにより製造した。例えば国際特許出願公開ＷＯ９２／１００９２号または米国特許第５，１４３，８５４号；第５，３８４，２６１号；第５，４０５，７８３

号；第５，４１２，０８７号；第５，４２４，１８６号；第５，４４５，９３４号；第５，７４４，３０５号；第５，８００，９９２号；第６，０４０，１３８号；および第６，０４０，１９３号（あらゆる目的のために、本明細書中に参考として全て組み込まれる）を参照されたい。一連の写真平版マスクを用いてガラス基板（ウェハ）上に露光部位を画定した後、特定の化学合成ステップを行い、このプロセスで、オリゴヌクレオチドプローブの高密度領域をアレイ上に作製した（各プローブは所定の位置にある）。多数のプローブ領域を同時且つ平行して合成した。

【０１４７】この合成プロセスでは、写真平版マスクに光を通過させて非保護領域中の化学基をこの光により活性化させることにより、光保護されたガラス基板を選択的に照射した。次に、選択的に活性化された基板ウェハを選択したヌクレオチドと共にインキュベートし、ウェハ上の活性化位置において化学結合を生じさせた。結合が起こったら、新しいマスクパターンをあてがい、他の選択されたヌクレオチドを用いて結合ステップを繰返した。所望のプローブセットが得られるまでこのプロセスを繰返した。１つの具体的な実施例において、１３番目の塩基が問い合わせしようとする塩基である場合、２５－ｍｅｒオリゴヌクレオチドプローブを用いた。４つのプローブを用いて、各配列中に存在する各ヌクレオチドに問い合わせしたところ、１つのプローブが該配列に相補的であり、３つのミスマッチプローブは、１３番目の塩基を除いて該相補的プローブと同じであった。幾つかのケースにおいて、少なくとも10×10^{6}個のプローブが各アレイ上に存在した。

【０１４８】アレイを製造したら、このアレイを、ハムスター／ヒト細胞ハイブリッドに対して行った長領域ＰＣＲ反応により得た生成物にハイブリダイズさせた。分析しようとするサンプルを標識し、該アレイと共にインキュベートして、該サンプルをウェハ上のプローブにハイブリダイズさせた。

【０１４９】ハイブリダイゼーション後、アレイを共焦高速スキャナ（ｃｏｎｆｏｃａｌｈｉｇｈ　ｐｅｒｆｏｒｍａｎｃｅ　ｓｃａｎｎｅｒ）に入れて、ハイブリダイゼーションのパターンを検出した。そのサンプルのＰＣＲ産物中に既に組み込まれた蛍光レポーター基（プローブに結合した）から発せられる光として、ハイブリダイゼーションデータを収集した。サンプル中に存在するウェハ上のプローブに相補的な配列は、ミスマッチを含むこれらの配列に比べ、より強力にウェハにハイブリダイズし、またより強力なシグナルを生成した。アレイ上の各プローブの配列および位置は分かっているので、相補性により、プローブアレイに加えたサンプル核酸における変異の正体を同定した。本発明で用いられるスキャナおよび走査手法は、当業者には公知であり、例えば米国特許第５，９８１，９５６号（マイクロアレイチップについて）、米国特許第６，２６２，８３８号および米国特許第５，４５９，３２５に開示されている。さらに、２０００年８月３日に出願された米国特許仮出願番号第６０／２２３，２７８号および米国特許仮出願番号第６０／２２３，２７８号を基に優先権主張し２００１年８月３日に出願された非仮出願（全ウェハ走査のためのスキャナおよび手法について）もまた、全ての目的のために参考として本明細書中に全て組み込まれる。

【０１５０】（実施例５）　ヒト第２１番染色体上のＳＮＰハプロタイプの決定

アフリカ人、アジア人およびカフカス人の染色体の第２１番染色体の２０個の独立したコピーを、ＳＮＰ発見のためおよびハプロタイプ構造について分析した。各個体から得た第２１番染色体の２つのコピーを、げっ歯類／ヒト体細胞ハイブリッド手法（図１０）を用いて物理的に分離した（上記記載）。この分析のための基準配列は、３２，３９７，４３９塩基からなるヒト第２１番染色体ゲノムＤＮＡ配列からなるものであった。この基準配列の反復配列をマスキングし、得られた２１，６７６，８６８塩基（６７％）のユニーク配列を、高密度オリゴヌクレオチドアレイを用いて変異について分析した。８つのユニークオリゴヌクレオチド（各々は２５塩基長）を用いて、ユニークサンプル第２１番染色体塩基の各々（合計で1.7×10^{8}個の異なるオリゴヌクレオチド）を問い合わせした。これらのオリゴヌクレオチドを、過去に記載されたタイリング法（ｔｉｌｉｎｇ　ｓｔｒａｔｅｇｙ）（Ｃｈｅｅら、Ｓｃｉｅｎｃｅ　２７４：６１０（１９９６））を用いて、全部で８個の異なる設計のウェハに分配した。Ａｆｆｙｍｅｔｒｉｘ，Ｉｎｃ．（Ｓａｎｔａ　Ｃｌａｒａ，　ＣＡ）より購入した５インチ四方のガラスウェハ上で、オリゴヌクレオチドの光誘導化学合成を行った。

【０１５１】３２．４Ｍｂの第２１番染色体の連続的ＤＮＡにまたがる平均１０ｋｂ長の、重複が最少である３２５３個の長領域ＰＣＲ（ＬＲＰＣＲ）産物を作成するために、ユニークオリゴヌクレオチドを設計し、上記のように調製した。各ウェハのハイブリダイゼーション毎に、対応するＬＲＰＣＲ産物をプールし、Ｑｉａｇｅｎチップ５００（Ｑｉａｇｅｎ）を用いて精製した。３７μｌの１０×Ｏｎｅ－Ｐｈｏｒ－Ａｌｌ緩衝液ＰＬＵＳ（Ｐｒｏｍｅｇａ）および全量３７０μｌのＤＮＡアーゼ（Ｌｉｆｅ　Ｔｅｃｈｎｏｌｏｇｉｅｓ／Ｉｎｖｉｔｒｏｇｅｎ）１ユニットを用いて、３７℃にて１０分間かけて、全部で２８０μｇの精製ＤＮＡを断片化した後、９９℃にて１０分間加熱により不活化した。５００ユニットのＴｄｔ（ベーリンガー・マンハイム社）および２０ｎｍｏｌのビオチン－Ｎ６－ｄｄＡＴＰ（ＤｕＰｏｎｔ　ＮＥＮ）を用いて３７℃にて９０分間かけて断片化生成物の末端を標識し、９５℃にて１０分間加熱することにより不活化した。標識したサンプルを、１０ｍ

55

Ｍ　Ｔｒｉｓ－ＨＣＬ（ｐＨ８）、３Ｍ塩化テトラメチルアンモニウム、０．０１％　Ｔｘ－１００、１０μｇ／ｍｌ変性ニシン精子ＤＮＡを含むウェハ（ウェハ１つあたり全量１４ｍｌ）に５０℃にて１４～１６時間ハイブリダイズさせた。ウェハを４×ＳＳＰＥ中で手早く濯ぎ、６×ＳＳＰＥで１０分間ずつ３回洗浄し、ストレプトアビジンＲ－フィコエリトリン（ＳＡＰＥ、５ｎｇ／ｍｌ）を用いて室温にて１０分間染色した。ストレプトアビジンに対する抗体（１．２５ｎｇ／ｍｌ）で染色することにより、およびＳＡＰＥを用いた染色ステップを繰返すことにより、シグナルを増幅した。

【０１５２】１つのウェハ上に存在する塩基に対応するＰＣＲ産物をプールし、１回の反応としてウェハにハイブリダイズさせた。１６０個のウェハ上で合計３．４×10^9個のオリゴヌクレオチドを合成し、ヒト第２１番染色体の２０個の独立コピーをＤＮＡ配列変異について走査した。長範囲ＰＣＲを用いてげっ歯類／ヒトハイブリッド細胞系から各ユニーク第２１番染色体を増幅した。ＬＲＰＣＲアッセイは、Ｏｌｉｇｏ６．２３プライマー設計ソフトウェアと高～中程度のストリンジェンシーパラメータとを用いて設計した。得られたプライマーは、典型的には３０ヌクレオチド長であり、融解温度は６５℃を超えるものであった。アンプリコンサイズの範囲は３ｋｂ～１４ｋｂであった。その染色体全体のプライマーデータベースを作製し、ソフトウェア（ｐＰｃｉｋｅｒ）を用いて、隣同士のアンプリコンの間の重複が最少であり第２１番染色体配列を最大限にカバーする非冗長プライマーの最少セットを選択した。あるいは、本明細書中の実施例３に記載したプライマー選択法を用いた。若干の変更を加えたＥｘｐａｎｄ　Ｌｏｎｇ　Ｔｅｍｐｌａｔｅ　ＰＣＲ　Ｋｉｔ（ベーリンガー・マンハイム社）を用いて、ＬＲＰＣＲ反応を行った。特注の共焦スキャナを用いてウェハを走査した。

【０１５３】パターン認識アルゴリズムを用いてハイブリダイゼーションの変化としてＳＮＰを検出した。過去に記載されたアルゴリズムの組合せ（Ｗａｎｇら、Ｓｃｉｅｎｃｅ　２８０：１０７７（１９９８））を用いて、変化したハイブリダイゼーションパターンに基づいてＳＮＰを検出した。２０個の染色体からなる該サンプルにおいて、全部で３５，９８９個のＳＮＰが同定された。これらのヒト多型の位置および配列は、ＧｅｎＢａｎｋのＳＮＰｄｂに寄託されている。ジデオキシ配列決定を用いて、元のＤＮＡサンプル中のこれらのＳＮＰのうちの２２７個の無作為に選んだサンプルを評価し、分析したＳＮＰのうち２２０個（９７％）を確認した。この３％という低い偽陽性ＳＮＰ率を達成するためには、ウェハ上でのＳＮＰ検出のために、高い偽陰性率をもたらすストリンジェント閾値が必要であった。ウェハ上に存在する全ての塩基の約６５％は、ＳＮＰ検出で使用するのに十分高品質なデータを生成し、３５％は偽陰性として破棄した。分析した全てのサンプルにおいて一貫して失敗した長範囲ＰＣＲは、この３５％の偽陰性率のうちの１５％を占める。残りの２０％の偽陰性は、高品質データを生成しない塩基（１０％）と分析した第２０番染色体の画分のみに高品質データを生成する塩基（１０％）との間に分布している。一般に、ある塩基が高品質データを生成するか否かを指令するのはその塩基の配列における前後関係である。全塩基のうちの約２０％が一貫して質の悪いデータをもたらすという知見は、５００塩基の１回のジデオキシ配列決定読取りにおいて塩基の約３０％が信頼性の高いＳＮＰ検出にとっては低過ぎる品質スコアを有するという知見に非常によく似ている（Ａｌｔｓｃｈｕｌｅｒら、Ｎａｔｕｒｅ　４０７：５１３（２０００））。分析されるサンプルのうちの限られた数しか所与の塩基についての高品質データを生成しない場合、より頻繁に見られるＳＮＰに比べて稀少なＳＮＰを発見するための能力は格段に低下する。その結果、この方法によるＳＮＰ発見は、共通ＳＮＰにとって有利なものである。

【０１５４】図１３Ａは、全体的に多様な染色体のサンプル中で発見された３５，９８９個全てのＳＮＰのマイナー対立遺伝子頻度の分布を表す。ヌクレオチド多様性の２つの測度（π＝１つの部位あたりの平均ヘテロ接合性；およびθ＝集団突然変異パラメータ（ｔｈｅ　ｐｏｐｕｌａｔｉｏｎ　ｍｕｔａｔｉｏｎ　ｐａｒａｍｅｔｅｒ））を用いて、遺伝学的変異（サンプル中の染色体の数について正規化した）を推定した（ＨａｒｔｌおよびＣｌａｒｋ，　Ｐｒｉｎｃｉｐｌｅｓ　ｏｆＰｏｐｕｌａｔｉｏｎ　Ｇｅｎｅｔｉｃｓ　（Ｓｉｎａｕｅｒ，　Ｍａｓｓａｃｈｕｓｅｔｔｓ，　１９９７）を参照されたい）。配列決定が完了したゲノム第２１番染色体ＤＮＡの３２，３９７，４３９塩基を２００，０００塩基対セグメントに分け、各セグメントにおけるＳＮＰ発見に使用される高品質塩基対を調べた。これらの塩基の観察されたヘテロ接合性を用いて、各セグメント毎に平均ヌクレオチド多様性（π）を算出した。全データセットの平均ヌクレオチド多様性の推定（π＝０．０００７２３およびθ＝０．０００７９８）ならびにヌクレオチド多様性の分布（第２１番染色体の連続的な２００，０００塩基対ｂｉｎｓで測定）（図１３）は、過去に記載された値の範囲内であった（Ｔｈｅ　Ｉｎｔｅｒｎａｔｉｏｎａｌ　ＳＮＰ　Ｍａｐ　Ｗｏｒｋｉｎｇ　Ｇｒｏｕｐ，　Ｎａｔｕｒｅ　４０９：９２８－３３（２００１））。

【０１５５】Ｔｈｅ　ＳＮＰ　Ｃｏｎｓｏｒｔｉｕｍ（ＴＳＣ）により発見された第２１番染色体の１５，５４９個のＳＮＰの重複の度合いを、この調査で発見されたＳＮＰと比較した。ＴＳＣ　ＳＮＰのうち、５，０８７個は反復ＤＮＡの中にあることが分かり、ウェハ上でタイリングされなかった。残りの１０，４６２個のＴＳ

Ｃ　ＳＮＰのうち、４７０５個（４５％）を同定した。θの推定値は、分析した連続的ＤＮＡ配列の１６２　２００－ｋｂ　ｂｉｎｓの１２９についてπ□の推定値よりも大きいことが分かった。この差は、近年のヒト人口の増大と一致し、ヒト遺伝子におけるヌクレオチド多様性の最近の研究結果と類似している（Ｓｔｅｐｈｅｎら、Ｓｃｉｅｎｃｅ　２９３：４８９（２００１））。この発見された量のヌクレオチド多様性の場合、４３％のシングルトンが得られるというニュートラルモデルの期待値（ＦｕおよびＬｉ，　Ｇｅｎｅｔｉｃｓ　１３　３：６９３（１９９３））と比較して、ＳＮＰのうち１１，６０３個（３２％）が、サンプル中において１回観察されるマイナー対立遺伝子（シングルトン）を有していたことが分かった。観測値と期待値との差は、上記のようなこの調査における共通ＳＮＰに比べて稀少ＳＮＰを同定する能力が低下したことによるものと思われる。

【０１５６】全部で、３２．４Ｍｂのヒトゲノム中に存在すると推定される１０％以上の対立遺伝子頻度を有する５３，０００個の共通ＳＮＰのうち４７％が同定された。これは、Ｉｎｔｅｒｎａｔｉｏｎａｌ　ＳＮＰ　Ｍａｐｐｉｎｇ　Ｗｏｒｋｉｎｇ　Ｇｒｏｕｐおよびｔｈｅ　ＳＮＰ　Ｃｏｎｓｏｒｔｉｕｍにより作製されたコレクション中に存在する全てのこのような共通ＳＮＰのうちの１８～２０％の推定に匹敵する。網羅度の差は、本調査がＳＮＰ発見のためにより多数の染色体を用いたことにより説明される。この知見の反復可能性（ｒｅｐｌｉｃａｂｉｌｉｔｙ）を評価するために、元のサンプルセットと同じ多様性パネルから得た第２１番染色体のさらに１９個のコピーを含む１つのウェハ設計について、ＳＮＰ発見を行った。２つのサンプルセットを用いて、全部で７１８８個のＳＮＰを同定した。サンプルの一方のセットで発見された全てのＳＮＰのうち平均で６６％が第２セットで発見されたが、これは、過去の知見と一致するものであった（Ｍａｒｔｈら、Ｎａｔｕｒｅ　Ｇｅｎｅｔ．　２７：３７１（２００１）およびＹａｎｇら、Ｎａｔｕｒｅ　Ｇｅｎｅｔ．　２６：１３（２０００））。予想通り、第２のサンプルセットにおけるＳＮＰの反復（ｒｅｐｌｉｃａｔｅ）の失敗は、対立遺伝子頻度に大きく依存する。一方のサンプルセットにおいて２回以上存在するマイナーな対立遺伝子を有するＳＮＰの８０％は第２のサンプルセットにおいても発見されたのに対して、１回だけ存在するマイナー対立遺伝子を有するＳＮＰは第２サンプルセットにおいてその３２％しか発見されないことが分かった。これらの知見は、２回以上現れるマイナー対立遺伝子を有する該コレクション中の２４，０４７個のＳＮＰは、異なる総合的サンプルにおいて反復率が高いこと、およびこのＳＮＰセットは共通総合ハプロタイプ（ｃｏｍｍｏｎ　ｇｌｏｂａｌ　ｈａｐｌｏｔｙｐｅ）を定義するのに有用であることを示唆する。ＳＮＰ発見の過程において、３以上の対立遺伝子を有すると思われる３３９個のＳＮＰを同定した。これらのＳＮＰはこの分析に含めなかった。

【０１５７】異なるサンプル中におけるＳＮＰの反復可能性に加え、ＳＮＰのあるコレクションの中の連続的ＳＮＰ間の距離は、意味深いハプロタイプ構造を画定するために非常に重要である。ハプロタイプブロック（数ｋｂという短いものであってもよい）は、あるコレクションの中の連続的ＳＮＰ間の距離が実際のハプロタイプブロックのサイズに比べて大きい場合、認識されない場合がある。ＳＮＰ発見プロセスに反復配列を含めなかったが、この調査におけるＳＮＰのコレクションは、染色体全体に非常に均一に分布されていた。図１３Ｃは、完了した第２１番染色体ＤＮＡ配列の３２，３９７，４３９塩基にまたがるＳＮＰ網羅度の分布を示す。インターバルは連続的ＳＮＰ間の距離である。全ＳＮＰセットでは全部で３５，９８８個のインターバルがあり、共通ＳＮＰ（すなわちサンプル中に２回以上存在するマイナー対立遺伝子を有するＳＮＰ）セットでは全部で２４，０４６個のインターバルがある。連続的ＳＮＰ間の平均距離は、全てのＳＮＰを考慮した場合は９００塩基であり、２４，０４７個の共通ＳＮＰのみを考慮した場合は１３００塩基であった。この共通ＳＮＰセットの場合、ゲノムＤＮＡ（反復ＤＮＡを含む）中の連続的ＳＮＰ間のインターバルのうち９３％は４０００塩基以下であった（再び図１３Ｃを参照されたい）。

【０１５８】二倍体データからのハプロタイプブロックまたはパターンの作製は、任意の２つのヘテロ接合性ＳＮＰの対立遺伝子間の関係を直接観察することができないため、複雑になっている。第２１番染色体の２つのコピーおよび２つの対立遺伝子ＡおよびＧを１つの第２１番染色体ＳＮＰに、ならびに２つの対立遺伝子ＡおよびＧを第２の第２１番染色体ＳＮＰに有する個体を考える。このようなケースでは、第２１番染色体の一方のコピーが第１ＳＮＰに対立遺伝子Ａおよび第２ＳＮＰに対立遺伝子Ａを含み且つ第２１番染色体の他方のコピーが第１ＳＮＰに対立遺伝子Ｇおよび第２ＳＮＰに対立遺伝子Ｇを含むか否か、あるいは第２１番染色体の一方のコピーが第１ＳＮＰに対立遺伝子Ａおよび第２ＳＮＰに対立遺伝子Ｇを含み且つ第２１番染色体の他方のコピーが第１ＳＮＰに対立遺伝子Ｇおよび第２ＳＮＰに対立遺伝子Ａを含むか否か、は明らかではない。この問題を回避するために用いられる現在の方法は、ハプロタイプ頻度の統計的推測、ファミリーデータからの直接的な妨害、および短いセグメントに対する対立遺伝子特異的ＰＣＲ増幅を含む。

【０１５９】これらの複雑性を回避するために、本発明は、げっ歯類／ヒト体細胞ハイブリッド中で単離された第２１番染色体の一倍体コピー上のＳＮＰを特徴付けて、これらの染色体の完全ハプロタイプを直接決定できるようにした。データセットの中で２回以上現れるマイ

ナー対立遺伝子を有する２４，０４７個のＳＮＰからなるセットを用いて、図１４に示すハプロタイプ構造を画定した。１４７個の共通ヒト第２１番染色体ＳＮＰにより定義される２０個の独立した全体的に多様な染色体のハプロタイプパターンが示されている。１４７個のＳＮＰがゲノムＤＮＡ配列の１０６ｋｂにまたがっている。色付きのボックスの各行は単一のＳＮＰを表す。各行の黒いボックスはそのＳＮＰのメジャーな対立遺伝子を表し、白いボックスはマイナー対立遺伝子を表す。行の中のどこにもボックスがなければ、欠測データであることを示す。色付きボックスの各列は、単一の染色体を表し、その染色体上の物理的な順番に従ってＳＮＰが並んでいる。連続的ＳＮＰ間の不変塩基は図中に表わされていない。１４７個のＳＮＰを１８個のブロック（黒い水平のラインにより画定されている）に分ける。第２１番染色体ゲノムＤＮＡ配列の中の、１つのブロックの始まりとそれに隣接するブロックの終りとを画定する塩基の位置は、垂直な黒いラインの左側の数字で示されている。図の右側の拡大ボックスは、ゲノムＤＮＡの１９ｋｂにまたがる２６個の共通ＳＮＰにより画定されるＳＮＰブロックを表す。サンプル中に現れる７つの異なるハプロタイプパターンのうち、４つの最も共通するパターンは、サンプリングした２０個の染色体のうちの１６個（すなわちそのサンプルの８０％）を含む。黒丸および白丸は、２つの情報提供ＳＮＰの対立遺伝子パターン（これはこのブロックの中の４つの共通ハプロタイプ間を明白に区別する）を示す。どの２つの染色体も、これら１４７個のＳＮＰについて同一ハプロタイプパターンを共有していなかったが、多数の染色体が共通パターンを共有する領域が沢山ある。より詳細に分析するために、１つのこのような領域（１９ｋｂにまたがる２６個のＳＮＰにより画定される）を拡大する（再び図１４の拡大領域を参照されたい）。このブロックは、第２０番染色体中の７つのユニークなハプロタイプパターンを画定する。データ品質の閾値を合格しなかったためにあるデータが欠けているにもかかわらず、全てのケースにおいて、所与の染色体はこの７つのハプロタイプのうちの１つに明白に割り当てることができる。４つの最も頻繁なハプロタイプ（各々は３以上の染色体により表される）はそのサンプル中の全ての染色体の８０％を占める。全部で２６個のうちの２つの「情報提供」ＳＮＰのみが、該４つの最も頻繁なハプロタイプを区別するのに必要である。この例では、これら２つの情報提供ＳＮＰのみから得た情報を用いることにより、頻度の低いハプロタイプを有する４つの染色体が共通ハプロタイプとして誤って分類される。にもかかわらず、全ての総合サンプルのハプロタイプ構造の８０％がそのブロックの中の全ＳＮＰの１０％未満により定義されるということは注目すべきことである。４つの共通ハプロタイプの各々が単一のＳＮＰにより唯一定義されるように３つの情報提供ＳＮＰを選択することができる、幾つかの異なる可能性がある。これら「３つのＳＮＰ」の選択肢のうち１つは、プールしたサンプルの遺伝子タイピングを含む実験において２つのＳＮＰ組合せよりも好ましいであろう。なぜなら、このような状況において、この２つのＳＮＰの組合せでは４つの共通ハプロタイプの頻度が決定できないからである。従って、本発明は、無作為なＳＮＰマッピングの選択方法に比べて大幅な進歩を提供する。

【０１６０】まとめると、この特定のアプリケーションは、ハプロタイプ情報を捕捉するために情報提供ＳＮＰの選択を命令し得るが、そのサンプルの中のハプロタイプ情報の大半は、全てのＳＮＰのある非常に小さなサブセットの中に含まれることは明白である。また、このＳＮＰブロックからの２つまたは３つの情報提供ＳＮＰを無作為に選択しても、多くの場合は、該４つの共通ハプロタイプのうちの１つに染色体を唯一割り当てるための十分な情報が得られない。

【０１６１】１つの問題は、そのハプロタイプ構造を画定するのに必要なＳＮＰの合計数を最小限に抑えながら第２１番染色体の全３２．４Ｍｂに広がる連続的なＳＮＰブロックのセットをどのように画定するかということである。１つの実施形態において、この問題を解決するために「欲張り」法に基づく最適化アルゴリズムを用いた。サイズが１ＳＮＰ以上である物理的に連続するＳＮＰからなる全ての可能なブロックについて考慮した。曖昧なハプロタイプパターンは欠測データとして扱い、網羅度のパーセンテージを算出する際には含めなかった。残りの重複ブロックを同時に考慮し、そのブロックの中に２回以上現れるハプロタイプを唯一識別するのに必要なＳＮＰの最小数に対するそのブロックの中の全ＳＮＰの比が最大であるブロックを選択した。選択されたブロックと物理的に重複する残りのブロックを捨て、ギャップを持たず且つ全てのＳＮＰがブロックに割り当てられた、第２１番染色体の３２．４Ｍｂをカバーする連続的な非重複ブロックのセットが選択されるまで、このプロセスを繰返した。サンプルのサイズを染色体２０個とすると、このアルゴリズムは、ブロック１つあたり最大で１０個の共通ハプロタイプパターン（各々は２つの独立染色体により表される）を生成する。

【０１６２】このアルゴリズムを２４，０４７個の共通ＳＮＰからなるデータセットに適用し、第２１番染色体にまたがる４，１３５個のＳＮＰブロックを画定した。全部で５８９個のブロック（全ブロックの１４％を占める）は、ブロック１つあたり１１以上のＳＮＰを含み、全３２．４Ｍｂの４４％を含む。これに対し、２，１３８ブロック（全ブロックの５２％を占める）は、ブロック１つあたり３個未満のＳＮＰを含み、該染色体の物理的長さのたった２０％しか構成しない。最も長いブロックは１１４個の共通ＳＮＰを含み、ゲノムＤＮＡの１１５ｋｂにまたがる。全般的にみて、１つのブロックの平

61

均物理的サイズは７．８ｋｂである。ブロックのサイズは染色体上におけるその順序には関係が無く、染色体の全長に沿って、大きなブロックの間には小さなブロックが散在している。ブロック１つあたり平均２．７個の共通ハプロタイプパターン（複数の染色体上で観察されるハプロタイプパターンとして定義される）がある。平均で、あるブロックの中の最も頻度が高いハプロタイプパターンは、サンプル中の２０個の染色体のうち９．６個の染色体によって表され、２番目に頻度の高いハプロタイプパターンは、４．２個の染色体によって表され、（あれば）３番目に頻度の高いハプロタイプパターンは、２．１個の染色体によって表される。全体的に多様な染色体のこのような大きな割合は、このように限定されたハプロタイプの多様性によって表されるという事実は、注目に値する。この知見は、ハプロタイプパターン頻度を考慮したときに、３１３個のヒト遺伝子からなるコレクションの中で観察されるハプロタイプパターンの８２％が全ての人種グループにおいて観察される一方で、ハプロタイプの８％のみは集団特異的である、という観察結果と一致する（Ｓｔｅｐｈｅｎｓら、Ｓｃｉｅｎｃｅ　２９３：４８９－９３（２００１））。得られたブロックパターンに対して及ぼす該ハプロタイプアルゴリズムのパラメータの影響を測定するために幾つかの実験を行った。共通ハプロタイプにより網羅される必要がある染色体の割合は様々であり、最初は８０％から、７０％および９０％であった。予測した通り、より完全な網羅度を必要とすると、多少大きな数のより短いブロックができる。そのサンプル中のマイナー対立遺伝子の頻度が少なくとも２０％である１６，５０３個のＳＮＰのみを用いたところ、ある程度長いブロックになったが、ブロック１つあたりのＳＮＰの数はそれほど変わらなかった。約３Ｍｂの１つの領域について、この２０個染色体分析に匹敵させるために、少なくとも１０％の頻度を有する共通ハプロタイプおよびＳＮＰについての３８個の染色体のより奥行きのあるサンプルを分析した。得られたブロックサイズの分布は、最初の結果とほぼ一致していた。また、各ＳＮＰにある非曖昧対立遺伝子の順序を変えた（ｐｅｒｍｕｔｅ）後にハプロタイプブロック発見に使用するという無作為なテストを行った。この分析では、ブロックの９４％が３個未満のＳＮＰを含み、１つのブロックのみが６個以上のＳＮＰを含んでいた。これは、データ中に見られる大きなブロックは、偶然（ｂｙ　ｃｈａｎｃｅ　ａｓｓｏｃｉａｔｉｏｎｓ）、または本発明のブロック選択法のアーチファクトとしては生成され得ない、ということを立証するものである。

【０１６３】大きなブロックおよび小さなブロックの両方において遺伝子が比例的に現れるか否かを決定するために、１１個以上のＳＮＰ、３～１０個のＳＮＰ、および３個未満のＳＮＰをそれぞれ含むブロックにおけるエキソン塩基の数で決定を行った。エキソン塩基は、３～１０個のＳＮＰを含むブロック中の全塩基に比べてある程度過剰に表れる（ｏｖｅｒ－ｒｅｐｒｅｓｅｎｔｅｄ）（並べ替えテスト（ｐｅｒｍｕｔａｔｉｏｎ　ｔｅｓｔ）で測定したところ、ｐ＜０．０５）。

62

【０１６４】共通ハプロタイプ情報（全３２．４Ｍｂにわたる該サンプルの８０％を超える部分を含み且つ該サンプル内に２回以上存在するハプロタイプの完全な情報として定義される）の所望の割合を捕捉するために、ブロック内のハプロタイプ構造の知識に基づいて、２４，０４７個の共通ＳＮＰのサブセットを選択することができる。図１５は、第２１番染色体の３２．４Ｍｂについての共通ハプロタイプ情報を捕捉するために必要なＳＮＰの数を表す。各ＳＮＰブロック毎に、２回以上存在するそのブロックの中のハプロタイプを明白に識別するのに必要なＳＮＰの最小数（すなわち共通ハプロタイプ情報）を決定した。これらのＳＮＰは、そのブロックにより画定される合計物理的距離の割合についての共通ハプロタイプ情報を提供する。最も大きな物理的距離についての共通ハプロタイプ情報を提供するＳＮＰから始めて、物理的網羅度（すなわちカバーされる割合）の累進的増加を、追加したＳＮＰ（すなわち必要なＳＮＰ）の数に関してプロットする。遺伝子ＤＮＡ（ｇｅｎｉｃ　ＤＮＡ）は各既知の第２１番染色体遺伝子の最初のエキソンの５’側の１０ｋｂから始まりその遺伝子の最後のエキソンの３’側の１０ｋｂに延びる全てのゲノムＤＮＡを含む。例えば、全ての共通ハプロタイプ情報を捕捉するには最低でも４，５６３個のＳＮＰが必要とされるが、３個以上のＳＮＰを含むブロック内の共通ハプロタイプ情報（３２．４Ｍｂの８１％をカバーする）を捕捉するためには２７９３個のＳＮＰしか必要としない。遺伝子ＤＮＡ中の全ての共通ハプロタイプ情報（約２２０個の異なる遺伝子を表す）を捕捉するためには、合計１７９４個のＳＮＰが必要である。

【０１６５】本発明は、共通疾患遺伝子（ｃｏｍｍｏｎ　ｄｉｓｅａｓｅ　ｇｅｎｅ）等の表現型をマッピングする全ゲノムの関連付け調査に特に適している。このアプローチは、共通する遺伝子変異体が共通する疾患に対する罹患性に関与するという仮説に基づく（ＲｉｓｃｈおよびＭｅｒｉｋａｎｇａｓ，　Ｓｃｉｅｎｃｅ　２７３：１５１６（１９９６），　Ｌａｎｄｅｒ，　Ｓｃｉｅｎｃｅ　２７４：５３６（１９９６））。非関連ケースおよび対照における遺伝子変異体の頻度を比較することにより、遺伝学的関連付け調査は、疾患において重要な役割を果たすヒトゲノム中の特定のハプロタイプを同定することができる。このアプローチは単一候補遺伝子を疾患に関連付けることに成功したが（Ａｌｔｓｃｈｕｌｅｒら、Ｎａｔｕｒｅ　Ｇｅｎｅｔ．　２６：７６（２０００））、ヒトＤＮＡ配列が近年入手可能となったことにより、全ゲノムを調査することが可能となり、

遺伝学的関連付け分析の能力を大幅に飛躍させた（Ｋｒｕｇｌｙａｋ，　Ｎａｔｕｒｅ　Ｇｅｎｅｔ．　２２：１３９（１９９９））。この方法の実施を制限する主な要因は、ヒトゲノムのハプロタイプ構造の知識（これは、分析用の適当な遺伝子変異体を選択するために必要である）に乏しいことであった。本発明は、高密度オリゴヌクレオチドアレイと体細胞の遺伝子サンプルの調製とを組み合わせることにより、ヒトゲノムの共通ハプロタイプ構造を経験的に画定する高分解能アプローチ（ｈｉｇｈ－ｒｅｓｏｌｕｔｉｏｎ　ａｐｐｒｏａｃｈ）を提供することを示す。

【０１６６】単純なハプロタイプ構造を有するゲノム領域の長さは非常にまちまちであるが、共通ＳＮＰの稠密セットにより、世界人口の８０％がたった３つの共通ハプロタイプにより描写されるヒトゲノムのブロックを定義する体系的なアプローチを可能とする。一般に、この実施形態に用いられる特定のアルゴリズムを適用する場合、任意のブロックの中で最も一般的なハプロタイプは、個体の５０％において見られ、２番目に一般的なハプロタイプは個体の２５％において見られ、および３番目に一般的なハプロタイプは個体の１２．５％において見られる。ブロックはその遺伝子情報の内容に基づいて定義されるのであって、この情報がどのように生じたのかということや何故存在するのかということについての知識に基づくものではない、ということに留意することは重要である。従って、ブロックは絶対的な境界を持たず、特定の用途に応じて様々な方法で画定することができる。この実施形態におけるアルゴリズムは、沢山の可能なアプローチのうちのたった１つを提供するに過ぎない。これらの結果は、全ての共通ハプロタイプ情報を捕捉するために、ＳＮＰの非常に稠密なセットが必要であることを示している。しかし一方では、この方法を用いて、総合的な全ゲノムの関連付け調査に有用なＳＮＰのもっと小さな部分集合を同定することができる。

【０１６７】当業者であれば、ヒト第２１番染色体に適用された手法をヒトゲノムの全ての染色体に適用することができることが容易に分かるであろう。本発明の好適な実施形態において、ヒト種（ｈｕｍａｎ　ｓｐｅｃｉｅｓ）を多様な集団の代表の多数の全ゲノムを用いて、ヒト種の全てまたは大部分のメンバーに共通するＳＮＰハプロタイプブロックを同定する。幾つかの実施形態において、ＳＮＰハプロタイプブロックは、低い頻度で現れるＳＮＰを除外することにより、古来のＳＮＰ（ａｎｃｉｅｎｔ　ＳＮＰ）に基づく。古来のＳＮＰは、そのＳＮＰを保有する生物に何らかの選択的利益を与えるため、ゲノム中で保存されているので重要であると思われる。

【０１６８】（実施例６）　遺伝子治療および薬剤発見に関連遺伝子を用いる

本発明の方法を用いるための１つの例を、この予想実施例（ｐｒｏｐｈｅｔｉｃ　ｅｘａｍｐｌｅ）で概説する。２０個の一倍体ゲノムに対してＳＮＰ発見を行い、ＳＮＰハプロタイプブロック、ＳＮＰハプロタイプパターン、情報提供ＳＮＰおよび各情報提供ＳＮＰのマイナー対立遺伝子の頻度を決定するための本発明の方法により、５０個の一倍体ゲノムを分析する。これら５０個の一倍体ゲノムは、この調査の対照ゲノムである（図１３のステップ１３００を参照されたい）。

