【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】平成27年8月20日(2015.8.20)

【公表番号】特表2014-508994(P2014-508994A)
【公表日】平成26年4月10日(2014.4.10)
【年通号数】公開・登録公報2014-018
【出願番号】特願2013-549922(P2013-549922)
【国際特許分類】
Ｇ０６Ｆ　19/28　(2011.01)
Ｃ１２Ｑ　1/68　(2006.01)
【ＦＩ】
Ｇ０６Ｆ　19/28　ＺＮＡ
Ｃ１２Ｑ　1/68　Ｚ

【手続補正書】
【提出日】平成27年7月3日(2015.7.3)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
対象のゲノムデータを処理する方法であって、当該方法は、
（ａ）対象のゲノム配列情報を取得するステップ、
（ｂ）前記ゲノム配列情報の複雑性及び量を低減するステップであり、疾患又は障害に関連するシグネチャーデータ以外の前記ゲノム配列情報を切り取ることを含む、ステップ、並びに
（ｃ）ステップ（ｂ）の前記ゲノム配列情報を迅速に検索可能な形で保存するステップ、
を含む方法。
【請求項２】
請求項１に記載の方法であり、前記ゲノム配列が対象のサンプル、好ましくは、組織、臓器、細胞及び／又はそれらの断片の混合物から、又は膣組織、舌、膵臓、肝臓、脾臓、卵巣、筋肉、関節組織、神経組織、胃腸組織、腫瘍組織からの組織生検などの組織特異的若しくは臓器特異的サンプル、体液、血液、血清、唾液、又は尿から取得される、方法。
【請求項３】
請求項１又は２に記載の方法であり、前記ステップ（ａ）が、対象のゲノム配列の繰り返しの取得を含み、及び最初の取得で得られたゲノム配列情報と２回目以降の取得で得られたゲノム配列情報との間の比較が実施される、方法。
【請求項４】
請求項３に記載の方法であり、追加のステップにおいて、最初に得られたゲノム配列情報及び第２回目以降で得られたゲノム配列情報間で異なる情報を含む増加データが迅速に検索可能な形で保存される、方法。
【請求項５】
請求項１又は２に記載の方法であり、ステップ（ｂ）が、対象のゲノム配列と、疾患又は障害に関連するシグネチャーデータを含む参照配列と整列させることで実施され、及び前記整列が、逆相補的配列を用いて実施される、方法。
【請求項６】
請求項５に記載の方法であり、前記シグネチャーデータが、疾患又は障害に特異的な少

なくとも１つの変異であり、該変異は、ミスセンス変異、ナンセンス変異、一塩基多型（ＳＮＰ）、コピー数多型（ＣＮＶ）、スプライシング変異、制御配列の変異、小欠失、小挿入、小インデル、総欠失、総挿入、複雑な遺伝子再配列、染色体間再配列、染色体内再配列、ヘテロ接合性消失、反復配列の挿入、及び反復配列の欠失を含む群から選択される、方法。

【請求項７】

請求項１乃至６のいずれか一項に記載の方法であり、当該方法がさらに、（ｄ）前記対象の機能的遺伝子情報を取得するステップ、（ｅ）機能的遺伝子情報の複雑性及び／又は量を低減させるステップ、及び、（ｆ）前記機能的遺伝子情報を迅速に検索可能な形で保存するステップを含み、前記機能的遺伝子情報の複雑性及び／又は量を低減させるステップが、疾患又は障害に関連するシグネチャーデータ以外の前記機能的遺伝子情報を切り取ることで実施される、方法。

【請求項８】

請求項７に記載の方法であり、前記機能的遺伝子情報が、

（ｉ）遺伝子発現の情報、好ましくは１以上のＲＮＡ種、１以上のタンパク質種、前記対象のトランスクリプトーム若しくはその一部、前記対象のプロテオーム若しくはその一部、又は、これらの混合物の存在に対する情報；及び／又は、

（ｉｉ）メチル化配列情報、好ましくはそれぞれ個別のヌクレオチド（Ｃ又はＡ）のメチル化配列情報；及び／又は、

（ｉｉｉ）活性遺伝子及び／又はサイレント遺伝子を示すヒストンマーク、好ましくはＨ３Ｋ４メチル化及び／又はＨ３Ｋ２７メチル化を示すヒストンマークの情報、

を含む、方法。

【請求項９】

請求項１又は８に記載の方法であり、ゲノム及び／又は機能的遺伝子情報の変化が行列内にエンコードされ、及び遺伝子の状態、ゲノム領域、制御領域、プロモーター、エクソン又は経路、好ましくは疾患又は障害に関連する情報がデコードされ、マルコフ連鎖処理に基づき表現される、方法。

【請求項１０】

請求項１乃至９に記載の方法により取得及び／又は保存されるゲノム配列情報の、場合により遺伝子発現情報と組み合わせた、

（ｉ）全ゲノム、レギュローム、又は前記ゲノムの制御状態、ゲノム領域、遺伝子、プロモーター、又はイントロン、エクソン、経路、経路成分又は所定の期間にわたるメチル化状態に対する情報を捕捉することで、種々の分子プロファイルモダリティの形で対象の分子履歴を作るための；及び／又は

（ｉｉ）疾患を診断、検出、モニター又は予後判定するための；

使用。

【請求項１１】

請求項５乃至９のいずれか一項に記載の方法又は請求項１０に記載される使用であり、前記疾患が癌性疾患、好ましくは乳癌、卵巣癌又は前立腺癌である、方法又は使用。

【請求項１２】

臨床判断サポート及び保存システムであり：

対象のゲノム配列情報、好ましくは対象の機能的遺伝子情報と組み合わせて提供するための入力装置；

プロセッサーに、請求項１乃至９又は請求項１１のいずれか一項に記載の方法のステップ（ｂ）及び場合によりステップ（ｅ）を実施させることができるコンピュータプログラム；

所定の期間にわたって対象のゲノム変異、増加ゲノム変化又は遺伝子発現変異パターンを出力するための出力装置；及び

前記出力された情報を保存する媒体；

を含むシステム。

【請求項１３】
請求項１２に記載のシステムであり、前記システムが、電子画像／データ保存記録及び通信システムである、システム。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【発明の詳細な説明】
【発明の名称】ゲノムデータ処理方法
【技術分野】
【０００１】
本発明は、対象のゲノムデータを処理する方法に関し、（ａ）対象のゲノム配列を取得し；（ｂ）上記ゲノム配列情報の複雑性及び／又は量を低減させ；及び（ｃ）ステップ（ｂ）の上記ゲノム配列情報を、迅速に検索可能な形で記憶するステップを含む。本発明はさらに、上記ゲノム配列情報の複雑性及び／又は量を低減するステップが、疾患又は障害に関連するシグネチャーデータを除く上記ゲノム配列情報を切り取ることで、又は対象のゲノム配列を、疾患又は障害に関連するシグネチャーデータを含む基準配列と整列させることで実行する、方法に関する。さらに、本発明は、対象の、特に遺伝子発現データでの機能性遺伝子情報の使用が含まれる方法に関し、同様に上記情報がマトリクス中にエンコード及びデコードされ、かつマルコフ連鎖過程に基づき表される、方法に関する。得られる情報はまた、疾患の診断、検出、モニター又は予後判定をするため及び／又は対象の分子履歴を作るために使用され得る。加えて、対応する臨床判断支援及び記憶システムが、好ましくは電子画像／データ保存記録及び通信システムの形で提供される。
【背景技術】
【０００２】
新たな又は次世代の配列決定技術の導入で、配列情報の取得のコスト及びこの情報の提供のために必要な時間は劇的に少なくなっており、将来さらに下がるものと考えられる。従って、全ゲノム配列決定は、現在の生化学的遺伝学的試験及びアッセイに代えて、費用対効果の優れたものとなるであろう。さらに、患者の全ゲノム配列決定は、ひとつの疾患の分析だけでなく、全集団の疾患遺伝子型を評価するために使用され、さらには全ての可能な第２のマーカーの自動的決定により治療見通しを結論することを可能にする。しかしながら、ゲノム配列データは、非常に大量の記憶容量を必要とする巨大なものであり、また、その分析には最高度のコンピュータ装置を必要とする。Ｓｃｈｕｓｔｅｒらは、「Ｎａｔｕｒｅ　４６３（１８）、９４３－９４７、２０１０」で、またＦｕｊｉｍｏｔｏらは、「Ｎａｔｕｒｅ　Ｇｅｎｅｔｉｃｓ、４２、９３１－９３６、２０１０」で、例えば、アフリカからの狩猟採集人、及び日本人個人の完全なゲノムの情報を提供する。これらの分析は、人の集団間での、一塩基多様性の存在、集団間の差について、対立遺伝子頻度同様、大量の新たな情報を提供する。遭遇するゲノム差及び類似性は遺伝子分野での基礎研究において基本的に重要なものとなり得る。しかし、これらは専門家に対しては主要な興味ではなく、専門家は具体的な臨床的質問に関心を持ち、症状又は疑われる疾患に関連する情報に焦点をあてることを望む。この関連で、全ゲノム配列決定の際に得られたゲノム配列データの大部分は、専門家の診断可能性を改善するというよりはむしろ阻害するものであり得る。
【０００３】
従って、利用可能な時間及び資源（リソース）で、患者の遺伝子データ処理を維持することを可能にする要求が存在する。
【先行技術文献】
【非特許文献】
【０００４】

【非特許文献１】Ｓｃｈｕｓｔｅｒら、２０１０、Ｎａｔｕｒｅ　４６３（１８）、９４３－９４７
【非特許文献２】Ｆｕｊｉｍｏｔｏら、２０１０、Ｎａｔｕｒｅ　Ｇｅｎｅｔｉｃｓ、４２、９３１－９３６
【発明の概要】
【発明が解決しようとする課題】
【０００５】
本発明は、この必要性に鑑み、対象のゲノム配列の複雑性及び／又は量を低減し、かつ迅速に検索可能にそれを保存（記憶）する方法を提供する。
【課題を解決するための手段】
【０００６】
上記課題は特に、対象のゲノムデータを処理する方法で達成され、上記方法は：
（ａ）対象のゲノム配列を取得し；
（ｂ）上記ゲノム配列情報の複雑性及び／又は量を低減し；及び
（ｃ）ステップ（ｂ）でのゲノム配列情報を、迅速に検索可能に保存する、ステップを含む。
【０００７】
この方法は、ゲノム情報に、専門家又は医者が集中して処理された形で容易にアクセスすることを可能にし、即ち、上記ゲノム情報を管理可能とし、必要な事実に限定されており、従って、時間及び資源が、非常に大量の元の配列データの処理を維持することを可能にする、という利点を持ち、迅速に検索可能な形で保存されることで、迅速に、いつでもかつどこででも、困難なく独立して利用することを可能とし、これにより例えば問題の臨床環境、移動病院又は患者の側で利用することを可能とする。
【０００８】
本発明の好ましい実施態様では、上記ゲノム配列は患者のサンプルから取得される。
【０００９】
さらに好ましい実施態様では、上記分析サンプルは組織、臓器、細胞の混合物である。上記サンプルはまた、これに代えて組織、臓器又は細胞の断片であり得る。さらなる実施態様では、上記サンプルは組織又は臓器特異的サンプルであり得る。特に好ましくは、サンプルは、膣組織、舌、膵臓、肝臓、脾臓、卵巣、筋肉、関節組織、神経組織、胃腸組織、腫瘍組織、体液、血液、血清、唾液、または尿からの生検サンプルであり得る。
【００１０】
本発明のさらに特に好ましい実施態様では、対象ゲノム配列を得るためのステップは繰り返され、例えばある一定時間後に繰り返される。
【００１１】
本発明のさらに好ましい実施態様では、患者のゲノム配列の取得の繰り返しは、データ追加（増加データ）又は変更を与え、既に得られたゲノム配列情報に比較して上記増加データが保存され、好ましくは迅速に検索可能な形で保存される。
【００１２】
本発明のさらに好ましい実施態様では、上記ゲノム配列情報の複雑性及び／又は量の低減は、上記ゲノム配列情報を切り取ることで実施され得る。かかる切り取り又は低減ステップは、好ましくは、疾患又は障害に関連するシグネチャーデータ以外のゲノム配列の全ての部分で実施される。
【００１３】
本発明のさらなる特に好ましい実施態様では、上記ゲノム配列情報の複雑性及び／又は量の低減は、疾患又は障害に関連するシグネチャーデータを含む参照配列（疾患参照配列）と整列させることで実施され得る。
【００１４】
本発明の他の好ましい実施態様では、上記シグネチャーデータは、ミスセンス変異、ナンセンス変異、一塩基多型（ＳＮＰ）、コピー数多型（ＣＮＶ）、スプライシング変異、