【０１６９】次に、肥満表現型を有する５００個の個体から得たゲノムＤＮＡを、上記のように長距離（ｌｏｎｇ　ｄｉｓｔａｎｃｅ）ＰＣＲおよびマイクロアッセイを用いて、変異体について分析し（Ｌｉｐｓｈｕｔｚらに付与された米国特許第６，３００，０６３号、およびＣｈｅｅらに発行された米国特許第５，８３７，８３２号も参照されたい）、各情報提供ＳＮＰのマイナー対立遺伝子の頻度をこの臨床集団について決定する（図１３のステップ１３１０を参照されたい）。これら２つの集団の情報提供ＳＮＰのマイナー対立遺伝子頻度を比較し、対照集団および臨床集団が、３つの情報提供ＳＮＰ位置において統計的に有意な差を有することを決定する（ステップ１３２０および１３３０）。対照集団および臨床集団のマイナー対立遺伝子頻度の差が最も大きなＳＮＰ位置を選択して分析する。

【０１７０】選択された情報提供位置は、レプチン遺伝子のコード領域（４ｋｂ）および該コード領域の５’側の非コード配列（１ｋｂ）にまたがることが判明したＳＮＰハプロタイプブロックの中に含まれる（ステップ１３４０）。この領域内に含まれる変異の分析は、この領域の中のあるＳＮＰ位置にあるＧがレプチン遺伝子のプロモーターの破壊に関与しており、レプチンタンパク質の発現がこれに比例して低下していることを示す。

【０１７１】皮膚生検により被験者から繊維芽細胞を得る。得られた組織を組織培養培地中に入れ、小片に分ける。この組織小片を、培地を含む組織培養フラスコの湿った表面の底に入れる。室温にて２４時間後、新しい培地（例えば１０％　ＦＢＳ、ペニシリンおよびストレプトマイシンを加えたＨａｍ’ｓ　Ｆ１２培地）を加える。次に組織を３７℃にて約１週間インキュベートする。このとき、新しい培地を加え、続いて数日おきに新しい培地に換える。さらに２週間培養した後、繊維芽細胞の単層が現れる。この単層をトリプシン処理し、大きなフラスコに移す。

【０１７２】モロニーネズミ白血病ウイルスから得たベクター（このベクターはカナマイシン耐性遺伝子を含む）を制限酵素で消化し、発現させる断片をクローニングする。消化されたベクターを仔ウシ腸ホスファターゼで処理し、自己ライゲーションを防ぐ。脱リン酸化した線状ベクターをアガロースゲル上で分離・精製する。レプチンｃＤＮＡ（活性レプチンタンパク質産物を発現することができる）を単離する。断片の末端を修飾し、必

要であればベクター中にクローニングする。等モル量のモロニーネズミ白血病ウイルスの線状主鎖（ｂａｃｋｂｏｎｅ）およびレプチン遺伝子断片を混ぜ合わせ、Ｔ４ＤＮＡリガーゼを用いてつなげる。このライゲーション混合物を用いて大腸菌を形質転換した後、この細菌を、カナマイシン含有寒天上に接種する。カナマイシンの表現型および制限分析により、このベクターの中にレプチン遺伝子がきちんと挿入されたか否かを確かめる。

【０１７３】１０％仔ウシ血清、ペニシリンおよびストレプトマイシンを加えたダルベッコ改変イーグル培地（ＤＭＥＭ）中で組織培養を行い、パッケージング細胞をコンフルエントな密度になるまで増殖させる。レプチン遺伝子を含むベクターを標準的手法によりパッケージング細胞中に導入する。このパッケージング細胞に新しい培地を加え、適当な時間インキュベートした後、コンフルエントなパッケージング細胞のプレートから培地を回収する。感染性ウイルス粒子を含む培地をＭｉｌｌｉｐｏｒｅフィルタで濾過して剥離したパッケージング細胞を除去した後、線維芽細胞を感染させるために用いる。線維芽細胞が準コンフルエント状態のプレートから培地を除去し、濾過した培地に素早く取り替える。形質導入を容易とするためにポリブレン（Ａｌｄｒｉｃｈ）を培地に入れても良い。適当な時間インキュベートを行った後、培地を除去して新しい培地に取り替える。ウイルス力価が高ければ、事実上全ての線維芽細胞が感染しており、選択の必要はない。力価が低ければ、選択マーカー（例えばｎｅｏやｈｉｓ等）を有するレトロウイルスベクターを用いて、増殖するための形質導入細胞を選択する必要がある。

【０１７４】次に、遺伝子操作した線維芽細胞を、単独で、またはマイクロキャリアビーズ（例えばｃｙｔｏｄｅｘ　３ビーズなど）上でコンフルエントになるまで増殖した後、個体中に導入する。注入された線維芽細胞はレプチン産物を生成し、該タンパク質の生物学的作用がその宿主に伝達される。

【０１７５】代替的にまたは更に、レプチン遺伝子を単離し、発現ベクター中にクローニングして、レプチンポリペプチドを産生するために使用する。発現ベクターは上記に開示したように、適切な転写および翻訳開始領域ならびに転写および翻訳停止領域を含む。単離したレプチンタンパク質をこのように産生して、該タンパク質に結合する物質を同定するか、あるいは遺伝子操作されたレプチン遺伝子およびタンパク質を発現する細胞を、物質を同定するアッセイで用いる。このような物質は、例えば候補物質を単離したレプチンポリペプチドに、ポリペプチド／化合物複合体を形成するのに十分な時間接触させ、そして該複合体を検出することによって同定される。ポリペプチド／化合物複合体が検出されたら、レプチンポリペプチドに結合する化合物を同定する。この方法によって同定された物質は、レプチンの活性をモジュレートする化合物を含み得る。このようにスクリーニングされた物質は、ペプチド、炭水化物、ビタミン誘導体、および他の小分子または医薬物質である。物質を同定するための生物学的アッセイに加え、物質を、レプチンタンパク質の立体構造に基づいて、タンパク質モデリング手法を用いて選択された候補物質を選択することによって予めスクリーニングしてもよい。

【０１７６】レプチンタンパク質に結合する物質の同定に加え、レプチン遺伝子に結合して遺伝子発現を制御する配列特異的またはエレメント特異的物質も同定される。核酸結合物質の１つのクラスは、レプチンｍＲＮＡにハイブリダイズして翻訳を遮断する塩基残基を含む物質である（例えばアンチセンスオリゴヌクレオチド等）。核酸結合物質の他のクラスは、ＤＮＡと３重らせんを形成して転写を遮断するものである（３本鎖オリゴヌクレオチド（ｔｒｉｐｌｅｘ　ｏｌｉｇｏｎｕｃｌｅｏｔｉｄｅｓ））。このような物質は通常２０～４０個の塩基を含み、古典的なホスホジエステル、リボ核酸主鎖に基づくものであるか、或いは塩基結合能を有する種々のスルフヒドリル誘導体またはポリマー誘導体であってもよい。

【０１７７】更に、レプチン遺伝子に特異的にハイブリダイズする対立遺伝子特異的オリゴヌクレオチドおよび／または変異体レプチンタンパク質に特異的に結合する物質（例えば変異体特異的抗体）を診断薬として用いることができる。対立遺伝子特異的オリゴヌクレオチドを調製および使用するための方法、ならびに抗体の調製方法は上記に記載済みであり、当分野において公知である。

【０１７８】この明細書中に記載した全ての特許および出版物は、本発明が属する分野において通常の知識を有する者のレベルを示す。全ての特許および出版物は、各々の出版物が特に且つ個々に本明細書中に組み込まれるものとする旨を注記したものとして、本明細書中に組み込まれる。

【０１７９】本発明は、個々の変異を同定し、ＳＮＰハプロタイプブロックを決定し、ハプロタイプパターンを決定し、そしてさらにこのＳＮＰハプロタイプパターンを用いて情報提供ＳＮＰを同定することによってゲノム全域の関連付け調査を行うための非常に進歩した方法を提供する。従来公知でない実用的且つコストが低い方法で、該情報提供ＳＮＰを用いて疾患および薬物応答の遺伝的根拠を詳細に分析することができる。上記説明は例示的なものであって限定的なものではないことを理解されたい。上記説明を読めば当業者には多くの実施形態が自明であろう。従って本発明の範囲は、上記説明を参照して決定されるのではなく、特許請求の範囲を参照して決定されるべきものであり、このような特許請求の範囲の権利が及ぶ同等物の全ての範囲を含むものとする。

【０１８０】

【配列表】

SEQUENCE LISTING

<110> Perlegen Sciences, Inc.
PATIL, Nila
COX, David R.
BERNO, Anthony J.
HINDS, David A.
FODOR, Stephen P. A.
<120> Methods for Genomic Analy
sis
<130> 054801-5001
<150> US 60/280,530
<151> 2001-03-30
<150> US 60/313,264
<151> 2001-08-17
<150> US 60/327,006
<151> 2001-10-05
<150> US 60/332,550
<151> 2001-11-26
<160> 7
<170> PatentIn version 3.1

<210> 1
<211> 13
<212> DNA
<213> Artificial sequence
<220>
<223> Sample SNP Haplotype: W
<400> 1
agattcgata acg
13
<210> 2
<211> 13
<212> DNA
<213> Artificial sequence
<220>
<223> Sample SNP Haplotype: X
<400> 2
agactacata acg
13
<210> 3
<211> 13
<212> DNA
<213> Artificial sequence

<220>
<223> Sample SNP Haplotype: Y
<400> 3
tatttcgata acg
13
<210> 4
<211> 13
<212> DNA

<220>
<223> Sample SNP Haplotype: Z
<400> 4
tatctacaat cac
13
<210> 5
<211> 13
<212> DNA
<213> Artificial sequence
<220>
<223> SNP sequence
<400> 5
agtaacccct ttt
13

<210> 6
<211> 13
<212> DNA
<213> Artificial sequence
<220>
<223> SNP sequence
<400> 6
actgacccct ttt
13
<210> 7
<211> 13
<212> DNA
<213> Artificial sequence
<220> 68
<223> SNP sequence
<400> 7 69

【図面の簡単な説明】

【図１】本発明の方法の１つの実施形態の概略図であり、変異体の位置の同定と変異体と表現型との関連付け、薬剤発見標的を同定するためまたは診断マーカーとしてこの関連付けの使用を示している。

【図２】本発明に従ったサンプルＳＮＰハプロタイプブロックおよびＳＮＰハプロタイプパターンを示す。

【図３】ＳＮＰハプロタイプブロックの選択方法の一実施形態を示す概略図である。

【図４】図３に示した方法の一実施形態の単純な使用を示す。

【図５Ａ】ＳＮＰハプロタイプブロックの最終セットを選択するための方法の一実施形態の概略図である。

【図５Ｂ】図５Ａに示した方法の簡単な使用であり、図中に示した「文字：数字」は、各ブロックの「ハプロタイプブロックＩＤ：情報提供値」を示す。

【図６】本発明の一実施形態に従って情報提供ＳＮＰがどのように選択されるのかを示す例である。

【図７Ａ】変異体の曖昧性および／またはＳＮＰハプロタイプパターンの曖昧性を解決する一実施形態を示す概略図である。

【図７Ｂ】図７Ａに示した方法の簡単な使用を示す。

【図８】関係付けの調査における本発明の方法の使用の一実施形態の概略図である。

【図９】本発明の幾つかの実施形態を実行するのに適した例示的なコンピュータネットワークシステムを示す。

【図１０】体細胞ハイブリッドの構築を示す概略図である。

【図１１】Ａｆｆｙｍｅｔｒｉｘ，　Ｉｎｃ社のＨｕＳＮＰ遺伝子チップを用いたハムスター／ヒト細胞ハイブリッドのスクリーニングにより得た結果の一部を示す表である。

【図１２】長範囲ＰＣＲを用いたヒト第２２番染色体およびヒト第１４番染色体のゲノムＤＮＡの様々な増幅ゲノム領域の例を示す。

【図１３Ａ】ＳＮＰのマイナー対立遺伝子（変異体）の頻度に対してプロットしたＳＮＰのパーセンテージを示す棒グラフである。

【図１３Ｂ】２００ｋｂインターバルにおけるヌクレオチドの多様性の関数として該インターバルのパーセンテージを示すグラフである。

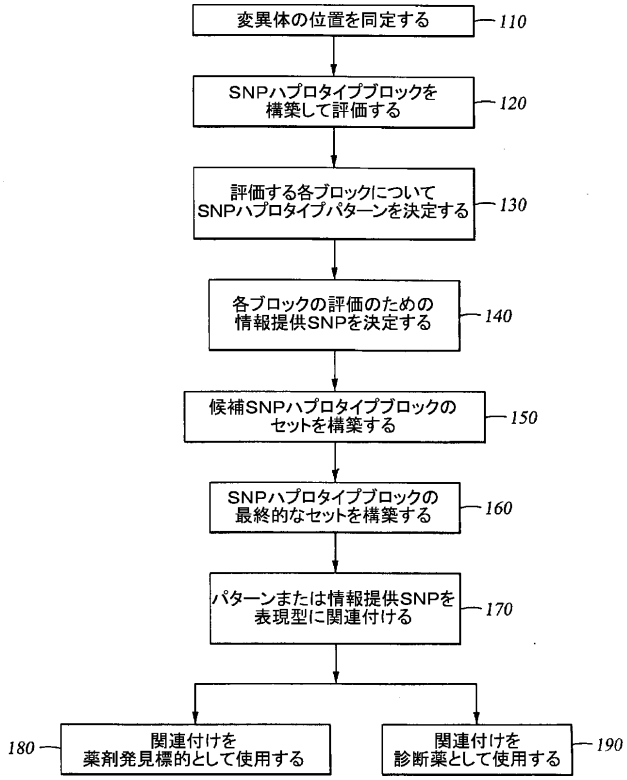
【図１３Ｃ】インターバルの長さに対してプロットした

全てのインターバルのパーセンテージを示す棒グラフである。

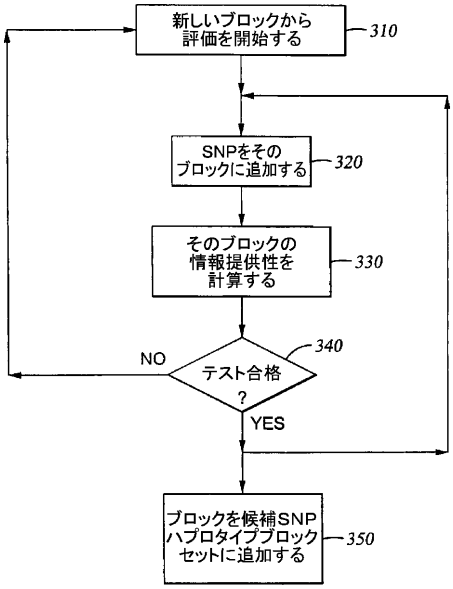
【図１４】１４７個の共通ヒト第２１番染色体ＳＮＰにより定義される２０個の独立した全体的に多様な染色体*のハプロタイプパターンを示す。

【図１５】網羅される染色体の割合を、その網羅度に必要なＳＮＰの数の関数としたプロットである。

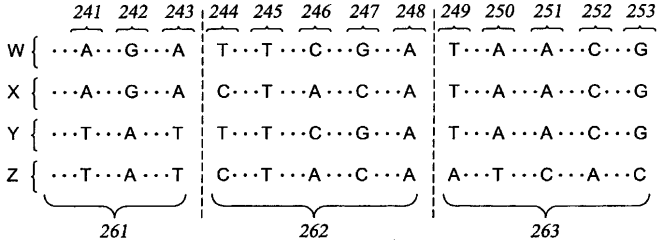
【図１】



【図３】



【図２】

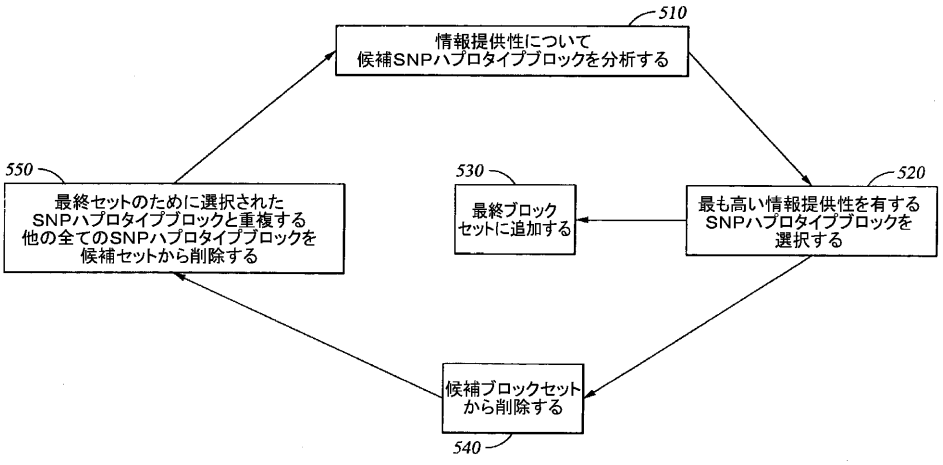


【図４】

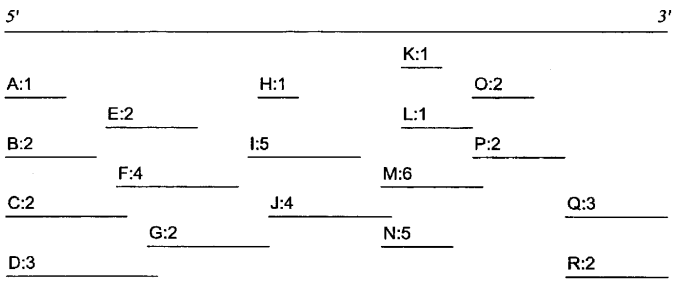
評価するブロック	SNP位置 1	2	3	4	5	6	情報提供性を有するか？
A	1						有り
B	1	2					有り
C	1	2	3				有り
D	1	2	3	4			無し
E		2					有り
F		2	3				有り
G		2	3	4			有り
H		2	3	4	5		無し
I			3				有り
J			3	4			無し
K				4			有り
L				4	5		有り
M				4	5	6	有り

候補セットとして選択されたブロック：ABCEFGIKLM

【図５Ａ】

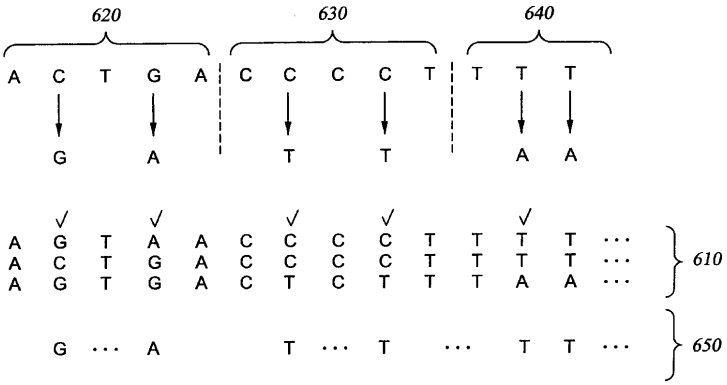


【図５Ｂ】

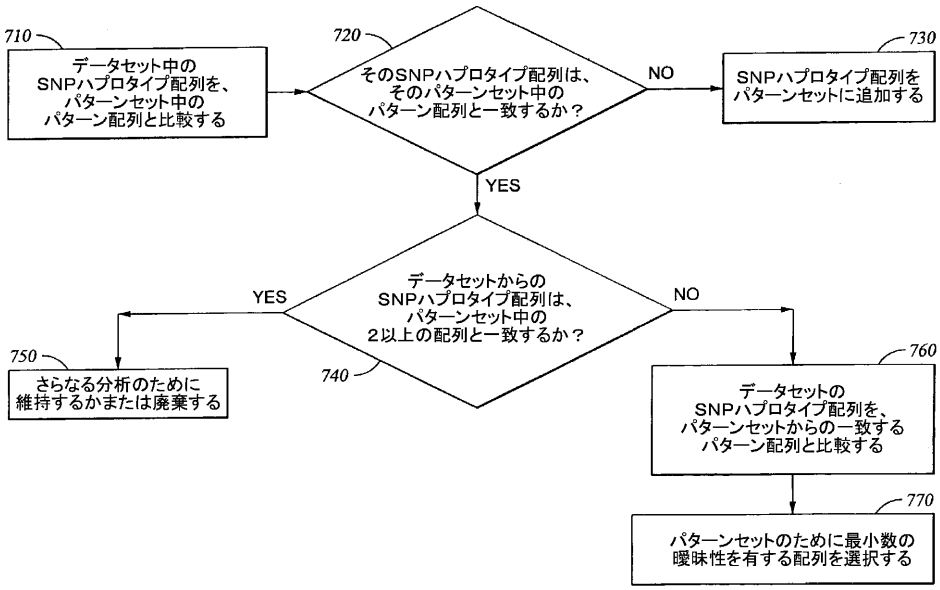


M　廃棄　J、N、K、L、OおよびP
I　廃棄　H
F　廃棄　E、G、CおよびD
Q　廃棄　R
B　廃棄　A

【図６】



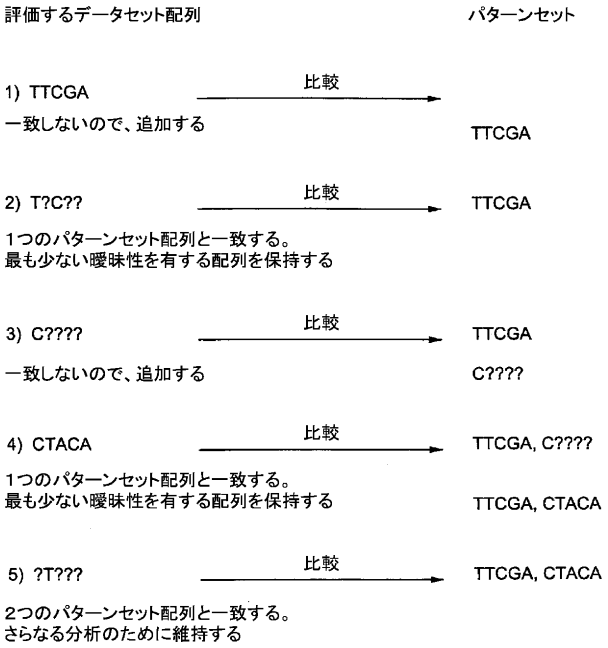
【図７Ａ】



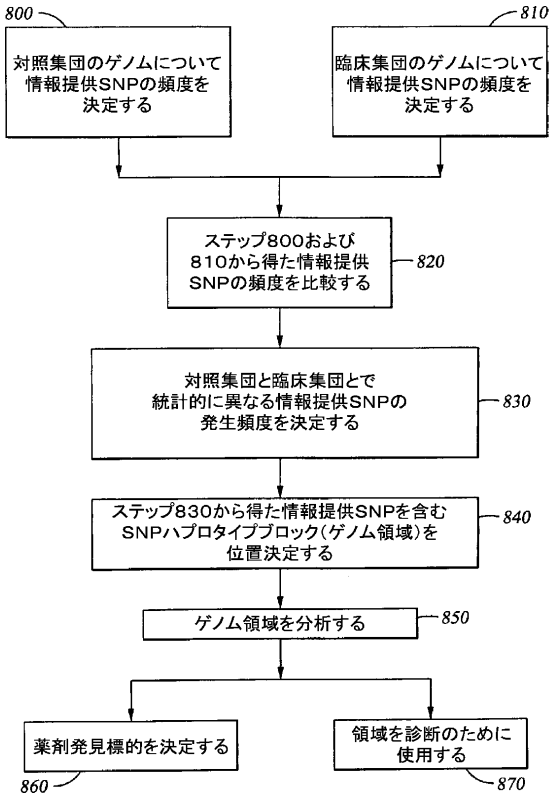
【図１１】

第21番染色体 HuSNPマーカー	ハムスター	CPD17	ハイブリッド1	ハイブリッド2
WIAF-3497	シグナル無し	A	A	A
WIAF-3498	シグナル無し	AB	A	B
WIAF-599	シグナル無し	A	A	A
WIAF-3562	シグナル無し	シグナル無し	A	B
WIAF-559	シグナル無し	AB	B	A
WIAF-4546	シグナル無し	AB	B	A
WIAF-3508	シグナル無し	B	B	B
WIAF-624	シグナル無し	B	B	B
WIAF-1500	シグナル無し	A	A	A
WIAF-3496	シグナル無し	AB	A	B
WIAF-1943	シグナル無し	A	A	A
WIAF-2477	シグナル無し	シグナル無し	シグナル無し	A
WIAF-1538	シグナル無し	B	シグナル無し	B
WIAF-3479	シグナル無し	A	A	シグナル無し
WIAF-2436	シグナル無し	A	A	A
WIAF-1857	シグナル無し	AB	B	A
WIAF-899	シグナル無し	AB	A	B
WIAF-1682	シグナル無し	B	B	B
WIAF-2214	シグナル無し	AB	A	B
WIAF-2643	シグナル無し	シグナル無し	A	シグナル無し
WIAF-4514	シグナル無し	B	B	B

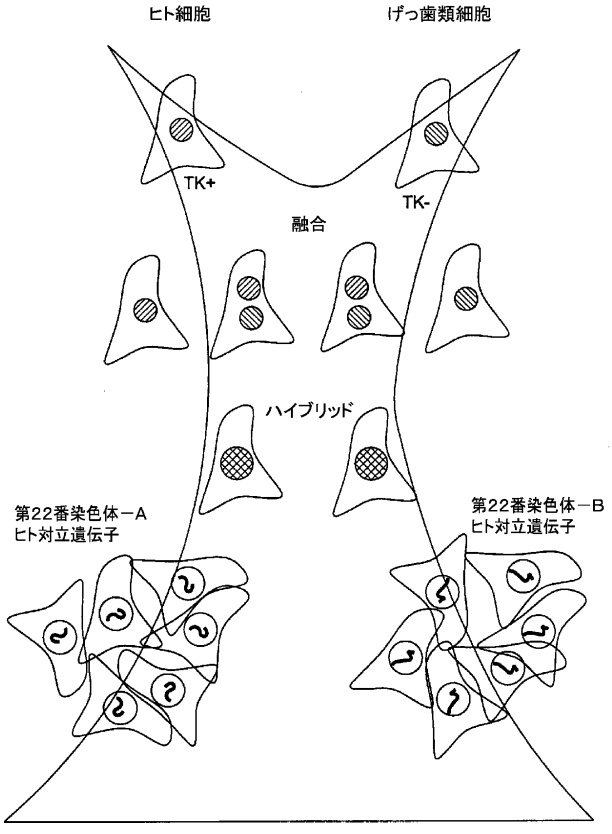
【図７Ｂ】


評価するデータセット配列
パターンセット
1) TTCGA
比較
一致しないので、追加する
TTCGA
2) T?C??
比較
TTCGA
1つのパターンセット配列と一致する。
最も少ない曖昧性を有する配列を保持する
3) C????
比較
TTCGA
一致しないので、追加する
C????
4) CTACA
比較
TTCGA, C????
1つのパターンセット配列と一致する。
最も少ない曖昧性を有する配列を保持する
TTCGA, CTACA
5) ?T???
比較
TTCGA, CTACA
2つのパターンセット配列と一致する。
さらなる分析のために維持する

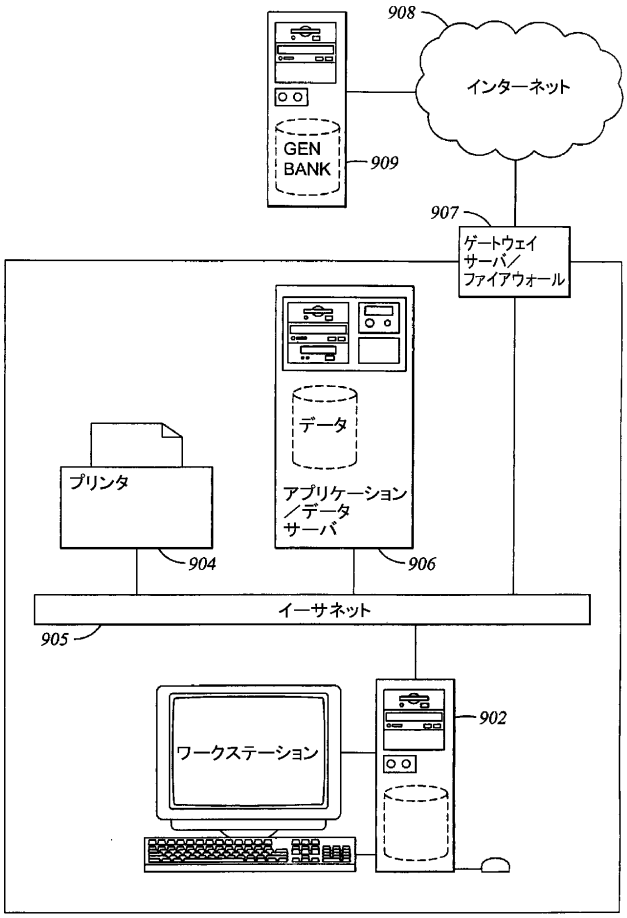
【図８】


800
対照集団のゲノムについて情報提供SNPの頻度を決定する
810
臨床集団のゲノムについて情報提供SNPの頻度を決定する
ステップ800および810から得た情報提供SNPの頻度を比較する
820
対照集団と臨床集団とで統計的に異なる情報提供SNPの発生頻度を決定する
830
ステップ830から得た情報提供SNPを含むSNPハプロタイプブロック(ゲノム領域)を位置決定する
840
ゲノム領域を分析する
850
薬剤発見標的を決定する
860
領域を診断のために使用する
870

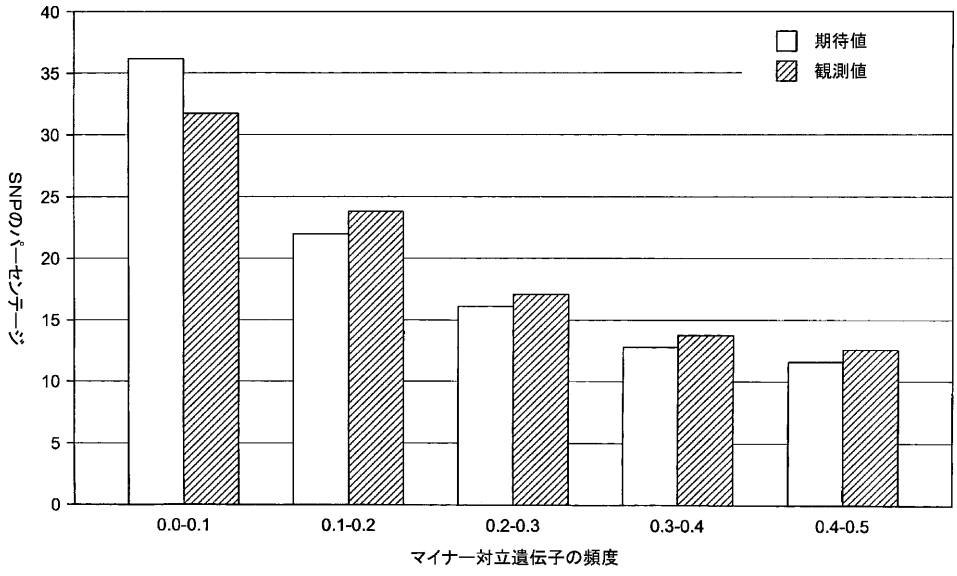
【図１０】


ヒト細胞
げっ歯類細胞
TK+
TK-
融合
ハイブリッド
第22番染色体－A
ヒト対立遺伝子
第22番染色体－B
ヒト対立遺伝子

【図９】

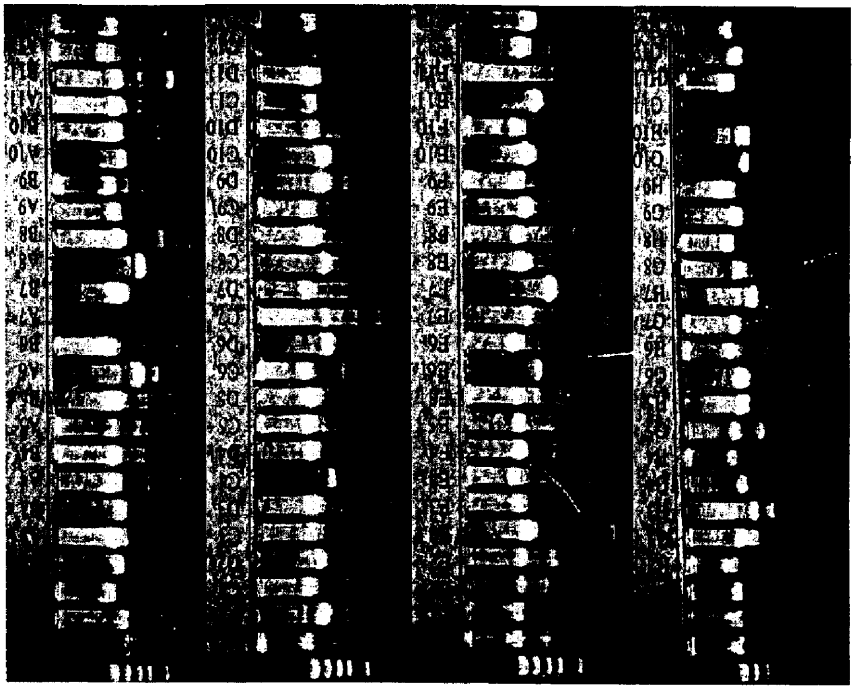

908
インターネット
GEN
BANK
909
907
ゲートウェイ
サーバ／
ファイアウォール
データ
プリンタ
アプリケーション
／データ
サーバ
904
906
イーサネット
905
ワークステーション
902

【図１３Ａ】

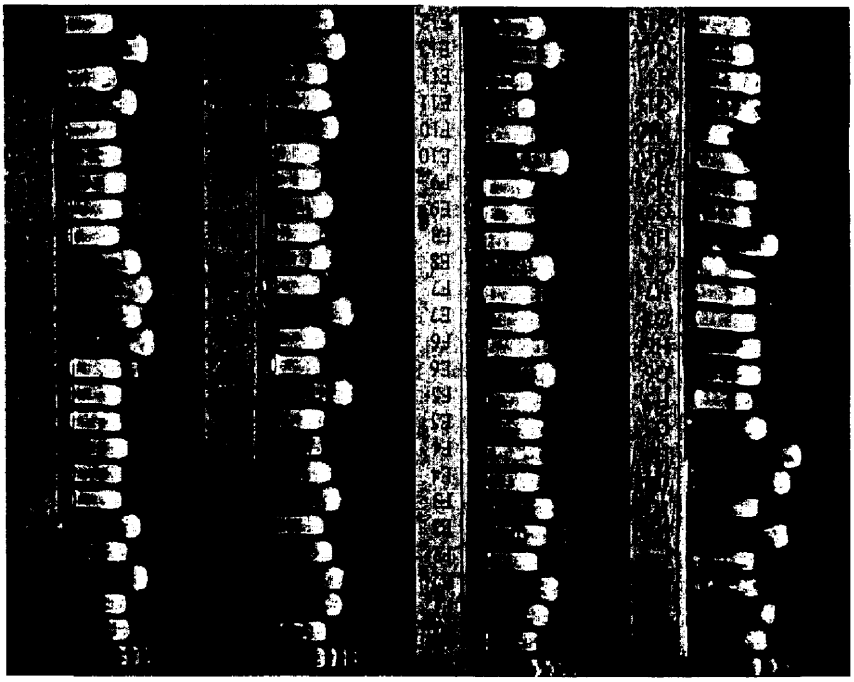

40
35
30
25
20
15
10
5
0
SNPのパーセンテージ
期待値
観測値
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
マイナー対立遺伝子の頻度

【図１２】

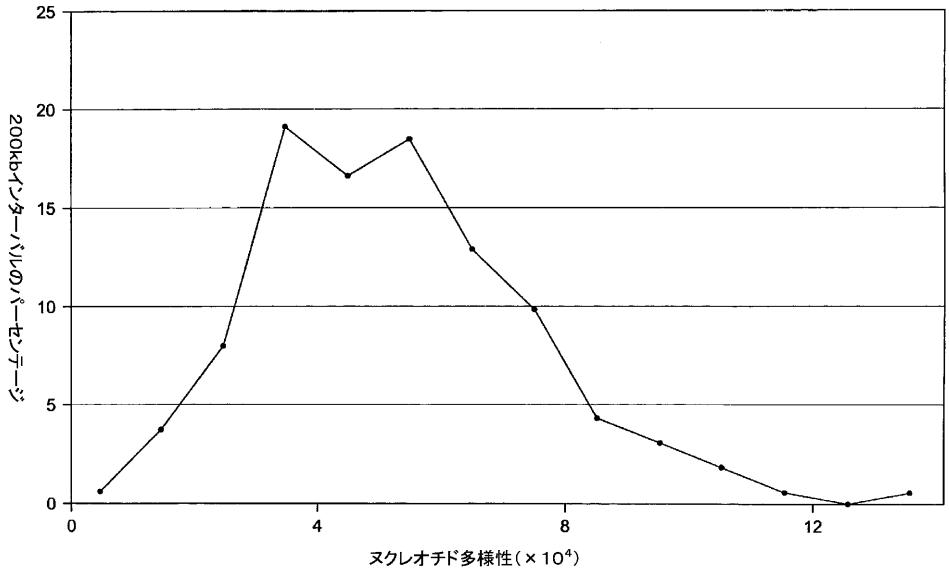
第22番染色体



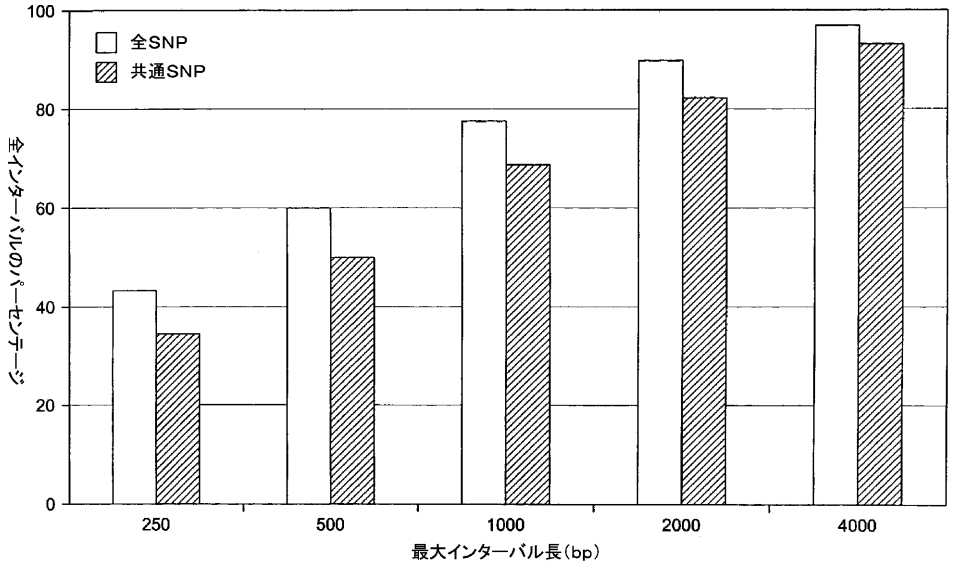
第14番染色体



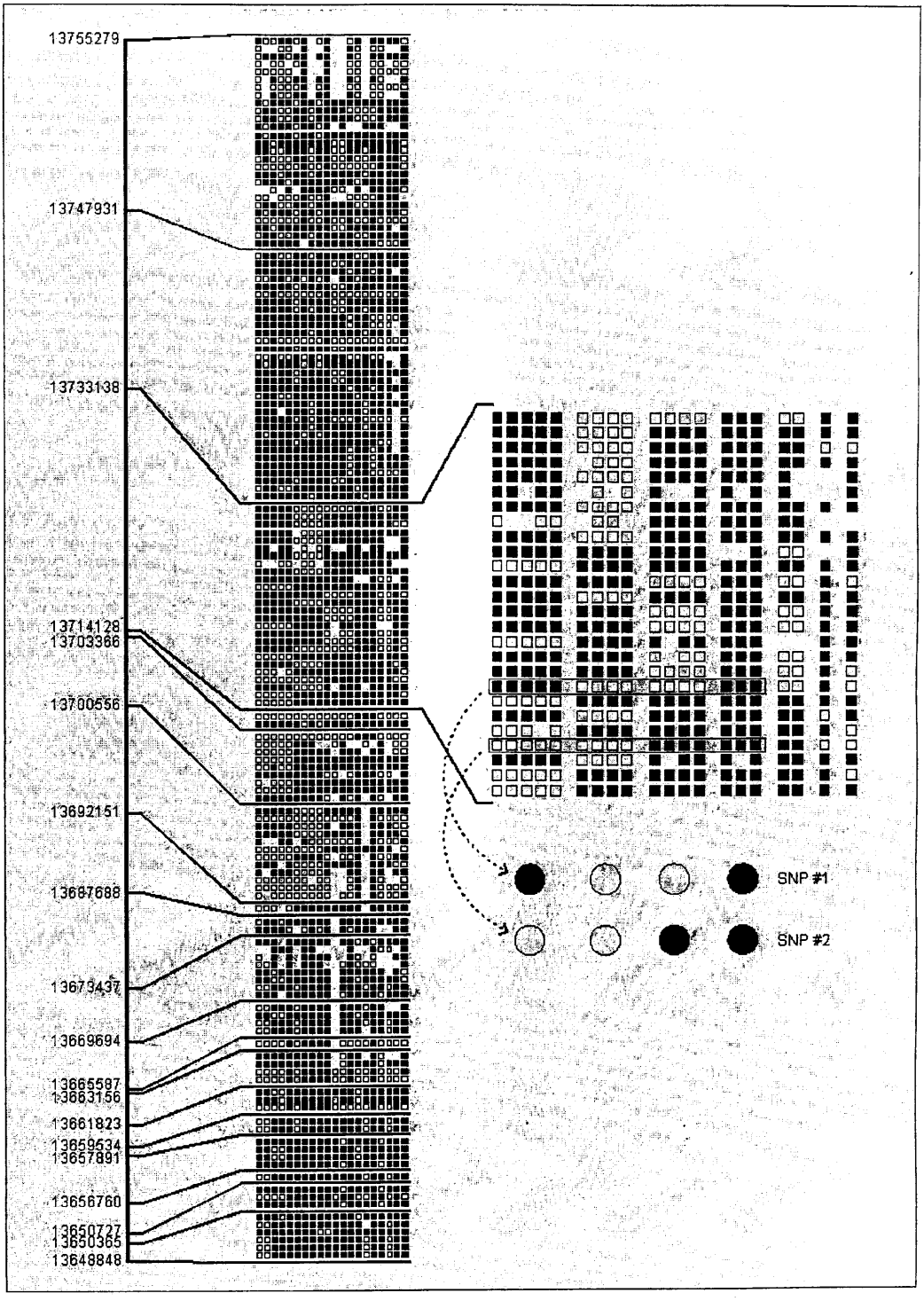
【図１３Ｂ】


25
20
15
10
5
0
200kbインターバルのパーセンテージ
0
4
8
12
ヌクレオチド多様性（×10⁴）

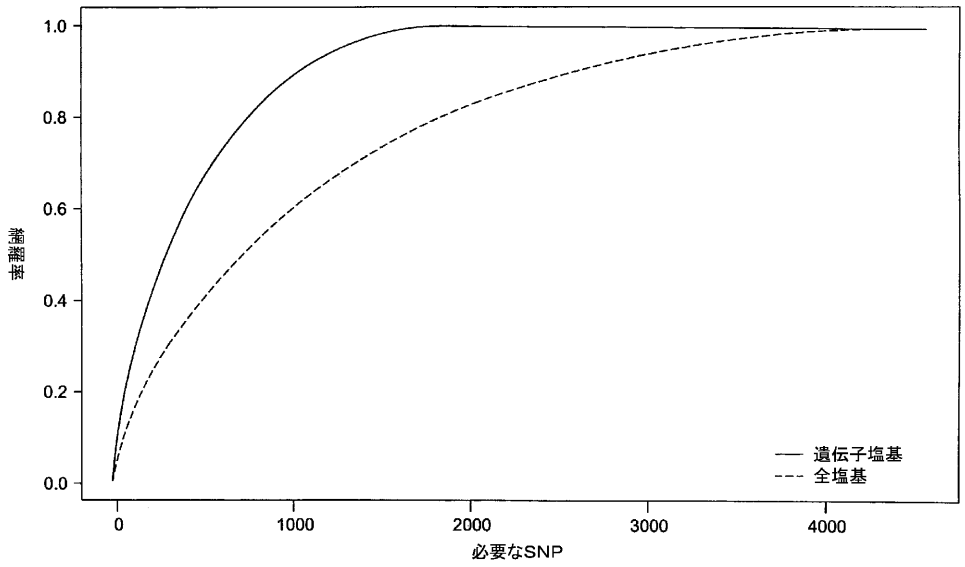
【図１３Ｃ】


100
80
60
40
20
0
全SNP
共通SNP
全インターバルのパーセンテージ
250
500
1000
2000
4000
最大インターバル長（bp）

【図１４】


13755279
13747931
13733138
13714128
13703366
13700556
13692151
13687688
13673437
13669694
13665587
13663156
13661823
13659534
13657891
13656760
13650727
13650365
13648848
SNP #1
SNP #2

【図１５】



フロントページの続き

(31)優先権主張番号　６０／３３２５５０
(32)優先日　　平成13年11月26日(2001．11．26)
(33)優先権主張国　　米国（ＵＳ）
(72)発明者　ニラ　パティル
アメリカ合衆国，　カリフォルニア州，
マウンテン　ヴュー，　スティアリン　コート　2021　パーレジェン　サイエンシーズ　インコーポレイテッド内
(72)発明者　デヴィッド　アール．　コックス
アメリカ合衆国，　カリフォルニア州，
マウンテン　ヴュー，　スティアリン　コート　2021　パーレジェン　サイエンシーズ　インコーポレイテッド内
(72)発明者　アンソニー　ジェイ．　バーノ
アメリカ合衆国，　カリフォルニア州，
マウンテン　ヴュー，　スティアリン　コート　2021　パーレジェン　サイエンシーズ　インコーポレイテッド内
(72)発明者　デヴィッド　エー．　ハインズ
アメリカ合衆国，　カリフォルニア州，
マウンテン　ヴュー，　スティアリン　コート　2021　パーレジェン　サイエンシーズ　インコーポレイテッド内

【外国語明細書】

1 Title of Invention

METHODS FOR GENOMIC ANALYSIS

2 Claims

1. A method for selecting SNP haplotype patterns, comprising:

isolating a substantially identical nucleic acid strand from a plurality of different origins for analysis;

determining more than one SNP location in each nucleic acid strand;

identifying SNP locations in said nucleic acid strands that are linked, wherein said linked SNP locations form a SNP haplotype block;

identifying isolate SNP haplotype blocks;

identifying SNP haplotype patterns that occur in each SNP haplotype block and isolate SNP haplotype block; and

selecting each identified SNP haplotype pattern that occurs in at least two of said substantially identical nucleic acid strands from different origins.

2. The method of claim 1, wherein said first identifying step is determined by a greedy algorithm or a shortest-paths algorithm.

3. The method of claim 1, wherein said SNP haplotype blocks are non-overlapping.

4. The method of claim 1, wherein said substantially identical nucleic acid strands are from at least between about 10 to about 100 different origins.

5. The method of claim 4, wherein said substantially identical nucleic acid strands are from at least about 16 different origins.

6. The method of claim 5, wherein said substantially identical nucleic acid strands are from at least about 25 different origins.

7. The method of claim 6, wherein said substantially identical nucleic acid strands are from at least about 50 different origins.

8. The method of claim 1, wherein said substantially identical nucleic acid strands are genomic DNA strands.

9. The method of claim 1, wherein at least ten percent of genomic DNA from an organism is isolated and analyzed.

10. The method of claim 1, wherein at least 1×10^8 bases from said substantially identical nucleic acid strands are isolated and analyzed.

11. The method of claim 1, wherein selected repeat regions from said substantially identical nucleic acid strands are not analyzed.

12. The method of claim 1, further comprising:

after said determining step, identifying which SNP locations occur only once in said plurality of identical nucleic acid strands; and

excluding said once-occuring SNP locations from analysis.

13. The method of claim 1, further comprising:

selecting a SNP haplotype pattern that occurs most frequently in said substantially identical nucleic acid strands; and

selecting a SNP haplotype pattern that occurs next most frequently in said substantially identical nucleic acid strands; and

repeating said second selecting step until said selected SNP haplotype patterns identify a portion of said substantially identical nucleic acid strands.

14. The method of claim 13, wherein said portion is between about 70% and 99% of said substantially identical nucleic acid strands.

15. The method of claim 14, wherein said portion is at least about 80% of said substantially identical nucleic acid strands.

16. The method of claim 13, wherein no more than about three SNP haplotype patterns are selected.

17. A method for selecting a data set of SNP haplotype blocks for data analysis, comprising:

comparing SNP haplotype blocks for informativeness;

selecting a first SNP haplotype block with a high informativeness;

adding said first SNP haplotype block to said data set;

selecting a second SNP haplotype block with a high informativeness;

adding said second selected SNP haplotype block to said data set; and

repeating said selecting and adding steps until a region of interest of a nucleic acid strand is covered.

18. The method of claim 17, wherein said selected SNP haplotype blocks are nonoverlapping.

19. The method of claim 17, wherein a greedy algorithm is used to perform said selecting steps.

20. A method for determining an informative SNP in a SNP haplotype pattern, comprising:

determining SNP haplotype patterns for a SNP haplotype block;

comparing each SNP haplotype pattern of interest in said SNP haplotype block to other SNP haplotype patterns of interest in said SNP haplotype block;

selecting at least one SNP in a first SNP haplotype pattern of interest that distinguishes such first SNP haplotype pattern of interest from other SNP haplotype patterns of interest in said SNP haplotype block, wherein said selected at least one SNP is an informative SNP for said first SNP haplotype pattern in said SNP haplotype block.

21. The method of claim 20, further comprising repeating said selecting step until a sufficient number of informative SNPs are selected to distinguish a portion of SNP haplotype patterns in a SNP haplotype block.

22. The method of claim 21, wherein said selected portion of SNP haplotype patterns is about 70% to about 99% of SNP haplotype patterns in said SNP haplotype block.

23. The method of claim 21, wherein said selected protion of SNP haplotype patterns allows identification of a disease of interest.

24. A method of determining informativeness of a SNP haplotype block, comprising:

determining a number of SNP locations in said SNP haplotype block;

determining a number of informative SNPs required to distinguish SNP haplotype patterns of interest in said SNP haplotype block; and

dividing said number of SNP locations by said number of informative SNPs to produce a quotient, wherein said quotient is said informativeness of said SNP haplotype block.

25. A method of determining informativeness of a SNP haplotype block, comprising:

determining a number of SNP locations in said SNP haplotype block;

determining a number of informative SNPs required to distinguish SNP haplotype patterns of interest in said SNP haplotype block from each other, wherein said number of informative SNPs required to distinguish SNP haplotype patterns of interest is said informativeness of said SNP haplotype block.

26. A method for determining disease-related genetic loci without a priori knowledge of a sequence or location of said disease-related genetic loci, comprising:

determining SNP haplotype patterns from at least 16 individuals in a control population;

determining SNP haplotype patterns from individuals in a diseased population; and

comparing frequencies of said SNP haplotype patterns of said control population with frequencies of said SNP haplotype patterns of said diseased population, wherein differences in said frequencies indicate locations of disease-related genetic loci.

27. The method of claim 26, wherein said SNP haplotype patterns are determined in at least 50 individuals in a control population.

28. The method of claim 26, wherein said SNP haplotype patterns from said populations are determined using informative SNPs.

29. A method of constructing a SNP haplotype block map using multiple whole genomes comprising:

arranging SNPs found in at least about ten percent of said whole genomes into SNP haplotype blocks.

30. A method of making associations between SNP haplotype patterns and a phenotypic trait of interest comprising:

building baseline of SNP haplotype patterns by the methods of the present invention;

pooling whole genomic DNA from a population having a common phenotypic trait of interest; and

identifying said SNP haplotype patterns that are associated with said phenotypic trait of interest.

31. The method of claim 30, wherein informative SNPs are used for said building and said identifying steps.

32. A method of identifying diagnostic markers comprising;

identifying informative SNPs according to claim 20, wherein said informative SNPs are diagnostic markers based on associations.

33. A method for identifying drug discovery targets comprising:

associating SNP haplotype patterns with a disease;

identifying a chromosomal location of said associated SNP haplotype patterns;

determining a nature of said association of said chromosomal location and said disease; and

selecting a chromosomal location or a product of expression of that chromosomal location that is associated with said disease; wherein said selected chromosomal location or a product of expression of that chromosomal location that is associated with said disease is a drug discovery target.

34. The method of claim 33, wherein said associated chromosomal locations are prioritized for drug discovery targets based on a set of criteria that includes location in a highly conserved region and location in an intergenic region.

35. The method of claim 33, wherein informative SNPs are used in said associating step.

36. A method of determining a SNP haplotype pattern of an individual comprising:

assaying for at least one informative SNP.

37. A method for defining SNP haplotype patterns of a species or subset of species comprising:

identifying SNPs present in genomes of multiple organisms of said species;

arranging said SNPs into SNP haplotype blocks by iteratively selecting for SNP haplotype patterns having few ambiguous positions.

38. A database comprising SNP haplotype blocks derived from genomes of multiple organisms, wherein said database identifies at least one informative SNP and wherein said database is on computer-readable medium.

39. A database on a computer-readable medium comprising SNP haplotype patterns identified as associated with one or more specific phenotypic traits.

40. A database on a computer-readable medium comprising informative SNPs identified as associated with one or more specific phenotypic traits.

41. The database of claim 38, 39 or 40, further comprising information on one or more factors selected from a group consisting of environmental factors, other genetic factors, related factors, including but not limited to biochemical markers, behaviors,

and/or other polymorphisms, including but not limited to low frequency SNPs, repeats, insertions and deletions.

42. A kit for diagnosis of a disease, disease susceptibility, or therapy response comprising means for detecting a presence or absence of SNP haplotype patterns or informative SNPs in a sample of genomic DNA from a patient and a data set of associations of said SNP haplotype patterns or informative SNPs with one or more specific phenotypic traits on a computer-readable medium.

43. An isolated nucleic acid comprising at least one informative SNP, wherein said informative SNP indicates a SNP haplotype pattern as determined in accordance with the methods of the invention, wherein said informative SNP is associated with a phenotypic trait.

44. A method comprising:

identifying genetic variations in a plurality of individuals;

identifying at least some of said genetic variations in individuals that occur with at least some other of said genetic variations; and

using some, but not all, of said variations that occur with at least some others of said genetic variations in correlation with a phenotypic state.

45. A method comprising:

determining a sequence of an organism;

scanning additional individuals of said organism for variants from said sequence;

identifying some of said variants that occur with others of said variants in a first group;

identifying some of said variants that occur with others of said variants in a second group; and

using some, but not all, of said variants in said first and second groups to correlate said groups with a phenotypic state.

46. A method for selecting a SNP haplotype block useful in genomic analysis, comprising:

isolating a substantially identical DNA strand from at least about five different origins for analysis;

analyzing at least about 1×10^6 bases from each of said substantially identical DNA strand from at least about five different origins;

determining more than one SNP location in each DNA strand;

identifying SNP locations in said DNA strands that are linked, wherein said linked SNP locations form a SNP haplotype block;

identifying SNP haplotype patterns that occur in each SNP haplotype block; and

selecting each identified SNP haplotype pattern that occurs in any of said substantially identical DNA strands from different origins.

47. A method for determining pharmacogenomic-related genetic loci without a priori knowledge of a sequence or location of said pharmacogenomic-related genetic loci, comprising:

determining SNP haplotype patterns from at least 16 individuals in a control population;

determining SNP haplotype patterns from individuals that react in an altered manner to administration of a substance; and

comparing frequencies of said SNP haplotype patterns of said control population with frequencies of said SNP haplotype patterns of said individuals that react in an altered manner to administration of a substance, wherein differences in said frequencies indicate locations of pharmacogenomic-related genetic loci.

48. The method of claim 47, wherein said SNP haplotype patterns are determined in at least 50 individuals in a control population.

49. The method of claim 47, wherein said SNP haplotype patterns from said populations are determined using informative SNPs.

3 Detailed Description of Invention

CROSS-REFERENCE TO RELATED APPLICATIONS

The present application claims priority to United States provisional patent application serial number 60/280,530, filed March 30, 2001, to United States provisional patent application serial number 60/313,264 filed August 17, 2001, to United States provisional patent application serial number 60/327,006, filed October 5, 2001, all entitled "Identifying Human SNP Haplotypes, Informative SNPs and Uses Thereof", and provisional patent application serial number 60/332,550 filed 11/26/01, entitled "Methods for Genomic Analysis", the disclosures all of which are specifically incorporated herein by reference.

BACKGROUND OF THE INVENTION

The DNA that makes up human chromosomes provides the instructions that direct the production of all proteins in the body. These proteins carry out the vital functions of life. Variations in the sequence of DNA encoding a protein produce variations or mutations in the proteins encoded, thus affecting the normal function of cells. Although environment often plays a significant role in disease, variations or mutations in the DNA of an individual are directly related to almost all human diseases, including infectious disease, cancer, and autoimmune disorders. Moreover, knowledge of genetics, particularly human genetics, has led to the realization that many diseases result from either complex interactions of several genes or their products or from any number of mutations within one gene. For example, Type I and II diabetes have been linked to multiple genes, each with its own pattern of mutations. In contrast, cystic fibrosis can be caused by any one of over 300 different mutations in a single gene.

Additionally, knowledge of human genetics has led to a limited understanding of variations between individuals when it comes to drug response—the field of pharmocogenetics. Over half a century ago, adverse drug responses were correlated with amino acid variations in two drug-metabolizing enzymes, plasma cholinesterase and glucose-6-phosphate dehydrogenase. Since then, careful genetic analyses have linked sequence polymorphisms (variations) in over 35 drug metabolism enzymes, 25 drug targets and 5 drug transporters with compromised levels of drug efficacy or safety (Evans

and Relling, *Science* 296:487-91 (1999)). In the clinic, such information is being used to prevent drug toxicity; for example, patients are screened routinely for genetic differences in the thiopurine methyltransferase gene that cause decreased metabolism of 6-mercaptopurine or azathiopurine. Yet only a small percentage of observed drug toxicities have been explained adequately by the set of pharmacogenetic markers validated to date. Even more common than toxicity issues may be cases where drugs demonstrated to be safe and/or efficacious for some individuals have been found to have either insufficient therapeutic efficacy or unanticipated side effects in other individuals.

In addition to the importance of understanding the effects of variations in the genetic make up of humans, understanding the effects of variation in the genetic makeup of other non-human organisms–particularly pathogens–is important in understanding their effect on or interaction with humans. For example, the expression of virulence factors by pathogenic bacteria or viruses greatly affects the rate and severity of infection in humans that come into contact with such organisms. In addition, a detailed understanding of the genetic makeup of experimental animals, *i.e.*, mice, rats, etc., is also of great value. For example, understanding the variations in the genetic makeup of animals used as model systems for evaluation of therapeutics is important for understanding the test results obtained using these systems and their predictive value for human use.

Because any two humans are 99.9% similar in their genetic makeup, most of the sequence of the DNA of their genomes is identical. However, there are variations in DNA sequence between individuals. For example, there are deletions of many-base stretches of DNA, insertion of stretches of DNA, variations in the number of repetitive DNA elements in non-coding regions, and changes in single nitrogenous base positions in the genome called "single nucleotide polymorphisms" (SNPs). Human DNA sequence variation accounts for a large fraction of observed differences between individuals, including susceptibility to disease.

Although most SNPs are rare, it has been estimated that there are 5.3 million common SNPs, each with a frequency of 10-50%, that account for the bulk of the DNA sequence difference between humans. Such SNPs are present in the human genome once every 600 base pairs (Kruglyak and Nickerson, *Nature Genet.* 27:235 (2001)). Alleles (variants) making up blocks of such SNPs in close physical proximity are often correlated, resulting in reduced genetic variability and defining a limited number of "SNP

haplotypes", each of which reflects descent from a single, ancient ancestral chromosome (Fullerton, *et al.*, *Am. J. Hum. Genet.* 67:881 (2000)).

The complexity of local haplotype structure in the human genome--and the distance over which individual haplotypes extend--is poorly defined. Empiric studies investigating different segments of the human genome in different populations have revealed tremendous variability in local haplotype structure. These studies indicate that the relative contributions of mutation, recombination, selection, population history, and stochastic events to haplotype structure vary in an unpredictable manner, resulting in some haplotypes that extend for only a few kilobases (kb), and others that extend for greater than 100 kb (A. G. Clark *et al.*, *Am. J. Hum. Genet.* 63:595 (1998)).

These findings suggest that any comprehensive description of the haplotype structure of the human genome, defined by common SNPs, will require empirical analysis of a dense set of SNPs in many independent copies of the human genome. Such whole-genome analyses would provide a fine degree of genetic mapping and pinpoint specific regions of linkage. Until the present invention, however, the practice and cost of genotyping over 3,000,000 SNPs across each individual of a reasonably sized population has made this endeavor impractical. The present invention allows for, among a wide variety of applications, whole-genome association analysis of populations using SNP haplotypes.

SUMMARY OF THE INVENTION

The present invention relates to methods for identifying variations that occur in the human genome and relating these variations to the genetic bases of phenotype such as disease resistance, disease susceptibility or drug response. "Disease" includes but is not limited to any condition, trait or characteristic of an organism that it is desirable to change. For example, the condition may be physical, physiological or psychological and may be symptomatic or asymptomatic. The methods allow for identification of variants, identification of SNPs, determination of SNP haplotype blocks, determining SNP haplotype patterns, and further, identification of informative SNPs for each pattern, which affords genetic data compression.

Thus, one aspect of the present invention provides methods for selecting SNP haplotype patterns useful in data analysis. Such selection can be accomplished by

isolating substantially identical (homologous) nucleic acid strands from a plurality of individuals; determining SNP locations in each nucleic acid strand; identifying the SNP locations in the nucleic acid strands that are linked, where the linked SNP locations form a SNP haplotype block; identifying isolate SNP haplotype blocks; identifying SNP haplotype patterns that occur in each SNP haplotype block; and selecting the identified SNP haplotype patterns that occur in at least two of the substantially identical nucleic acid strands. In one preferred embodiment, nucleic acid strands from at least about 10 different individuals or origins are used. In a more preferred embodiment, nucleic acid strands from at least 16 different origins are used. In an even more preferred embodiment, nucleic acid strands from at least 25 different origins are used, and in a yet more preferred embodiment, nucleic acid strands from at least 50 different origins are used. Further, a more preferred embodiment would determine SNP locations in at least about 100 nucleic acid strands from different origins. In addition, this method may further comprise selecting the SNP haplotype pattern that occurs most frequently in the substantially identical nucleic acid strands; selecting the SNP haplotype pattern that occurs next most frequently in the substantially identical nucleic acid strands; and repeating the selecting until the selected SNP haplotype patterns identify a portion of interest of the substantially identical nucleic acid strands. In a preferred embodiment, the portion of interest is between 70% and 99% of the substantially identical nucleic acid strands, and, in a more preferred embodiment, the portion of interest is about 80% of the substantially identical nucleic acid strands. Alternatively, one may wish to limit the selection of SNP haplotype patterns to no more than about three SNP haplotype patterns per SNP haplotype block.

In addition, the present invention provides a method for selecting a data set of SNP haplotype blocks for data analysis, comprising comparing SNP haplotype blocks for informativeness; selecting a first SNP haplotype block with high informativeness; adding the first SNP haplotype block to the data set; selecting a second SNP haplotype block with high informativeness; adding the second selected SNP haplotype block to the data set; and repeating the selecting and adding steps until the region of interest of a DNA strand is covered. In preferred embodiments, the SNP haplotype blocks selected are non-overlapping.

The present invention further provides methods for determining at least one informative SNP in a SNP haplotype pattern, comprising first determining SNP haplotype patterns for a SNP haplotype block, then comparing each SNP haplotype pattern of interest in the SNP haplotype block to the other SNP haplotype patterns of interest in the SNP haplotype block, and selecting at least one SNP in each SNP haplotype pattern that distinguishes this SNP haplotype pattern of interest from the other SNP haplotype patterns of interest in the SNP haplotype block. The selected SNP (or SNPs) is an informative SNP for the SNP haplotype pattern.

Also, the present invention allows for rapid scanning of genomic regions and provides a method for determining disease-related genetic loci or pharmacogenomic-related loci without a priori knowledge of the sequence or location of the disease-related genetic loci or pharmacogenomic-related loci. This can be done by determining SNP haplotype patterns from individuals in a control population, then determining SNP haplotype patterns from individuals in a experimental population, such as individuals in a diseased population or individuals that react in a particular manner when administered a drug. The frequencies of the SNP haplotype patterns of the control population are compared to the frequencies of the SNP haplotype patterns of the experimental population. Differences in these frequencies indicate locations of disease-related genetic loci or pharmacogenomic-related loci.

An additional aspect of the present invention provides a method of making associations between SNP haplotype patterns and a phenotypic trait of interest comprising: building baseline of SNP haplotype patterns of control individuals by the methods of the present invention; pooling whole genomic DNA from a clinical population having a common phenotypic trait of interest; and identifying the SNP haplotype patterns that are associated with the phenotypic trait of interest. Thus, the present invention allows for genome scanning to identify multiple haplotype blocks associated with a phenotype, which is particularly useful when studying polygenic traits.

Also, the present invention provides a method for identifying drug discovery targets comprising: associating SNP haplotype patterns with a disease; identifying a chromosomal location of the associated SNP haplotype patterns; determining the nature of the association of the chromosomal location and said disease; and using the gene or gene product of the chromosomal location as a drug discovery target.

DETAILED DESCRIPTION OF THE INVENTION

It readily should be apparent to one skilled in the art that various embodiments and modifications may be made to the invention disclosed in this application without departing from the scope and spirit of the invention. All publications mentioned herein are cited for the purpose of describing and disclosing reagents, methodologies and concepts that may be used in connection with the present invention. Nothing herein is to be construed as an admission that these references are prior art in relation to the inventions described herein.

As used in the specification, "a" or "an" means one or more. As used in the claim(s), when used in conjunction with the word "comprising", the words "a" or "an" mean one or more. As used herein, "another" means at least a second or more.

As used herein, when the term "different origins" is used, it refers to the fact DNA strands from different organisms come from a different origin. Further, each DNA strand in a single organism's genome come from different origins. In a diploid organism, an

individual organism's genome is made up of a set of pairs of substantially identical DNA strands. That is, a single individual would have substantially identical DNA strands from two different origins--one DNA strand of the pair is of maternal origin and one DNA strand of the pair is of paternal origin. Two or more nucleic acid sequences--for example, two or more DNA strands--are considered to be substantially identical if they exhibit at least about 70% sequence identity at the nucleotide level, preferably about 75%, more preferably about 80%, still more preferably about 85%, yet more preferably about 90%, even more preferably about 95% and even more preferably nucleic acid sequences are considered to be substantially identical if they exhibit at least about 98% sequence identity at the nucleotide level. The extent of sequence identity that is relevant between two or more nucleic acid sequences will depend on the host source of the nucleic acids. For example, a greater than 95% sequence identity may be relevant when looking at same species comparisons, whereas a sequence identity of 70% or even less may be relevant when making cross species comparisons. Of course, when one refers to DNA herein such reference may include derivatives of DNA such as amplicons, RNA transcripts, nucleic acid mimetics, etc.

As used herein, "individual" refers to a specific single organism, such as a single animal, human insect, bacterium, etc.

As used herein, "informativeness" of a SNP haplotype block is defined as the degree to which a SNP haplotype block provides information about genetic regions.

As used herein, the term "informative SNP" refers to a genetic variant such as a SNP or subset (more than one) of SNPs that tends to distinguish one SNP haplotype pattern from other SNP haplotype patterns within a SNP haplotype block.

As used herein, the term "isolate SNP block" refers to a SNP haplotype block that consists of one SNP.

As used herein, the term "linkage disequilibrium", "linked" or "LD" refers to genetic loci that tend to be transmitted from generation to generation together; *e.g.*, genetic loci that are inherited non-randomly.

As used herein, the term "singleton SNP haplotype" or "singleton SNP" refers to a specific SNP allele or variant that occurs in less than a certain portion of the population.

As used herein, the term "SNP" or "single nucleotide polymorphism" refers to a genetic variation between individuals; *e.g.*, a single nitrogenous base position in the DNA

of organisms that is variable. As used herein, "SNPs" is the plural of SNP. Of course, when one refers to DNA herein such reference may include derivatives of DNA such as amplicons, RNA transcripts, etc.

As used herein, the term "SNP haplotype block" means a group of variant or SNP locations that do not appear recombine independently and that can be grouped together in blocks of variants or SNPs.

As used herein, the term "SNP haplotype pattern" refers to the set of genotypes for SNPs in a SNP haplotype block in a single DNA strand.

As used herein, the term "SNP location" is the site in a DNA sequence where a SNP occurs.

As used herein a "SNP haplotype sequence" is a DNA sequence in a DNA strand that contains at least one SNP location.

Preparation of Nucleic Acids for Analysis

Nucleic acid molecules may be prepared for analysis using any technique known to those skilled in the art. Preferably such techniques result in the production of a nucleic acid molecule sufficiently pure to determine the presence or absence of one or more variations at one or more locations in the nucleic acid molecule. Such techniques may be found, for example, in Sambrook, *et al.*, Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, New York) (1989), and Ausubel, *et al.*, Current Protocols in Molecular Biology (John Wiley and Sons, New York) (1997), incorporated herein by reference.

When the nucleic acid of interest is present in a cell, it may be necessary to first prepare an extract of the cell and then perform further steps–i.e., differential precipitation, column chromatography, extraction with organic solvents and the like–in order to obtain a sufficiently pure preparation of nucleic acid. Extracts may be prepared using standard techniques in the art, for example, by chemical or mechanical lysis of the cell. Extracts then may be further treated, for example, by filtration and/or centrifugation and/or with chaotropic salts such as guanidinium isothiocyanate or urea or with organic solvents such as phenol and/or $HCCl_3$ to denature any contaminating and potentially interfering proteins. When chaotropic salts are used, it may be desirable to remove the salts from the

nucleic acid-containing sample. This can be accomplished using standard techniques in the art such as precipitation, filtration, size exclusion chromatography and the like.

In some instances, it may be desirable to extract and separate messenger RNA from cells. Techniques and material for this purpose are known to those skilled in the art and may involve the use of oligo dT attached to a solid support such as a bead or plastic surface. Suitable conditions and materials are known to those skilled in the art and may be found in the Sambrook and Ausubel references cited above. It may be desirable to reverse transcribe the mRNA into cDNA using, for example, a reverse transcriptase enzyme. Suitable enzymes are commercially available from, for example, Invitrogen, Carlsbad CA. Optionally, cDNA prepared from mRNA may then be amplified.

One approach particularly suitable for examining haplotype patterns and blocks is using somatic cell genetics to separate chromosomes from a diploid state to a haploid state. In one embodiment, a human lymphoblastoid cell line that is diploid may be fused to a hamster fibroblast cell line that is also diploid such that the human chromosomes are introduced into the hamster cells to produce cell hybrids. The resulting cell hybrids are examined to determine which human chromosomes were transferred, and which, if any, of the transferred human chromosomes are in a haploid state (see, e.g., Patterson, *et al.*, *Annal. N.Y. Acad. Of Sciences*, 396:69-81 (1982)).

A schematic of the procedure is shown in Figure 10. Figure 10 shows a diploid human lymphoblastoid cell line that is wildtype for the thymidine kinase gene being fused to a diploid hamster fibroblast cell line containing a mutation in the thymidine kinase gene. In a sub-population of the resulting cells, human chromosomes are present in hybrids. Selection for the human DNA-containing hybrid cells is achieved by utilizing HAT medium (selective medium). Only hybrid cells that have a stably-incorporated human DNA strand having the wildtype human thymidine kinase gene grow in cell culture medium containing HAT. Of the resulting hybrids, some hybrids may contain both copies of some human chromosomes, only one copy of a human chromosome or no copies of a particular human chromosome. For example, for a human chromosome 22 having a locus with either an A or a B allele, the resulting hybrid cells may contain one human chromosome 22 variant (e.g., the "A" variant) or a portion thereof, some may contain the other human chromosome 22 variant (the "B" variant) or a portion thereof, some may contain both human chromosome 22 variants or portions thereof, and some

hybrids may not contain any portion of a human chromosome 22 at all. In Figure 10, only two of the resulting hybrid populations are shown. Once the appropriate hybrids are selected, the nucleic acids from these hybrids may be isolated by, for example, the techniques described above and then subjected to SNP discovery, and haplotype block and pattern analyses of the present invention.