制御配列の変異、小欠失、小挿入、小インデル、総欠失、総挿入、複雑な遺伝子再配列、染色体間再配列、染色体内再配列、ヘテロ接合性消失、反復配列の挿入及び反復配列の欠失を含む群から選択される、疾患又は障害に特異的な少なくとも１つの変異である。

【００１５】

本発明の他の好ましい実施態様では、対象のゲノムデータを処理するための方法がさらに、ステップ（ｄ）を含み、ここで対象の機能性遺伝子情報を得ること、ステップ（ｅ）を含み、ここでこの情報の複雑性及び／又は量を低減させ、かつステップ（ｆ）を含み、ここで上記機能的ゲノム情報が迅速に検索可能に保存する。

【００１６】

本発明の他の特に好ましい実施態様では、上記機能的ゲノム情報が、（ｉ）遺伝子発現の情報、好ましくは、１以上のＲＮＡ種、１以上のタンパク質、上記対象のトランスクリプトーム又はその部分、対象のプロテオーム又はその部分、又はこれらの混合物；及び／又は（ｉｉ）メチル化配列情報、好ましくは、それぞれ個別ヌクレオチド（Ｃ又はＡ）についてのメチル化配列情報；及び／又は、（ｉｉｉ）活性化遺伝子及び／又はサイレント化遺伝子を示すヒストンマーク、好ましくはＨ３Ｋ４メチル化及び／又はＨ３Ｋ２７メチル化のヒストンマークについての情報を含む。

【００１７】

他の好ましい実施態様では、上記情報の複雑性及び／又は量を低減するステップが、上記機能的遺伝子情報を切り取ることで実施される。かかる切り取り又は低減ステップは、好ましくは、疾患又は障害に関連するシグネチャーデータ（疾患参照配列）について以外の機能的ゲノム情報の全ての部分で実施される。

【００１８】

本発明のさらなる実施態様では、ゲノム情報の及び／又は機能的ゲノム情報の変化が行列内でエンコード（符号化）される。なお他の好ましい実施態様では、遺伝子状態、ゲノム領域、調節領域、プロモーター、エクソン又は、特に疾患又は障害に関する経路に関連する、ゲノム情報及び／又は機能的ゲノム情報がデコードされ、マルコフ連鎖過程に基づき表される。特に好ましい実施態様では、上記表現は可視化表現である。

【００１９】

他の側面では、本発明は、対象の分子履歴を作るためのゲノム配列情報の使用に関する。本発明の好ましい実施態様では、ここで定められる方法により得られ及び／又は保存されたような機能的ゲノム情報とゲノム配列情報との組合せが対象の分子履歴を作るために使用される。

【００２０】

特に好ましい実施態様では、上記分子履歴は、上記全ゲノムの機能的側面、レギュローム、又は上記ゲノムの制御状態、ゲノム領域、遺伝子、プロモーター、イントロン、エクソン、経路、経路成分又は所定時間のわたるメチル化状態などを捕捉することで生成される。

【００２１】

他の側面では、本発明は、ここで定められた方法により得られ及び／又は保存されたゲノム配列情報を、疾患の診断、検出、モニター又は予後のために使用することに関する。本発明の特に好ましい実施態様では、ここで定められた方法により得られ及び／又は保存された機能的遺伝情報と、ゲノム配列情報との組合せが、疾患の診断、検出、モニター又は予後のために使用され得る。

【００２２】

本発明の特に好ましい実施態様では、ここで記載される方法又は使用に関して説明される疾患又は障害は、癌疾患、腫瘍疾患又は新生物であり得る。本発明のさらに特に好ましい実施態様では、癌性疾患が、乳癌、卵巣癌又は前立腺癌であり得る。

【００２３】

他の側面では、本発明は臨床判断サポート及び保存システムに関し、上記システムは、対象のゲノム配列情報の入力；プロセッサーに、上で定められた上記ゲノムの配列情報の

複雑性及び／又は量を低減させ得る、コンピュータプログラム製品、対象の遺伝子変異、増加された遺伝子変異又は遺伝子発現変異パターンを出力するための出力、及び上記出力情報を保存するための媒体を含む。特に好ましい実施態様では、上記臨床サポート及び保存システムは、対象のゲノム配列情報を、対象の機能的遺伝子情報、好ましくは遺伝子発現情報と組み合わせて提供するための入力；プロセッサーに、上記ゲノム発現情報の複雑性及び／又は量を低減するステップ及び／又は上記機能的遺伝子情報、好ましくはここで定めた遺伝子発現情報の複雑性及び／又は量を低減するステップとを実施させるコンピュータプログラム製品、対象の遺伝子変異、増加された遺伝子変異又は好ましくは機能的遺伝子変異パターン、好ましくは遺伝子発現変異パターンを出力するための出力、及び上記出力情報を保存するための媒体を含む。

【００２４】

本発明の好ましい実施態様では、上記システムは、電子画像／データ保存記録及び通信システムであり得る。

【図面の簡単な説明】

【００２５】

【図１】図１は、従来の全ゲノム配列決定（ＷＧＳ）手順（パイプライン）の完全な手順を示す。

【図２】図２は、対象のゲノム配列の複雑性及び量を低減するための比較及び整列ステップの概要を示す。

【図３】図３は、本発明による、参照配列と疾患参照配列間の比較を示し、上記疾患の関連するヌクレオチドは染色体１で強調表示されている。

【図４】図４は、変異が互いに近い状況を示す。かかる状況では全ての変異をカバーするより長い配列長さが準備される。

【図５】図５は、対象の時間経過進展についてモニターする方法の通常のステップを示す。

【図６】図６は、疾患発症後及び治療後の、遺伝子コピー数（ＧＣＮ）多型の変化を示す。特定の遺伝子の状態（アップレギュレーション又はダウンレギュレーション）が、有限マルコフ連鎖過程に基づきグラフとして表される。マルコフ連鎖は連続的に動く一組の状態を介して動く過程であることから、状態Ａから状態Ｂへの動きがある確率で起こり得る。これらの確率は、遷移行列の形で表される。この遷移行列内で、イタリックス文字での数値は、疾患進展の際に変化した状態を表し、太字での数値は完全に回復されなかった状態を表す。

【図７】図７は、疾患進展の際の遺伝子コピー数（ＧＣＮ）の変化を示す。この図は、配列決定を用いて得られたサンプルの中間データが、図６の最初の遺伝子コピー数が、疾患の進展に際し変更されたことを示す（即ち図６の行列２への行列１）。これらの増分変化は、上記疾患進展を研究し、所定の遺伝的集団での疾患進展パターンを判断するためのキーとなる。このように表されるそれぞれの行列は、上記疾患の異なる状態を表す。

【発明を実施するための形態】

【００２６】

本発明者は、対象のゲノム配列の複雑性及び／又は量を低減させ、それを迅速に検索可能な形で保存し得る、手段及び方法を開発した。

【００２７】

本発明は、具体的な実施態様により説明されるが、この説明はなにかを限定することを意図するものではない。

【００２８】

本発明の詳細に例示的実施態様を説明する前に、本発明を理解するために重要な定義を与えることとする。

【００２９】

本明細書及び特許請求の範囲で使用される、単数を示す「ひとつの」、「１つの」などは特に記載されない限り複数を含むことを意味する。

【００３０】
本発明の文脈で、用語「約」及び「略」は、当業者が、問題の構成による技術的効果が保証されると理解する精度を意味する。上記用語は通常は、±２０％、好ましくは±１５％、より好ましくは±１０％、さらに好ましくは±５％である。
【００３１】
理解されるべきことは、用語「含む」は限定的な意味ではない、ということである。本発明の目的において、用語「からなる」は、「を含む」の好ましい実施態様と考えられる。以下、群が、少なくともある数の実施態様を含むように定義される場合、これはまた、これらの実施態様のみからなる群を含むことを意味する。
【００３２】
さらに明細書中及び特許請求の範囲中での用語「第１の」、「第２の」、「第３の」又は「（ａ）」、「（ｂ）」、「（ｃ）」、「（ｄ）」などは、類似の要素を区別するためであり、この順序に又は時間的に記載れることは必要ではない。理解されるべきことは、使用される用語は適切な場合には交互に使用できることであり、ここで説明される本発明の実施態様は、ここで説明される順序以外の他の順序でも実施され得る、ということである。
【００３３】
用語「第１」、「第２」、「第３」又は「（ａ）」、「（ｂ）」、「（ｃ）」、「（ｄ）」などが方法又は使用に関連する場合には、このステップ間の時間又は間隔には一貫性はなく、即ち、上記ステップは同時に実施されてよく、又は特に記載されない限り、ステップの間にある時間間隔があってよく、例えば、秒、分、時間、日、週、月又は年であり得る。
【００３４】
理解されるべきことは、本発明は、ここで記載される具体的な方法論、手順、試薬などに限定されるものではなく、変更され得るものである、ということである。また理解されるべきことは、ここで使用される用語は、具体的な実施態様を説明するためであり、本発明を限定する意図はなく、本発明は添付された特許請求の範囲でのみ限定されるものである、ということである。特に記載されない限り、ここで使用される全ての技術的科学的用語は、当業者が共通に理解するものと同じ意味を持つ。
【００３５】
上で説明したように、本発明はひとつの側面で、対象のゲノム配列を処理するための方法に関し、
（ａ）対象のゲノム配列を取得し；
（ｂ）上記ゲノム配列状態の複雑性及び／又は量を低減し；及び
（ｃ）ステップ（ｂ）のゲノム配列状態を迅速に検索可能な形で保存することを含む。
【００３６】
上記方法の第１のステップでは、対象のゲノム配列が取得される。ここで使用される用語「対象」とは、ゲノムを持つ全ての有機体であり得る。好ましくは上記対象は人である。又は、動物のゲノム配列、例えば犬、猫などのペット、ウシ、馬、豚など、又は植物のゲノム配列が得られ得る。本発明の方法は、しかし、これらの有機体の群に限定されるものではなく、一般に、遺伝的、特にゲノム状態を含む全ての対象又は有機体で使用され得る。
【００３７】
ここで使用される用語「対象のゲノム配列を取得する」とは、対象のゲノム配列を決定することを意味する。配列決定の方法は当業者に知られている。好ましくは、次世代配列決定方法又はハイスループット配列決定方法である。例えば、対象のゲノム配列は、多量平行シグネチャー配列決定方法（Ｍａｓｓｉｖｅｌｙ　Ｐａｒａｌｌｅｌ　Ｓｉｇｎａｔｕｒｅ　Ｓｅｑｕｅｎｃｉｎｇ　（ＭＰＳＳ））を用いることで得られ得る。想定される配列決定方法の一例は、パイロシーケンシングで、特に４５４パイロシーケンシング、例えばＲｏｃｈｅの４５４Ｇｅｎｏｍｅ　Ｓｅｑｕｅｎｃｅｒである。この方法は、油溶液

中の水滴内部のＤＮＡを増幅する方法であり、それぞれの液滴は単一のＤＮＡをテンプレートとして含み、これは単一のプライマーコーティングされたビーズに結合され、次にクローン化コロニーを形成する、という方法である。パイロシーケンシング方法はルシフェラーゼを用いて、上記最初のＤＮＡに結合された個別のヌクレオチドの検出のために光発生させ、上記組み合わせデータが配列読み取り出力を生成するために使用される。他の想定される例はＩｌｌｕｍｉｎａ又はＳｏｌｅｘａ配列決定方法であり、例えば、Ｉｌｌｕｍｉｎａ　Ｇｅｎｏｍｅ　Ａｎａｌｙｚｅｒ技術を用いるものであり、これは可逆的色素ターミネータに基づく。ＤＮＡ分子は通常はスライド上のプライマーに結合して増幅され、従って局所的クローンコロニーが形成される。続いて、１つのタイプのヌクレオチドが一度に添加され、取り込まれないヌクレオチドが洗浄で除去される。続いて、蛍光ラベル化ヌクレオチドの画像が取得され、上記色素がＤＮＡにから化学的に除去され、次のサイクルを可能にする。さらに可能な想定される対象のゲノム配列の取得方法は、Ａｐｐｌｉｅｄ　ＢｉｏｓｙｓｔｅｍｓのＳＯＬｉＤ技術を用いる方法であり、これはライゲーションにより配列を決定する方法である。この方法は、固定長さの全ての可能なオリゴヌクレオチドの集団を使用することに基づき、これらは配列位置によりラベル化されている。かかるオリゴヌクレオチドをアニールしてライゲートさせる。続いて、マッチング配列に有利なＤＮＡリガーゼによるライゲーションは、上記位置にあるヌクレオチドのシグナル情報を与える結果となる。ＤＮＡは通常懸濁ＰＣＲにより増幅されることから、得られるビーズは、それぞれ上記同じＤＮＡ分子の１つだけのコピーを含み、ガラススライド上に蓄積され得るものであり、Ｉｌｌｕｍｉｎａ配列決定と同程度の配列量及び長さを与える結果となる。さらなる想定される方法は、ＨｅｌｉｃｏｓのＨｅｌｉｓｃｏｐｅ技術に基づく方法であり、断片がポリＴオリゴマーにより捕捉されアレイに繋げられる。それぞれの配列決定サイクルで、ポリメラーゼ及び単一の蛍光ラベル化ヌクレオチドが添加されて上記アレイを画像化する。上記蛍光タグが続いて除去され上記サイクルが繰り返される。本発明の方法に含まれるさらなる配列決定技術は、ハイブリダイゼーションによる配列決定方法であり、ナノポア、ミクロサイズ配列決定技術、マイクロ流体サンガー配列決定方法、又はマイクロチップ配列決定方法を用いる方法である。本発明はまた、さらに、これらの技術の発展を想定しており、例えばさらに配列決定の精度の改善又は有機体などのゲノム配列決定のために必要な時間の改善などである。

【００３８】

上記ゲノム配列決定は任意の好適な品質、精度及び／又は範囲で得られる。ゲノム配列取得はまた、既に行われた又は独立して得られた配列情報を適用することを含み、例えばデータベース、データリポジトリ、配列決定プロジェクトなどである。

【００３９】

好ましくは、得られるゲノム配列は、１００００塩基、５００００塩基、７５０００塩基、さらには１０００００塩基につき１つ以下のエラーを持つものである。より好ましくは、得られるゲノム配列は、１５００００塩基、２０００００塩基又は２５００００塩基につき１つ以下のエラーを持つものである。

【００４０】

さらには、具体的な実施態様では得られるゲノム配列は、カバーする範囲が、少なくとも９０％、９１％、９２％、９３％、９４％、９５％、９６％、９７％、９８％、９９％、９９．１％、９９．２％、９９．３％、９９．４％、９９．５％、９９．６％、９９．７％、９９．８％、９９．９％、９９．９９％、９９．９９９％又は１００％である。さらに具体的な実施態様では、得られるゲノム配列は、半数体ゲノム当たりの平均リード深さが、少なくとも約１５ｘ、２０ｘ、２５ｘ、３０ｘ、３５ｘ、４０ｘ以上、又は１５ｘから５０ｘ以上の他の任意の平均リード深さを持ち得る。本発明はまた、配列決定技術の改良によるより高いカバー範囲を持つ配列を作るか用いることを想定する。本発明は、従って、いかなるエラー幅又はカバー範囲限界に縛られるものではなく、むしろ、好適な現代的配列決定技術により利用可能な、作られ及び得られる配列情報を実装することに焦点を合わせている。

【００４１】
本発明の好ましい実施態様では、半数体ゲノム当たり、約１５ｘ、２０ｘ、２５ｘ、３５ｘ、４０ｘ以上の得られたゲノム配列の平均リード深さが、上記ゲノムの１以上のサブ領域、例えば、制御領域、オープンリーディングフレーム、１以上のプロモーター領域、１以上のエンハンサー要素、制御ネットワーク部分又は任意のその他の好適なゲノム領域のサブセット、例えば疾患又は障害に関連するシグネチャーデータにより定められる領域に限定され得る。本発明の特に好ましい実施態様では、制御領域又は疾患又は障害に関連するシグネチャーデータで定められる領域では、それぞれの塩基は、少なくとも約１５、２０、２５、３０、４０以上の配列リード数でカバーされており、又は１５から５０の任意のリード数でカバーされている。本発明はまた、配列決定技術の改善によるより高いリード深さを持つ配列の調製及び使用を想定する。本発明は、従って、いかなるエラー幅又はリード深さの限界に縛られるものではなく、むしろ、現在好適な配列決定技術により得られる利用可能な、調製され得られる配列情報の実装に焦点を合わせている。
【００４２】
対象のゲノム配列は、任意の好適なインビトロ及び／又はインビボでの方法により得られる。特に好ましくは、対象から得られるサンプル、例えば以下定められるサンプルからのゲノム配列を得ることである。本発明の具体的な実施態様では、対象のゲノムデータを処理するための方法は、生検サンプルを得ること又は実施をすることを含む。
【００４３】
さらなる実施態様では、対象のゲノム配列は、また、データリポジトリから、例えば対象のゲノム配列を含む１以上のデータベースから、又は対象のゲノム配列を再構成することによる１以上のデータベースから得られる。
【００４４】
得られたゲノム配列は、当業者に知られる任意の好適なフォーマットで表現され得る。例えば、上記配列は、生（元）データとして、ＦＡＳＴＡフォーマットとして、単純なテキストデータとして、ユニコードテキストとして、ｘｍｌフォーマットとして、ｈｔｍｌフォーマットとして表され得る。好ましくは、得られるゲノム配列は、バリアントコールフォーマット（ＶＣＦ）、ゼネラルフィーチャーフォーマット（ＧＦＦ）、ＢＥＤフォーマット、ＡＶＬＩＳＴ又はアノバア（Ａｎｎｏｖａｒ）フォーマットで表され得る。
【００４５】
本発明の第２のステップは、上記ゲノム配列情報の複雑性及び／又は量を低減することである。ここで使用される用語「複雑性」とは、上記ゲノム配列に存在する情報の多様性、上記ゲノム配列に存在する配列情報の冗長性、既知の変異が起こりやすい染色体領域の範囲、遺伝子又は点など、同じく当業者に知られる遺伝子変異のさらなるパラメータなどを意味する。ここで使用される用語「ゲノム配列の量」とは、配列情報の範囲を意味し、例えば染色体の範囲、染色体領域、遺伝子、遺伝子要素、イントロン、エクソン、疾患関連領域また遺伝子などを意味する。上記ゲノム配列の複雑性及び／又は量を低減することで、上記第１のステップで得られた全ゲノム配列データは、異なる好適なパラメータ、例えば遺伝子間領域、イントロン又はエクソンの存在、転写因子の存在、繰り返し領域の存在、知られた変異の点又は領域の存在などのパラメータにより選別される。例えば、エクソン（エクソーム）の配列のみが得られ、又は上記エクソンのあるサブグループのみが得られ得る。同様に、イントロンの配列又はイントロンのサブグループ又はイントロン－エクソン境界領域のなどの配列が得られ得る。さらに、選別パラメータは染色体に局所化することもできる。例えば、上記データは、１、２、３などの染色体へ低減されたり、又は色素化又は発現パターンにより染色体腕や染色領域に低減され得る。さらに、想定される選別パラメータは、例えば生化学的経路、転写因子経路、成長因子又はリガンド活性化による発現パターン、特定の栄養学的状況による発現パターンから導かれる、知られた発現パターンであり得る。さらに一組の選別パラメータは、ゲノム全体の知られた多型、特定の染色体の知られた多型、遺伝子の知られた多型、遺伝子間領域の知られた多型、プロモーター領域の知られた多型であり得る。さらに選別パラメータは、疾患、疾患群、疾患の

素因の知られたデータと連携され得るものであり、例えば選別パラメータは、特定の疾患、疾患群又は疾患の素因に関連する遺伝子変異についての全ての情報を含み得る。

【００４６】

本発明の具体的な実施態様では、上記ゲノム配列は、ゲノム領域、全遺伝子、エクソン（エクソーム配列）、転写因子結合サイト、ＤＮＡメチル化結合タンパク質結合サイト、短い又は長い非コードＲＮＡなどを含み得る遺伝子間領域であって、臨床的に関連し又は重要であり、及び変異可能であるか高変異性であることが知られ又は疑われている、人間、人種間又は集団間、人又は動物の性間、人の年齢集団、例えば新生児及び成人間、人及び他の生物などの間、同じ種の動物間、異なる種、族又はクラス間の動物、植物品種、植物種などの間、又は疾患又は障害において変異可能又は高変異性であることが知られているか又は疑われている遺伝子間領域に低減され得る。かかるゲノム領域、遺伝子、エクソン、結合サイトなどは当業者に知られており、又は好適な教科書又は情報リポジトリ、例えばＵＣＳＣゲノムブラウザ又はＮＣＢＩから導き出せる。

【００４７】

ゲノム配列の複雑性及び／又は量の低減は、１以上のステップで実施され、例えば比較方法又はアルゴリズム、モチーフ検索方法又はアルゴリズム、反復プロセスなどでありこれらは当業者に知られている。例えば、上記低減は、適切な教科書又は科学文献に基づき実行でき、例えば、Ｓ．Ｋｕｒｔｚ、Ａ．Ｐｈｉｌｌｉｐｐｙ、Ａ．Ｌ．Ｄｅｌｃｈｅｒ、Ｍ．Ｓｍｏｏｔ、Ｍ．Ｓｈｕｍｗａｙ、Ｃ．Ａｎｔｏｎｅｓｃｕ、及びＳ．Ｌ．Ｓａｌｚｂｅｒｇらの「Ｖｅｒｓａｔｉｌｅ ａｎｄ ｏｐｅｎ ｓｏｆｔｗａｒｅ ｆｏｒ ｃｏｍｐａｒｉｎｇ ｌａｒｇｅ ｇｅｎｏｍｅｓ、（Ｇｅｎｏｍｅ Ｂｉｏｌｏｇｙ、５：Ｒ１２、Ｓｃｈｕｓｔｅｒ ｅｔ ａｌ．、２０１０、Ｎａｔｕｒｅ ４６３（１８）、９４３－９４７（２０００））」又はＦｕｊｉｍｏｔｏらの「Ｎａｔｕｒｅ Ｇｅｎｅｔｉｃｓ、４２、９３１－９３６（２０１０）」が挙げられ、これらの内容は参照されて本明細書に援用される。

【００４８】

さらにゲノム配列の複雑性及び／又は量を低減するために想定される方法は、Ａｓｈｌｅｙらの「Ｔｈｅ Ｌａｎｃｅｔ、３７５、１５２５－１５３５、２０１０」から導き出せ、この内容はまた参照されて本明細書に援用される。特に上記刊行物の図１に与えられるゲノム変異に関する分子情報に基づき上記複雑性の低減は本発明の範囲内である。

【００４９】

さらなる具体的な実施態様では、医薬－応答表現型、遺伝子座特異的変異データベース（ＬＳＭＤ）又は人ミトコンドリア遺伝子多型データベース（ｍｔＳＮＰ）に関する医薬品知識ベース（ＰｈａｒｍＧＫＢ）により提供される情報に基づく、上記ゲノム配列の複雑性及び／又は量の低減が想定される。