Amplification Techniques

It may be desirable to amplify one or more nucleic acids of interest before determining the presence or absence of one or more variations in the nucleic acid. Nucleic acid amplification increases the number of copies of the nucleic acid sequence of interest. Any amplification technique known to those of skill in the art may be used in conjunction with the present invention including, but not limited to, polymerase chain reaction (PCR) techniques. PCR may be carried out using materials and methods known to those of skill in the art.

PCR amplification generally involves the use of one strand of a nucleic acid sequence as a template for producing a large number of complements to that sequence. The template may be hybridized to a primer having a sequence complementary to a portion of the template sequence and contacted with a suitable reaction mixture including dNTPs and a polymerase enzyme. The primer is elongated by the polymerase enzyme producing a nucleic acid complementary to the original template.

For the amplification of both strands of a double stranded nucleic acid molecule, two primers may be used, each of which may have a sequence which is complementary to a portion of one of the nucleic acid strands. Elongation of the primers with a polymerase enzyme results in the production of two double-stranded nucleic acid molecules each of which contains a template strand and a newly synthesized complementary strand. The sequences of the primers typically are chosen such that extension of each of the primers results in elongation toward the site in the nucleic acid molecule where the other primer hybridizes.

The strands of the nucleic acid molecules are denatured–for example, by heating–and the process is repeated, this time with the newly synthesized strands of the preceding step serving as templates in the subsequent steps. A PCR amplification protocol may

involve a few to many cycles of denaturation, hybridization and elongation reactions to produce sufficient amounts of the desired nucleic acid.

Although PCR methods typically employ heat to achieve strand denaturation and allow subsequent hybridization of the primers, any other means that results in making the nucleic acids available for hybridization to the primers may be used. Such techniques include, but are not limited to, physical, chemical, or enzymatic means, for example, by inclusion of a helicase, (see Radding, *Ann. Rev. Genetics* 16: 405-436 (1982)) or by electrochemical means (see PCT Application Nos. WO 92/04470 and WO 95/25177).

Template-dependent extension of primers in PCR is catalyzed by a polymerase enzyme in the presence of at least 4 deoxyribonucleotide triphosphates (typically selected from dATP, dGTP, dCTP, dUTP and dTTP) in a reaction medium which comprises the appropriate salts, metal cations, and pH buffering system. Suitable polymerase enzymes are known to those of skill in the art and may be cloned or isolated from natural sources and may be native or mutated forms of the enzymes. So long as the enzymes retain the ability to extend the primers, they may be used in the amplification reactions of the present invention.

The nucleic acids used in the methods of the invention may be labeled to facilitate detection in subsequent steps. Labeling may be carried out during an amplification reaction by incorporating one or more labeled nucleotide triphosphates and/or one or more labeled primers into the amplified sequence. The nucleic acids may be labeled following amplification, for example, by covalent attachment of one or more detectable groups. Any detectable group known to those skilled in the art may be used, for example, fluorescent groups, ligands and/or radioactive groups. An example of a suitable labeling technique is to incorporate nucleotides containing labels into the nucleic acid of interest using a terminal deoxynucleotidyl transferase (TdT) enzyme. For example, a nucleotide–preferably a dideoxy nucleotide–containing a label is incubated with the nucleic acid to be labeled and a sufficient amount of TdT to incorporate the nucleotide. A preferred nucleotide is a dideoxynucleotide–i.e., ddATP, ddGTP, ddCTP, ddTTP, etc–having a biotin label attached.

Techniques to optimize the amplification of long sequences may be used. Such techniques work well on genomic sequences. The methods disclosed in pending US patent applications USSN 60/317,311, filed 9/5/01; USSN [unassigned], attorney docket

number 1011N-1, filed 01/09/02 entitled "Algorithms for Selection of Primer Pairs"; and USSN [assigned], attorney docket number 1011N1D1, filed 01/09/02, entitled "Methods for Amplification of Nucleic Acids" are particularly suitable for amplifying genomic DNA for use in the methods of the present invention.

Amplified sequences may be subjected to other post amplification treatments either before or after labeling. For example, in some cases, it may be desirable to fragment the amplified sequence prior to hybridization with an oligonucleotide array. Fragmentation of the nucleic acids generally may be carried out by physical, chemical or enzymatic methods that are known in the art. Suitable techniques include, but are not limited to, subjecting the amplified nucleic acids to shear forces by forcing the nucleic acid containing fluid sample through a narrow aperture or digesting the PCR product with a nuclease enzyme. One example of a suitable nuclease enzyme is Dnase I. After amplification, the PCR product may be incubated in the presence of a nuclease for a period of time designed to produce appropriately sized fragments. The sizes of the fragments may be varied as desired, for example, by increasing the amount of nuclease or duration of incubation to produce smaller fragments or by decreasing the amount of nuclease or period of incubation to produce larger fragments. Adjusting the digestion conditions to produce fragments of the desired size is within the capabilities of a person of ordinary skill in the art. The fragments thus produced may be labeled as described above.

Methods for the Detection of SNPs (SNP Discovery)

Determination of the presence or absence of one or more variations in a nucleic acid may be made using any technique known to those of skill in the art. Any technique that permits the accurate determination of a variation can be used. Preferred techniques will permit rapid, accurate determination of multiple variations with a minimum of sample handling required. Some examples of suitable techniques are provided below.

Several methods for DNA sequencing are well known and generally available in the art and may be used to determine the location of SNPs in a genome. See, for example, Sambrook, *et al.*, Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, New York) (1989), and Ausubel, *et al.*, Current Protocols in Molecular Biology (John Wiley and Sons, New York) (1997), incorporated herein by

reference. Such methods may be used to determine the sequence of the same genomic regions from different DNA strands where the sequences are then compared and the differences (variations between the strands) are noted. DNA sequencing methods may employ such enzymes as the Klenow fragment of DNA polymerase I, Sequenase (US Biochemical Corp, Cleveland, Ohio.), Taq polymerase (Perkin Elmer), thermostable T7 polymerase (Amersham, Chicago, Ill.), or combinations of polymerases and proofreading exonucleases such as those found in the Elongase Amplification System marketed by Gibco/BRL (Gaithersburg, Md.). Preferably, the process is automated with machines such as the Hamilton Micro Lab 2200 (Hamilton, Reno, Nev.), Peltier Thermal Cycler (PTC200; MJ Research, Watertown, Mass.) and the ABI Catalyst and 373 and 377 DNA Sequencers (Perkin Elmer, Wellesley, MA).

In addition, capillary electrophoresis systems which are commercially available may be used to perform variation or SNP analysis. In particular, capillary sequencing may employ flowable polymers for electrophoretic separation, four different fluorescent dyes (one for each nucleotide) which are laser activated, and detection of the emitted wavelengths by a charge coupled device camera. Output/light intensity may be converted to electrical signal using appropriate software (e.g. Genotyper and Sequence Naviagator, Perkin Elmer, Wellesley, MA) and the entire process from loading of samples to computer analysis and electronic data display may be computer controlled. Again, this method may be used to determine the sequence of the same genomic regions from different DNA strands where the sequences are then compared and the differences (variations between the strands) are noted.

Optionally, once a genomic sequence from one reference DNA strand has been determined by sequencing, it is possible to use hybridization techniques to determine variations in sequence between the reference strand and other DNA strands. These variations may be SNPs. An example of a suitable hybridization technique involves the use of DNA chips (oligonucleotide arrays), for example, those available from Affymetrix, Inc. Santa Clara, CA. For details on the use of DNA chips for the detection of, for example, SNPs, see United States Patent No. 6,300,063 issued to Lipshultz, *et al.*, and United States Patent No. 5,837,832 to Chee, *et al.*, HuSNP Mapping Assay, reagent kit and user manual, Affymetrix Part No. 90094 (Affymetrix, Santa Clara, CA), all incorporated by reference herein.

In preferred embodiments, more than 10,000 bases of a reference sequence and the other DNA strands are scanned for variants. In more preferred embodiments, more than $1x10^6$ bases of a reference sequence and the other DNA strands are scanned for variants, even more preferably more than $2x10^6$ bases of a reference sequence and the other DNA strands are scanned, even more preferably $1x10^7$ bases are scanned, and more preferably more than $1x10^8$ bases are scanned, and more preferably more than $1x10^9$ bases of a reference sequence and the other DNA strands are scanned for variants. In preferred embodiments at least exons are scanned for variants, and in more preferred embodiments both introns and exons are scanned for variants. In an even more preferred embodiment, introns, exons and intergenic sequences are scanned for variants. In preferred embodiments the scanned nucleic acids are genomic DNA, including both coding and noncoding regions. In most preferred embodiments, such DNA is from a mammalian organism such as a human. In preferred embodiments, more than 10% of the genomic DNA from the organism is scanned, in more preferred embodiments more than 25% of the genomic DNA from the organism is scanned, in more preferred embodiments, more than 50% of the genomic DNA from the organism is scanned, and in most preferred embodiments, more than 75% of the genomic DNA is scanned. In some embodiments of the present invention, known repetitive regions of the genome are not scanned, and do not count toward the percentage of genomic DNA scanned. Such known repetitive regions may include Single Interspersed Nuclear Elements (SINEs, such as alu and MIR sequences), Long Interspersed Nuclear Elements (LINEs, such as LINE1 and LINE2 sequences), Long Terminal Repeats (LTRs such as MaLRs, Retrov and MER4 sequences), transposons, and MER1 And MER2 sequences.

Briefly, in one embodiment, labeled nucleic acids in a suitable solution are denatured–for example, by heating to 95 °C–and the solution containing the denatured nucleic acids is incubated with a DNA chip. After incubation, the solution is removed, the chip may be washed with a suitable washing solution to remove un-hybridized nucleic acids, and the presence of hybridized nucleic acids on the chip is detected. The stringency of the wash conditions may be adjusted as necessary to produce a stable signal. Detecting the hybridized nucleic acids may be done directly, for example, if the nucleic acids contain a fluorescent reporter group, fluorescence may be directly detected. If the label on the nucleic acids is not directly detectable, for example, biotin, then a solution

containing a detectable label, for example, streptavidin coupled to phycoerythrin, may be added prior to detection. Other reagents designed to enhance the signal level may also be added prior to detection, for example, a biotinylated antibody specific for streptavidin may be used in conjunction with the biotin, streptavidin-phycoerythrin detection system. In some embodiments, the oligonucleotide arrays used in the methods of the present invention contain at least 1 x 10^6 probes per array. In a preferred embodiment, the oligonucleotide arrays used in the methods of the present invention contain at least 10 x 10^6 probes per array. In a more preferred embodiment, the oligonucleotide arrays used in the methods of the present invention contain at least 50 x 10^6 probes per array.

Once variant locations have been determined (SNP discovery) by using, for example, sequencing or microarray analysis, it is necessary to genotype the SNPs of control and sample populations. The hybridization methods just described work well for this purpose, providing an accurate and rapid technique for detecting and genotyping SNPs in multiple samples. In addition, a technique suitable for the detection of SNPs in genomic DNA–without amplification–is the Invader technology available from Third Wave Technologies, Inc., Madison, WI. Use of this technology to detect SNPs may be found, *e.g.*, in Hessner, *et al.*, *Clinical Chemistry* 46(8):1051-56 (2000); Hall, *et al.*, *PNAS* 97(15):8272-77 (2000); Agarwal, *et al.*, *Diag. Molec. Path.* 9(3):158-64 (2000); and Cooksey, *et al.*, *Antimicrobial and Chemotherapy* 44(5):1296-1301 (2000). In the Invader process, two short DNA probes hybridize to a target nucleic acid to form a structure recognized by a nuclease enzyme. For SNP analysis, two separate reactions are run—one for each SNP variant. If one of the probes is complementary to the sequence, the nuclease will cleave it to release a short DNA fragment termed a "flap". The flap binds to a fluorescently-labeled probe and forms another structure recognized by a nuclease enzyme. When the enzyme cleaves the labeled probe, the probe emits a detectable fluorescence signal thereby indicating which SNP variant is present.

An alternative to Invader technology, rolling circle amplification utilizes an oligonucleotide complementary to a circular DNA template to produce an amplified signal (see, for example, Lizardi, *et al.*, *Nature Genetics* 19(3):225-32 (1998); and Zhong, *et al.*, *PNAS* 98(7):3940-45 (2001)). Extension of the oligonucleotide results in the production of multiple copies of the circular template in a long concatemer. Typically, detectable labels are incorporated into the extended oligonucleotide during the extension

reaction. The extension reaction can be allowed to proceed until a detectable amount of extension product is synthesized.

In order to detect SNPs using rolling circle amplification, three probes and two circular DNA templates may be used. The first probe–the target specific probe–may be constructed to be complementary to a target nucleic acid molecule such that the 5'-terminus of the probe hybridizes to the nucleotide immediately adjacent 5' to the SNP site in the target nucleic acid. The site of the SNP is not base paired to the first probe.

The other two probes–rolling circle probes–are constructed to have two 3'-terminals. This can be accomplished in various ways, for example, by introducing a 5'-5' linkage in the central portion of the probes resulting in a reversal of polarity of the nucleotide sequence at that point. One end of each of the probes has a sequence that is complementary to a portion of a different circular template molecule while the other end is complementary to a portion of the target nucleic acid sequence. The target-sequence-complementary terminal is constructed such that the 3'-most nucleotide aligns with the nucleotide at the SNP site. One of the probes may contain a nucleotide complementary to the nucleotide at the SNP site in the target nucleic acid while the other contains a nucleotide that is not complementary. In the instance where two or more variants of the SNP are present in the population, probes may be constructed to have 3'-nucleotides complementary to the variants to be detected.

The probes–both target specific and rolling circle–may be hybridized to the target sequence and contacted with a ligase enzyme. When the 3'-most nucleotide of the rolling circle probe forms a base pair with the nucleotide at the SNP site, the two probes–the target specific and the rolling circle–are efficiently ligated together. When the 3'-most nucleotide of the rolling circle probe is not capable of base pairing with the nucleotide at the SNP site in the target, the probes are not ligated. The unligated probe is washed away and the sample is contacted with the template circles, polymerase and labeled nucleoside triphosphates.

Another technique suitable for the detection of SNPs makes use of the 5'-exonuclease activity of a DNA polymerase to generate a signal by digesting a probe molecule to release a fluorescently labeled nucleotide. This assay is frequently referred to as a Taqman assay (see, *e.g.*, Arnold, *et al.*, *BioTechniques* 25(1):98-106 (1998); and Becker, *et al.*, *Hum. Gene Ther.* 10:2559-66 (1999)). A target DNA containing a SNP is

amplified in the presence of a probe molecule that hybridizes to the SNP site. The probe molecule contains both a fluorescent reporter-labeled nucleotide at the 5'-end and a quencher-labeled nucleotide at the 3'-end. The probe sequence is selected so that the nucleotide in the probe that aligns with the SNP site in the target DNA is as near as possible to the center of the probe to maximize the difference in melting temperature between the correct match probe and the mismatch probe. As the PCR reaction is conducted, the correct match probe hybridizes to the SNP site in the target DNA and is digested by the Taq polymerase used in the PCR assay. This digestion results in physically separating the fluorescent labeled nucleotide from the quencher with a concomitant increase in fluorescence. The mismatch probe does not remain hybridized during the elongation portion of the PCR reaction and is, therefore, not digested and the fluorescently labeled nucleotide remains quenched.

Denaturing HPLC using a polystyrene-divinylbenzene reverse phase column and an ion-pairing mobile phase can be used to identify SNPs. A DNA segment containing a SNP is PCR amplified. After amplification, the PCR product is denatured by heating and mixed with a second denatured PCR product with a known nucleotide at the SNP position. The PCR products are annealed and are analyzed by HPLC at elevated temperature. The temperature is chosen to denature duplex molecules that are mismatched at the SNP location but not to denature those that are perfect matches. Under these conditions, heteroduplex molecules typically elute before homoduplex molecules. For an example of the use of this technique see Kota, *et al.*, *Genome* 44(4):523-28 (2001).

SNPs can be detected using solid phase amplification and microsequencing of the amplification product. Beads to which primers have been covalently attached are used to carry out amplification reactions. The primers are designed to include a recognition site for a Type II restriction enzyme. After amplification–which results in a PCR product attached to the bead–the product is digested with the restriction enzyme. Cleavage of the product with the restriction enzyme results in the production of a single stranded portion including the SNP site and a 3'-OH that can be extended to fill in the single stranded portion. Inclusion of ddNTPs in an extension reaction allows direct sequencing of the product. For an example of the use of this technique to identify SNPs see Shapero, *et al.*, *Genome Research* 11(11):1926-34 (2001).

Data Analysis

Figure 1 is a schematic showing the steps of one embodiment of the methods of the present invention. Once SNPs (variants) have been located or discovered by, *e.g.*, the methods described *supra* (step 110 of Figure 1), SNP haplotype blocks, SNP haplotype patterns within each SNP haplotype block, and informative SNPs for the SNP haplotype patterns may be determined. One may use all SNPs or variants located; alternatively, one may focus the analysis on only a portion of the SNPs located. For example, the set of SNPs analyzed may exclude transition SNPs of the form Cg<-> Tg or cG <-> cA. In addition, in one embodiment of the present invention, the focus is on common SNPs. Common SNPs are those SNPs whose less common form is present at a minimum frequency in a given population. For example, common SNPs are those SNPs that are found in at least about 2% to 25% of the population. In a preferred embodiment, common SNPs are those SNPs that are found in at least about 5% to 15% of the population. In a more preferred embodiment, common SNPs are those that are found in at least about 10% of the population. Common SNPs likely result from mutations that occurred early in the evolution of humans. Focusing on common SNPs minimizes systematic allele or variant differences between control and experimental populations that appear as disease or drug-response associated, yet result only from migratory history or mating practices; *i.e.*, focusing on common SNPs decreases the false positives that result from recent population anomalies. Moreover, common SNPs are relevant to a larger proportion of the human population, making the present invention more broadly applicable to disease and drug response studies. Along the same line, SNPs in which an variant is observed only once may be eliminated from analysis in some embodiments of the present invention (for example, singleton SNPs). However, certain analyses may be performed including some or all of these singleton SNPs, particularly when looking at specific sub-populations or populations that have been influenced by migratory practices and the like.

In step 120 of Figure 1, the variants or SNPs of interest are assigned to haplotype blocks for evaluation. Variants or SNPs from a whole genome or chromosome may be analyzed and assigned to SNP haplotype blocks. Alternatively, variants from only a focused genomic region specific to some disease or drug response mechanism may be assigned to the SNP haplotype blocks.

Figure 2 provides one illustration of showing how variants, usually SNPs, occur in haplotype blocks in a genome, and that more than one haplotype pattern can occur within each haplotype block. If SNP haplotype patterns were completely random, it would be expected that the number of possible SNP haplotype patterns observed for a SNP haplotype block of N SNPs would be 2^N. However, it was observed in performing the methods of the present invention that the number of SNP haplotype patterns in each SNP haplotype block is smaller than 2^N because the SNPs are linked (not 4^N, as the variants will most commonly be biallelic, *i.e.*, occur in only one of two forms, not all four nucleotide base possibilities). Certain SNP haplotype patterns were observed at a much higher frequency than would be expected in a non-linkage case. Thus, SNP haplotype blocks are chromosomal regions that tend to be inherited as a unit, with a relatively small number of common patterns. Each line in Fig. 2 represents portions of the haploid genome sequence of different individuals. As shown therein, individual W has an "A" at position 241, a "G" at position 242, and an "A" at position 243. Individual X has the same bases at positions 241, 242, and 243. Conversely, individual Y has a T at positions 241 and 243, but an A at position 242. Individual Z has the same bases as individual Y at positions 241, 242, and 243. Variants in block 261 will tend to occur together. Similarly, the variants in block 262 will tend to occur together, as will those variants in block 263. Of course, only a few bases in a genome are shown in Figure 2. In fact, most bases will be like those at position 245 and 248, and will not vary from individual to individual.

The assignment of SNPs to SNP haplotype blocks, step 120 of Figure 1, is, in one case, an iterative process involving the construction of SNP haplotype blocks from the SNP locations along a genomic region of interest. In one embodiment, once the initial SNP haplotype blocks are constructed, SNP haplotype patterns present in the constructed SNP haplotype blocks are determined (step 130 of Figure 1). In some specific embodiments, the number of SNP haplotype patterns selected per SNP haplotype block in step 130 is no greater than about five. In another specific embodiment, the number of SNP haplotype patterns selected per SNP haplotype block is equal to the number of SNP haplotype patterns necessary to identify SNP haplotype patterns in greater than 50% of the DNA strands being analyzed. In other words, enough SNP haplotype patterns are selected, for example, four patterns per block are selected, such that at least half of the DNA strands analyzed will have a SNP haplotype pattern that matches one of the four

patterns selected in each SNP haplotype block. In a preferred embodiment, the number of SNP haplotype patterns selected per SNP haplotype block is equal to the number of SNP haplotype patterns necessary to identify SNP haplotype patterns in greater than 70% of the DNA strands being analyzed. In one preferred embodiment, the number of SNP haplotype patterns selected per SNP haplotype block is equal to the number of SNP haplotype patterns necessary to identify SNP haplotype patterns in greater than 80% of the DNA strands being analyzed. In addition, in some embodiments of the present invention, SNP haplotype patterns that occur in less than a certain portion of DNA strands being analyzed are eliminated from analysis. For example, in one embodiment, if ten DNA strands are being analyzed, SNP haplotype patterns that are found to occur in only one sample out of ten are eliminated from analysis.

Once the SNP haplotype patterns of interest are selected, informative SNPs for these SNP haplotype patterns are determined (step 140 of Figure 1). From this initial set of blocks, a set of candidate SNP blocks that fit certain criteria for informativeness is constructed (step 150 of Figure 1). Figures 4 and 5 illustrate steps 120, 130, 140 and 150 in more detail.

In Figure 3, step 310 provides that a new block of SNPs is chosen for evaluation. In one embodiment, the first block chosen contains only the first SNP in a SNP haplotype sequence; thus at step 320, the first, single, SNP is added to the block. At step 330, informativeness of this block is determined.

"Informativeness" of a SNP haplotype block is defined in one embodiment as the degree to which the block provides information about genetic regions. For example, in one embodiment of the present invention, informativeness could be calculated as the ratio of the number of SNP locations in a SNP haplotype block divided by the number of SNPs required to distinguish each SNP haplotype pattern under consideration from other SNP haplotype patterns under consideration (number of informative SNPs) in that block. Another measure of informativeness might be the number of informative SNPs in the block. One skilled in the art recognizes that informativeness may be determined in any number of ways.

Referring again to Figure 2, SNP haplotype block 261 contains three SNPs and two SNP haplotype patterns (AGA and TAT). Any one of the three SNPs present can be used to tell the patterns apart; thus, any one of these SNPs can be chosen to be the

informative SNP for this SNP haplotype pattern. For example, if it is determined that a sample nucleic acid contains a T at the first position, the same sample will contain an A at the second position and a T at the third position. If it is determined in a second sample that the SNP in the second position is a G, the first and third SNPs will be A's. Thus, by one measure of informativeness, the informativeness value for this first block is 3: 3 total SNPs divided by 1 informative SNP needed to distinguish the patterns from each other. Similarly, SNP haplotype block 262 contains three SNPs (two positions do not have variants) and two haplotype patterns (TCG and CAC). As with the previously-analyzed block, any one of the three SNPs can be evaluated to tell one pattern from the other; thus, the informativeness of this block is 3: 3 total SNPs divided by 1 informative SNP needed to distinguish the patterns. SNP haplotype block 263 contains five SNPs and two SNP patterns (TAACG and ATCAC). Again, any one of the five SNPs can be used to tell one pattern from the other; thus, the informativeness of this block is 5: 5 total SNPs divided by 1 informative SNP needed to distinguish the patterns.

Figure 2 provides a simple example of genetic analysis. When several SNP haplotype patterns are present in a block, it may be necessary to use more than one SNP as informative SNPs. For example, in a case where a block contains, for example, six SNPs and two SNPs are needed to distinguish the patterns of interest, the informativeness of the block is 3: 6 total SNPs divided by 2 SNPs needed to distinguish the patterns. Generally speaking, as many as 2^N distinct SNP haplotype patterns can be distinguished by using the genotypes of N suitably selected SNPs. Therefore, if there exist only two SNP haplotype patterns in the SNP haplotype block, a single SNP should be able to differentiate between the two. If there are three or four patterns, at least two SNPs would likely be required, etc.

In step 340 of Figure 3, once the informativeness of a SNP haplotype block is determined, a test is performed. The test essentially evaluates the SNP haplotype blocks based on selected criteria (for example, whether a block meets a threshold measure of informativeness), and the result of the test determines whether, for example, another SNP will be added to the block for analysis or whether the analysis will proceed with a new block starting at a different SNP location. Figure 4 illustrates one embodiment of this process.

In Figure 4, assume there is a DNA sequence with six SNP locations. The analysis of SNP haplotype blocks described above might be performed in the following manner: SNP haplotype block A is selected containing only the SNP at SNP position 1 (steps 310 and 320 of Figure 3). The informativeness of this block is calculated (step 330), and it is determined whether the informativeness of this block meets a threshold measure of informativeness (step 340). In this case, it "passes" and two things happen. First, this block of one SNP (SNP position 1) is added to the set of candidate SNP haplotype blocks (step 350). Second, another SNP (here, SNP position 2) is added to this block (step 320) to create a new block, B, containing SNP positions 1 and 2, which is then analyzed. In this illustration block B also meets the threshold measure of informativeness (step 340), so it would be added to the set of candidate SNP haplotype blocks (step 350), and another SNP (here, SNP position 3) is added to this block (step 320) to create new block C, containing SNP positions 1, 2 and 3, which is then analyzed. In this illustration, C also meets the threshold measure of informativeness and it is added to the set of candidate SNP haplotype blocks (step 350), and another SNP (here, SNP position 4) is added to this block (step 320) to create new block D, containing SNP positions 1, 2, 3, and 4, which is then analyzed. In the Figure 4 illustration, SNP block D does not meet the threshold measure of informativeness. SNP block D is not added to the set of candidate SNP haplotype blocks (step 350), nor does another SNP get added to block D for analysis. Instead, a new SNP location is selected for a round of SNP block evaluations.

In Figure 4, after block D fails to meet the threshold measure of informativeness, a new block, E, is selected that contains only the SNP at position 2. Block E is evaluated for informativeness, is found to meet the threshold measure, is added to the set of candidate SNP haplotype blocks (step 350), and another SNP (here, SNP position 3) is added to this block (step 320) to create new block F, containing SNP positions 2 and 3, which is then analyzed, and so on. Note that block H fails to meet the threshold measure of informativeness, is not added to the set of candidate SNP haplotype blocks (step 350), nor does another SNP get added to block H for analysis. Instead, a new block, I, is selected that contains only the SNP at position 3, and so on.

Once a set of candidate SNP blocks is constructed (step 350 of Figure 3), analysis is performed on the set to select a final set of SNP blocks (step 160 of Figure 1). The

selection of the final set of SNP blocks can performed in a variety of ways. For example, referring back to Figure 4, one could select the largest block containing SNP position 1 that passes the threshold test (block C, containing SNPs 1, 2 and 3), discard the smaller blocks that contain the same SNPs (blocks A and B). Then the next block selected might be the next block starting with SNP position 4 that is the largest block that meets the threshold test for informativeness (block G) and the smaller blocks that contain the same SNPs (blocks E and F) would be discarded. Such a method would give a set of final, non-overlapping SNP haplotype blocks that span the genomic region of interest, contain the SNPs of interest and that have a high level of informativeness. Thus, once all candidate SNP haplotype blocks are evaluated, the result may be, in a preferred embodiment, a set of non-overlapping SNP haplotype blocks that encompasses all the SNPs in the original set. Some groups, called isolates, may consist of only a single SNP, and by definition have an informativeness of 1. Other groups may consist of a hundred or more SNPs, and have an informativeness exceeding 30.

An alternative method for selecting a final set of SNP haplotype blocks is shown in Figures 5A and 5B. Looking first at Figure 5A, in a first step 510, the candidate SNP haplotype block set (generated, for example, by the methods described in Figures 3 and 4 herein) is analyzed for informativeness. In step 520, the candidate SNP haplotype block with the highest informativeness in the entire candidate set is chosen to be added to the final SNP haplotype block set (step 530). Once this candidate SNP haplotype block is chosen to be a member of the final SNP haplotype block set, it is deleted from the candidate block set (step 540), and all other candidate SNP haplotype blocks that overlap with the chosen block are deleted from the candidate SNP haplotype block set (step 550). Next, the candidate SNP haplotype blocks remaining in the candidate set are analyzed for informativeness (step 510), and the candidate SNP haplotype block with the highest informativeness is chosen to be added to the final SNP haplotype block set (steps 520 and 530). As before, once this SNP haplotype block is chosen to be a member of the final SNP haplotype block set, it is deleted from the candidate block set (step 540), and all other candidate SNP haplotype blocks that overlap with the chosen block are deleted from the candidate SNP haplotype block set (step 550). The process continues until a final set of non-overlapping SNP haplotype blocks that encompasses all the SNPs in the original set is constructed.

Figure 5B illustrates a simple employment of the method of selecting a final set of SNP haplotype blocks described in Figure 5A. In figure 5B, a sequence 5' to 3' is analyzed for SNPs, SNP haplotype patterns and candidate SNP haplotype blocks according to the methods of the present invention. Candidate SNP haplotype blocks contained within this sequence are indicated by their placement under the sequence, and are designated by a letter. In addition, after the letter, the informativeness of each block is indicated. For example, candidate SNP haplotype block A is located at the extreme 5' end of the sequence, and has an informativeness of 1. Candidate SNP haplotype block R is located at the extreme 3' end of the sequence, and has an informativeness of 2.

According to figure 5A, in a first step 510, the candidate SNP haplotype blocks are analyzed for informativeness, and in step 520, the SNP haplotype block with the highest informativeness is chosen to be added to the final SNP haplotype block set (steps 520 and530). In the case of figure 5B, candidate SNP haplotype block M with an informativeness of 6 would be the first candidate SNP haplotype block selected to be added to the final SNP haplotype block set. Once SNP haplotype block M is selected, it is deleted or removed from the candidate set of SNP haplotype blocks (step 540), and all other candidate SNP haplotype blocks that overlap with SNP haplotype block M (blocks J, N, K, L, O and P) are deleted from the candidate SNP haplotype block set (step 550). Next, the remaining blocks of the candidate SNP haplotype block set, namely SNP haplotype blocks A, B, C, D, E, F, G, H, I, Q and R are analyzed for informativeness, and in step 520, the remaining SNP haplotype block with the highest informativeness, I, with an informativeness of 5, is chosen to be added to the final SNP haplotype block set (530) and deleted or removed from the candidate set of SNP haplotype blocks (step 540). Next, in step 550, all other candidate SNP haplotype blocks that overlap with SNP haplotype block I, here, only block H, is deleted from the candidate SNP haplotype block set. Again, the remaining blocks of the candidate SNP haplotype block set, namely SNP haplotype blocks A, B, C, D, E, F, G, Q and R are analyzed for informativeness. In step 520, the remaining SNP haplotype block with the highest informativeness, block F, with an informativeness of 4, is chosen to be added to the final SNP haplotype block set (530) and deleted or removed from the candidate set of SNP haplotype blocks (step 540). Next, all other candidate SNP haplotype blocks that overlap with SNP haplotype block F--here, blocks E, G, C and D--are deleted from the candidate SNP haplotype block set, and the

remaining blocks of the candidate SNP haplotype block set, namely SNP haplotype blocks A, B, Q and R, are analyzed for informativeness, and so on.

Other methods can be employed to select a final set of SNP haplotype blocks for analysis from the set of candidate SNP haplotype blocks (step 160 of Figure 1). For example, algorithms known in the art may be applied for this purpose. For example, shortest-paths algorithms may be used (see, generally, Cormen, Leiserson, and Rivest, Introduction to Algorithms (MIT Press) pp. 514-78 (1994). In a shortest-paths problem, a weighted, directed graph $G=(V,E)$, with weight function $w : E \rightarrow \mathbf{R}$ mapping edges to real-valued weights is given. The weight of path $p = (v_0, v_1, \ldots v_k)$ is the sum of the weights of its constituent edges:

$$w(p) = \sum_{i=1}^{k} w(v_{i-1}, v_i).$$

The shortest-path weight from u to v is defined by $\delta(u,v)$ being equal to min $w(p):u \rightarrow v$ if there is a path from u to v; otherwise, $\delta(u,v)$ is equal to infinity. A shortest path from vertex u to vertex v is then defined as any path p with weight $w(p) = \delta(u,v)$. Edge weights can be interpreted as various metrics: for example, distance, time, cost, penalties, loss, or any other quantity that accumulates linearly along a path that one wishes to minimize. In the embodiment of the shortest path algorithm used in applications of this invention, each SNP haplotype block would be considered a "vertex" with an "edge" defined for each boundary of the block. Each SNP haplotype block has a relationship to each other SNP haplotype block, with a "cost" for each edge. Cost is determined by parameters of choice, such as overlap (or the extent thereof) of the vertices or gaps between the vertices.

Single-source shortest-paths problems focus on a given graph $G=(V,E)$, where a shortest path from a given source vertex $s \in V$ to every vertex $v \in V$ is determined. Additionally, variants of the single source algorithm may be applied. For example, one may apply a single-destination shortest-paths solution where a shortest path to a given destination vertex t from every vertex v is found. Reversing the direction of each edge in the graph reduces this problem to a single-source problem. Alternatively, one may apply a single-pair shortest-path problem where the shortest path from u to v for given vertices u and v is found. If the single-source problem with source vertex u is solved, the single-source shortest path problem is solved as well. Also, the all-pairs shortest-paths approach

may be employed. In this case, a shortest path from u to v for every pair of vertices u and v is found--a single-source algorithm is run from each vertex.

One single-source shortest-path algorithm that may be employed in the methods of the present invention is Dijkstra's algorithm. Dijkstra's algorithm solves the single-source shortest-paths problem on a weighted, directed graph $G=(V,E)$ for the case in which all edge weights are nonnegative. Dijkstra's algorithm maintains a set of vertices, S, whose final shortest-path weights from a source s have already been determined. That is, for all vertices v being elements of S, $d[v]=\delta(s,v)$. The algorithm repeatedly selects the vertex u as an element of V-S with the minimum shortest-path estimate, inserts u into S, and relaxes all edges radiating from u. In one implementation, a priority queue Q that contains all the vertices in V-S, keyed by their d values, is maintained. This implementation assumes that graph G is represented by adjacency lists.

Dijkstra (G, w, s)
1 INITIALIZE-SINGLE SOURCE (G,s)
2 $S \leftarrow \emptyset$
3 $Q \leftarrow V[G]$
4 **while** $Q \neq \emptyset$
5 **do** $u \leftarrow$ EXTRACT-MIN (Q)
6 $S \leftarrow S \cup \{u\}$
7 **for** each vertex $v \in Adj[u]$
8 **do** RELAX (u,v,w)

Thus, G in this case is the graph of linear coverage of the genomic sequence being analyzed and S is the set of vertices selected. Once one vertex is selected that covers a particular area of the genomic sequence, other vertices that overlap this sequence can be discarded.