【００５０】

特に好ましくは、上記得られるゲノム情報について集団系選別を適用することである。例えば、ゲノム配列変異、特にＳＮＰはここで定めた比較方法で検出され、さらに患者の集団、人種又は祖先の内容に沿って比較又は分析され得る。従って、例えば、特定の集団、人種、年齢群などについてひとつの変異ＳＮＰが存在する場合、この変異は本発明の目的において、関連すると報告され識別されず又は選別されて除去される。具体的な実施態様では、かかる変異が－ある集団、人種、年齢群などに特異的又は典型的であっても－上記変異が重要な／臨床的機能的意味を示す場合には本発明の目的において関連あるとして考慮され識別される。全集団で見出される機能的重要なＳＮＰのクラスとしての一例はＣＹＰ関連遺伝子であり、これは上記医薬を代謝し排泄することを助ける。ある医薬は、（非白人などの）異なる集団では、容量が異なる、例えば低容量であることが知られており、ＣＹＰ－関連遺伝子での変異は、患者の集団所属又は患者の人種により、選別、ソート、クラス分け及び／又は評価される。かかる選別は、例えば上記ＰｈａｒｍＧＫＢデータベースに提供される情報に基づき実施され得る。

【００５１】

選別され又は低減されたゲノム配列は任意の好適なフォーマットで表され得る。好ましくは、上記配列は、ＦＡＳＴＡフォーマット、単純なテキストフォーマット、ユニコードテキスト、ｘｍｌフォーマット、ｈｔｍｌフォーマット、バリアントコールフォーマット（ＶＣＦ）、ゼネラルフィーチャーフォーマット（ＧＦＦ）、ＢＥＤフォーマット、ＡＶＬＩＳＴフォーマット又はアノバールフォーマット（Ａｎｎｏｖａｒ）で表され得る。さらに、上記ゲノム配列は、デリバティブフォーマットで表されてよく、例えば、データベースエントリーとして、注釈付きデータベースエントリーとして、ゲノム／遺伝子的変異の点のリストとして表されてよく、好ましくは発生、例えば集団などでの発生の関連性又は数で並べ替えられる。

【００５２】

上記方法の第３のステップでは、上記第２のステップで得られたゲノム配列情報が迅速に検索可能な形で保存される。保存されるべき情報は、任意の好適な形又はフォーマットでよく、例えば上で説明したフォーマットが挙げられる。上記ゲノム情報の保存は、好ましくは、好適な保存媒体、例えばコンピュータハードディスク・ドライブ、モバイル保存装置などの利用可能な空間に限定される。特に好ましい保存構造は、（１）階層的及び／又は（２）時間情報をエンコードし及び／又は（３）患者データ、画像、報告などにリンクするものである。より好ましくは、差分ＤＮＡ保存構造（ＤＤＳＳ）などの構造である。

【００５３】

ここで使用される用語「迅速に検索可能」とは、上記ゲノム情報が、容易に情報にアクセスでき、及び／又は上記保存データ情報の複雑でない抽出を可能にする形で提供される、ということを意味する。本発明で想定される保存の形は、好適なデータベース保存、リストでの保存、数字付け文書及び／又はグラフの形での保存、例えば絵文字、グラフ配列、比較図などである。本発明の具体的な実施態様では、上記情報は、保存媒体から取り出され、続いて、例えば好適なモニター上に、ハンドヘルド装置、コンピュータ装置などで表示される。

【００５４】

本発明の具体的な実施態様では、対象のゲノム配列を処理するための方法は、ステップ（ａ）で、上で定めた上記ゲノム配列情報の複雑性及び／又は量を低減させることを含み；かつステップ（ｂ）でステップ（ａ）のゲノム配列情報をここで説明したように迅速に検索可能な形で保存することを含む。

【００５５】

本発明の好ましい実施態様では、対象のゲノム配列を得るための分析されるサンプルは、対象の身体又は器官の任意の好適な部又は部分から誘導され得る。上記サンプルは、ひとつの実施態様では、純粋な組織又は臓器から又は細胞型から誘導され、又は非常に特異的な位置、例えば１つのタイプの組織、細胞又は臓器のみを含む位置から誘導され得る。さらなる実施態様では、上記サンプルは組織、臓器、細胞又はそれらの断片の混合物から誘導され得る。サンプルは、好ましくは、臓器又は組織から得られ得るものであり、例えば消化管、膣、胃、心臓、舌、膵臓、肝臓、肺、腎臓、皮膚、脾臓、卵巣、筋肉、関節、脳、前立腺、リンパシステムまたは臓器または当業者に知られている組織が含まれる。本発明のさらなる実施態様では、上記サンプルは身体液、例えば血液、血清、唾液、尿、糞便、精液、リンパ液などの体液から誘導され得る。

【００５６】

特に好ましくは、腫瘍組織の適用又は癌性として知られる臓器から誘導されるサンプルの使用である。また、疾患、感染、障害などに関連した、又は影響されると診断された任意の他の臓器又は組織又は細胞又は細胞型から誘導されるサンプルの使用が想定されている。本発明の具体的な実施態様では、上記サンプルは固体腫瘍、腫瘍又は癌性の疑いがある組織切除、疾患臓器又は組織からの生検、例えば感染又は癌性臓器や組織などから得られる細胞を含む。上記感染は、例えば細菌性又はウイルス性感染である。

【００５７】

上記サンプルは１以上の細胞、例えば組織学的又は形態的に同一の細胞、又は組織学的又は形態的に異なる細胞を含み得る。好ましくは、組織学的に同一又は類似の細胞、例えば上記身体の１つの閉鎖領域から生じる細胞の使用である。

【００５８】

さらに、異なる時点での、同じ対象から、同じ対象の異なる臓器又は組織から、又は同じ対象の異なる時点での、異なる臓器又は組織から得られるサンプルの使用が想定されている。例えば、腫瘍組織のサンプル又は、同じ組織又は臓器の近隣の非癌性領域の腫瘍組織及び１以上のサンプルが取得され、対象のゲノム配列を得るために使用され得る。

【００５９】

非人又は非動物対象の場合には、サンプルは他の組織型、例えば使用される特定の植物組織などから誘導され、これには例えば葉、根組織、分裂組織、発光組織、植物種から誘導される組織などを含み得る。

【００６０】

対象のゲノム配列は、従って、取得されたサンプルに依存し、ゲノム配列情報の混合物を含み、例えば対象の異なる組織、臓器及び／又は細胞の混合物であり、又は対象の特定の単一ソースから誘導されるゲノム情報、例えば１つの臓器や臓器型、１つの組織や組織型、１つの細胞や細胞型であり、従って対応する臓器、組織又は細胞を表すものである。癌性臓器や組織の場合、組織学的方法及び手法での生検のサポートと同じく、特定して選択されたサンプルはまた、本発明で想定されるものである。

【００６１】

本発明のさらなる実施態様では、対象のゲノム配列は最初に取得され、続いて上記取得ステップが繰り返される。好ましくは対象のゲノム配列の取得は、１回、２回、３回、４回、５回、６回以上繰り返される。上記第２の又はそれ以上の取得はある一定期間後に実施され、例えば１週間後、２週間後、３週間後、４週間後、２、３、４、５、６、７、８、９、１０、１１、１２ヶ月後、１．５年後、２年後、３年後、４年後、５年後、６年後など、又はずっと後の時点、又はこれらの時点間での任意の期間後であり得る。対象のゲノム配列の、第１回と第２回取得との間の時間、及び第２回と続く取得との時間は同じ、本質的に同じ又は異なっていてもよく、例えば増加又は減少も可能である。例えば、治療モニターの間、対象のゲノム配列は、等間隔、又はより長い間隔又はより短い間隔で取得され得る。

【００６２】

通常は、対象のゲノム配列が最初の取得後のさらなる取得の場合、同じ臓器、組織、細胞、臓器型、組織型、細胞型で、また、同じサンプルタイプ、例えば尿、血液、血清、唾液サンプルなど上記最初の取得で使用されたもので、取得される。又は、非同一の臓器、組織、細胞、臓器型、組織型、細胞型又はサンプルタイプなどが、対象のゲノム配列の続く取得の対象とされ得る。さらに、組織、臓器、細胞などの混合物から対象のゲノム配列を最初に取得し、続いて、決まった特定のソース、例えばここで定められた特定の臓器、組織、細胞、臓器型、組織型また細胞型からの対象のゲノム配列の取得がなされることが想定される。又は、最初に、特定のソース、例えばここで定められた特定の臓器、組織、細胞、臓器型、組織型また細胞型から対象のゲノム配列を取得し、続いて組織、臓器、細胞などの混合物から対象のゲノム配列を取得する。例えば、疾患、例えば癌の治療の間、後者の方法が取られ、変性又は異常細胞、細胞型又は組織部分の残渣の存在をカバーする。

【００６３】

本発明のさらなる実施態様では、対象のゲノム配列を、２以上の異なる位置、臓器、組織、細胞、組織型、細胞型などから同時に又は平行して取得し、それに対応して得られるゲノム配列情報を、また上で記載されたように処理する。

【００６４】

対象のゲノム配列を最初に及び続いて取得するための方法は、また並行して配列が取得される場合の方法は、同じであってもよく、異なっていてもよい。

【００６５】
上記配列決定技術、及び／又は得られる結果のフォーマットなどが本質的に同じであることが好ましい。
【００６６】
対象のゲノム配列が、第２の取得、又は細胞の最初の取得後さらに時間経過後取得された後、又は一回で１以上のゲノム配列が取得された場合、得られたゲノム配列情報間、例えば最初の取得及び第２の取得又はさらなる取得された情報間の比較が実施される。好ましくは、かかる比較は、上記最初に得られたゲノム配列と、続いて得られたゲノム配列の間の、又は異なる位置、臓器、組織、細胞で得られたゲノム配列間との変化、変性又は差を明らかにするために実施される。ここで使用される用語「比較」とは、２つのゲノム配列をマッチングするための任意の好適な方法又は技術に関連する。通常は、当業者に知られる整列アルゴリズムが適用されて、２つのゲノム配列間の相違を検出する。かかるアルゴリズムの例は、Ｓ．Ｋｕｒｔｚ、Ａ．Ｐｈｉｌｌｉｐｐｙ、Ａ．Ｌ．Ｄｅｌｃｈｅｒ、Ｍ．Ｓｍｏｏｔ、Ｍ．Ｓｈｕｍｗａｙ、Ｃ．Ａｎｔｏｎｅｓｃｕ及びＳ．Ｌ．Ｓａｌｚｂｅｒｇの「Ｖｅｒｓａｔｉｌｅ ａｎｄ ｏｐｅｎ ｓｏｆｔｗａｒｅ ｆｏｒ ｃｏｍｐａｒｉｎｇ ｌａｒｇｅ ｇｅｎｏｍｅｓ．」、Ｇｅｎｏｍｅ Ｂｉｏｌｏｇｙ、５：Ｒ１２、２００４」から導き出される方法を含む。好適な及び想定されるアルゴリズムのさらなる例は、ベースコールのＵＭＫＡアルゴリズム（Ｐｕｓｈｋａｒｅｖ ｅｔ ａｌ．、Ｎａｔ Ｂｉｏｔｅｃｈｎｏｌｏｇｙ、２００９、２７：８４７－５２）であり、この内容は参照されて本明細書に援用され、及びＡｓｈｌｅｙらによるアルゴリズム「Ｔｈｅ Ｌａｎｃｅｔ、３７５、１５２５－１５３５、２０１０」である。
【００６７】
本発明の１つの実施態様では、比較は、最初の取得と第２の又は続く取得で得られた全ゲノム配列間で、又は同時に得られたゲノム配列間で実施される。これにより、全ての変異、変化及び差が全ゲノム配列を通じて完全な概要を与える。
【００６８】
本発明の他の実施態様では、比較は、選別された又は低減されたゲノム配列又は上記のゲノム配列情報間で実施される。好ましくは上記最初に得られたゲノム配列又は同時に得られたゲノム配列であって、ゲノム領域、全遺伝子、エクソン（エクソーム配列）、転写因子結合サイト、ＤＮＡメチル化結合タンパク質結合サイト、短い又は長い非コードＲＮＡなどを含み得る遺伝子間領域であって、臨床的に関連し又は重要であり、及び変異可能であるか高変異性であることが知られ又は疑われている、人間、人種間又は集団間、人種間又は集団間、人の年齢集団、例えば新生児及び成人間、人及び他の生物などの間、同じ種の動物間、異なる種、族又はクラス間の動物、植物品種、植物種などの間、又は疾患又は障害において変異可能な又は高変異性であることが知られているか又は疑われている遺伝子間領域に低減されるゲノム配列が、第２の又は続いて取得されるゲノム配列と比較されるために使用され得る。
【００６９】
他の実施態様では、比較はさらに、試験、例えば遺伝データ解釈、データ標準化、データクラスタ化、ｋ－平均クラスタ化、階層クラスタ化、主因子分析、教師方法などを含む。かかる追加の試験は当業者に知られており、好適なソース、例えばＴｊａｄｅｎらの「Ａｐｐｌｉｅｄ Ｍｙｃｏｌｏｇｙ ａｎｄ Ｂｉｏｔｅｃｈｎｏｌｏｇｙ：Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ、６、２００６」から導入され、この内容はここで参照されて本明細書に援用される。
【００７０】
さらなる実施態様では、最初の取得後、得られた第３、４、５などの続いて得られた対象のゲノム配列が比較され、この比較は、最初に得られたゲノム配列及び／又は続いて得られたゲノム配列と共に実行され得る。かかる比較は、全ゲノム配列間、又は上で説明された低減又は選別されたサブセット間で実行され得る。
【００７１】