Other algorithms that may be used for selecting SNP haplotype blocks include a greedy algorithm (again, see, Cormen, Leiserson, and Rivest, Introduction to Algorithms (MIT Press) pp. 329-55 (1994)). A greedy algorithm obtains an optimal solution to a problem by making a sequence of choices. For each decision point in the algorithm, the choice that seems best at the moment is chosen. This heuristic strategy does not always produce an optimal solution. Greedy algorithms differ from dynamic programming in that in dynamic programming, a choice is made at each step, but the choice may depend on the solutions to subproblems. In a greedy algorithm, whatever choice seems best at the moment is chosen and then subproblems arising after the choice is made are solved. Thus, the choice made by a greedy algorithm may depend on the choices made thus far,

but cannot depend on any future choices or on the solutions to subproblems. One variation of greedy algorithms is Huffman codes. A Huffman greedy algorithm constructs an optimal prefix code and the algorithm builds a tree T corresponding to the optimal code in a bottom-up manner. It begins with a set of $|C|$ leaves and performs a sequence of $|C|$-1 "merging" operations to create the final tree. For example, assuming C is a set of n characters and that each character $c \in C$ is an object with a defined frequency $f[c]$; a priority queue Q, keyed on f, is used to identify the two least-frequent objects to merge together. The result of the merger of two objects is a new object whose frequency is the sum of the frequencies of the two objects that were merged. For example:

```
1.   n←|C|
2.   Q←C
3.   for i←1 to n-1
4.   do z←ALLOCATE-NODE()
5.   x←left[z] ←EXTRACT-MIN(Q)
6.   y←right[z] ←EXTRACT-MIN(Q)
7.   f[z] ←f[x] + f[y]
8.   INSERT (Q,z)
9.   return EXTRACT-MIN(Q)
```

Line 2 initializes the priority queue Q with the characters in C. The **for** loop in lines 3-8 repeatedly extracts the two nodes x and y of lowest frequency from the queue, and replaces them in the queue with a new node z representing their merger. The frequency of z is computed as the sum of the frequencies of x and y in line 7. The node z has x as its left child and y as its right child. After n-1 mergers, the one node left in the queue—the root of the code tree—is returned in line 9.

Again, these methods result in a set of final, non-overlapping SNP haplotype blocks that encompasses all SNPs evaluated in a particular genomic region. An important result of selecting SNPs, SNP haplotype blocks and SNP haplotype patterns according to the methods of the present invention, is that in some embodiments during the calculation of informativeness of SNP haplotype blocks, informative SNPs for each SNP haplotype block and pattern are determined. Informative SNPs allow for data compression.
In one embodiment of the present invention, the selection of at least $\log_2 p$ SNPs from each group containing p patterns (rounding up to the nearest integer) provides one set of informative SNPs which are unusually powerful for predicting genotype/phenotype associations. One skilled in the art recognizes that in other analyses it is not necessary to use spatially contiguous groups to determine such a subset. For example, in some

embodiments of the present invention, it may be desirable to identify sets of non-adjacent SNPs that statistically are passed on in a fashion analogous to that of SNP haplotype blocks even though they are not spatially contiguous on the DNA strand.

In order to determine SNP haplotype blocks that will be used in association studies accurately (build an accurate baseline of SNPs and SNP haplotype blocks and patterns), it is necessary to examine more than a few individual DNA strands. Figure 6 illustrates the importance of examining at least about five different DNA strands for determining SNP haplotype blocks and for the selection of informative SNPs. The top portion of Figure 6 illustrates the sequence of a hypothetical stretch of DNA, with the variant positions indicated and variant block boundaries drawn; however, SNP haplotype block boundaries would not be known ab initio. Sequencing results 610 show the results of sequencing haploid DNA of three individuals. As shown, in general it is possible to have identified a large fraction of the common SNPs after a relatively small number of individuals have been sequenced. In the case in Figure 6, the SNPs at each location shown in the top portion of Figure 6 have been identified, as indicated by check marks.

If, however, further individuals are not evaluated, the block boundaries would not be correctly identified at this stage. For example, while one could at this stage draw block boundaries between blocks 620 and 630 (note that the first C →G variant predicts the first G→A variant, and the first C→T variant predicts the second C→T variant), it is not possible to distinguish between the blocks 630 and 640 at this stage. At this stage it appears that the first C→T variant would predict the first and second T→A variants. Accordingly, a more statistically significant sample set is required to draw the block boundaries. For example, in the methods of the present invention, the number of DNA strands analyzed to determine SNP haplotype blocks, SNP haplotype patterns, and/or informative SNPs is a plurality, for example, at least about five or at least about 10. In preferred embodiments, the number of DNA stands is at least 16. In more preferred embodiments, the number of DNA strands analyzed to determine SNP haplotype blocks, SNP haplotype patterns, and/or informative SNPs is at least 25. However, once relevant SNPs have been identified (*i.e.*, SNP discovery has been performed), it is possible to genotype only the variant positions in the remaining samples to complete the process of identifying block boundaries without sequencing the entire stretch of genomic DNA. For examples of such methods, see USSN 10/042,819, filed 01/06/02, attorney docket number

1016N-1, entitled "Whole Genome Scanning".

The results of performing a genotyping process on only the SNPs in another hypothetical genomic sample are shown in Fig. 6 at 650. As shown, by performing this additional genotyping step, it is now possible to see that blocks 630 and 640 are distinguishable. Specifically, it is now possible to see that the first C→T variant does not track with the first and second T→A variants, but instead, the first C→T variant can be used to predict only the second C→T variant (and vice versa) and the first T→A variant can be used only to predict the second T→A variant (and vice versa).

In addition to the aspects of the present invention described above, a specific embodiment of the present invention is that it can be employed to resolve ambiguous SNP haplotype sequences for data analysis. For example, a SNP may be ambiguous because data from a gel sequencing operation or array hybridization experiment does not give a clear result. "Resolving" in this case may mean, *e.g.*, resolving ambiguous SNP locations in a SNP haplotype sequence by matching the SNP haplotype sequence to the SNP haplotype pattern to which the SNP haplotype sequence most closely relates. Additionally, "resolving" may mean removing an ambiguous SNP haplotype sequence from data analysis.

In one embodiment of resolving ambiguous SNP haplotype sequences, SNP haplotype sequences are placed in a data set for possible addition to a pattern set. The data set will contain all SNP haplotype sequences that are to be evaluated for possible assignment to a SNP haplotype pattern. Referring now to Figure 7A, in step 710, the SNP haplotype sequences in the data set are compared, one by one, to the pattern sequences in the pattern set. In some cases, there will be no patterns in the pattern set initially, though in other cases some or all pattern sequences may be known beforehand. In step 720, a query is made: is the SNP haplotype sequence from the data set consistent with a pattern sequence in the pattern set? If the answer is no, step 730 provides the SNP haplotype sequence being evaluated will be added to the pattern set. If the answer is yes, another query is made (740): is the SNP haplotype sequence from the data set consistent with more than one pattern sequence in the pattern set?

If the answer is yes, the SNP sequence from the data set may be discarded or, in some embodiments, held for further or different analyses (step 750). If the answer to the second query is no, then, in step 760, the SNP sequence from the data set is compared to

the pattern sequence from the pattern set with which it is consistent. From these two sequences, the SNP sequence with the least number of ambiguities is selected and placed in the pattern set (770). The SNP sequence containing the more ambiguities may be discarded, or, in some embodiments, held for further or different types of analyses.

The resolving process may be understood further by referring to Figures 7A and 7B. In Figure 7B, a first SNP sequence, TTCGA, is compared to the sequences contained in the pattern set (step 710). At this point, there are no pattern sequences contained in the pattern set, thus TTCGA is not consistent with any pattern sequence in the pattern set. This occurrence of SNP sequence TTCGA is then removed from the data set (or is retained for different analyses), and added to the pattern set (730). The pattern set now has one pattern sequence, TTCGA.

Looking again at Figure 7B, the second SNP sequence in the data set, T?C??, is compared to the sequence contained in the pattern set (step 710). Now there is one pattern sequence in the pattern set, TTCGA, and T?C?? is consistent with sequence (step 720). The answer to the second query (740), whether SNP sequence T?C?? is consistent with more than one pattern sequence in the pattern set, is no, as currently there is only one pattern sequence, TTCGA, in the pattern set. In step 760, T?C?? is compared to TTCGA to determine which sequence has the more ambiguities. T?C?? clearly does; thus, TTCGA is retained in the pattern set (770) and T?C?? may be discarded or held for further analyses.

The third sequence of the data set in Figure 7B is C????. C???? first is compared to TTCGA (step 710), is found not to be consistent with TTCGA (720), and is thus added to the pattern set (730). The fourth sequence in Figure 7B is CTACA. CTACA is compared to TTCGA and C???? (the pattern sequences in the pattern set, step 710), and is found to be consistent with C???? (720). The second query (740) now is made: is CTACA consistent with both C???? and TTCGA? The answer is no, so C???? and CTACA are then compared (760) and the sequence with the least number of ambiguities, in this case, CTACA, is held in the pattern set and C???? is discarded (removed from analysis), or held for further analyses (770).

The fifth SNP sequence in the data set in Figure 7B is ?T??A. This SNP sequence is compared to pattern sequences TTCGA and CTACA (710) and is found to be consistent with both TTCGA and CTACA. Thus, the answer to query 740 is yes: ?T??A

is consistent with more than one pattern sequence in the pattern set. In step 750, SNP sequence ?T??A is held for further analysis or discarded (removed from analysis). Another approach to resolving allows that if, for example, one pattern sequence is CCATT? and a SNP sequence from the data set is C?ATTG, the sequences are "combined" to solve the ambiguities (CCATTG), and the "combined" sequence is added to the pattern set. Additional array hybridizations, sequencing or other techniques known in the art may be employed to analyze ambiguous SNP nucleotide positions.

Association of Phenotypes with SNP Haplotypes Blocks and Patterns

The SNP haplotype blocks, SNP haplotype patterns and/or informative SNPs identified may be used for a variety of genetic analyses. For example, once informative SNPs have been identified, they may be used in a number of different assays for association studies. For example, probes may be designed for microarrays that interrogate these informative SNPs. Other exemplary assays include, e.g., the Taqman assays and Invader assays described *supra*, as well as conventional PCR and/or sequencing techniques.

In some embodiments, as shown in step 170 of Figure 1, the haplotype patterns identified may be used in the above-referenced assays to perform association studies. This may be accomplished by determining haplotype patterns in individuals with the phenotype of interest (for example, individuals exhibiting a particular disease or individuals who respond in a particular manner to administration of a drug) and comparing the frequency of the haplotype patterns in these individuals to the haplotype pattern frequency in a control group of individuals. Preferably, such SNP haplotype pattern determinations are genome-wide; however, it may be that only specific regions of the genome are of interest, and the SNP haplotype patterns of those specific regions are used. In addition to the other embodiments of the methods of the present invention disclosed herein, the methods additionally allow for the "dissection" of a phenotype. That is, a particular phenotype may result from two or more different genetic bases. For example, obesity in one individual may be the result of a defect in Gene X, while the obesity phenotype in a different individual may be the result of mutations in Gene Y and Gene Z. Thus, the genome scanning capabilities of the present invention allow for the dissection of varying genetic bases for similar phenotypes. Once specific regions of the

genome are identified as being associated with a particular phenotype, these regions may be used as drug discovery targets (step 180 of Figure 1) or as diagnostic markers (step 190 of Figure 1).

As described in the previous paragraph, one method of conducting association studies is to compare the frequency of SNP haplotype patterns in individuals with a phenotype of interest to the SNP haplotype pattern frequency in a control group of individuals. In a preferred method, informative SNPs are used to make the SNP haplotype pattern comparison. The approach of using informative SNPs has tremendous advantage over other whole genome scanning or genotyping methods known in the art to date, for instead of reading all 3 billion bases of each individual's genome–or even reading the 3-4 million common SNPs that may be found–only informative SNPs from a sample population need to be determined. Reading these particular, informative SNPs provides sufficient information to allow statistically accurate association data to be extracted from specific experimental populations, as described above.

Figure 8 illustrates an embodiment of one method of determining genetic associations using the methods of the present invention. In step 800, the frequency of informative SNPs is determined for genomes of a control population. In step 810, the frequency of informative SNPs is determined for genomes of a clinical population. Steps 800 and 810 may be performed by using the aforementioned SNP assays to analyze the informative SNPs in a population of individuals. In step 820, the informative SNP frequencies from steps 800 and 810 are compared. Frequency comparisons may be made, for example, by determining the minor allele frequency (number of individuals with a particular minor allele divided by the total number of individuals) at each informative SNP location in each population and comparing these minor allele frequencies. In step 830, the informative SNPs displaying a difference between the frequency of occurrence in the control versus clinical populations are selected for analysis. Once informative SNPs are selected, the SNP haplotype blocks that contain the informative SNPs are identified, which in turn identifies the genomic region of interest (step 840). The genomic regions are analyzed by genetic or biological methods known in the art (step 850), and the regions are analyzed for possible use as drug discovery targets (step 860) or as diagnostic markers (step 870), as described in detail below.

Uses of Identified Genomic Sequences

Once a genetic locus or multiple loci in the genome are associated with a particular phenotypic trait--for example, a disease susceptibility locus--the gene or genes or regulatory elements responsible for the trait can be identified. These genes or regulatory elements may then be used as therapeutic targets for the treatment of the disease, as shown in step 180 of Figure 1 or step 860 of Figure 8. The genomic sequences identified by the methods of the present invention may be genic or nongenic sequences. The term "gene" intended to mean the open reading frame (ORF) encoding specific polypeptides, intronic regions, as well as adjacent 5' and 3' non-coding nucleotide sequences involved in the regulation of expression of the gene up to about 10 kb beyond the coding region, but possibly further in either direction. The ORFs of an identified gene may affect the disease state due to their effect on protein structure. Alternatively, the noncoding sequences of the identified gene or nongenic sequences may affect the disease state by impacting the level of expression or specificity of expression of a protein. Generally, genomic sequences are studied by isolating the identified gene substantially free of other nucleic acid sequences that do not include the genic sequence. The DNA sequences are used in a variety of ways. For example, the DNA may be used to detect or quantify expression of the gene in a biological specimen. The manner in which cells are probed for the presence of particular nucleotide sequences is well established in the literature and does not require elaboration here, however, see, *e.g.*, Sambrook, *et al.*, Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, New York) (1989)

In addition, the sequence of the gene, including flanking promoter regions and coding regions, may be mutated in various ways known in the art to generate targeted changes in expression level, or changes in the sequence of the encoded protein, etc. The sequence changes may be substitutions, insertions, translocations or deletions. Deletions may include large changes, such as deletions of an entire domain or exon. Techniques for *in vitro* mutagenesis of cloned genes are known. Examples of protocols for site specific mutagenesis may be found in Gustin, *et al.*, *Biotechniques* 14:22 (1993); Barany, *Gene* 37:111-23 (1985); Colicelli, *et al.*, *Mol. Gen. Genet.* 199:537-9 (1985); Prentki, *et al.*, *Gene* 29:303-13 (1984); Sambrook, *et al.*, Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Press) pp. 15.3-15.108 (1989); Weiner, *et al.*, *Gene* 126:35-41

(1993); Sayers, *et al.*, *Biotechniques* 13:592-6 (1992); Jones and Winistorfer, *Biotechniques* 12:528-30 (1992); and Barton, *et al.*, *Nucleic Acids Res.* 18:7349-55 (1990). Such mutated genes may be used to study structure/function relationships of the protein product, or to alter the properties of the protein that affect its function or regulation.

The identified gene may be employed for producing all or portions of the resulting polypeptide. To express a protein product, an expression cassette incorporating the identified gene may be employed. The expression cassette or vector generally provides a transcriptional and translational initiation region, which may be inducible or constitutive, where the coding region is operably linked under the transcriptional control of the transcriptional initiation region, and a transcriptional and translational termination region. These control regions may be native to the identified gene, or may be derived from exogenous sources.

The peptide may be expressed in prokaryotes or eukaryotes in accordance with conventional methods, depending upon the purpose for expression. For large scale production of the protein, a unicellular organism, such as *E. coli, B. subtilis, S. cerevisiae*, insect cells in combination with baculovirus vectors, or cells of a higher organism such as vertebrates, particularly mammals, e.g. COS 7 cells, may be used as the expression host cells. In many situations, it may be desirable to express the gene in eukaryotic cells, where the gene will benefit from native folding and post-translational modifications. Small peptides also can be synthesized in the laboratory. With the availability of the protein or fragments thereof in large amounts, the protein may be isolated and purified in accordance with conventional ways. A lysate may be prepared of the expression host and the proteins or fragments thereof purified using HPLC, exclusion chromatography, gel electrophoresis, affinity chromatography, or other purification techniques.

An expressed protein may be used for the production of antibodies, where short fragments induce the expression of antibodies specific for the particular polypeptide (monoclonal antibodies), and larger fragments or the entire protein allow for the production of antibodies over the length of the polypeptide (polyclonal antibodies). Antibodies are prepared in accordance with conventional ways, where the expressed polypeptide or protein is used as an immunogen, by itself or conjugated to known immunogenic carriers, *e.g.* KLH, pre-S HBsAg, other viral or eukaryotic proteins, or the

like. Various adjuvants may be employed, with a series of injections, as appropriate. For monoclonal antibodies, after one or more booster injections, the spleen is isolated, the lymphocytes are immortalized by cell fusion and screened for high affinity antibody binding. The immortalized cells, *i.e*, hybridomas, producing the desired antibodies may then be expanded. For further description, see Monoclonal Antibodies: A Laboratory Manual, Harlow and Lane, eds. (Cold Spring Harbor Laboratories, Cold Spring Harbor, N.Y.) (1988). If desired, the mRNA encoding the heavy and light chains may be isolated and mutagenized by cloning in *E. coli*, and the heavy and light chains mixed to further enhance the affinity of the antibody. Alternatives to in vivo immunization as a method of raising antibodies include binding to phage "display" libraries, usually in conjunction with in vitro affinity maturation.

The identified genes, gene fragments, or the encoded protein or protein fragments may be useful in gene therapy to treat degenerative and other disorders. For example, expression vectors may be used to introduce the identified gene into a cell. Such vectors generally have convenient restriction sites located near the promoter sequence to provide for the insertion of nucleic acid sequences in a recipient genome. Transcription cassettes may be prepared comprising a transcription initiation region, the target gene or fragment thereof, and a transcriptional termination region. The transcription cassettes may be introduced into a variety of vectors, e.g. plasmid; retrovirus, *e.g.* lentivirus; adenovirus; and the like, where the vectors are able to be transiently or stably maintained in the cells. The gene or protein product may be introduced directly into tissues or host cells by any number of routes, including viral infection, microinjection, or fusion of vesicles. Jet injection may also be used for intramuscular administration, as described by Furth, *et al.*, *Anal. Biochem*, 205:365-68 (1992). Alternatively, the DNA may be coated onto gold microparticles, and delivered intradermally by a particle bombardment device, or "gene gun" as described in the literature (see, for example, Tang, *et al.*, *Nature*, 356:152-54 (1992)).

Antisense molecules can be used to down-regulate expression of the identified gene in cells. The antisense reagent may be antisense oligonucleotides, particularly synthetic antisense oligonucleotides having chemical modifications, or nucleic acid constructs that express such antisense molecules as RNA. A combination of antisense molecules may be administered, where a combination may comprise multiple different

sequences.

As an alternative to antisense inhibitors, catalytic nucleic acid compounds, *e.g.*, ribozymes, anti-sense conjugates, etc., may be used to inhibit gene expression. Ribozymes may be synthesized in vitro and administered to the patient, or may be encoded on an expression vector, from which the ribozyme is synthesized in the targeted cell (for example, see International patent application WO 9523225, and Beigelman, *et al., Nucl. Acids Res.* 23:4434-42 (1995)). Examples of oligonucleotides with catalytic activity are described in WO 9506764. Conjugates of antisense oligonucleotides with a metal complex, e.g. terpyridylCu(II), capable of mediating mRNA hydrolysis are described in Bashkin, *et al., Appl. Biochem. Biotechnol.* 54:43-56 (1995).

In addition to using the identified sequences for gene therapy, the identified nucleic acids can be used to generate genetically modified non-human animals to create animal models of diseases or to generate site-specific gene modifications in cell lines for the study of protein function or regulation. The term "transgenic" is intended to encompass genetically modified animals having an exogenous gene that is stably transmitted in the host cells where, for example, the gene may be altered in sequence to produce a modified protein, or may be a reporter gene operably linked to an exogenous promoter. Transgenic animals may be made through homologous recombination, where the endogenous gene locus is altered, replaced or otherwise disrupted. Alternatively, a nucleic acid construct may be randomly integrated into the genome. Vectors for stable integration include plasmids, retroviruses and other animal viruses, YACs, and the like. Of interest are transgenic mammals, *e.g.*, cows, pigs, goats, horses, etc., and, particularly, rodents, *e.g.*, rats, mice, etc.

Investigation of genetic function may also utilize non-mammalian models, particularly using those organisms that are biologically and genetically well-characterized, such as *C. elegans, D. melanogaster and S. cerevisiae*. The subject gene sequences may be used to knock-out corresponding gene function or to complement defined genetic lesions in order to determine the physiological and biochemical pathways involved in protein function. Drug screening may be performed in combination with complementation or knock-out studies, *e.g.*, to study progression of degenerative disease, to test therapies, or for drug discovery.

In addition, the modified cells or animals are useful in the study of protein function and

regulation. For example, a series of small deletions and/or substitutions may be made in the identified gene to determine the role of different domains in enzymatic activity, cell transport or localization, etc. Specific constructs of interest include, but are not limited to, antisense constructs to block gene expression, expression of dominant negative genetic mutations, and over-expression of the identified gene. One may also provide for expression of the identified gene or variants thereof in cells or tissues where it is not normally expressed or at abnormal times of development. In addition, by providing expression of a protein in cells in which it is not normally produced, one can induce changes in cellular behavior that provide information regarding the normal function of the protein.

Protein molecules may be assayed to investigate structure/function parameters. For example, by providing for the production of large amounts of a protein product of an identified gene, one can identify ligands or substrates that bind to, modulate or mimic the action of that protein product. Drug screening identifies agents that provide, *e.g.*, a replacement or enhancement for protein function in affected cells, or for agents that modulate or negate protein function. The term "agent" as used herein describes any molecule, e.g. protein or small molecule, with the capability of altering, mimicking or masking, either directly or indirectly, the physiological function of an identified gene or gene product. Generally a plurality of assay mixtures are run in parallel with different concentrations of the agent to obtain a differential response to the various concentrations. Typically, one of these concentrations serves as a negative control, *i.e.*, at zero concentration or below the level of detection.

A wide variety of assays may be used for this purpose, including labeled in vitro protein-protein binding assays, protein-DNA binding assays, electrophoretic mobility shift assays, immunoassays for protein binding, and the like. Also, all or a fragment of the purified protein may be used for determination of three-dimensional crystal structure, which can be used for determining the biological function of the protein or a part thereof, modeling intermolecular interactions, membrane fusion, etc.

Candidate agents encompass numerous chemical classes, though typically they are organic molecules or complexes, preferably small organic compounds, having a molecular weight of more than 50 and less than about 2,500 daltons. Candidate agents comprise functional groups necessary for structural interaction with proteins, particularly

hydrogen bonding, and typically include at least an amine, carbonyl, hydroxyl or carboxyl group, and frequently at least two of the functional chemical groups. The candidate agents often comprise cyclical carbon or heterocyclic structures and/or aromatic or polyaromatic structures substituted with one or more of the above functional groups. Candidate agents are also found among biomolecules including, but not limited to: peptides, saccharides, fatty acids, steroids, purines, pyrimidines, derivatives, structural analogs or combinations thereof.

Candidate agents are obtained from a wide variety of sources including libraries of synthetic or natural compounds. For example, numerous means are available for random and directed synthesis of a wide variety of organic compounds and biomolecules, including expression of randomized oligonucleotides and oligopeptides. Alternatively, libraries of natural compounds in the form of bacterial, fungal, plant and animal extracts are available or readily produced. Additionally, natural or synthetically produced libraries and compounds are readily modified through conventional chemical, physical and biochemical means, and may be used to produce combinatorial libraries. Known pharmacological agents may be subjected to directed or random chemical modifications, such as acylation, alkylation, esterification, amidification, etc., to produce structural analogs.

Where the screening assay is a binding assay, one or more of the molecules may be coupled to a label, where the label can directly or indirectly provide a detectable signal. Various labels include radioisotopes, fluorescers, chemiluminescers, enzymes, specific binding molecules, particles, *e.g.*, magnetic particles, and the like. Specific binding molecules include pairs, such as biotin and streptavidin, digoxin and antidigoxin, etc. For the specific binding members, the complementary member would normally be labeled with a molecule that provides for detection, in accordance with known procedures.

A variety of other reagents may be included in the screening assay. These include reagents like salts, neutral proteins, *e.g.*, albumin, detergents, etc that are used to facilitate optimal protein-protein binding and/or reduce non-specific or background interactions. Reagents that improve the efficiency of the assay, such as protease inhibitors, nuclease inhibitors, anti-microbial agents, etc., may be used.

Agents may be combined with a pharmaceutically acceptable carrier, including any and all solvents, dispersion media, coatings, anti-oxidant, isotonic and absorption delaying

agents and the like. The use of such media and agents for pharmaceutically active substances is well known in the art. Except insofar as any conventional media or agent is incompatible with the active ingredient, its use in the therapeutic compositions and methods described herein is contemplated. Supplementary active ingredients can also be incorporated into the compositions.

The formulation may be prepared for use in various methods for administration. The formulation may be given orally, by inhalation, or may be injected, e.g. intravascular, intratumor, subcutaneous, intraperitoneal, intramuscular, etc. The dosage of the therapeutic formulation will vary widely, depending upon the nature of the disease, the frequency of administration, the manner of administration, the clearance of the agent from the host, and the like. The initial dose may be larger, followed by smaller maintenance doses. The dose may be administered as infrequently as once, weekly or biweekly, or fractionated into smaller doses and administered daily, semi-weekly, etc., to maintain an effective dosage level. In some cases, oral administration will require a different dose than if administered intravenously. Identified agents of the invention can be incorporated into a variety of formulations for therapeutic administration. More particularly, the complexes can be formulated into pharmaceutical compositions by combination with appropriate, pharmaceutically acceptable carriers or diluents, and may be formulated into preparations in solid, semi-solid, liquid or gaseous forms, such as tablets, capsules, powders, granules, ointments, solutions, suppositories, injections, inhalants, gels, microspheres, and aerosols. As such, administration of the agents can be achieved in various ways. Agents may be systemic after administration or may be localized by the use of an implant that acts to retain the active dose at the site of implantation.

The following methods and excipients are merely exemplary and are in no way limiting. For oral preparations, an agent can be used alone or in combination with appropriate additives to make tablets, powders, granules or capsules, for example, with conventional additives, such as lactose, mannitol, corn starch or potato starch; with binders, such as crystalline cellulose, cellulose derivatives, acacia, corn starch or gelatins; with disintegrators, such as corn starch, potato starch or sodium carboxymethylcellulose; with lubricants, such as talc or magnesium stearate; and if desired, with diluents, buffering agents, moistening agents, preservatives and flavoring agents.

Additionally, agents may be formulated into preparations for injections by

dissolving, suspending or emulsifying them in an aqueous or nonaqueous solvent, such as vegetable or other similar oils, synthetic aliphatic acid glycerides, esters of higher aliphatic acids or propylene glycol; and if desired, with conventional additives such as solubilizers, isotonic agents, suspending agents, emulsifying agents, stabilizers and preservatives. Further, agents may be utilized in aerosol formulation to be administered via inhalation. The agents identified by the present invention can be formulated into pressurized acceptable propellants such as dichlorodifluoromethane, propane, nitrogen and the like. Alternatively, agents may be made into suppositories by mixing with a variety of bases such as emulsifying bases or water-soluble bases. Further, identified agents of the present invention can be administered rectally via a suppository. The suppository can include vehicles such as cocoa butter, carbowaxes and polyethylene glycols, which melt at body temperature, yet are solidified at room temperature.

Implants for sustained release formulations are well-known in the art. Implants are formulated as microspheres, slabs, etc. with biodegradable or non-biodegradable polymers. For example, polymers of lactic acid and/or glycolic acid form an erodible polymer that is well-tolerated by the host. The implant containing identified agents of the present invention may be placed in proximity to the site of action, so that the local concentration of active agent is increased relative to the rest of the body. Unit dosage forms for oral or rectal administration such as syrups, elixirs, and suspensions may be provided wherein each dosage unit, for example, teaspoonful, tablespoonful, gel capsule, tablet or suppository, contains a predetermined amount of the compositions of the present invention. Similarly, unit dosage forms for injection or intravenous administration may comprise the compound of the present invention in a composition as a solution in sterile water, normal saline or another pharmaceutically acceptable carrier. The specifications for the novel unit dosage forms of the present invention depend on the particular compound employed and the effect to be achieved, and the pharmacodynamics associated with each active agent in the host.

The pharmaceutically acceptable excipients, such as vehicles, adjuvants, carriers or diluents, are readily available to the public. Moreover, pharmaceutically acceptable auxiliary substances, such as pH adjusting and buffering agents, tonicity adjusting agents, stabilizers, wetting agents and the like, are readily available to the public.

A therapeutic dose of an identified agent is administered to a host suffering from a

disease or disorder. Administration may be topical, localized or systemic, depending on the specific disease. The compounds are administered at an effective dosage such that over a suitable period of time the disease progression may be substantially arrested. It is contemplated that the composition will be obtained and used under the guidance of a physician for *in vivo* use. The dose will vary depending on the specific agent and formulation utilized, type of disorder, patient status, etc., such that it is sufficient to address the disease or symptoms thereof, while minimizing side effects. Treatment may be for short periods of time, *e.g.*, after trauma, or for extended periods of time, *e.g.*, in the prevention or treatment of schizophrenia.

The SNPs identified by the present invention may be used to analyze the expression pattern of an associated gene and the expression pattern correlated to a phenotypic trait of the organism such as disease susceptibility or drug responsiveness. The expression pattern in various tissues can be determined and used to identify ubiquitous expression patterns, tissue specific expression patterns, temporal expression patterns and expression patterns induced by various external stimuli such as chemicals or electromagnetic radiation. Such determinations would provide information regarding function of the gene and/or its protein product.

The newly identified sequences also may be used as diagnostic markers, *i.e.*, to predict a phenotypic characteristic such as disease susceptibility or drug responsiveness. In addition, the methods of the present invention may be used to stratify populations for clinical studies. As such, the genes or fragments thereof may be used as probes to determine whether the same nucleic acid sequence is present in the genome of an organism being tested. In addition, the probes may be used to monitor RNA or mRNA levels within the organism to be tested or a part thereof, such as a specific tissue or organ, so as to determine the expression level of the marker where the expression level can be correlated to a particular phenotypic characteristic of the organism. Likewise, the marker may be assayed at the protein level using any customary technique such as immunological methods–Western blots, radioimmune precipitation and the like–or activity based assays measuring an activity associated with the gene product. Moreover, when a phenotype cannot clearly distinguish between similar diseases having different genetic bases, the methods of the present invention can be used to identify correctly the disease.

Also, it should be apparent that the methods of the present invention can be used on organisms aside from humans. For example, when the organism is an animal, the methods of the invention may be used to identify loci associated, *e.g.*, with disease resistance/ or susceptibility, environmental tolerance, drug response or the like, and when the organism is a plant, the method of the invention may be used to identify loci associated with disease resistance/ or susceptibility, environmental tolerance and or herbicide resistance.

It is to be understood that this invention is not limited to the particular methodology, protocols, cell lines, animal species or genera, and reagents described, as such may vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention, which will be limited only by the appended claims.

Databases

The present invention includes databases containing information concerning variations, for instance, information concerning SNPs, SNP haplotype blocks, SNP haplotype patterns and informative SNPs. In some embodiments, the databases of the present invention may comprise information on one or more haplotype patterns associated with one or more phenotypic traits. Databases may also contain information associated with a given variation such as descriptive information about the general genomic region in which the variation occurs, such as whether the variation is located in a known gene, whether there are known genes, gene homologs or regulatory regions nearby and the like.

Other information that may be included in the databases of the present invention include, but are not limited to, SNP sequence information, descriptive information concerning the clinical status of a tissue sample analyzed for SNP haplotype patterns, or the clinical status of the patient from which the sample was derived. The database may be designed to include different parts, for instance a variation database, a SNP database, a SNP haplotype block or SNP haplotype pattern database and an informative SNP database. Methods for the configuration and construction of databases are widely available, for instance, see Akerblom *et al.*, (1999) U.S. Patent 5,953,727, which is herein incorporated by reference in its entirety.

The databases of the invention may be linked to an outside or external database.

Figure 9 shows an exemplary computer network that is suitable for the databases and executing the software of the present invention. A computer workstation 902 is connected with the application/data server(s) 906 through a local area network (LAN), such as an ethernet 905. A printer 904 may be connected directly to the workstation or to the Ethernet 905. The LAN may be connected to a wide area network (WAN), such as the internet 908 via a gateway server 907 which may also serve as a firewall between the WAN 908 and the LAN 905. In preferred embodiments, the workstation may communicate with outside data sources, such as The SNP Consortium (TSC) or the National Center for Biotechnology Information 909, through the internet 908.

Any appropriate computer platform may be used to perform the necessary comparisons between SNP haplotype blocks or patterns, associated phenotypes, any other information in the database or information provided as an input. For example, a large number of computer workstations are available from a variety of manufacturers, such has those available from Silicon Graphics. Client-server environments, database servers and networks are also widely available and are appropriate platforms for the databases of the invention.

The databases of the invention may also be used to present information identifying the SNP haplotype pattern in an individual and such a presentation may be used to predict one or more phenotypic traits of the individual. Such methods may be used to predict the disease susceptibility/resistance and/or drug response of the individual. Further, the databases of the present invention may comprise information relating to the expression level of one or more of the genes associated with the variations of the invention.

The following examples describe specific embodiments of the present invention and the materials and methods are illustrative of the invention and are not intended to limit the scope of the invention.

Example 1: Preparation of Somatic Cell Hybrids

Standard procedures in somatic cell genetics were used to separate human DNA strands (chromosomes) from a diploid state to a haploid state. In this case, a diploid human lymphoblastoid cell line that was wildtype for the thymidine kinase gene was fused to a diploid hamster fibroblast cell line containing a mutation in the thymidine kinase gene. A sub-population of the resulting cells were hybrid cells containing human

chromosomes. Hamster cell line A23 cells were pipetted into a centrifuge tube containing 10 ml DMEM in which 10% fetal bovine serum (FBS) + 1X Pen/Strep + 10% glutamine were added, centrifuged at 1500 rpm for 5 minutes, resuspended in 5 ml of RPMI and pipetted into a tissue culture flask containing 15 ml RPMI medium. The lymphoblastoid cells were grown at 37° C to confluence. At the same time, human lymphoblastoid cells were pipetted into a centrifuge tube containing 10 ml RPMI in which 15% FBCS + 1x Pen/Strep + 10% glutamine were added, centrifuged at 1500 rpm for 5 minutes, resuspended in 5 ml of RPMI and pipetted into a tissue culture flask containing 15 ml RPMI. The lymphoblastoid cells were grown at 37 °C to confluence.

To prepare the A23 hamster cells, the growth medium was aspirated and the cells were rinsed with 10 ml PBS. The cells were then trypsinized with 2 ml of trypsin, divided onto 3-5 plates of fresh medium (DMEM without HAT) and incubated at 37 °C. The lymphoblastoid cells were prepared by transferring the culture into a centrifuge tube and centrifuging at 1500 rpm for 5 minutes, aspirating the growth medium, resuspending the cells in 5 ml RPMI and pipetting 1 to 3 ml of cells into 2 flasks containing 20 ml RPMI.

To achieve cell fusion, approximately 8-10 x 10^6 lymphoblastoid cells were centrifuged at 1500 rpm for 5 min. The cell pellet was then rinsed with DMEM by resuspending the cells, centrifuging them again and aspirating the DMEM. The lymphoblastoid cells were then resuspended in 5 ml fresh DMEM. The recipient A23 hamster cells had been grown to confluence and split 3-4 days before the fusion and were, at this point, 50-80% confluent. The old media was removed and the cells were rinsed three times with DMEM, trypsinized, and finally suspended in 5 ml DMEM. The lymphoblastoid cells were slowly pipetted over the recipient A23 cells and the combined culture was swirled slowly before incubating at 37 °C for 1 hour. After incubation, the media was gently aspirated from the A23 cells, and 2 ml room temperature PEG 1500 was added by touching the edge of the plate with a pipette and slowly adding PEG to the plate while rotating the plate with the other hand. It took approximately one minute to add all the PEG in one full rotation of the plate. Next, 8 ml DMEM was added down the edge of the plate while rotating the plate slowly. The PEG/DMEM mixture was aspirated gently from the cells and then 8 ml DMEM was used to rinse the cells. This DMEM was removed and 10 ml fresh DMEM was added and the cells were incubated for 30 min. at

37 °C. Again the DMEM was aspirated from the cells and 10 ml DMEM in which 10% FBCS and 1x Pen/Strep were added, was added to the cells, which were then allowed to incubate overnight.