好ましい実施態様では、比較は、連続したゲノム配列情報の組み間で、例えば最初に得られたゲノム配列情報とゲノム配列取得の第１回目の繰り返しで得られたゲノム配列情報間で実施され；上記ゲノム配列取得の第１回目の繰り返しで得られたゲノム配列情報と、ゲノム配列取得の上記２回目の繰り返しで得られたゲノム配列情報間で実施され；上記ゲノム配列取得の第２回目の繰り返しで得られたゲノム配列情報と、ゲノム配列取得の上記３回目の繰り返しで得られたゲノム配列情報間で実施され得る。

【００７２】

又は、比較は次のように実施され得る：例えば、最初に得られたゲノム配列情報とゲノム配列取得の第２回目の繰り返しで得られたゲノム配列情報との間；最初に得られたゲノム配列情報とゲノム配列取得の第３回目の繰り返しで得られたゲノム配列情報との間である。さらなる実施態様では、例えば上記ゲノム配列情報はよりしばしば得られる場合においては、それぞれの組みのゲノム配列情報間の全てのタイプの比較が実施され得る。

【００７３】

特に好ましい実施態様では、対象のゲノム配列が第２又は続く時間で得られる場合には、すでに保存されたゲノム配列情報のゲノム配列情報と比較して上記増加データが保存される。ここで使用される「増加データ」とは、与えられた２つの組みのゲノム配列情報間で異なるか又は変化した情報を意味する。

【００７４】

例えば、保存されるデータは、変化のあった位置又は特質を含む。加えて、さらなるパラメータが保存され、例えば配列伸長、取得時間、取得間隔などである。かかる保存は、任意の好適なフォーマット又は形で実施され、例えばデータベースエントリーの形で、グラフ化情報として、テキスト又は携帯可能な資料として、又は専門家のために音声として検索可能な音声又は会話フォーマットで保存され得る。特に好ましくは、（１）階層的及び／又は（２）時間情報をエンコードする及び／又は（３）患者データ、画像、報告などとリンクする、保存構造である。さらに好ましくは、差ＤＮＡ保存構造（ＤＤＳＳ）などの保存構造である。

【００７５】

具体的な実施態様では、例えば、対象のゲノム配列が２回以上得られる場合、上記データが上記２回目に表される場合、上記遺伝データでの変化は識別され（即ち、G^2及びG^1間の差）かつ変更された部分のみが保存される（δG^2）。上記遺伝データは、第ｎ回時（G^n）につき表される場合、前回の遺伝データ（G^{n-1}）は次のように再構成される。

【００７６】

【数１】

$$G^{n-1}=G^1+\sum_{i=2}^{n-1}\delta G^i$$

G^nとG^{n-1}の間に変化があることが検出されるとこの変化がδG^nとして保存される。かかるプロセスの利点は、遺伝情報を保存するためのメモリ及び保存スペースが劇的に低減できるということである。

【００７７】

本発明の好ましい実施態様では、G^n及びG^{n-1}間で変化がある場合にはこの変化は上記疾患状態に対応し得るものであり、好ましくはエンコードされ行列に記載される（例えば図６で示されるように）。ある遺伝子の状態（例えば、増幅又は削減された状態であり、これはそれぞれの遺伝子がアップレギュレーション又はダウンレギュレーションされている結果である）が、例えばデコードされ得る。

【００７８】

本発明は、従って、次の方法を想定し、上記方法は、ゲノム及び／又は機能的遺伝子情報での変化が行列内にエンコードされ、及び好ましくは疾患又は障害との関連で、遺伝子、ゲノム領域、制御領域、プロモーター、エクソン又は経路の状態を保持する情報がデコ

ードされ、好適なプロセスで表される。
【００７９】
好ましい実施態様では、好ましくは疾患又は障害との関連で、遺伝子、ゲノム領域、制御領域、プロモーター、エクソン又は経路の状態が、かかる行列からエンコードされるか、濃縮されて表され、及び好適なグラフモデルで可視的に表現され得る。
【００８０】
好ましくは、かかるグラフモデルは有限マルコフ連鎖過程に基づく。マルコフ連鎖は、一組の状態が連続的に動き、状態Ａから状態Ｂへの動きがある確率を持っている過程である。この確率は、行列として、好ましくは遷移行列の形で表され得る。図７は、連続的な一組の状態を示し、患者のプロファイルをマッチングさせ、患者への意思決定がある確率を持って状態Ａから状態Ｂへ遷移することを示す。かかるプロセスの利点は、（ｉ）上記遺伝情報を保存するための必要なメモリ及び保存スペースが劇的に低減されることであり、（ｉｉ）上記表現が、疾患の進展（又は後退）の状態を表す行列とマッチングするための助けとなる、ということである。この方法で、上記保存された表現は、容易に臨床判断サポートソフトウェアに準拠することが可能となり、これは遷移状態をマッチングさせ、診断判断を行う上で助けとなる。
【００８１】
本発明の具体的な実施態様では、上記ゲノム配列及び／又は上記機能的遺伝情報の複雑性及び／又は量を低減及び／又はゲノム及び／又は機能的型遺伝情報での変化のエンコード又は分析は、確率ブーリアンネットワーク（ＰＢＮ）で、又はこれに基づき実施され得る。かかるＰＢＮは、モデル化方法についての規則ベースのパラダイムとして、使用され得る、例えば制御ネットワーク、又はここで説明したデータ又は情報の選別やリンクのために使用され得る。本発明はまた、従って、例えばここで説明されたマルコフ連鎖過程に含まれるマルコフ遺伝子制御ネットワークのサブクラスとしてかかるネットワークを採用することを想定する。ひとつの実施態様では、上記ＰＢＮは、異なる遺伝子、経路、疾患状態、疾患因子、分子疾患症状又はその他の当業者に知られる好適な情報を表すために使用され得る。ＰＢＮの好適な実装及び形式化は当業者に知られており、又は高品質科学的資料、例えばＨａｍｉｄ　Ｂｏｌｏｕｒｉの「Ｃｏｍｐｕｔａｔｉｏｎａｌ　Ｍｏｄｅｌｌｉｎｇ　Ｏｆ　Ｇｅｎｅ　Ｒｅｇｕｌａｔｏｒｙ　Ｎｅｔｗｏｒｋｓ、２００８、Ｉｍｐｅｒｉａｌ　Ｃｏｌｌｅｇｅ　Ｐｒｅｓｓ」から導入することが可能である。
【００８２】
かかる表現は、臨床判断サポートソフトウェアでの実装での対応と同じく本発明において想定されている。
【００８３】
本発明のさらなる実施態様では、ここで定められる方法はまた、時間経過にわたり変化又は差をモニターするステップを含む。さらに又はこれに代えて、本方法は傾向を予想するステップを含み、例えば治療の進行中又は疾患の進展中の改善傾向又は悪化傾向などである。
【００８４】
他の実施態様では、本発明はさらに、例えば（δG^n）に基づく関連するリスク因子の計算を含む。遺伝データの変化（δG^n）が、上記人が影響され得るリスクを示唆しないか、直接示唆しない場合において、１以上の（δG^2、δG^3、．．．δG^{n-1}）と組み合わせて（δG^n）がリスク因子の計算のために使用され得る。ここで使用される用語「リスク因子」とは、疾患を発症する可能性及び／又は疾患が悪化して次の段階へ進む可能性、又は疾患の素因が疾患へ向かう可能性を意味する。
【００８５】
特に好ましい実施態様では、増加データの全ての可能な組合せが上記リスクを導くために分析され得る。従って、リスクのための上記遺伝子データを分析する際の複雑性は、それが大量のデータ（G^1、G^2、．．．G^n）を処理するものではないことから大きく低減され得る。具体的な実施態様では、上記保存された表現が疾患防止ステップを作るため

に使用され得る。さらなる実施態様では、上記保存表現は、より頻繁なスクリーニング、好ましくは画像化又はその他の診断モダリティを用いることで実行され得る。

【００８６】

さらに具体的な実施態様では、上記保存ゲノム配列データは、これらのデータが専門家に使用されるために十分であることから、上記増加データ即ち（δG^2、δG^3、．．．δG^n）のみがアクセス許容される選択肢と共に提供される。かかる可能性は、上記対象が彼の遺伝データ又はゲノムデータを開示することから秘匿することを可能にする、という利点を持つ。

【００８７】

本発明のさらに好ましい具体的な実施態様では、ゲノム配列情報の複雑性及び／又は量を低減することは、疾患又は障害に関連するシグネチャーデータ以外のデータを上記ゲノム配列情報から切り取ることで実施され得る。ここで使用される用語「ゲノム配列情報を切り取る」とは、ゲノム配列の最初又は続く取得で得られるゲノム配列セットにおいて実施される、集中化又は削除手順を意味する。従って、非関連及び／又は冗長なゲノム配列情報は、最初のゲノム情報から削除されるか除去され得る。かかる集中化又は切り取りステップは通常は、遺伝子的症状、障害、疾患のシグネチャーデータ、障害又は疾患の予兆、疾患などの進展へのリスク因子などに基づく。

【００８８】

ここで使用される用語「シグネチャーデータ」とは、遺伝子又がゲノム変異についての情報を意味する。好ましくは、かかるシグネチャーデータは、疾患、障害に特異的、疾患又は障害の予兆に特異的、疾患などの進展へのリスク因子へ特異的な遺伝子的又はゲノム変異であり得る。又はシグネチャーデータは、それ自体が疾患や障害に関連しているものではなく、対象の適合性、丈夫さ、特定の状態への適合性、適合可能性、変異の履歴に基づく情報、又は対象の又は対象の識別に必要な情報、例えば犯罪捜査、指紋手法、父性試験などに基づく情報を提供する。

【００８９】

好ましい実施態様では、シグネチャーデータは、疾患、障害、疾患や障害の予兆、疾患進展へのリスク因子に特異的な情報であり、又は提供するものであり、これらは、ミスセンス変異、ナンセンス変異、一塩基多型（ＳＮＰ）、コピー数多型（ＣＮＶ）、スプライシング変異、制御配列の変異、小欠失、小挿入、小インデル、総欠失、総挿入、複雑な遺伝子再配列、染色体間再配列、染色体内再配列、ヘテロ接合性消失、反復配列の挿入及び／又は反復配列の欠失、及び／又はこれらのシグネチャーのいずれかの組み合わせ、から選択される。さらに、好適な上記ゲノム又は対象の遺伝子配列や、当業者に知られる症状やシグネチャーデータが本発明の範囲に含まれる。

【００９０】

本発明のさらなる実施態様では、上記シグネチャーデータは、特異的疾患に関連することが知られる特異的遺伝子又は遺伝子座であり、例えばＨＥＲ２、ＥＦＧＲ、ＫＲＡＳ、ＢＲＡＦ、Ｂｃｒ－ａｂｌ、ＰＴＥＮ、ＰＩ３Ｋ、ＢＲＣＡ１、ＢＲＣＡ２、ＧＡＴＡ４、ＣＤＫＮ２Ａ、ＰＡＲＰ、ｐ５３などである。かかるマーカーシグネチャーは、もちろんまた、追加パラメータ又は追加の遺伝子情報、例えばＳＮＰ、コピー数変異などと組合せることが可能である。

【００９１】

特に好ましい実施態様では、シグネチャーデータは、一塩基多型（ＳＮＰ）及び／又はコピー数変動（ＣＮＶ）、又は遺伝子コピー数多型（ＧＣＮ）、即ち、対象の遺伝子型での特定に遺伝子のコピー数の変異であるか、又はこれらを与えるものである。上記ＧＣＮは、例えば、癌性細胞で複雑に変性させ得る。対応する遺伝子発現情報は、さらに具体的な実施態様で得られる。

【００９２】

対応する遺伝子又はゲノム変異は、例えば同様に疾患や障害に関連して当業者には知られており、及び／又は好適なデータリポジトリから導き出せ、これらは例えば、

「ｔｈｅ Ｎａｔｉｏｎａｌ Ｃｅｎｔｅｒ ｆｏｒ Ｂｉｏｔｅｃｈｎｏｌｏｇｙ Ｉｎｆｏｒｍａｔｉｏｎ（ＮＣＢＩ）、ＮＩＨ、ＵＳＡ、ｗｗｗ．ｎｃｂｉ．ｎｌｍ．ｎｉｈ．ｇｏｖからアクセス可能」や「ｔｈｅ Ｅｕｒｏｐｅａｎ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ Ｉｎｓｔｉｔｕｔｅ（ＥＢＩ） ｏｆ ｔｈｅ ＥＭＢＬ、ｗｗｗ．ｅｂｉ．ａｃ．ｕｋからアクセス可能」であり、特に特異的なデータ収集は「ｔｈｅ ＳＮＰ ｄａｔａｂａｓｅ、ＯＭＩＭ、ＲｅｆＳｅｑ」や「ｔｈｅ Ｈｕｍａｎ Ｇｅｎｏｍｅ Ｍｕｔａｔｉｏｎ Ｄａｔａｂａｓｅ」などからのデータリポジトリである。

【００９３】

特に好ましい実施態様では、上記シグネチャーデータは、遺伝子又はゲノム領域のパネルに基づくものであり、これらは少なくとも対象又は症状の２つの群を識別し得るものであり、例えば、腫瘍状態対正常／健常状態間；又は悪性腫瘍状態対良性状態間；又は医薬組成物例えば制癌剤への化学的感受性対医薬組成物、例えば制癌剤への化学的抵抗性の状態間、などである。対象の遺伝子データを処理する本発明の具体的な実施態様では、ここで定められるようにまた、遺伝子データの変性がさらなる続く変化の結果となり得る状態も含む。従って、遺伝子データの変化（$\delta G^{n'}$）が、（δG^{2}、δG^{3}、．．．δG^{n-1}）から、知られる遺伝疾患のシグネチャーデータを用いることで予期され得る。例えば、上記予期される変化$\delta G^{n'}$が実際の変化δG^{n}に等しい場合は、対象は上記疾患に影響を受けやすいと考えられる。さらなる実施態様では、δG^{n}が、これまでの遺伝子変化を用いて計算され得るものであり、従って保存されるか保存されなくてもよい。また、上記得られたデータは保存又は一時的に保存され得る。

【００９４】

本発明の他の好ましい実施態様では、対象の遺伝子データを処理するための本発明のゲノム配列情報の複雑性及び／又は量を低減するステップは、対象のゲノム配列をシグネチャーデータを含む標準シグネチャーデータと整列させることで実施され得る。好ましくは、参照配列（ＲｅｆＳｅｑ）は疾患又は障害に関連するシグネチャーデータを含み得る、例えば、障害、疾患、障害又は疾患の予兆、疾患の進展のリスク因子の基づく情報であり、ミスセンス変異、ナンセンス変異、一塩基多型（ＳＮＰ）、コピー数多型（ＣＮＶ）、スプライシング変異、制御配列の変異、小欠失、小挿入、小インデル、総欠失、総挿入、複雑な遺伝子再配列、染色体間再配列、染色体内再配列、ヘテロ接合性消失、反復配列の挿入及び／又は反復配列の欠失、及び／又はこれらのシグネチャーのいずれかの組み合わせ、から選択される。特に好ましくは、１又は全てのゲノムシグネチャーについての全ての可能な配列が存在する参照配列に基づくシグネチャーの提供である。さらなる実施態様では、これらのシグネチャーは、上記ゲノム変異の上流又は下流又は上記ゲノム変異の上流又は下流のいずれかの、特定の長さ、例えば１００ｂｐ、２００ｂｐ、５００ｂｐ、１ｋｂｐ、２ｋｂｐ、５ｋｂｐ、１０ｋｂｐのフランキング配列での情報と組合せることが可能である。

【００９５】

本発明によるこれらのシグネチャー参照配列は、任意の好適なフォーマット又は形で提供される。好ましくはＦＡＳＴＡ又はＦＡＳＴＱフォーマットである。さらに好ましくは、アライナ、好ましくはアライナ（ａｌｉｇｎｅｒ）のマルチタイプにより任意の認識されるフォーマットが好ましい。

【００９６】

本発明によるシグネチャー参照配列の具体的な実施態様では、通常の参照配列（例えばＮＣＢＩなどのデータリポジトリから導きさせるゲノム配列情報）を、例えば、疾患のデータ、遺伝子要素の位置及び／又は方向の情報、関連する遺伝子の情報、変異型及び／又は変異サイズの情報及び／又は変異の頻度の情報を含むゲノムシグネチャーと組み合わせることから導かれ得る。これらのデータはさらに、注釈付きデータベース、例えば遺伝子要素の位置及び／又は方向及び／又はこれらの要素のタイプ及びサイズに関連する注釈付きデータから導かれるデータと組合せ得る。例示的ワークフローは図２に与えられる。

【００９７】

他の実施態様では、本発明によるシグネチャー参照配列は、検出されるゲノム変異のタイプ及び／又は得られる又は得られ得るゲノム配列情報のタイプに適合され得る。これらのパラメータは組み合わせることができ、又は相互に排他的であり得る。

【００９８】

例えば、シグネチャー参照配列は、単一末端及び／又は対末端データとしてゲノム配列と比較するために与えられ得る。かかるシグネチャー参照配列は、置換、インデル、ＳＮＰ、ＣＮＶ、規則的変異、ミスセンス又はナンセンス変異などを含み得る。このシグネチャー参照配列に基づき、対象から得られるゲノム配列に存在する知られる置換、インデル、ＳＮＰ、ＣＮＶ、規則的変異、ミスセンス又はナンセンス変異が検出され得る。上記シグネチャー参照配列は、ＦＡＳＴＡファイル、例えばｓＲｅｆＳｅｑＩとして与えられ得る。

【００９９】

さらなる実施例では、シグネチャー参照配列は、対末端データとして存在するゲノム配列と比較するために与えられ得る。かかるシグネチャー参照配列は、総挿入、総欠失、染色体異常、染色体間、染色体内変異などの情報を含む。対象から得られる、知られた総挿入、総欠失、染色体異常、染色体間、染色体内変異などの知られるシグネチャー参照配列が削除され得る。上記シグネチャー参照配列は、ＦＡＳＴＡファイル、例えばｓＲｅｆＳｅｑＩＩなどのファイルとして与えられ得る。

【０１００】

さらなる例では、シグネチャー参照配列は、単一末端データ又は対末端データとして存在するゲノム配列と比較するために与えられ得る。かかるシグネチャー参照配列は、ゲノム領域又は興味領域の情報を含み、例えば、特定の疾患や障害、ホットスポット又は変異などの観点で変化又は変性されることが知られる領域である。このシグネチャー参照配列に基づき、対象から得られたゲノム配列に存在する知られた特定の疾患や障害、ホットスポット又は変異などの文脈で変化又は変性される領域が削除され得る。このシグネチャー参照配列は、ＦＡＳＴＡファイル、例えば、ｓＲｅｆＳｅｑＩＩＩとして与えられる。

【０１０１】

本発明の他の実施態様では、ここで定められたように対象から得られるゲノム配列はまた、参照配列として使用され得る。かかる参照配列では、知られる変異、例えばＳＮＰ又は置換が検索され得る。

【０１０２】

通常の実施態様では、置換、インデル、ＳＮＰ、ＣＮＶ、規則的変異、ミスセンス又はナンセンス変異など（ｓＲｅｆＳｅｑＩ）の検出のための上記説明されたシグネチャー参照配列は、以下の方法ステップで実施され得る：

（１）置換、インデル、ＳＮＰ、ＣＮＶ、規則的変異、ミスセンス又はナンセンス変異などに対応するシグネチャーのリストが作られる。

（２）シグネチャーのリストは、染色体、配位数及び方向により並べ替えられ得る。さらに識別コード、正常配列情報及び変異配列情報が含まれる。

（３）上記配列は、正常及び変異配列の両方で利用可能な配列情報に基づき拡張され得る。

例えば上記変異のいずれかの側の５０、１００、２００、３００、４００、５００、６００、７００、８００、９００、１０００塩基が含まれ得る。

通常は、上記変異側からの配列の拡張は、配列読み取りの数倍（１００塩基の読み取りにつき５００塩基）であり得る。

（４）正常及び変異拡張型の逆相補的配列が生成され得る。

（５）上記変異が互いに近い場合、上記配列は拡張された型であり、上記変異が末端に位置する。正常及び変異配列の両方の対応する逆相補的配列が作られる。

【０１０３】

さらなる実施態様では、総挿入、総欠失、染色体上で説明した染色体異常、染色体内又は染色体間変異などを検出するために上で説明したようなシグネチャー参照配列が、次の

方法ステップを実行するために作られる。
（１）総挿入、総欠失、染色体上で説明した染色体異常、染色体内又は染色体間変異などに対応するシグネチャーのリストが作られ得る。
（２）上記変異配列が、上記染色体変異の情報により与えられる。さらに、上記染色体の情報、上記変異の説明及び／又は識別コードが与えられる。
（３）上記変異配列の逆相補的配列が生成され得る。
【０１０４】
上記シグネチャー参照配列及び対象から得られるゲノム配列との整列は、好適な整列方法又は技術により実施され得る。かかる方法の例は好適な刊行物、特に、Ｌｉ Ｈ．及びＤｕｒｂｉｎ Ｒ．の「Ｆａｓｔ ａｎｄ ａｃｃｕｒａｔｅ ｓｈｏｒｔ ｒｅａｄ ａｌｉｇｎｍｅｎｔ ｗｉｔｈ Ｂｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ ｔｒａｎｓｆｏｒｍ（Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ、２５、１７５４－６０［ＰＭＩＤ：１９４５１１６８］２００９」；又はＬｉ及びＤｕｒｂｉｎ Ｒ．の「Ｆａｓｔ ａｎｄ ａｃｃｕｒａｔｅ ｌｏｎｇ－ｒｅａｄ ａｌｉｇｎｍｅｎｔ ｗｉｔｈ Ｂｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ ｔｒａｎｓｆｏｒｍ（Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ、２６；５８９－９５［ＰＭＩＤ：２００８０５０５］、２０１０」から導かれ、これらの内容は参照されて本明細書に援用される。
【０１０５】
好ましくは、上記整列は、逆相補的配列を用いることで実施される。これらの配列は、ここで説明した方法によるここで説明された又は与えられたシグネチャー参照配列にすでに存在し得る。従って、特に好ましくは、逆相補的配列を含むシグネチャー参照配列を用いることである。任意の逆相補的計算をバイパスすることで、分析時間が大きく低減され、本発明のさらなる利点を構成する。
【０１０６】
本発明のさらなる実施態様では、ここで説明した方法によるゲノム配列情報を、例えば上記配列をここで定めたシグネチャー参照配列と整列又は比較することで低減することは、続いて迅速に検索可能な形で保存され、例えばデータベースエントリーの形、好ましくは差ＤＮＡ保存構造（ＤＤＳＳ）フォーマット又はその誘導フォーマットで保存され得る。
【０１０７】
本発明の他の好ましい実施態様では、対象のゲノムデータを処理するための方法はさらに対象の機能的遺伝子情報を分析するステップを含む。好ましくは、上記方法は、対象の機能的遺伝子情報を得るステップ、この情報の複雑性又は量を低減するステップ及び上記機能的遺伝子情報を迅速に検索可能な形で保存するステップを含む。ここで使用される用語「機能的遺伝子情報」とは、上記プライマリ配列又は遺伝子配列の生物／生化学的機能を意味するか示唆する任意のタイプの分子データを意味する。機能的遺伝子情報は従って、特に、（ｉ）遺伝子発現の情報及び／又は、（ｉｉ）メチル化配列情報、好ましくはこのヌクレオチド（Ｃ又はＡ）のメチル化配列情報；及び／又は、（ｉｉｉ）活性遺伝子及び／又はサイレント遺伝子、好ましくはＨ３Ｋ４メチル化及び／又はＨ３Ｋ２７メチル化を示し得るヒストンマークの情報である。さらなる機能的情報は、変異に関連し、例えばタンパク質機能を変化させ及び／又は非コードＲＮＡの部分として制御的影響を持つ一塩基変異多型、又は患者の機能に伴い及び／又は非コードＲＮＡの部分としての制御的影響を持つ、増幅遺伝子又は削除遺伝子及び非コードＲＮＡとしてのコピー数変異である。
【０１０８】
本発明の特に好ましい実施態様では、対象のゲノムデータを処理するための方法はさらに、対象の遺伝子発現を分析するステップを含む。例えば、上記方法は、対象の遺伝子発現の情報を得るステップ、この情報の複雑性又は量を低減するステップ及び上記遺伝子発現情報を迅速に検索可能な形で保存するステップを含む。ここで用語「遺伝子発現」とは、遺伝子又は遺伝子要素の転写、翻訳及び／又は翻訳後変性に関する情報の任意のタイプに関連する。好ましくは、遺伝子発現の情報は、１以上のＲＮＡ種の存在又は不存在の情