After incubation, the media was aspirated and the cells were rinsed with PBS. The cells were then trypsinized and divided among plates containing selection media (DMEM in which 10% FBS + 1x Pen/Strep + 1x HAT were added) so that each plate received approximately 100,000 cells. The media was changed on the third day following plating. Colonies were picked and placed into 24-well plates upon becoming visible to the naked eye (day 9-14). If a picked colony was confluent within 5 days, it was deemed healthy and the cells were trypsinized and moved to a 6-well plate.

DNA and stock hybrid cell cultures were prepared from the cells from the 6-well plate cultures. The cells were trypsinized and divided between a 100 mm plate containing 10 ml selection media and an Eppendorf tube. The cells in the tube were pelleted, resuspended 200 µl PBX and DNA was isolated using a Qiagen DNA mini kit at a concentration of <5 million cells per spin column. The 100 mm plate was grown to confluence, and the cells were either continued in culture or frozen.

Example 2: Selecting Haploid Hybrids

Scoring for the presence, absence and diploid/haploid state of human chromosomes in each hybrid was performed using the Affymetrix, HuSNP genechip (Affymetrix, Inc,. of Santa Clara, CA, HuSNP Mapping Assay, reagent kit and user manual, Affymetrix Part No. 900194), which can score 1494 markers in a single chip hybridization. As controls, the hamster and human diploid lymphoblastoid cell lines were screened using the HuSNP chip hybridization assay. Any SNPs which were heterozygous in the parent lymphoblastoid diploid cell line were scored for haploidy in each fusion cell line. Assume that "A" and "B" are alternative variants at each SNP location. By comparing the markers that were present as "AB" heterozygous in the parent diploid cell line to the same markers present as "A" or "B" (hemizygous) in the hybrids, the human DNA strands which were in the haploid state in each hybrid line was determined.

Figure 11 shows results after two human/hamster cell hybrids (Hybrid 1 and Hybrid 2) are tested for selected markers on human chromosome 21. The first column lists the HuSNP chip marker designations. The second column reports whether a signal

was obtained when the hamster cell nucleic acid (no fusion) was used for hybridization with a HuSNP chip. As expected, there was no signal for any marker in the hamster cell sample. The third column reports which variants for each marker were detected ("A", "B" or "AB") in the diploid parent human lymphoblastoid cell line, CPD17. In some instances, only an A variant was present, in some instances only a B variant was present, and in some cases the CPD17 cells were heterozygous ("AB") for the variants. The last two columns report the result when nucleic acid samples from two human/hamster hybrids (Hybrid 1 and Hybrid 2) are hybridized with the HuSNP chip. Note in cases where only A variants were present in the parent CPD17 cell line, only A variants were transferred in the fusion. In cases where only B variants were present in the parent CPD17 cell line, only B variants were transferred in the fusion. In cases where the CPD17 cell line was heterozygous, an A variant was transferred to some fusion clones, and a B variant was transferred to other fusion clones. It should be understood, however, that often only portions of chromosomes are present in the hybrid cell lines resulting from this fusion process, that some hybrids may be diploid for some human chromosomes or portions thereof, that some hybrids may be haploid for other human chromosomes or portions thereof, and some hybrids may not have either variant of some chromosomes. Hybrids containing only one variant of a particular human chromosome (for instance, chromosome 21) were selected for analysis. Even more preferably, hybrids containing a whole chromosome (as opposed to only a portion thereof) were selected for analysis.

Example 3: Long Range PCR

DNA from the hamster/human cell hybrids was used to perform long-range PCR assays. Long range PCR assays are known generally in the art and have been described, for example, in the standard long range PCR protocol from the Boehringer Mannheim Expand Long Range PCR Kit, incorporated herein by reference or all purposes.

Primers used for the amplification reactions were designed in the following way: a given sequence, for example the 23 megabase contig on chromosome 21, was entered into a software program known in the art herein called "repeat masker" which recognizes sequences that are repeated in the genome (*e.g.*, Alu and Line elements)(see, A. F. A. Smit and P. Green, www.genome.washington.edu/uwgc/analysistools/repeatmask, incorporated herein by reference). The repeated sequences were "masked" by the

program by substituting each specific nucleotide of the repeated sequence (A, T, G or C) with "N". The sequence output after this repeat mask substitution was then fed into a commercially available primer design program (Oligo 6.23) to select primers that were greater than 30 nucleotides in length and had melting temperatures of over 65 °C. The designed primer output from Oligo 6.23 was then fed into a program which then "chose" primer pairs which would PCR amplify a given region of the genome but have minimal overlap with the adjacent PCR products. The success rate for long range PCR using commercially available protocols and this primer design was at least 80%, and greater than 95% success was achieved on some portions of human chromosomes.

An illustrative protocol for long range PCR uses the Expand Long Template PCR System from Boehringer Mannheim Cat.# 1681 834, 1681 842, or 1759 060. In the procedure each 50 μL PCR reaction requires two master mixes. In a specific example, Master Mix 1 was prepared for each reaction in 1.5 ml microfuge tubes on ice and includes a final volume of 19 μL of Molecular Biology Grade Water (Bio Whittaker, Cat.# 16-001Y); 2.5 μL 10 mM dNTP set containing dATP, dCTP, dGTP, and dTTP at 10 mM each (Life Technologies Cat.# 10297-018) for a final concentration of 400 μM of each dNTP; and 50 ng DNA template.

Master Mix 2 for all reactions was prepared and kept on ice. For each PCR reaction Master Mix 2 includes a final volume of 25 μL of Molecular Biology Grade Water (Bio Whittaker); 5 μL 10 x PCR buffer 3 containing 22.50 mM $MgCl_2$ (Sigma, Cat.# M 10289); 2.5 μL 10 mM $MgCl_2$ (for a final $MgCl_2$ concentration of 2.75 mM); and 0.75 μL enzyme mix (added last)

Six microliters of premixed primers (containing 2.5 μL of Master Mix 1) were added to appropriate tubes, then 25 μL of Master Mix 2 was added to each tube. The tubes were capped, mixed, centrifuged briefly and returned to ice. At this point, the PCR cycling was begun according to the following program: step 1: 94°C for 3 min to denature template; step 2: 94°C for 30 sec; step 3: annealing for 30 sec at a temperature appropriate for the primers used; step 4: elongation at 68°C for 1 min/kb of product; step 5: repetition of steps 2-4 38 times for a total of 39 cycles; step 6: 94°C for 30 sec; step 7: annealing for 30 sec; step 8: elongation at 68°C for 1 min/kb of product plus 5 additional minutes; and step 9: hold at 4°C. Alternatively, a two-step PCR would be performed: step 1: 94°C for 3 min to denature template; step 2: 94°C for 30 sec; step 3: annealing and elongation at

68°C for 1 min/kb of product; step 4: repetition of steps 2-3 38 times for a total of 39 cycles; step 5:94°C for 30 sec; step 6: annealing and elongation at 68°C for 1 min/kb of product plus 5 additional minutes; and step 7: hold at 4°C.

Results of the long range PCR amplification reaction for various regions on human chromosomes 14 and 22 were visualized on ethidium bromide-stained agarose gels (Figure 12). The long range PCR amplification methods of the present invention routinely produced amplified fragments having an average size of about 8 kb, and appeared to fail to amplify genomic regions in only rare cases (see G11 on the chromosome 22 gel).

Example 4: Wafer Design, Manufacture, Hybridization and Scanning

The set of oligonucleotide probes to be contained on an oligonucleotide array (chip or wafer) was defined based on the human DNA strand sequence to be queried. The oligonucleotide sequences were based on consensus sequences reported in publicly available databases. Once the probe sequences were defined, computer algorithms were used to design photolithographic masks for use in manufacturing the probe-containing arrays. Arrays were manufactured by a light-directed chemical synthesis processes which combines solid-phase chemical synthesis with photolithographic fabrication techniques. See, for example, WO 92/10092, or U.S. Patent Nos. 5,143,854; 5,384,261; 5,405,783; 5,412,087; 5,424,186; 5,445,934; 5,744,305; 5,800,992; 6,040,138; 6,040,193, all of which are incorporated herein by reference in their entireties for all purposes. Using a series of photolithographic masks to define exposure sites on the glass substrate (wafer) followed by specific chemical synthesis steps, the process constructed high-density areas of oligonucleotide probes on the array, with each probe in a predefined position. Multiple probe regions were synthesized simultaneously and in parallel.

The synthesis process involved selectively illuminating a photo-protected glass substrate by passing light through a photolithographic mask wherein chemical groups in unprotected areas were activated by the light. The selectively-activated substrate wafers were then incubated with a chosen nucleoside, and chemical coupling occurred at the activated positions on the wafer. Once coupling took place, a new mask pattern was applied and the coupling step was repeated with another chosen nucleoside. This process was repeated until the desired set of probes was obtained. In one specific example, 25-

mer oligonucleotide probes were used, where the thirteenth base was the base to be queried. Four probes were used to interrogate each nucleotide present in each sequence--one probe complementary to the sequence and three mismatch probes identical to the complementary probe except for the thirteenth base. In some cases, at least 10×10^6 probes were present on each array.

Once fabricated, the arrays were hybridized to the products from the long range PCR reactions performed on the hamster-human cell hybrids. The samples to be analyzed were labeled and incubated with the arrays to allow hybridization of the sample to the probes on the wafer.

After hybridization, the array was inserted into a confocal, high performance scanner, where patterns of hybridization were detected. The hybridization data were collected as light emitted from fluorescent reporter groups already incorporated into the PCR products of the sample, which was bound to the probes. Sequences present in the sample that are complimentary to probes on the wafer hybridized to the wafer more strongly and produced stronger signals than those sequences that had mismatches. Since the sequence and position of each probe on the array was known, by complementarity, the identity of the variation in the sample nucleic acid applied to the probe array was identified. Scanners and scanning techniques used in the present invention are known to those skilled in the art and are disclosed in, *e.g.*, U.S. Patent No. 5,981,956 drawn to microarray chips, U.S. Patent No. 6,262,838 and U.S. Patent No. 5,459,325. U.S.S.N. In addition, 60/223,278 filed on August 3, 2000, and non-provisional application claiming priority to USSN 60/223,278 filed on August 3, 2001, drawn to scanners and techniques for whole wafer scanning, are also incorporated herein by reference in their entireties for all purposes.

Example 5: Determination of SNP Haplotypes on Human Chromosome 21

Twenty independent copies of chromosome 21, representing African, Asian, and Caucasian chromosomes were analyzed for SNP discovery and haplotype structure. Two copies of chromosome 21 from each individual were physically separated using a rodent-human somatic cell hybrid technique (Figure 10), discussed *supra*. The reference sequence for the analysis consisted of human chromosome 21 genomic DNA sequence consisting of 32,397,439 bases. This reference sequence was masked for repetitive

sequences and the resulting 21,676,868 bases (67%) of unique sequence were assayed for variation with high density oligonucleotide arrays. Eight unique oligonucleotides, each 25 bases in length, were used to interrogate each of the unique sample chromosome 21 bases, for a total of 1.7 X 10^8 different oligonucleotides. These oligonucleotides were distributed over a total of eight different wafer designs using a previously described tiling strategy (Chee, *et al.*, *Science* 274:610 (1996)). Light-directed chemical synthesis of oligonucleotides was carried out on 5 inch x 5 inch glass wafers purchased from Affymetrix, Inc. (Santa Clara, CA).

Unique oligonucleotides were designed to generate 3253 minimally overlapping longe range PCR (LRPCR) products of 10 kb average length spanning 32.4 Mb of contiguous chromosome 21 DNA, and were prepared as described *supra*. For each wafer hybridization, corresponding LRPCR products were pooled and were purified using Qiagen tip 500 (Qiagen). A total of 280 µg of purified DNA was fragmented using 37 µl of 10X One-Phor-All buffer PLUS (Promega) and 1 unit of DNAase (Life Technolgies/Invitrogen) in 370 µl total volume at 37°C for 10 min followed by heat inactivation at 99°C for 10 min. The fragmented products were end labeled using 500 units of Tdt (Boehringer Manheim) and 20 nmoles of biotin-N6-ddATP (DuPont NEN) at 37°C for 90 min and heat inactivated at 95°C for 10 min. The labeled samples were hybridized to the wafers in 10 mM Tris-HCL (pH 8), 3M Tetramethylammonium chloride, 0.01% Tx-100, 10 µg/ml denatured herring sperm DNA in a total volume of 14 ml per wafer at 50°C for 14-16 hours. The wafers were rinsed briefly in 4X SSPE, washed three times in 6X SSPE for 10 min each, stained using streptavidin R-phycoerythrin (SAPE, 5 ng/ml) at room temp for 10 min. The signal was amplified by staining with an antibody against streptavidin (1.25 ng/ml) and by repeating the staining step with SAPE.

PCR products corresponding to the bases present on a single wafer were pooled and hybridized to the wafer as a single reaction. In total, 3.4 x 10^9 oligonucleotides were synthesized on 160 wafers to scan 20 independent copies of human chromosome 21 for DNA sequence variation. Each unique chromosome 21 was amplified from a rodent-human hybrid cell line by using long range PCR. LRPCR assays were designed using Oligo 6.23 primer design software with high-moderate stringency parameters. The resulting primers were typically 30 nucleotides in length with the melting temperature of

> 65°C. The range of amplicon size was from 3 kb-14 kb. A primer database for the entire chromosome was generated and software (pPicker) was utilized to choose a minimal set of non-redundant primers that yield maximum coverage of chromosome 21 sequence with a minimal overlap between adjacent amplicons. Alternatively, the primer selection method described in Example 3, herein, was employed. LRPCR reactions were performed using the Expand Long Template PCR Kit (Boehringer Mannheim) with minor modifications. The wafers were scanned using a custom built confocal scanner.

SNPs were detected as altered hybridization by using a pattern recognition algorithm. A combination of previously described algorithms (Wang, *et al.*, *Science* 280:1077 (1998)), was used to detect SNPs based on altered hybridization patterns. In total, 35,989 SNPs were identified in the sample of twenty chromosomes. The position and sequence of these human polymorphisms have been deposited in GenBank's SNPdb. Dideoxy sequencing was used to assess a random sample of 227 of these SNPs in the original DNA samples, confirming 220 (97%) of the SNPs assayed. In order to achieve this low rate of 3% false positive SNPs, stringent thresholds were required for SNP detection on wafers that resulted in a high false negative rate. Approximately 65% of all bases present on the wafers yielded data of high enough quality for use in SNP detection with 35% being discarded as being false negatives. Consistent failure of long range PCR in all samples analyzed accounts for 15% of the 35% false negative rate. The remaining 20% false negatives are distributed between bases that never yield high quality data (10%) and bases that yield high quality data in only a fraction of the 20 chromosomes analyzed (10%). In general, it is the sequence context of a base that dictates whether or not it will yield high quality data. The finding that approximately 20% of all bases give consistently poor data is very similar to the finding that approximately 30% of bases in single dideoxy sequencing reads of 500 bases have quality scores too low for reliable SNP detection (Altschuler, *et al.*, *Nature* 407:513 (2000)). The power to discover rare SNPs as compared to more frequent SNPs is disproportionately reduced in cases where only a limited number of the samples analyzed yield high quality data for a given base. As a result, SNP discovery by this method is biased in favor of common SNPs.

Figure 13A shows the distribution of minor allele frequencies of all 35,989 SNPs discovered in the sample of globally diverse chromosomes. Genetic variation, normalized for the number of chromosomes in the sample, was estimated with two

measures of nucleotide diversity: π the average heterozygosity per site and θ the population mutation parameter (see Hartl and Clark, Principles of Population Genetics (Sinauer, Massachusetts, 1997)). The 32,397,439 bases of finished genomic chromosome 21 DNA were divided into 200,000 base pair segments, and the high-quality base pairs used for SNP discovery in each segment were examined. The observed heterozygosity of these bases was used to calculate an average nucleotide diversity (π) for each segment. The estimates of average nucleotide diversity for the total data set ($\pi = 0.000723$ and $\theta = 0.000798$), as well as the distribution of nucleotide diversity, measured in contiguous 200,000 base pair bins of chromosome 21 (Fig. 13B), are within the range of values previously described (The International SNP Map Working Group, *Nature* 409:928-33 (2001)).

The extent of overlap of 15,549 chromosome 21 SNPs discovered by The SNP Consortium (TSC) was compared with the SNPs found in this study. Of the TSC SNPs, 5,087 were found to be in repeated DNA and were not tiled on the wafers. Of the remaining 10,462 TSC SNPs, 4705 (45%) were identified. The estimate of θ was observed to be greater than the estimate of π□for 129 of the 162 200-kb bins of contiguous DNA sequence analyzed. This difference is consistent with a recent expansion of the human population and is similar to the finding of a recent study of nucleotide diversity in human genes (Stephens, *et al.*, *Science* 293:489 (2001)). It was found that 11,603 of the SNPs (32%) had a minor allele observed a single time in the sample (singletons), as compared with the neutral model expectation of 43% singletons given the observed amount of nucleotide diversity (Fu and Li, *Genetics* 133:693 (1993)). The difference between the observed and expected values is likely attributable to the reduced power to identify rare as compared to common SNPs in this study as discussed above.

Over all, 47% of the 53,000 common SNPs with an allele frequency of 10% or greater estimated to be present in 32.4 Mb of the human genome were identified. This compares with an estimate of 18-20% of all such common SNPs present in the collection generated by the International SNP Mapping Working Group and the SNP Consortium. The difference in coverage is explained by the fact that the present study used larger numbers of chromosomes for SNP discovery. To assess the replicability of the findings, SNP discovery was performed for one wafer design with nineteen additional copies of

chromosome 21 derived from the same diversity panel as the original set of samples. A total of 7188 SNPs were identified using the two sets of samples. On average, 66% of all SNPs found in one set of samples were discovered in the second set, consistent with previous findings (Marth, *et al.*, *Nature Genet.* 27:371 (2001) and Yang, *et al.*, *Nature Genet.* 26:13 (2000)). As expected, failure of a SNP to replicate in a second set of samples is strongly dependent on allele frequency. It was found that 80% of SNPs with a minor allele present two or more times in a set of samples were also found in a second set of samples, while only 32% of SNPs with a minor allele present a single time were found in a second set of samples. These findings suggest that the 24,047 SNPs in the collection with a minor allele represented more than once are highly replicable in different global samples and that this set of SNPs is useful for defining common global haplotypes. In the course of SNP discovery, 339 SNPs which appeared to have more than two alleles were identified. These SNPs were not included in the present analysis.

In addition to the replicability of SNPs in different samples, the distance between consecutive SNPs in a collection of SNPs is critical for defining meaningful haplotype structure. Haplotype blocks, which can be as short as several kb, may go unrecognized if the distance between consecutive SNPs in a collection is large relative to the size of the actual haplotype blocks. The collection of SNPs in this study was very evenly distributed across the chromosome, even though repeat sequences were not included in the SNP discovery process. Figure 13C shows the distribution of SNP coverage across 32,397,439 bases of finished chromosome 21 DNA sequence. An interval is the distance between consecutive SNPs. There are a total of 35,988 intervals for the entire SNP set and a total of 24,046 intervals for the common SNP set (i.e. SNPs with a minor allele present more than once in the sample). The average distance between consecutive SNPs was 900 bases when all SNPs are considered, and 1300 bases when only the 24,047 common SNPs were considered. For this set of common SNPs, 93% of intervals between consecutive SNPs in genomic DNA, including repeated DNA, were 4000 bases or less (again, see Figure 13C).

The construction of haplotype blocks or patterns from diploid data is complicated by the fact that the relationship between alleles for any two heterozygous SNPs is not directly observable. Consider an individual with two copies of chromosome 21 and two alleles, A and G, at one chromosome 21 SNP, as well as two alleles, A and G, at a second chromosome 21 SNP. In such a case, it is unclear if one copy of chromosome 21 contains

allele A at the first SNP and allele A at the second SNP, while the other copy of chromosome 21 contains allele G at the first SNP and allele G at the second SNP, or if one copy of chromosome 21 contains allele A at the first SNP and allele G at the second SNP, while the other copy of chromosome 21 contains allele G at the first SNP and allele A at the second SNP. Current methods used to circumvent this problem include statistical estimation of haplotype frequencies, direct inference from family data, and allele-specific PCR amplification over short segments.

To avoid these complexities, the present invention characterized SNPs on haploid copies of chromosome 21 isolated in rodent-human somatic cell hybrids were characterized, allowing direct determination of the full haplotypes of these chromosomes. The set of 24,047 SNPs with a minor allele represented more than once in the data set was used to define the haplotype structure are shown in Figure 14. The haplotype patterns for twenty independent globally diverse chromosomes defined by 147 common human chromosome 21 SNPs is shown. The 147 SNPs span 106 kb of genomic DNA sequence. Each row of colored boxes represents a single SNP. The black boxes in each row represent the major allele for that SNP, and the white boxes represent the minor allele. Absence of a box at any position in a row indicates missing data. Each column of colored boxes represents a single chromosome, with the SNPs arranged in their physical order on the chromosome. Invariant bases between consecutive SNPs are not represented in the figure. The 147 SNPs are divided into eighteen blocks, defined by black horizontal lines. The position of the base in chromosome 21 genomic DNA sequence defining the beginning of one block and the end of the adjacent block is indicated by the numbers to the left of the vertical black line. The expanded boxes on the right of the figure represent a SNP block defined by 26 common SNPs spanning 19 kb of genomic DNA. Of the seven different haplotype patterns represented in the sample, the four most common patterns include sixteen of the twenty chromosomes sampled (i.e. 80% of the sample). The black and white circles indicate the allele patterns of two informative SNPs, which unambiguously distinguish between the four common haplotypes in this block. Although no two chromosomes shared an identical haplotype pattern for these 147 SNPs, there are numerous regions in which multiple chromosomes shared a common pattern. One such region, defined by 26 SNPs spanning 19 kb, is expanded for more detailed analysis (again, see the enlarged region of Figure 14). This block defines seven unique haplotype

patterns in 20 chromosomes. Despite the fact that some data is missing due to failure to pass the threshold for data quality, in all cases a given chromosome can be assigned unambiguously to one of the seven haplotypes. The four most frequent haplotypes, each of which is represented by three or more chromosomes, account for 80% of all chromosomes in the sample. Only two "informative" SNPs out of the total of twenty-six are required to distinguish the four most frequent haplotypes from one another. In this example, four chromosomes with infrequent haplotypes would be incorrectly classified as common haplotypes by using information from only these two informative SNPs. Nevertheless, it is remarkable that 80% of the haplotype structure of the entire global sample is defined by less than 10% of the total SNPs in the block. Several different possibilities exist in which three informative SNPs can be chosen so that each of the four common haplotypes is defined uniquely by a single SNP. One of these "three SNP" choices would be preferred over the two SNP combination in an experiment involving genotyping of pooled samples, since the two SNP combination would not permit determination of frequencies of the four common haplotypes in such a situation; thus, the present invention provides a dramatic improvement over the random selection method of SNP mapping.

In summary, while the particular application may dictate the selection of informative SNPs to capture haplotype information, it is clear that the majority of the haplotype information in the sample is contained in a very small subset of all the SNPs. It is also clear that random selection of two or three informative SNPs from this block of SNPs will often not provide enough information to uniquely assign a chromosome to one of the four common haplotypes.

One issue is how to define a set of contiguous blocks of SNPs spanning the entire 32.4 Mb of chromosome 21 while minimizing the total number of SNPs required to define the haplotype structure. In one embodiment, an optimization algorithm based on a "greedy" strategy was used to address this problem. All possible blocks of physically consecutive SNPs of size one SNP or larger were considered. Ambiguous haplotype patterns were treated as missing data and were not included when calculating percent coverage. Considering the remaining overlapping blocks simultaneously, the block with the maximum ratio of total SNPs in the block to the minimal number of SNPs required to uniquely discriminate haplotypes represented more than once in the block was selected.

Any of the remaining blocks that physically overlapped with the selected block were discarded, and the process was repeated until a set of contiguous, non-overlapping blocks that cover the 32.4Mb of chromosome 21 with no gaps, and with every SNP assigned to a block, was selected. Given the sample size of twenty chromosomes, the algorithm produces a maximum of ten common haplotype patterns per block, each represented by two independent chromosomes.

Applying this algorithm to the data set of 24,047 common SNPs, 4135 blocks of SNPs spanning chromosome 21 were defined. A total of 589 blocks, comprising 14% of all blocks, contain greater than ten SNPs per block and include 44% of the total 32.4 Mb. In contrast, 2138 blocks, comprising 52% of all blocks, contain less than three SNPs per block and make up only 20% of the physical length of the chromosome. The largest block contains 114 common SNPs and spans 115 kb of genomic DNA. Overall, the average physical size of a block is 7.8 kb. The size of a block is not correlated with its order on the chromosome, and large blocks are interspersed with small blocks along the length of the chromosome. There are an average of 2.7 common haplotype patterns per block, defined as haplotype patterns that are observed on multiple chromosomes. On average, the most frequent haplotype pattern in a block is represented by 9.6 chromosomes out of the twenty chromosomes in the sample, the second most frequent haplotype pattern is represented by 4.2 chromosomes, and the third most frequent haplotype patterns, if present, is represented by 2.1 chromosomes. The fact that such a large fraction of globally diverse chromosomes are represented by such limited haplotype diversity is remarkable. The findings are consistent with the observation that when haplotype pattern frequency is considered, 82% of the haplotype patterns observed in a collection of 313 human genes are observed in all ethnic groups, while only 8% of haplotypes are population specific (Stephens, *et al.*, *Science* 293:489-93 (2001)). Several experiments were performed to measure the influence of parameters of the haplotype algorithm on the resulting block patterns. The fraction of chromosomes required to be covered by common haplotypes was varied, from an initial 80%, to 70% and 90%. As would be expected, requiring more complete coverage results in somewhat larger numbers of shorter blocks. Using only the 16,503 SNPs with a minor allele frequency of at least 20% in the sample resulted in somewhat longer blocks, but the numbers of SNPs per block did not change significantly. For one region of about 3 Mb, a

deeper sample of 38 chromosomes for SNPs and common haplotype blocks with at least 10% frequency was analyzed, so as to be comparable with the 20 chromosome analysis. The resulting distribution of block sizes closely matched the initial results. Also, a randomization test was performed in which the non-ambiguous alleles at each SNP were permuted, and then used for haplotype block discovery. In this analysis, 94% of blocks contained fewer than three SNPs, and only one block contained more than five SNPs. This confirms that the larger blocks seen in the data cannot be produced by chance associations or as artifacts of the block selection methods of the present invention.

In an effort to determine if genes were proportionately represented in both large and small blocks, a determination was made of the number of exonic bases in blocks containing more than 10 SNPs, 3 to 10 SNPs, and less than 3 SNPs. Exonic bases are somewhat over-represented as compared to total bases in blocks containing 3 to 10 SNPs ($p<0.05$ as determined by a permutation test).

Based on knowledge of the haplotype structure within blocks, subsets of the 24,047 common SNPs can be selected to capture any desired fraction of the common haplotype information, defined as complete information for haplotypes present more than once and including greater than 80% of the sample across the entire 32.4 Mb. Figure 15 shows the number of SNPs required to capture the common haplotype information for 32.4 Mb of chromosome 21. For each SNP block, the minimum number of SNPs required to unambiguously distinguish haplotypes in that block that are present more than once (*i.e.*, common haplotype information) was determined. These SNPs provide common haplotype information for the fraction of the total physical distance defined by that block. Beginning with the SNPs that provide common haplotype information for the greatest physical distance, the cumulative increase in physical coverage (*i.e.*, fraction covered) is plotted relative to the number of SNPs added (*i.e.*, SNPs required). Genic DNA includes all genomic DNA beginning 10 kb 5' of the first exon of each known chromosome 21 gene and extending 10 kb 3' of the last exon of that gene. For example, while a minimum of 4563 SNPs are required to capture all the common haplotype information, only 2793 SNPs are required to capture the common haplotype information in blocks containing three or more SNPs that cover 81% of the 32.4 Mb. A total of 1794 SNPs are required to capture all the common haplotype information in genic DNA, representing approximately two hundred and twenty distinct genes.

The present invention has particular relevance for whole-genome association studies mapping phenotypes such as common disease genes. This approach relies on the hypothesis that common genetic variants are responsible for susceptibility to common diseases (Risch and Merikangas, *Science* 273:1516 (1996), Lander, *Science* 274:536 (1996)). By comparing the frequency of genetic variants in unrelated cases and controls, genetic association studies can identify specific haplotypes in the human genome that play important roles in disease. While this approach has been used to successfully associate single candidate genes with disease (Altschuler, *et al.*, *Nature Genet.* 26:76 (2000)), the recent availability of the human DNA sequence offers the possibility of surveying the entire genome, dramatically increasing the power of genetic association analysis (Kruglyak, *Nature Genet.* 22:139 (1999)). A major limitation to the implementation of this method has been lack of knowledge of the haplotype structure of the human genome, which is required in order to select the appropriate genetic variants for analysis. The present invention demonstrates that high-density oligonucleotide arrays in combination with somatic cell genetic sample preparation provide a high-resolution approach to empirically define the common haplotype structure of the human genome.

Although the length of genomic regions with a simple haplotype structure is extremely variable, a dense set of common SNPs enables the systematic approach to define blocks of the human genome in which 80% of the global human population is described by only three common haplotypes. In general, when applying the particular algorithm used in this embodiment, the most common haplotype in any block is found in 50% of individuals, the second most common in 25% of individuals, and the third most common in 12.5% of individuals. It is important to note that blocks are defined based on their genetic information content and not on knowledge of how this information originated or why it exists. As such, blocks do not have absolute boundaries, and may be defined in different ways, depending on the specific application. The algorithm in this embodiment provides only one of many possible approaches. The results indicate that a very dense set of SNPs is required to capture all the common haplotype information. Once in hand, however, this information can be used to identify much smaller subsets of SNPs useful for comprehensive whole-genome association studies.

Those skilled in the art will appreciate readily that the techniques applied to human chromosome 21 can be applied to all the chromosomes present in the human

genome. In a preferred embodiment of the present invention, multiple whole genomes of a diverse population representative of the human species are used to identify SNP haplotype blocks common to all or most members of the species. In some embodiments, SNP haplotype blocks are based on ancient SNPs by excluding SNPs that are represented at low frequency. The ancient SNPs are likely to be important as they have been preserved in the genome because they impart some selective benefit to organisms carrying them.

Example 6: Using Associated Genes for Gene Therapy and Drug Discovery

One example for using the methods of the present invention is outlined in this prophetic example. SNP discovery is performed on twenty haploid genomes, and fifty haploid genomes are analyzed by the methods of the present invention to determine SNP haplotype blocks, SNP haplotype patterns, informative SNPs and minor allele frequency for each informative SNP. These fifty haploid genomes comprise the control genomes of the present study (see step 1300 of Figure 13).

Next, genomic DNA from 500 individuals having an obesity phenotype are assayed for variants by using long distance PCR and microarrays as described *supra* (see also, United States Patent No. 6,300,063 issued to Lipshutz, *et al.*, and United States Patent No. 5,837,832 to Chee, *et al.*), and the frequency of the minor allele for each informative SNP is determined for this clinical population (see step 1310 of Figure 13). The minor allele frequencies of the informative SNPs for the two populations are compared, and the control and clinical populations are determined to have statistically significant differences in three informative SNP locations (steps 1320 and 1330). The SNP location with the largest difference in the minor allele frequency between the control and clinical populations is selected for analysis.

The informative location selected is contained within a SNP haplotype block that is found to span 1 kb of noncoding sequence 5' of the coding region and 4 kb of the coding region of the leptin gene (step 1340). Analysis of the variations contained within this region indicates that a G at one SNP position in this region is responsible for destruction of the promoter for the leptin gene, with a commensurate lack of expression of the leptin protein.

Fibroblasts are obtained from a subject by skin biopsy. The resulting tissue is placed in tissue-culture medium and separated into small pieces. Small pieces of the tissue are placed on the bottom of a wet surface of a tissue culture flask with medium. After 24 hours at room temperature, fresh media is added (*e.g.*, Ham's F12 media, with 10% FBS, penicillin and streptomycin). The tissue is then incubated at 37°C for approximately one week. At this time, fresh media is added and subsequently changed every several days. After an additional two weeks in culture, a monolayer of fibroblasts emerges. The monolayer is trypsinized and scaled into larger flasks.

The vector derived from the Moloney murine leukemia virus, which contains a kanamycin resistance gene, is digested with restriction enzymes for cloning a fragment to be expressed. The digested vector is treated with calf intestinal phosphatase to prevent self-ligation. The dephosphorylated, linear vector is fractionated on an agarose gel and purified. Leptin cDNA, capable of expressing active leptin protein product, is isolated. The ends of the fragment are modified, if necessary, for cloning into the vector. Equal molar quantities of the Moloney murine leukemia virus linear backbone and the leptin gene fragment are mixed together and joined using T4 DNA ligase. The ligation mixture is used to transform *E. coli* and the bacteria are then plated onto agar-containing kanamycin. Kanamycin phenotype and restriction analysis confirm that the vector has the properly inserted leptin gene.

Packaging cells are grown in tissue culture to confluent density in Dulbecco's Modified Eagles Medium (DMEM) with 10% calf serum, penicillin and streptomycin. The vector containing the leptin gene is introduced into the packaging cells by standard techniques. Fresh media is added to the packaging cells, and after an appropriate incubation period, media is harvested from the plates of confluent packaging cells. The media, containing the infectious viral particles, is filtered through a Millipore filter to remove detached packaging cells, then is used to infect fibroblast cells. Media is removed from a sub-confluent plate of fibroblasts and quickly replaced with the filtered media. Polybrene (Aldrich) may be included in the media to facilitate transduction. After appropriate incubation, the media is removed and replaced with fresh media. If the titer of virus is high, then virtually all fibroblasts will be infected and no selection is required. If the titer is low, then it is necessary to use a retroviral vector that has a selectable marker, such as *neo* or *his*, to select out transduced cells for expansion.

Engineered fibroblasts then are introduced into individuals, either alone or after having been grown to confluence on microcarrier beads, such as cytodex 3 beads. The injected fibroblasts produce leptin product, and the biological actions of the protein are conveyed to the host.

Alternatively or in addition, the leptin gene is isolated, cloned into an expression vector and employed for producing leptin polypeptides. The expression vector contains suitable transcriptional and translational initiation regions, and transcriptional and translational termination regions, as disclosed *supra*. Isolated leptin protein can be produced in this manner and used to identify agents which bind it; alternatively cells expressing the engineered leptin gene and protein are used in assays to identify agents. Such agents are identified by, for example, contacting a candidate agent with an isolated leptin polypeptide for a time sufficient to form a polypeptide/compound complex, and detecting the complex. If a polypeptide/compound complex is detected, the compound that binds to the leptin polypeptide is identified. Agents identified via this method can include compounds that modulate activity of leptin. Agents screened in this manner are peptides, carbohydrates, vitamin derivatives, and other small molecules or pharmaceutical agents. In addition to biological assays to identify agents, agents may be pre-screened by choosing candidate agents selected by using protein modeling techniques, based on the configuration of the leptin protein.

In addition to identifying agents that bind the leptin protein, sequence-specific or element-specific agents that control gene expression through binding to the leptin gene are also identified. One class of nucleic acid binding agents are agents that contain base residues that hybridize to leptin mRNA to block translation (e.g., antisense oligonucleotides). Another class of nucleic acid binding agents are those that form a triple helix with DNA to block transcription (triplex oligonucleotides). Such agents usually contain 20 to 40 bases, are based on the classic phosphodiester, ribonucleic acid backbone, or can be a variety of sulfhydryl or polymeric derivatives that have base attachment capacity.

Additionally, allele-specific oligonucleotides that hybridize specifically to the leptin gene and/or agents that bind specifically to the variant leptin protein (e.g., a variant-specific antibody) can be used as diagnostic agents. Methods for preparing and

using allele-specific oligonucleotides and for preparing antibodies are described *supra* and are known in the art.

All patents and publications mentioned in this specification are indicative of the levels of those skilled in the art to which the invention pertains. All patents and publications are herein incorporated by reference to the same extent as if each individual publication was specifically and individually indicated to be incorporated by reference.

The present invention provides greatly improved methods for conducting genome-wide association studies by identifying individual variations, determining SNP haplotype blocks, determining haplotype patterns and, further, using the SNP haplotype patterns to identify informative SNPs. The informative SNPs may be used to dissect the genetic bases of disease and drug response in a practical and cost effective manner unknown previously. It is to be understood that the above description is intended to be illustrative and not restrictive. Many embodiments will be apparent to those skilled in the art upon reviewing the above description. The scope of the invention should, therefore, be determined not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled.

SEQUENCE LISTING

<110> Perlegen Sciences, Inc.
PATIL, Nila
COX, David R.
BERNO, Anthony J.
HINDS, David A.
FODOR, Stephen P. A.

<120> Methods for Genomic Analysis

<130> 054801-5001

<150> US 60/280,530
<151> 2001-03-30

<150> US 60/313,264
<151> 2001-08-17

<150> US 60/327,006
<151> 2001-10-05

<150> US 60/332,550
<151> 2001-11-26

<160> 7

<170> PatentIn version 3.1

<210> 1
<211> 13
<212> DNA
<213> Artificial sequence

<220>
<223> Sample SNP Haplotype: W

<400> 1
agattcgata acg 13

<210> 2
<211> 13
<212> DNA
<213> Artificial sequence

<220>
<223> Sample SNP Haplotype: X

<400> 2
agactacata acg 13

<210> 3
<211> 13
<212> DNA
<213> Artificial sequence

<220>
<223> Sample SNP Haplotype: Y

<400> 3
tatttcgata acg 13

<210> 4
<211> 13
<212> DNA
<213> Artificial sequence

<220>
<223> Sample SNP Haplotype: Z

<400> 4
tatctacaat cac 13

<210> 5
<211> 13
<212> DNA
<213> Artificial sequence

<220>
<223> SNP sequence

<400> 5
agtaacccct ttt 13

<210> 6
<211> 13
<212> DNA
<213> Artificial sequence

<220>
<223> SNP sequence

<400> 6
actgaccccct ttt 13

<210> 7
<211> 13
<212> DNA
<213> Artificial sequence

<220>
<223> SNP sequence

<400> 7
agtgactctt taa 13

4 Brief Description of Drawings

The following figures and drawings form part of the present specification and are included to further demonstrate certain aspects of the patent invention. The invention may be better understood by reference to one or more of these drawings in combination with the detailed description of the specific embodiments presented herein.

Figure 1 is a schematic of one embodiment of the methods of the present invention from identifying variant locations to associating variants with phenotype, to using the associations to identify drug discovery targets or as diagnostic markers.

Figure 2 shows sample SNP haplotype blocks and SNP haplotype patterns according to the present invention.

Figure 3 is a schematic showing one embodiment of a method for selecting SNP haplotype blocks.

Figure 4 illustrates a simple employment of one embodiment of the method shown in Figure 3.

Figure 5A is a schematic of one embodiment of a method for choosing a final set of SNP haplotype blocks. Figure 5B is a simple employment of the method shown in Figure 5A. The "letter:number" designations in Figure 5B indicate "haplotype block ID:informativeness value" for each block.

Figure 6 shows an example of how informative SNPs may be selected according to one embodiment of the present invention.

Figure 7A is a schematic showing one embodiment for resolving variant ambiguities and/or SNP haplotype pattern ambiguities. Figure 7B illustrates a simple employment of the method shown in Figure 7A.

Figure 8 is a schematic of one embodiment of using the methods of the present invention in an association study.

Figure 9 shows an exemplary computer network system suitable for executing some embodiments of the present invention.

Figure 10 is a schematic of the construction of somatic cell hybrids.

Figure 11 is a table illustrating a portion of results obtained from screening hamster-human cell hybrids with the HuSNP genechip from Affymetrix, Inc.

Figure 12 shows an example of various amplified genomic regions of human chromosome 22 and human chromosome 14 genomic DNA using long range PCR.

Figure 13A is a bar graph showing the percentage of SNPs plotted against the frequency of the minor allele (variant) of the SNP. Figure 13B is a graph of the percentage of 200kb intervals as a function of the nucleotide diversity in the interval. Figure 13C is a bar graph showing the percentage of all intervals plotted against interval length.

Figure 14 shows the haplotype patterns for twenty independent globally diverse chromosomes defined by 147 common human chromosome 21 SNPs.

Figure 15 is a plot of the fraction of chromosome covered as a function of the number of SNPs required for that coverage.

The present invention relates to methods for identifying variations that occur in the human genome and relating these variations to the genetic basis of disease and drug response. In particular, the present invention relates to identifying individual SNPs, determining SNP haplotype blocks and patterns, and, further, using the SNP haplotype blocks and patterns to dissect the genetic bases of disease and drug response. The methods of the present invention are useful in whole genome analysis.

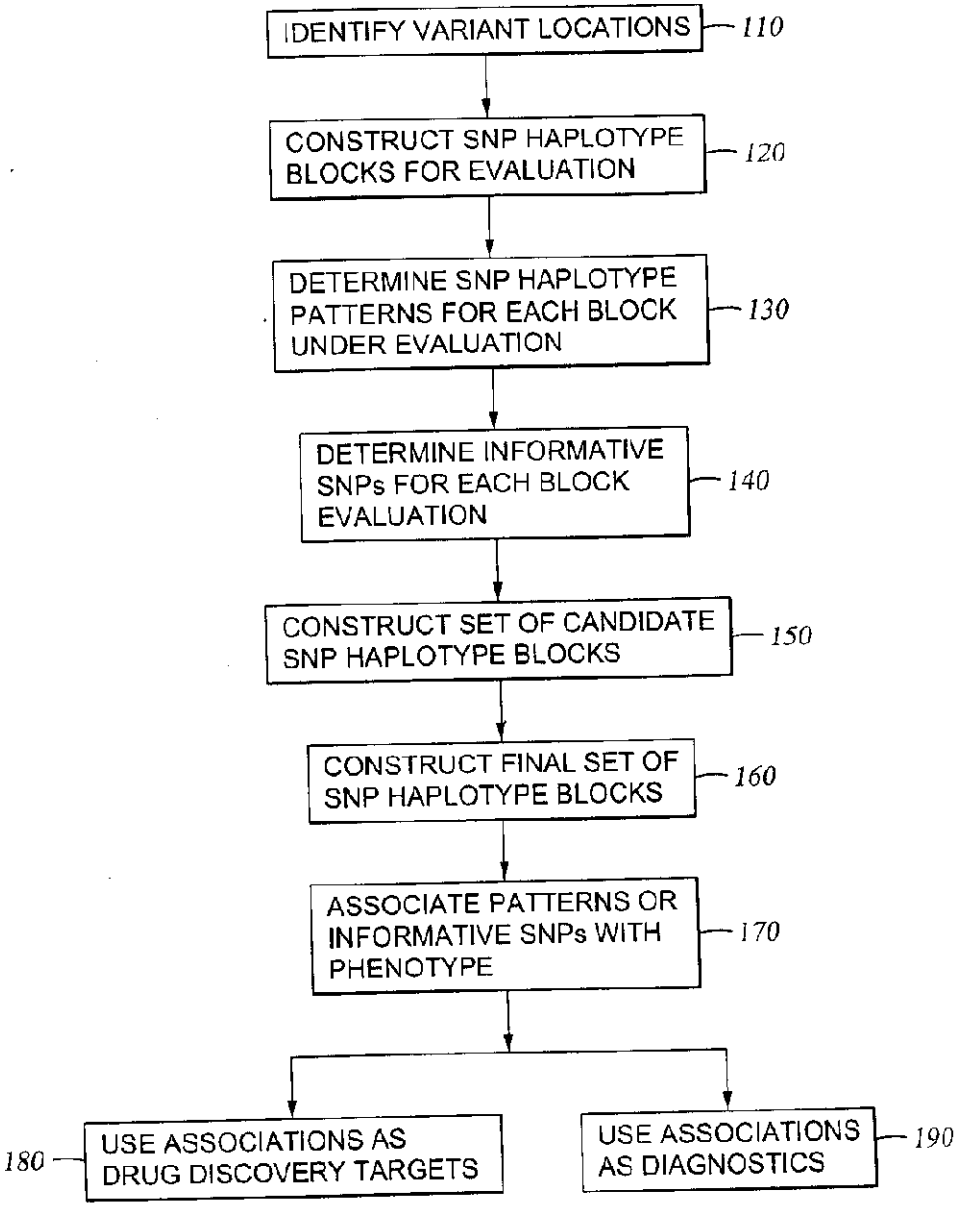

IDENTIFY VARIANT LOCATIONS
110
CONSTRUCT SNP HAPLOTYPE BLOCKS FOR EVALUATION
120
DETERMINE SNP HAPLOTYPE PATTERNS FOR EACH BLOCK UNDER EVALUATION
130
DETERMINE INFORMATIVE SNPs FOR EACH BLOCK EVALUATION
140
CONSTRUCT SET OF CANDIDATE SNP HAPLOTYPE BLOCKS
150
CONSTRUCT FINAL SET OF SNP HAPLOTYPE BLOCKS
160
ASSOCIATE PATTERNS OR INFORMATIVE SNPs WITH PHENOTYPE
170
180
USE ASSOCIATIONS AS DRUG DISCOVERY TARGETS
USE ASSOCIATIONS AS DIAGNOSTICS
190

Fig. 1

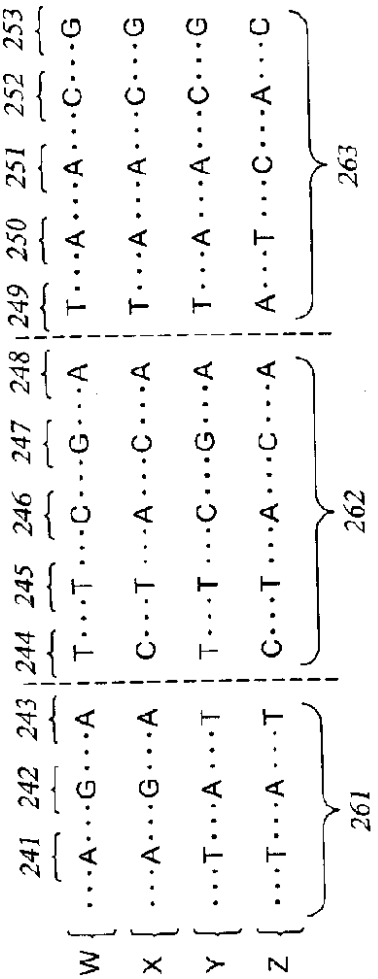

241 242 243 244 245 246 247 248 249 250 251 252 253
W ...A...G...A T...T...C...G...A T...A...A...C...G
X ...A...G...A C...T...A...C...A T...A...A...C...G
Y ...T...A...T T...T...C...G...A T...A...A...C...G
Z ...T...A...T C...T...A...C...A A...T...C...A...C
261
262
263

Fig. 2

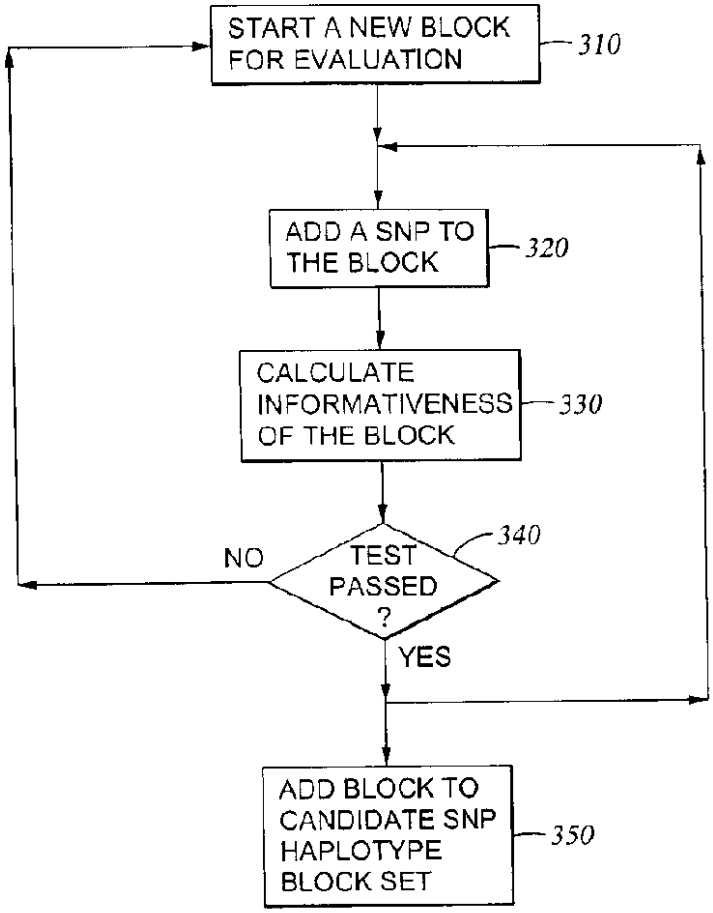

START A NEW BLOCK FOR EVALUATION
310
ADD A SNP TO THE BLOCK
320
CALCULATE INFORMATIVENESS OF THE BLOCK
330
TEST PASSED ?
340
NO
YES
ADD BLOCK TO CANDIDATE SNP HAPLOTYPE BLOCK SET
350

Fig. 3

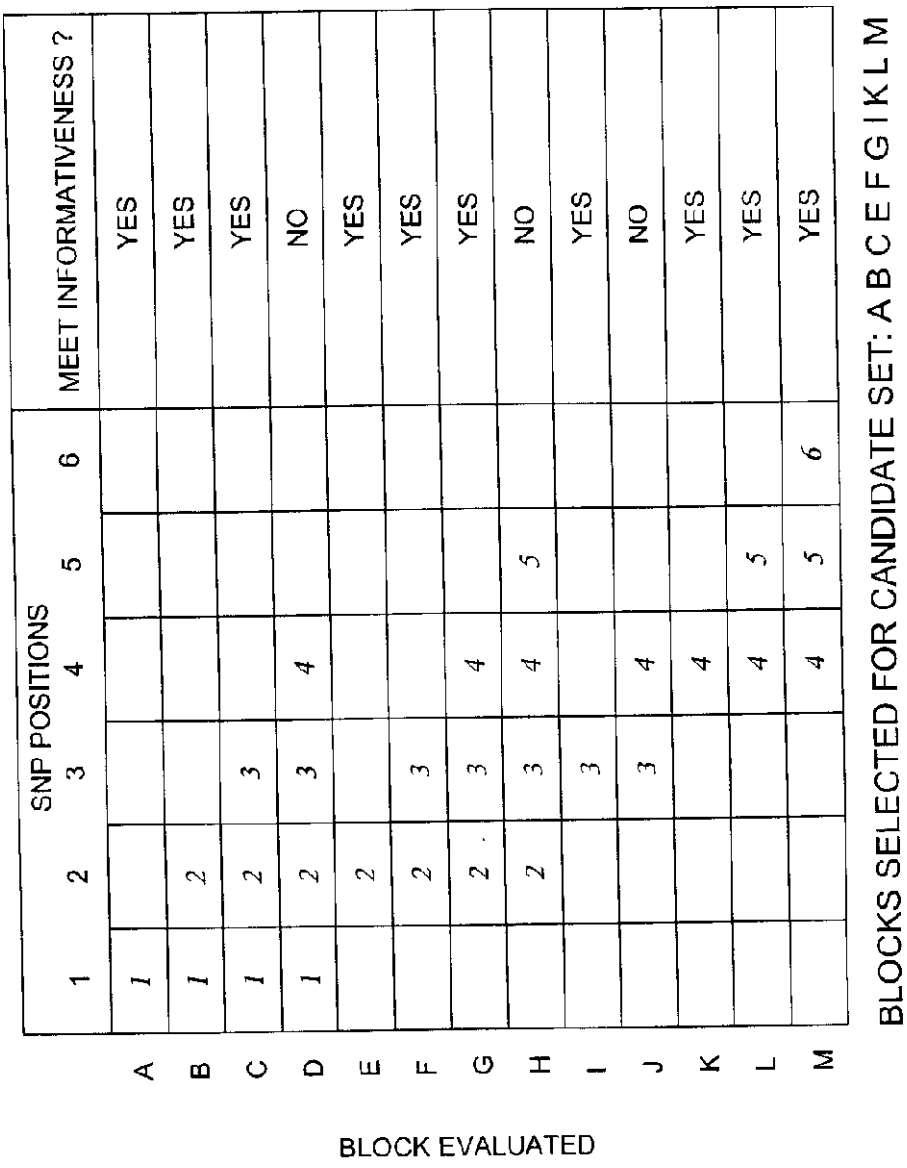

SNP POSITIONS
1 2 3 4 5 6
MEET INFORMATIVENESS ?
A 1 YES
B 1 2 YES
C 1 2 3 YES
D 1 2 3 4 NO
E 2 YES
F 2 3 YES
G 2 3 4 YES
H 2 3 4 5 NO
I 3 YES
J 3 4 NO
K 4 YES
L 4 5 YES
M 4 5 6 YES
BLOCK EVALUATED
BLOCKS SELECTED FOR CANDIDATE SET: A B C E F G I K L M

Fig. 4

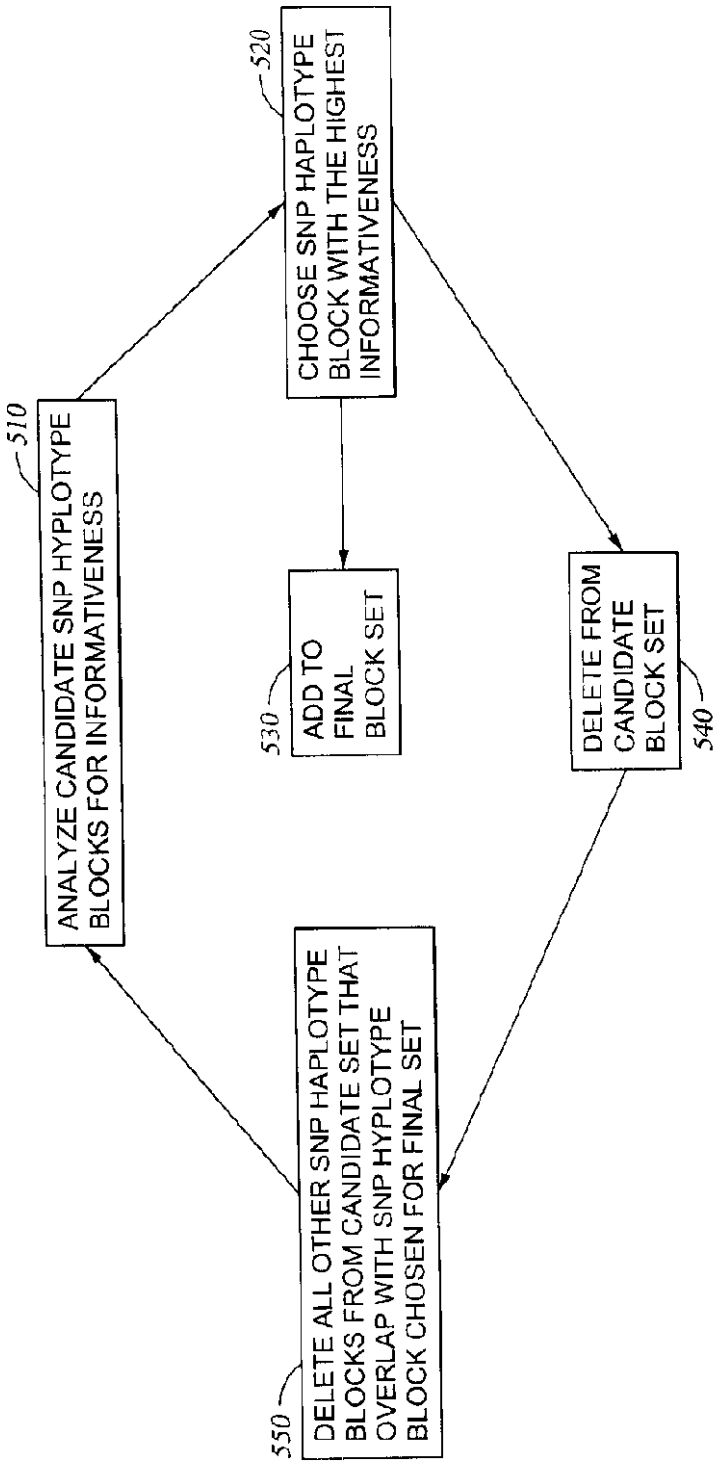

510
ANALYZE CANDIDATE SNP HYPLOTYPE BLOCKS FOR INFORMATIVENESS
520
CHOOSE SNP HAPLOTYPE BLOCK WITH THE HIGHEST INFORMATIVENESS
530
ADD TO FINAL BLOCK SET
540
DELETE FROM CANDIDATE BLOCK SET
550
DELETE ALL OTHER SNP HAPLOTYPE BLOCKS FROM CANDIDATE SET THAT OVERLAP WITH SNP HYPLOTYPE BLOCK CHOSEN FOR FINAL SET

Fig. 5A

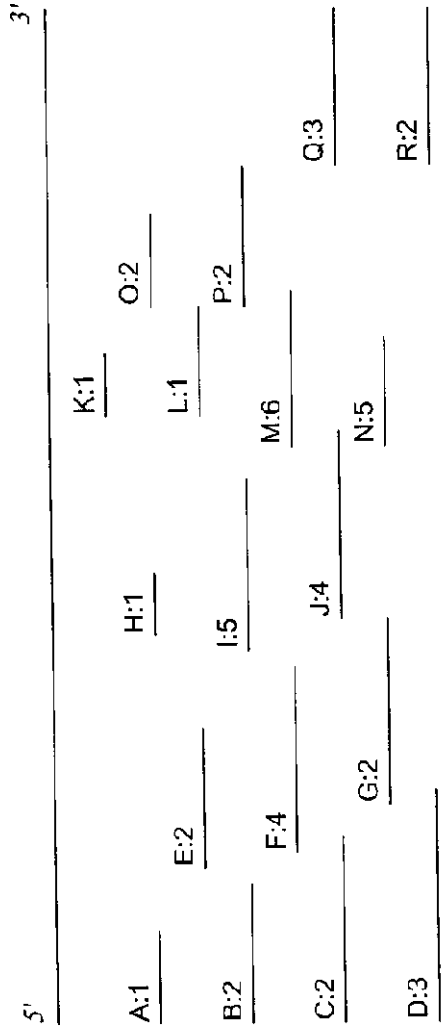

3'
5'
A:1
E:2
B:2
F:4
C:2
G:2
D:3
H:1
I:5
J:4
K:1
L:1
M:6
N:5
O:2
P:2
Q:3
R:2
M DISCARD J,N,K,L,O & P
I DISCARD H
F DISCARD E,G,C & D
Q DISCARD R
B DISCARD A

Fig. 5B

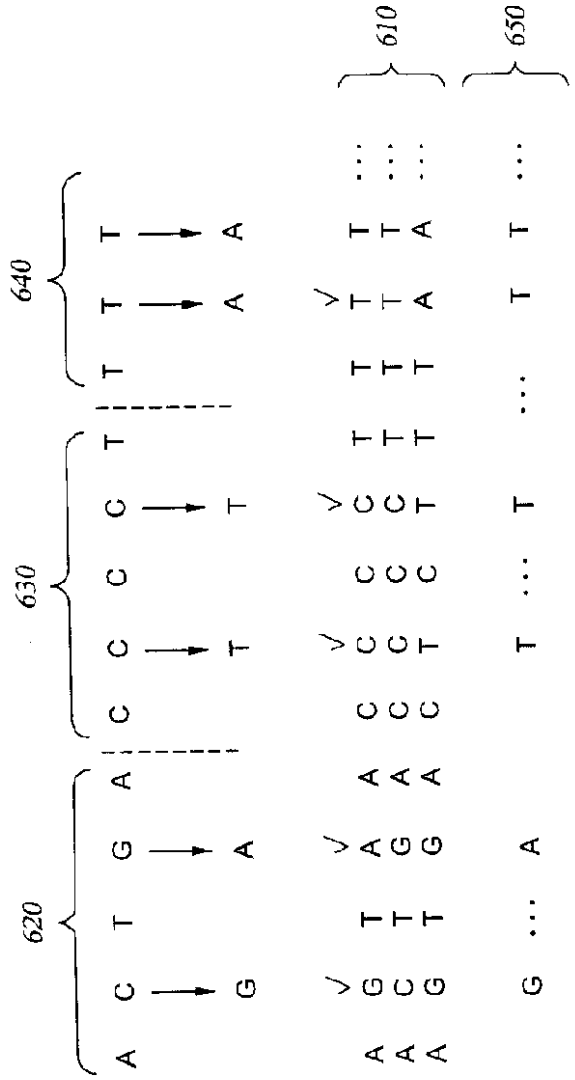


Fig. 6

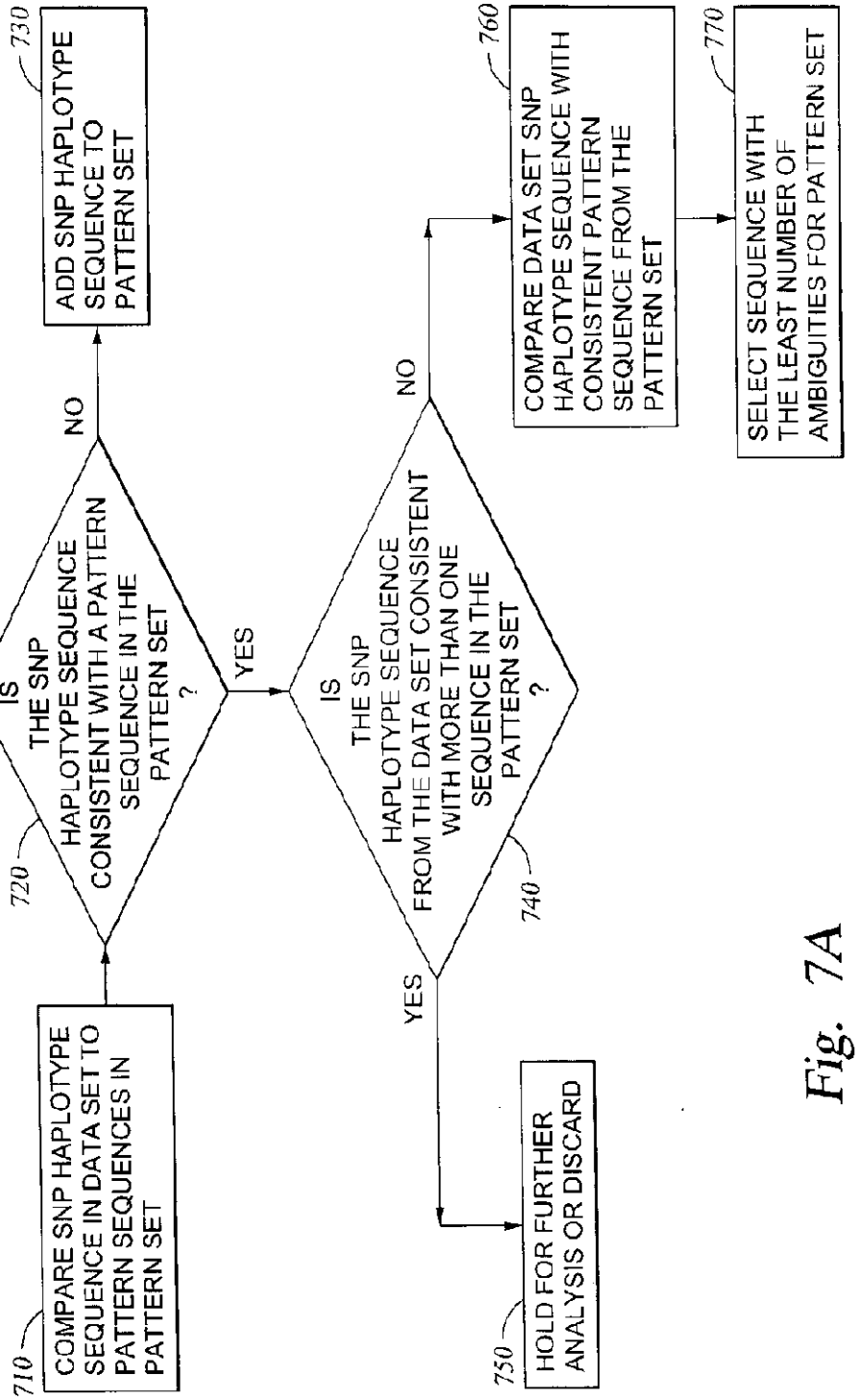

710
COMPARE SNP HAPLOTYPE SEQUENCE IN DATA SET TO PATTERN SEQUENCES IN PATTERN SET
720
IS THE SNP HAPLOTYPE SEQUENCE CONSISTENT WITH A PATTERN SEQUENCE IN THE PATTERN SET ?
NO
730
ADD SNP HAPLOTYPE SEQUENCE TO PATTERN SET
YES
740
IS THE SNP HAPLOTYPE SEQUENCE FROM THE DATA SET CONSISTENT WITH MORE THAN ONE SEQUENCE IN THE PATTERN SET ?
YES
750
HOLD FOR FURTHER ANALYSIS OR DISCARD
NO
760
COMPARE DATA SET SNP HAPLOTYPE SEQUENCE WITH CONSISTENT PATTERN SEQUENCE FROM THE PATTERN SET
770
SELECT SEQUENCE WITH THE LEAST NUMBER OF AMBIGUITIES FOR PATTERN SET

Fig. 7A

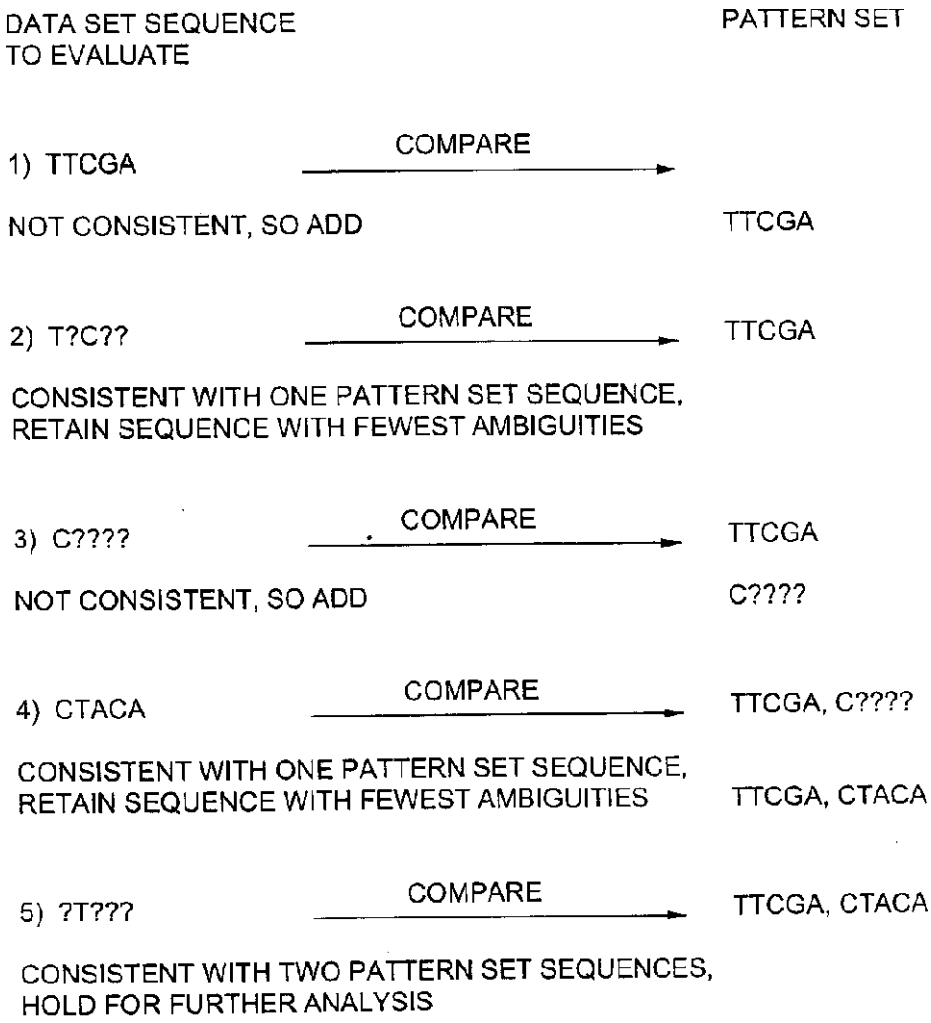

DATA SET SEQUENCE
TO EVALUATE
PATTERN SET
1) TTCGA
COMPARE
NOT CONSISTENT, SO ADD
TTCGA
2) T?C??
COMPARE
TTCGA
CONSISTENT WITH ONE PATTERN SET SEQUENCE,
RETAIN SEQUENCE WITH FEWEST AMBIGUITIES
3) C????
COMPARE
TTCGA
NOT CONSISTENT, SO ADD
C????
4) CTACA
COMPARE
TTCGA, C????
CONSISTENT WITH ONE PATTERN SET SEQUENCE,
RETAIN SEQUENCE WITH FEWEST AMBIGUITIES
TTCGA, CTACA
5) ?T???
COMPARE
TTCGA, CTACA
CONSISTENT WITH TWO PATTERN SET SEQUENCES,
HOLD FOR FURTHER ANALYSIS