報、１以上のタンパク質種の存在又は不存在の情報、対象のトランスクリプトームの情報、対象のプロテオームの情報又は対象のトランスクリプトーム又はプロテオームの部分の情報を含む。遺伝子発現データは、当業者に知られる全ての好適な方法により得ることが可能であり、例えば、マイクロアレイ分析、ＰＣＲ実施、特に定量的ＰＣＲ分析により、タンパク質検出アッセイ、２Ｄゲル電気泳動法、３Ｄゲル電気泳動法などで可能である。さらに好適な技術は、当業者に知られているか、適切な教科書から導かれ得る。対応する試験は、対象から誘導されるサンプルで、例えばここで定められたサンプルで実施され得る。好ましくは、上記ゲノム配列の取得のために使用されるサンプルと同じサンプル、又は同じ時間に及び／又は同じ場所又は位置で、同じ臓器、組織又は組織型で取得されたサンプルが、対象の遺伝子発現の分析のために使用され得る。又は遺伝子発現データはまた、情報リポジトリ、例えば疾患タイプ、性別、年齢群などに関連する対象の状態に関連する具体的な条件下で遺伝子発現パターンの情報を提供するデータベースから誘導することができる。さらに対象について得られる遺伝子発現データは、比較され、標準化され及び／又は、情報リポジトリ又は好適なデータベースから得られる情報に標準を用いて訂正され得る。

【０１０９】

さらに好ましい実施態様では、上記機能的遺伝子情報、例えば遺伝子発現の情報の複雑性及び／又は量が低減され得る。この低減手順は好ましくは、機能的遺伝子情報、例えば遺伝子発現情報を切り取ることで実施される。ここで用語「機能的遺伝子情報を切り取る」及び「遺伝子情報を切り取る」とは、利用可能な機能的遺伝子情報又は遺伝子発現情報の特定のパラメータに集中する手順を意味する。例えば、機能的遺伝子情報は、特定の遺伝子、遺伝子要素、生化学的経路の成分、特定の領域のメチル化、特定の制御的要素、特定の領域での特定の塩基などの情報に低減されることが可能である。同様に、遺伝子発現情報は、特定の遺伝子、特定の遺伝子要素、又は領域の発現、又は生化学的経路の成分の発現、転写因子、成長因子などによる上記経路の活性化の反応での発現の情報に低減され得る。好ましくは、上記機能的遺伝子情報及び特に遺伝子発現情報は、疾患又は障害に関連するシグネチャーデータへ低減され得る。例えば、機能的遺伝子情報、例えば特定の癌疾患に関連するとして知られる情報について以外の遺伝子発現情報を切り取ることが可能である。従って、例えばかかる疾患に関連するメチル化パターン又は発現パターンに関する従来技術から知られる情報に基づき、この観点から関連するマーカーの例えばＲＮＡ種、タンパク質種などの存在又は不存在などが決定される。

【０１１０】

加えて、対象の状態のさらなるパラメータ、例えば組織学的パラメータ、細胞サイズに関連するパラメータ、疾患などについて知られたタンパク質スコアに関するパラメータを決定され得る。

【０１１１】

本発明のさらなる実施態様では、対象の遺伝子発現の情報は、最初に得られ、続いて上記取得ステップを繰り返して得られ得る。好ましくは、対象の遺伝子発現情報の取得は、１回、２回、３回、４回、５回、又は６回以上繰り返され得る。上記第２の又はそれ以上の取得は、ある時間経過後、例えば１週間後、２週間後、３週間後、４週間後、２、３、４、５、６、７、８、９、１０、１１、１２ヶ月後、１．５年、２年、３年、４年、５年、６年後など、又はその期間よりも長い期間後、又はこれらの期間の任意の期間で取得され得る。対象のゲノム配列の１回目と２回目の取得期間、及び２回目と続く取得との期間は同じ、本質的に同じであってよく、又は例えばそれ以上又は以下の異なる期間であってよい。例えば、治療モニター期間では、対象の遺伝子発現情報が、等間隔又はより長い又はより短い期間で取得され得る。好ましくは、対象の遺伝子発現情報の取得は、対象のゲノム配列の取得と調整され又は協調してなされる。好ましくは、対象のゲノム配列の取得及び対象の遺伝子発現情報の取得は本質的に同時になされる。

【０１１２】

対象の遺伝子発現情報が、最初の取得後第２回目又はそれ以降で得られるか、又は１以

上の遺伝子発現情報の組み、例えば異なる組織や組織型で同時に与えられると、例えば最初の取得で得られた遺伝子発現情報と、第２回目又はそれ以降で得られた遺伝子発現情報間での比較がなされる。好ましくは、かかる比較は、上記最初に得られた遺伝子発現情報と続いて得られた遺伝子発現情報間の、又は異なる位置、臓器、組織、細胞などで得られた遺伝子発現情報間の変化、変性又は差を明らかにするために実施される。ここで「比較」とは、発現データを整合させる全ての好適な方法や技術を意味する。通常は、当業者に知られるクラスタアルゴリズムが適用され得る。かかるアルゴリズムの例は、階層クラスタ化又はｋ－平均クラスタ化を含む。さらなる例は、好適な刊行物から得られ、例えばＡ．Ｋ．Ｊａｉｎ及びＲ．Ｃ．Ｄｕｂｅｓの、「Ａｌｇｏｒｉｔｈｍｓ ｆｏｒ Ｃｌｕｓｔｅｒｉｎｇ Ｄａｔａ、Ｐｒｅｎｔｉｃｅ Ｈａｌｌ、１９８８」であり、この内容は参照されて本明細書に援用される。

【０１１３】

好ましい実施態様では、比較は、連続する機能的遺伝子情報の組みの間で実施され、特に、遺伝子発現情報について行われ、例えば機能的遺伝子情報間、例えば最初に得られた及び上記情報取得の第１回目の繰り返しで得られた遺伝子発現情報間での比較である。

【０１１４】

特に好ましい実施態様では、対象の機能的遺伝子情報、例えば対象の遺伝子発現情報が、第２回目又はそれ以降で得られた場合に、既に保存されている機能的遺伝子情報、例えば既に保存されている遺伝子発現情報との比較で増加されたデータが保存される。従って、２つの組みの機能的遺伝子情報間、例えば遺伝子発現情報間で変化した又は異なる情報が保存され得る。

【０１１５】

具体的な実施態様では、例えば対象の遺伝子発現情報が２回以上得られた場合、上記データが第２回目につき提示される際に、遺伝子発現データでの変化が識別され（即ち、E^2及びE^1との差）、及び上記変化した部分のみが保存される（δE^2）。遺伝子発現データが、ｎ番目（n^{th}）時間（E^n）につき得られる場合、以前の遺伝子データ（E^{n-1}）は次の形で再構成され得る。

【０１１６】

【数２】

$$E^{n-1}=E^1+\sum_{i=2}^{n-1}\delta E^i$$

E^n及びE^{n-1}間の変化が検出されると、δE^nとして保存される。かかる手順の利点は、機能的遺伝子情報、特に遺伝子発現情報を保存するために必要なメモリと保存空間が大きく低減され得る、ということである。

【０１１７】

本発明のさらなる実施態様では、ここで説明する対象の遺伝子発現などの対象の機能的遺伝子情報の情報は、（ｉ）上記ゲノム配列の情報と共に保存される、及び／又は（ｉｉ）上記ゲノム配列の情報とリンクされて保存させるかである。特に好ましくは、両方の情報の組みを組み合わせるステップであり、例えばゲノム配列情報と機能的遺伝子情報の情報であり、例えば遺伝子発現情報は特定の疾患や障害に集中された情報であり、これにより対象の健康状態を相互に影響する上記データの解釈により判断することを可能にする。

【０１１８】

さらに、時間を経過して増加したデータを取得することで、機能的遺伝子変異の進行経路、特にゲノム配列に状況に依存して遺伝子発現の進行が観察され得ることであり、例えば疾患治療の間、疾患が進行している間などである。この情報の組合せは、対象の治療への応答、疾患の進展、対象の見通しについてより詳細な判断を可能にするという利点を提供する。

【０１１９】

他の側面で本発明は、ここで説明される本発明の方法により、取得され、処理され及び

／又は保存されたゲノム配列情報を、疾患の診断、検出、モニター又は予後のために使用することに関する。具体的な実施態様では、ここで説明される本発明の方法により、取得され、処理され及び／又は保存されたゲノム配列情報を、機能的遺伝子情報、特にここで説明される本発明の方法により、取得され、処理され及び／又は保存された遺伝子発現情報と組み合わせることで、疾患の診断、検出、モニター又は予後のために使用することに関する。

【０１２０】

ここで用語「疾患を診断」とは、最初に得られたゲノム配列情報が、対象の遺伝子状態につき通常の既定の状態とは異なる場合に対象がある疾患を患っていると考えられことを意味する。「対象の遺伝子状態につき通常の既定の状態」とは、従来技術の知識、又は１以上の特定の遺伝子及び／又は機能的遺伝子状態、例えば遺伝子発現状態に基づき、健康であると考えられ、一方上記状態からの変化が疾患に関連すると仮定される、ことを意味する。用語「診断」はまた、かかる比較プロセスを通じて到達される結論を意味する。

【０１２１】

ここで使用される用語「疾患検出」とは、対象の疾患又は障害が、器官で識別され得ることを意味する。疾患又は障害の判断及び識別は、ゲノム配列変性の決定により達成され得る。より好ましくは、上記疾患又は障害の判断又は識別は、ゲノム配列の変性及び機能的遺伝子変化、例えばここで説明した遺伝子発現変化を決定することで達成され得る。

【０１２２】

ここで使用する用語「疾患をモニターする」とは、診断された又は検出された疾患又は障害に伴い、例えば治療手順の間、又はある期間、通常は１日、２日、５日、１週間、２週間、４週間、２ヶ月、３ヶ月、４ヶ月、５ヶ月、６ヶ月、１年、２年、３年、５年、１０年又はそれ以上の期間行われる。用語「伴い」とは、疾患のこれらの状態及び特に状態の変化が、本発明の方法により得られる増加情報に基づき又は対応するデータベース値に基づき、任意の時間周期間隔で検出され得ることを意味し、例えば毎週、２週間毎、毎月、２、３、４、５、６、７、８、９１９、１１１２ヶ月毎、１．５年毎、２、３、４、５、６、７、８、９、１０年毎、任意の期間例えばそれぞれ２週間、３週間、１、２、３、４、５、６、７、８、９、１０、１１、１２ヶ月、１．５年、２、３、４、５、６、７、８、９、１０、１５、２０年間である。