Fig. 7B

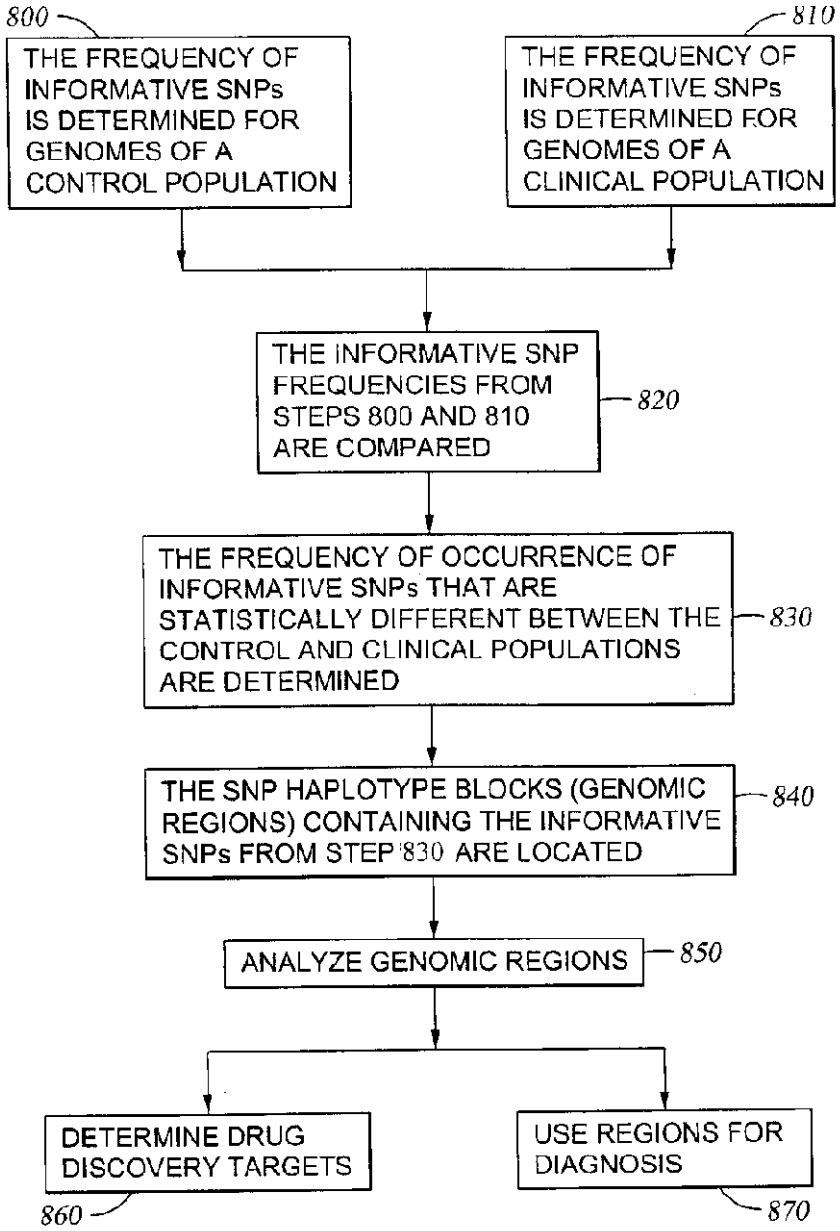
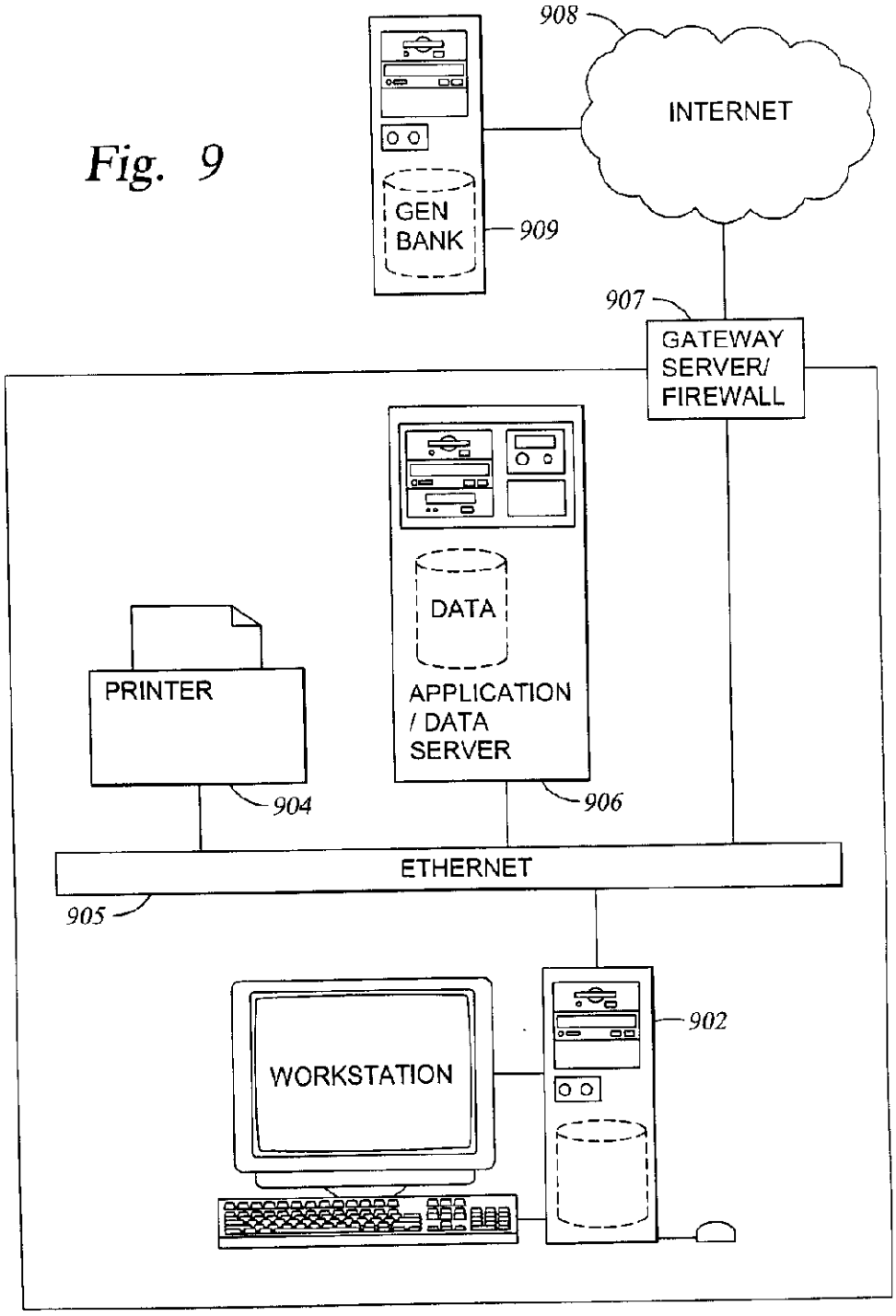

800
THE FREQUENCY OF INFORMATIVE SNPs IS DETERMINED FOR GENOMES OF A CONTROL POPULATION
810
THE FREQUENCY OF INFORMATIVE SNPs IS DETERMINED FOR GENOMES OF A CLINICAL POPULATION
THE INFORMATIVE SNP FREQUENCIES FROM STEPS 800 AND 810 ARE COMPARED
820
THE FREQUENCY OF OCCURRENCE OF INFORMATIVE SNPs THAT ARE STATISTICALLY DIFFERENT BETWEEN THE CONTROL AND CLINICAL POPULATIONS ARE DETERMINED
830
THE SNP HAPLOTYPE BLOCKS (GENOMIC REGIONS) CONTAINING THE INFORMATIVE SNPs FROM STEP 830 ARE LOCATED
840
ANALYZE GENOMIC REGIONS
850
DETERMINE DRUG DISCOVERY TARGETS
860
USE REGIONS FOR DIAGNOSIS
870

Fig. 8

Fig. 9


908
INTERNET
GEN
BANK
909
907
GATEWAY
SERVER/
FIREWALL
DATA
APPLICATION
/ DATA
SERVER
906
PRINTER
904
ETHERNET
905
WORKSTATION
902

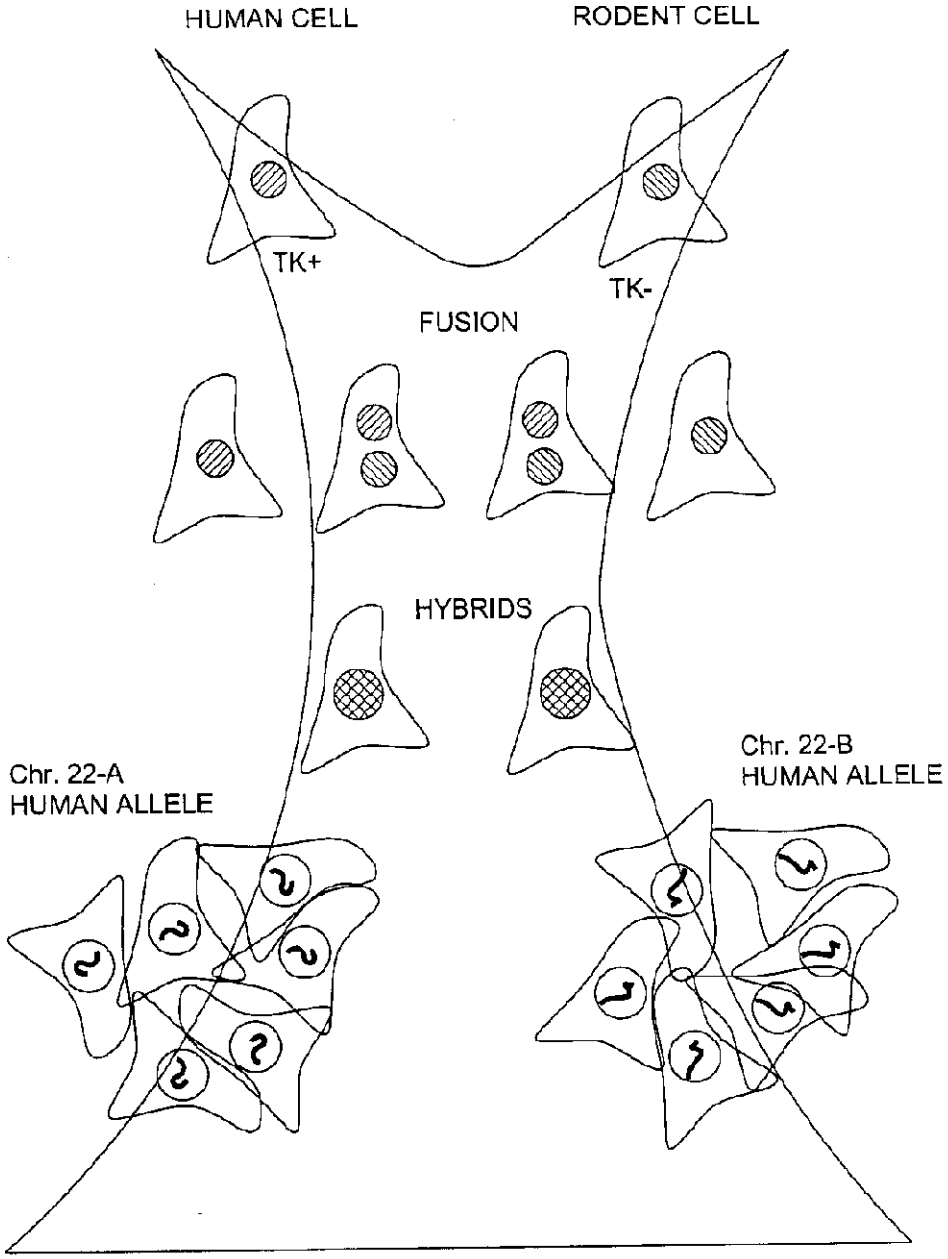

HUMAN CELL
RODENT CELL
TK+
TK-
FUSION
HYBRIDS
Chr. 22-A
HUMAN ALLELE
Chr. 22-B
HUMAN ALLELE

Fig. 10

Chr21 HuSNP MARKERS	HAMSTER	CPD17	HYBRID 1	HYBRID 2
WIAF-3497	NO SIGNAL	A	A	A
WIAF-3498	NO SIGNAL	AB	A	B
WIAF-599	NO SIGNAL	A	A	A
WIAF-3562	NO SIGNAL	NO SIGNAL	A	B
WIAF-559	NO SIGNAL	AB	B	A
WIAF-4546	NO SIGNAL	AB	B	A
WIAF-3508	NO SIGNAL	B	B	B
WIAF-624	NO SIGNAL	B	B	B
WIAF-1500	NO SIGNAL	A	A	A
WIAF-3496	NO SIGNAL	AB	A	B
WIAF-1943	NO SIGNAL	A	A	A
WIAF-2477	NO SIGNAL	NO SIGNAL	NO SIGNAL	A
WIAF-1538	NO SIGNAL	B	NO SIGNAL	B
WIAF-3479	NO SIGNAL	A	A	NO SIGNAL
WIAF-2436	NO SIGNAL	A	A	A
WIAF-1857	NO SIGNAL	AB	B	A
WIAF-899	NO SIGNAL	AB	A	B
WIAF-1682	NO SIGNAL	B	B	B
WIAF-2214	NO SIGNAL	AB	A	B
WIAF-2643	NO SIGNAL	NO SIGNAL	A	NO SIGNAL
WIAF-4514	NO SIGNAL	B	B	B

Fig. 11

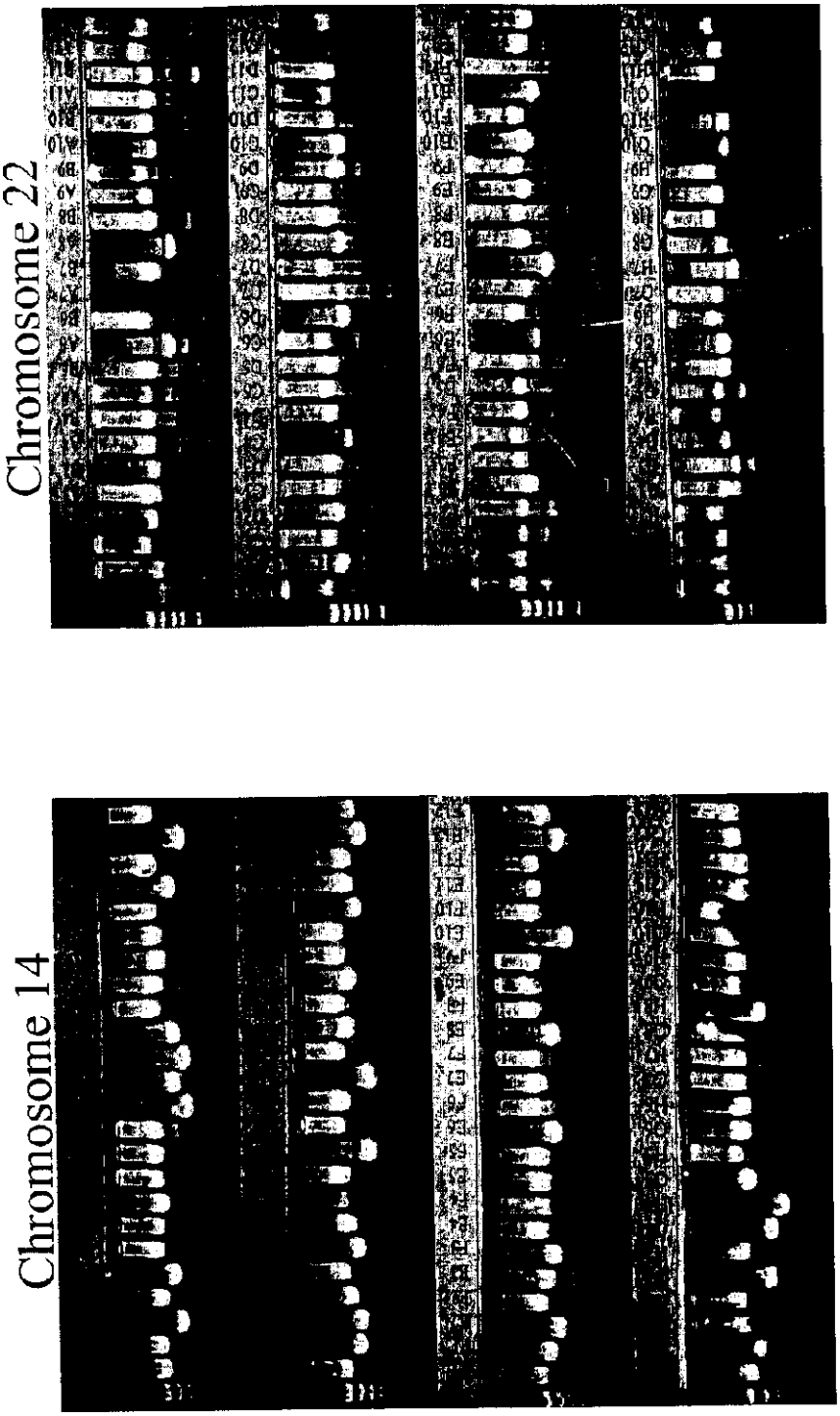


Fig. 12

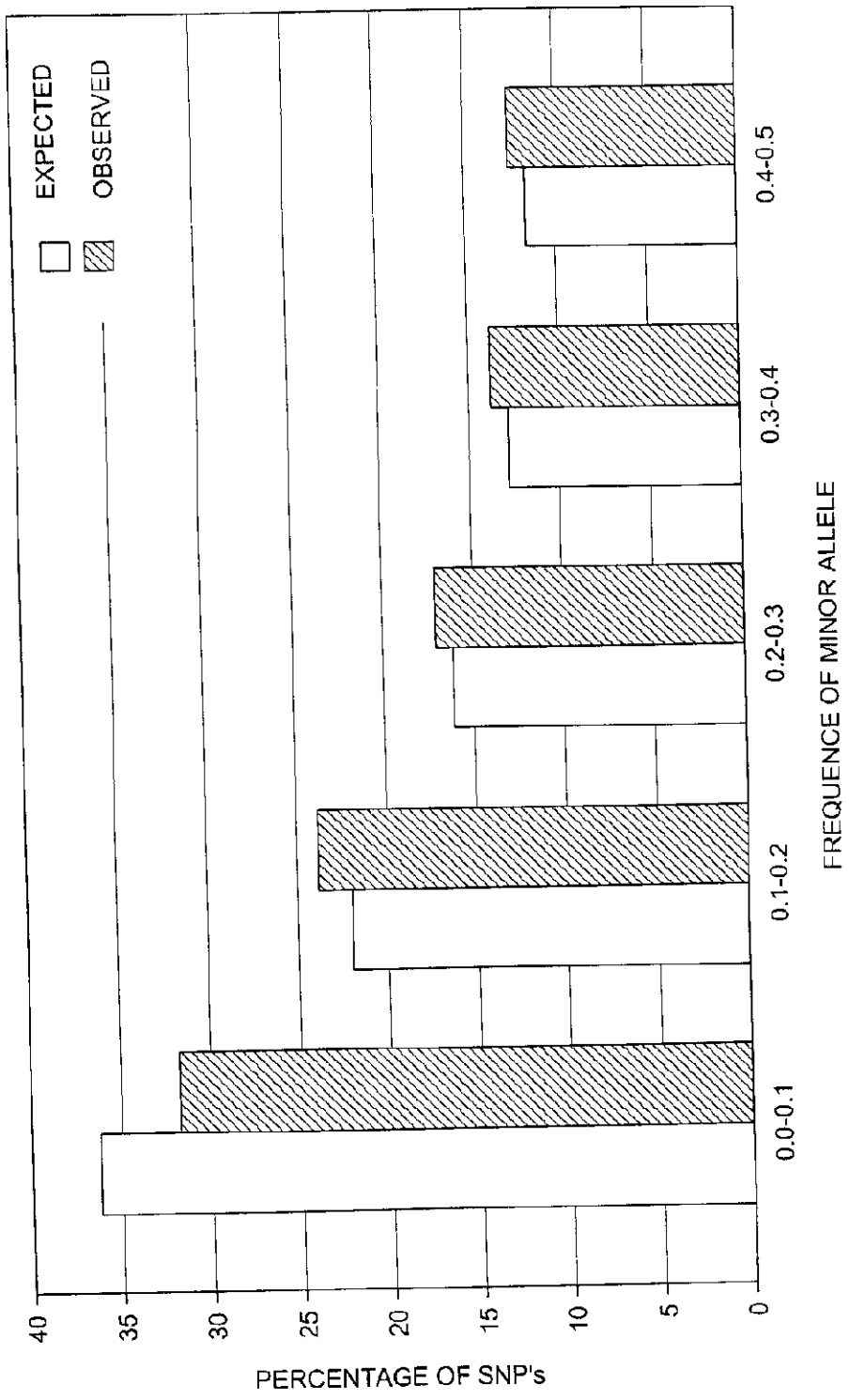

EXPECTED
OBSERVED
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
FREQUENCE OF MINOR ALLELE
0
5
10
15
20
25
30
35
40
PERCENTAGE OF SNP's

Fig. 13A

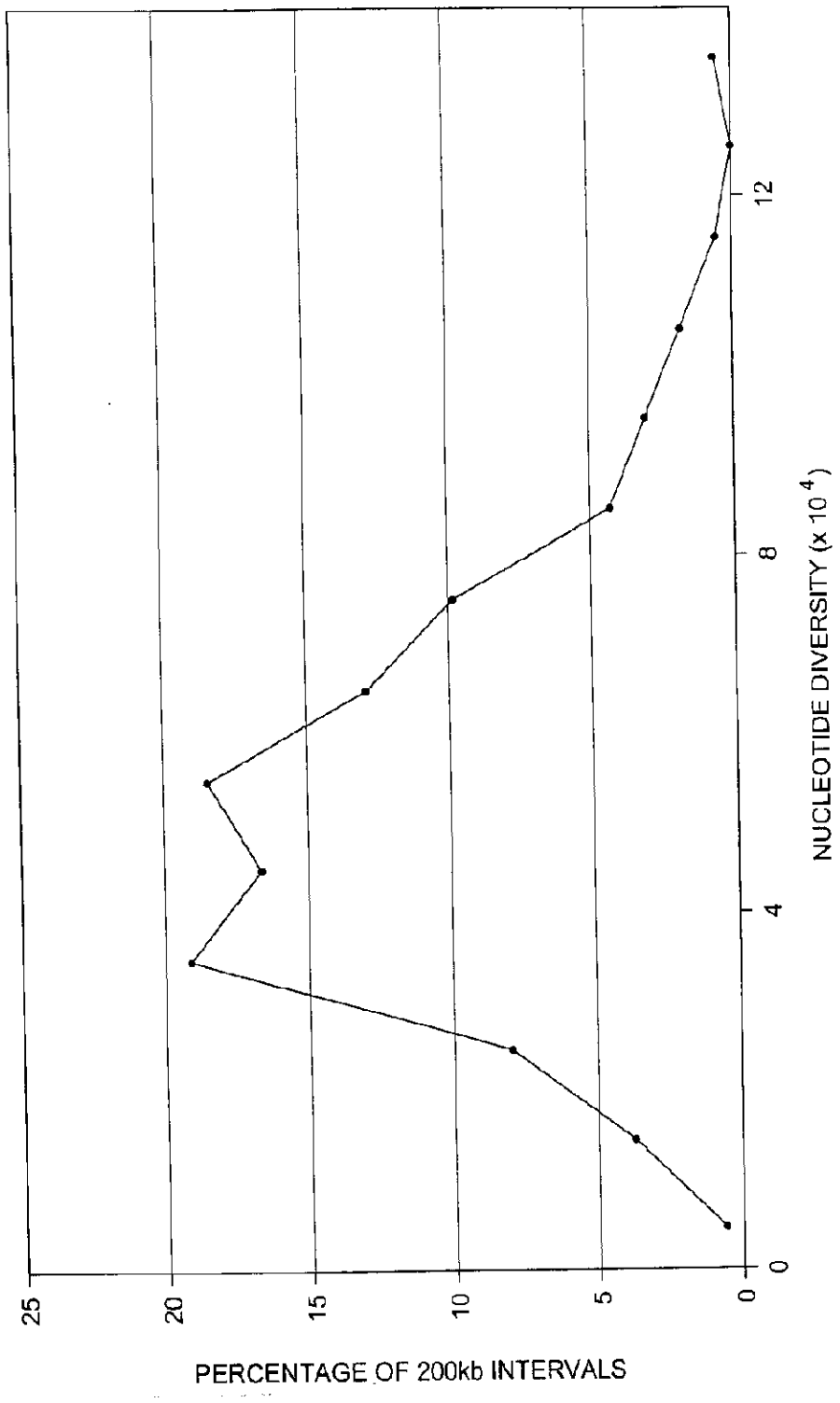

PERCENTAGE OF 200kb INTERVALS
25
20
15
10
5
0
0
4
8
12
NUCLEOTIDE DIVERSITY (x 10 4)

Fig. 13B

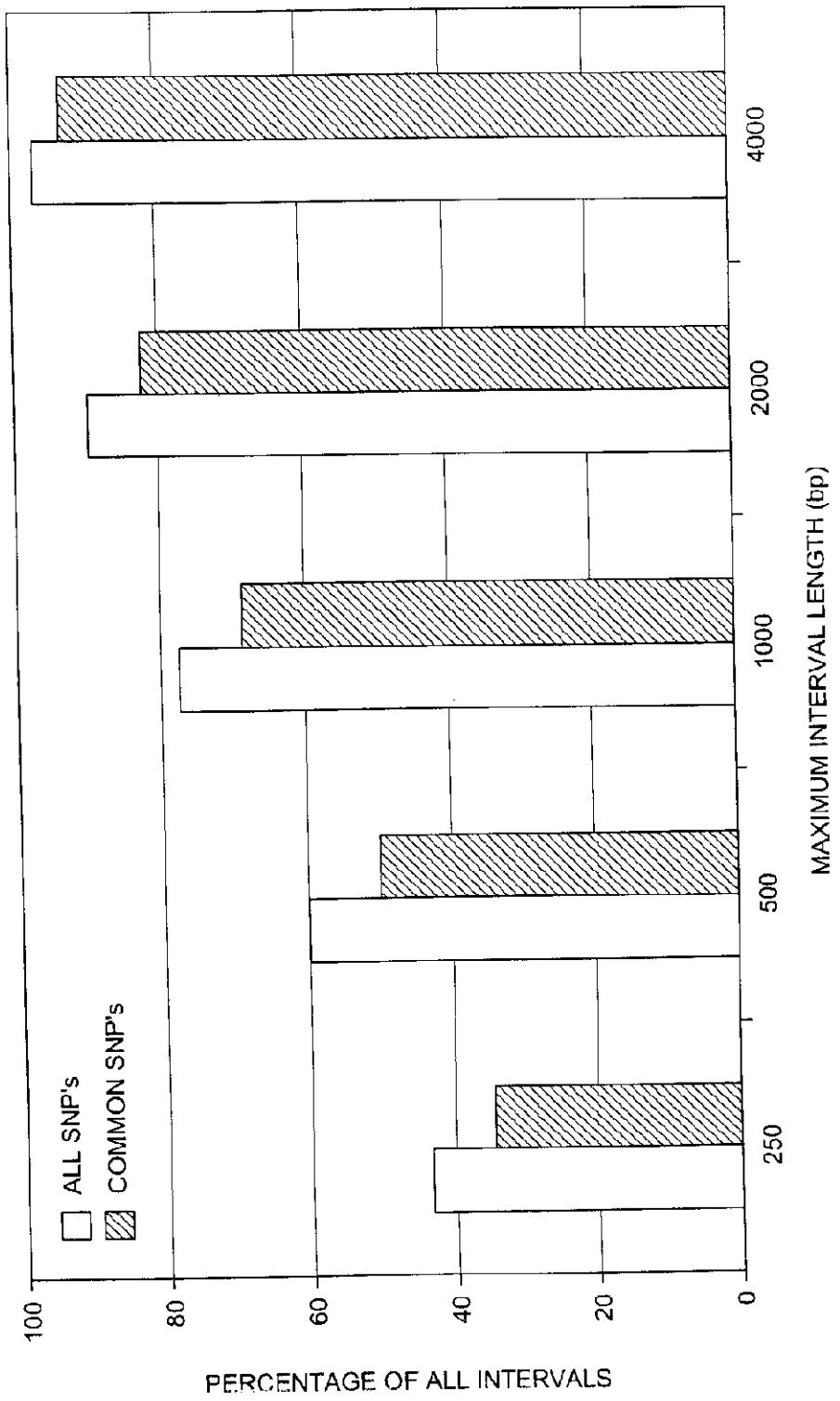

ALL SNP's
COMMON SNP's
100
80
60
40
20
0
PERCENTAGE OF ALL INTERVALS
250
500
1000
2000
4000
MAXIMUM INTERVAL LENGTH (bp)

Fig. 13C

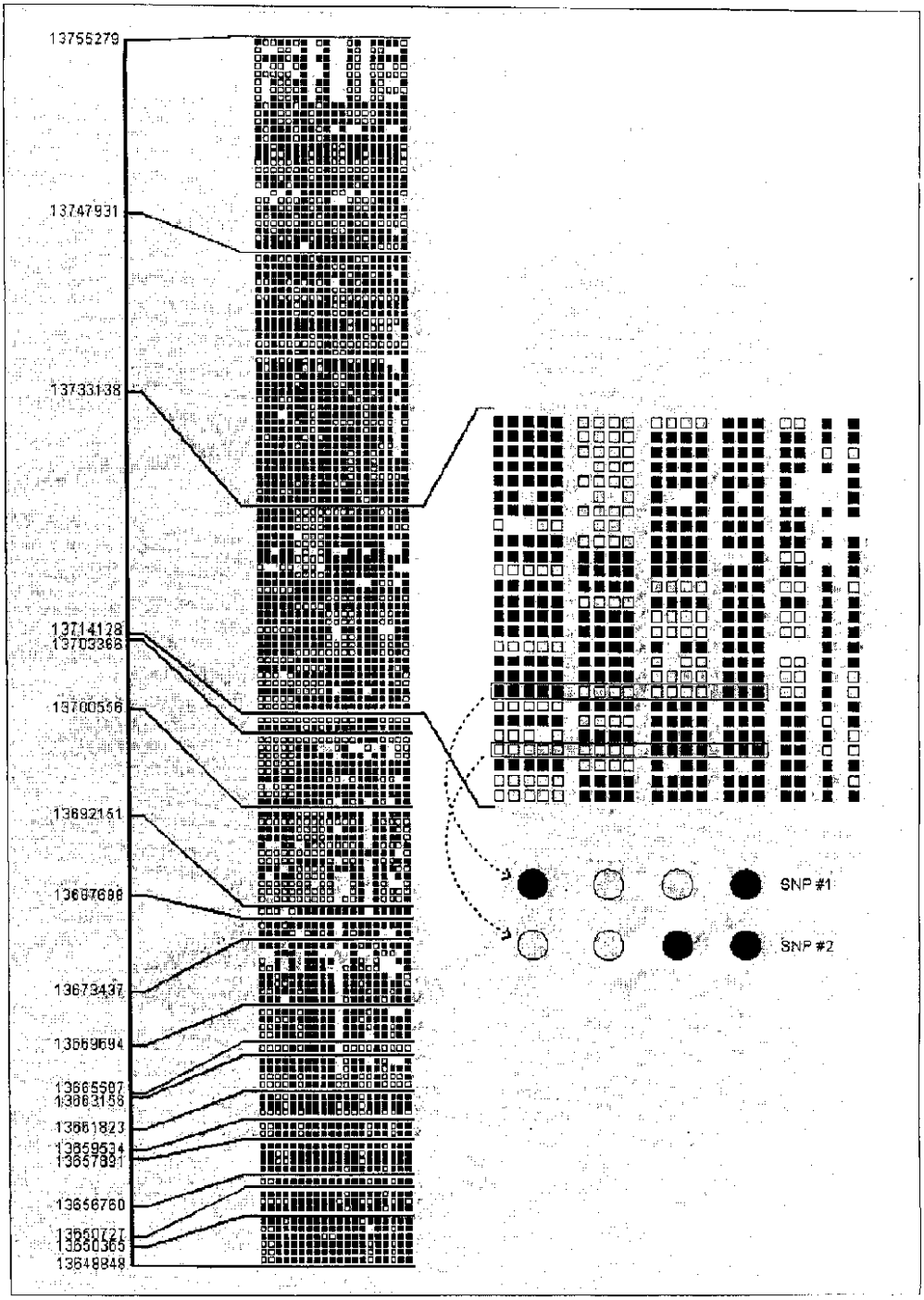

13755279
13747931
13733138
13714128
13703368
13700536
13692151
13687688
13673437
13669694
13665507
13663155
13661823
13659534
13657891
13656760
13650727
13650365
13648848
SNP #1
SNP #2

Figure 14

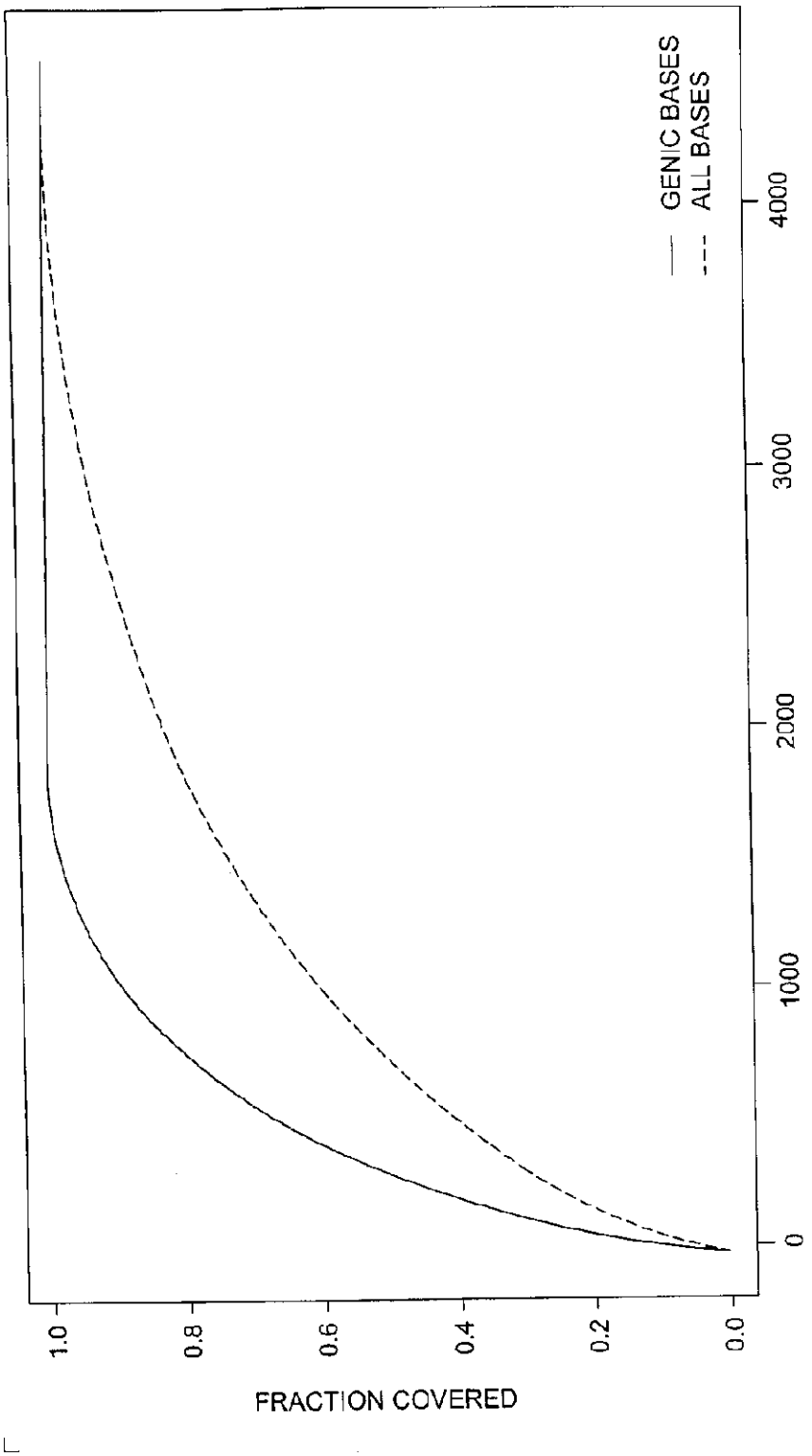

GENIC BASES
ALL BASES
FRACTION COVERED
SNP's REQUIRED
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000

Fig. 15

The present invention relates to methods for identifying variations that occur in the human genome and relating these variations to the genetic basis of disease and drug response. In particular, the present invention relates to identifying individual SNPs, determining SNP haplotype blocks and patterns, and, further, using the SNP haplotype blocks and patterns to dissect the genetic bases of disease and drug response. The methods of the present invention are useful in whole genome analysis.

2 Representative Drawing Fig. 1

专利名称(译)	<无法获取翻译>		
公开(公告)号	JP2003052383A5	公开(公告)日	2008-12-04
申请号	JP2002099196	申请日	2002-04-01
[标]申请(专利权)人(译)	每摄政科学公司		
申请(专利权)人(译)	Parejen科学公司		
[标]发明人	PATIL NILA COX DAVID R BERNO ANTHONY J HINDS DAVID A ニラパティル デヴィッドアールコックス アンソニージェイバーノ デヴィッドエーハインズ		
发明人	ニラ パティル デヴィッド アール. コックス アンソニー ジェイ. バーノ デヴィッド エー. ハインズ		
IPC分类号	C12N15/09 C12Q1/68 G01N33/53 G01N33/566		
CPC分类号	C12Q1/6827 G06F19/24 G06F19/22 G16B30/00 G16B40/00 Y10T436/143333		
FI分类号	C12N15/00.ZNA.A C12Q1/68.A G01N33/53.M G01N33/566		
F-TERM分类号	4B024/AA11 4B024/AA20 4B024/CA01 4B024/HA11 4B024/HA12 4B024/HA17 4B063/QA11 4B063/QQ08 4B063/QQ42 4B063/QR32 4B063/QR55 4B063/QR62 4B063/QR77 4B063/QR82 4B063/QS25 4B063/QS34		
优先权	60/332550 2001-11-26 US 60/327006 2001-10-05 US 60/313264 2001-08-17 US 60/280530 2001-03-30 US		
其他公开文献	JP2003052383A		

摘要(译)

（带更正）要解决的问题：提供一种用于人类基因组分析的方法。 本发明涉及鉴定人类基因组中发生的突变的方法，以及将这些突变与疾病和药物反应的遗传基础相关联的方法。 特别地，本发明涉及鉴定单个SNP，确定SNP单倍型模块和模式，并且进一步使用SNP单倍型模块和模式来解剖疾病和药物反应的遗传基础。 本发明的方法可用于全基因组分析。