【０１２３】

ここで使用される用語「疾患予後」とは、診断され検出された疾患の進展又は結果の予想を意味し、例えばある期間の間、治療の間又は治療後などである。上記用語はまた、上記疾患から生存又は回復の機会を決定することを意味し、同様に対象の予想生存時間の予想を意味する。予後は、特に、対象の将来の生存の可能性の期間を含み、例えば６ヶ月、１年、２年、３年５年、１０年又は任意の期間である。

【０１２４】

好ましくは、疾患の情報、例えば診断又は予後情報は迅速に検索可能な形で保存され得る。

【０１２５】

他の実施態様では、本発明は、ここで記載された方法を、対象の分子履歴又は上記分子履歴に記録化に使用することを含む。ここで使用される用語「分子履歴」とは、上記全ゲノムの機能的側面を捕捉すること、又はここで記載されるサブ部分の捕捉、又は上記レグローム（ｒｅｇｕｌｏｍｅ）又は上記ゲノム、ゲノム領域、遺伝子、プロモーター、イントロン、エクソン、経路、経路成分、メチル化状態など既定の期間にわたる制御状態の捕捉を意味する。上記履歴は、他の実施態様ではまた、種々の分子プロファイルモダリティを含む。好ましい実施態様では、上記分子履歴は、以下の時間間隔で生成され、例えば１から７日、例えば１、２、３、４、５、６、７、８、９、１０週間などの週、例えば１、２、３、４、５、６、７、８、９、１０、１１、１２ヶ月などの月、又は例えば１、２、３、４、５、６、７、８、９、１０、１５、２０、２５間年などの年である。ここで記載される全ゲノム又はその部分、又はレギュローム、又は上記ゲノム、ゲノム領域、遺伝子

、プロモーター、イントロン、エクソン、経路、経路成分、メチル化状態の制御状態、の機能的側面同じくそれらの変化が、任意の好適な時間間隔で捕捉され得る、例えば１から７日、１、２、３、４、５、６、７、８、９、１０週間、１、２、３、４、５、６、７、８、９、１０、１１、１２ヶ月間、１、２、３、４、５、６、７、８、９、１０年間などである。上記捕捉はまた、非定期的に実施され、例えば患者が医師又はゲノム専門家を訪れる際である。分子履歴は、迅速に検索可能な、容易にアクセス可能な形で提供されることが有利である。好ましくは、１つの疾患又は限られた群の疾患に関連する特定の分子シグネチャーに集中したフォーマットである。この情報は、さらなる実施態様では、また疾患とは直接は関連しないが、対象の健康状態の情報を提供する他の臨床的指標とリンクされ得る。

【０１２６】

本発明により判断され、検出され、診断され、モニターされ又は予後される疾患又は障害は、当業者に知られる全ての検出可能な疾患であり得る。特に好ましい実施態様では、上記疾患は遺伝子疾患又は障害、であり、特にゲノム配列情報の基づき検出され得る遺伝子障害である。かかる障害には、限定されるものではないが、上記障害を含み、例えば好適な科学文献、臨床又は医学刊行物、高い品質の教科書、公開情報リポジトリ、インターネットソース又はデータベースが含まれ、「ｈｔｔｐ：／／ｅｎ．ｗｉｋｉｐｅｄｉａ．ｏｒｇ／ｗｉｋｉ／Ｌｉｓｔ＿ｏｆ＿ｇｅｎｅｔｉｃ＿ｄｉｓｏｒｄｅｒｓ」で検索されるものが含まれる。

【０１２７】

本発明の特に好ましい実施態様では、上記疾患は癌性疾患であり、例えば当業者に知られる癌疾患又は腫瘍である。

【０１２８】

他の側面では、本発明は臨床判断サポート及び保存システムに関連し、対象のゲノム配列情報を与えるための入力及びその機能的読み出しを含み、例えば遺伝子、又は非コードＲＮＡ発現、又はタンパク質レベルであり；コンピュータプログラム製品を含み、これはプロセッサーに、ここで定義されたゲノム配列情報の複雑性及び／又は量を低減するステップを実行させ、対象のゲノム変異、増加ゲノム変異又は遺伝子発現変化パターンを出力するために出力を含み、及び上記出力された情報を保存する媒体を含む。具体的な実施態様では、上記臨床判断サポート及び保存システムは、対象のゲノム配列情報を対象の遺伝子発現情報と組み合わせて提供するための入力を持ち；コンピュータプログラム製品を含み、これはプロセッサーに、上記ゲノム配列情報の複雑性及び／又は量を低減するステップを実行させ、及びここで定めた上記対象の遺伝子発現情報の複雑性及び／又は量を低減するステップを実行させ、対象のゲノム変化、増加ゲノム変化又は遺伝子発現変化パターンを出力するための出力を含み、及び上記出力された情報を保存する媒体を含む。

【０１２９】

具体的な実施態様では、上記臨床判断サポート及び保存システムは、分子腫瘍学判断ワークステーションであり、好ましくは上記人又は患者の分子履歴を捕捉する時系列データであり得る。上記判断ワークステーションは、好ましくは、対象について癌治療を開始する及び／又は継続するかどうかにつき判断するために使用される。より好ましくは、上記判断ワークステーションは、治療の反応性の確率及び可能性について判断するために使用され得る。さらに、異なるタイプの疾患、例えば上で説明した疾患のいずれについても、同様の判断ワークステーションが想定される。

【０１３０】

さらなる実施態様では、本発明はまた、ここで説明した判断ワークステーションで使用されるソフトウェア又はコンピュータプログラムが含まれる。上記ソフトウェアは、ひとつの実施態様では、ここで説明したゲノム配列情報の分析に基づく。例えば、上記ソフトウェアは、ここで説明したゲノム配列情報の複雑性及び／又は量を低減するための方法ステップを実行し得る。さらなる実施態様では、上記ソフトウェアはさらに、ここで説明した遺伝子発現情報の複雑性及び／又は量を低減する方法ステップを実行し得る。なお他の

実施態様では、上記ソフトウェアはここで説明したシグネチャー参照配列に基づき比較のステップを実行し得る。他の実施態様では、上記ソフトウェアは、対象の分子履歴の記録化を実行し得る。

【０１３１】

出力される結果データは、従って、任意の好適な方法又はフォーマットで、好ましくは、（１）階層的及び／又は（２）時間情報をエンコードし、及び／又はさらに（３）患者データ、画像、報告などをリンクする保存構造で保存され得る。さらに好ましくは、保存構造が差ＤＮＡ保存構造（ＤＤＳＳ）としてである。

【０１３２】

なお他の具体的な本発明の実施態様では、上記臨床判断サポート及び保存システムは電子画像／データ取り出し及び通信システムである。かかる電子画像／データ保存記録及び通信システムの例は、ＰＡＣＳシステムである。特に好ましくは、ｉＳｉｔｅＰＡＣＳシステムであり、Ｐｈｉｌｉｐｓ社から提供される。これらのシステムは、本発明の方法の要求に適合させるため及び／又はここで記載されたコンピュータプログラム又はアルゴリズムを実行させるため、及び／又はここで説明したゲノム配列情報及び／又は機能的遺伝情報を保存するために、調節又は変更することが可能である。

【０１３３】

以下の実施例及び図面は、説明目的で与えられる。従って、理解されるべきことは、実施例及び図面は、なんらを限定するものではない、ということである。当業者が、ここで説明した原理のさらなる変更を想定することができることは明らかである。

【実施例】

【０１３４】

実施例１：　整列パラメータの比較

整列アルゴリズムで設定される現在の限界は通常は最大５ミスマッチ（例えば置換、ギャップ）及び最大３挿入又は削除である。一般的に２ｂｐミスマッチは、上記メモリ／プロセッサー利用及び実行時間を最適化するためのデフォルト入力パラメータとして使用される。目標の数がないとこれを超えるパラメータが膨大化する。しかし、これは、我々がより大きい挿入及び削除を検索する際に必要となるよりもずっと少ない。どのくらいの数のリードマッチ及び変異が、上記ＲｅｆＳｅｑから呼ばれるかは、直接表１に示される入力パラメータに比例する。表１は、それぞれ２ｂｐ及び３ｂｐミスマッチを用いるマウスｃｈｒ１９の１１ＭＲＮＡ－Ｓｅｑリードを示す。ここで、３ｂｐマッピングは、１８．５％より特異的なマップ化リードを与え、かつその４２％が従来のＲｅｆＳｅｑ遺伝子で注釈される転写領域内にあり、上記ゲノムの僅か２から３％を占めるにすぎないことが示される。

【０１３５】

表１：許容される異なるミスマッチを含むＲｅｆＳｅｑへのリード整列。

【表１】

マッピングパラメータ	特異的マップリード	転写領域へマップされたリード
２ｂｐミスマッチ	３０８０９５	１９５９８６
３ｂｐミスマッチ	３６５１７２	２２００５０

【０１３６】

本発明で説明したように、より小さい疾患／適用特異的焦点化参照配列（例えば、ｓＲｅｆＳｅｑＩ、ｓＲｅｆＳｅｑＩＩ、ｓＲｅｆＳｅｑＩＩＩ）を用いて、ミスマッチ及び

インデルの数が増加され、それによって、より大きなゲノム変異を検出可能となり、高い臨床的重要性を持つ。

【０１３７】

実施例２：治療への患者反応の経時的モニター

本発明の方法により得られる増加情報は、患者の治療への経時的反応をモニターするために使用され得る。患者が治療を開始した後計算される上記δＧｓが、どの程度迅速に彼／彼女が治療へ反応するかを見るようにチェックされ得る。上記変化が最小の場合、次に患者は、$Ｇ^{n}$が$Ｇ^{1}$に等しい場合、完全に回復したか、治療に十分反応していないかであり、いずれの場合も代わりの治療を適用されるべきである。

【０１３８】

実施例３：疾患傾向の予想

上記増加情報はまた、上記疾患の予想と同様に追跡するために使用され、疾患（例えば癌）の診断及び段階を知るために使用され得る。例えば、特定の疾患を患う患者の上記δＧｓ（診断相）が利用可能であれば、それらは上記疾患の進展の際のキーとなる遺伝子変化を検出するために使用され得る。この情報は、他の患者での上記疾患の初期発症を検出するために使用され得る。また、これらは疾患が進行する人の遺伝子的構造の影響を識別するために使用され得る。例えば、正常なプロファイル（図６）を持つ癌患者において、患者が結腸直腸癌を持つとして診断される変化が検出される。化学療法及び放射線治療を行った結果、上記疾患が診断される前の正常なプロファイルと非常に近いプロファイルが得られ得る。上記行列の値は、ＲＮＡシグナルのレベルを表し得る（遺伝子発現データー又は遺伝子コピー数多型の値）。

【０１３９】

上記疾患の進展の間は、図６に与えられるデータをさらに加える複数の分子データが関連するようになる。例えば、治療の全反応を見るために、それぞれの薬物治療の後３日連続して実験することがあり得る。それぞれの時点で、通常の診断画像（例えばＭＲＩ）が取得され、差分データが経時的に保存され得る。

【０１４０】

図６では、疾患進展段階で、６つの値が劇的に変化し、ついで治療後これらの値の３つが正常値に戻り、残る３つは最初の値に近くなる。従って、分子履歴保存では、$δＧ^{2}$は６つの値を持ち、$δＧ^{3}$が３つの値を持ち得る。上記$δＧ^{2}$は、上記疾患のこの段階での既知のプロファイルに対してマッチされるプロファイルを表す。実際の実験では、多くの数、例えば３１６４．７百万の化学的ヌクレオチド塩基（Ａ、Ｃ、Ｔ及びＧ）であり得る。

【０１４１】

実施例４：疾患の進展速度

患者は、疾患の進展の間、いくつかの遺伝子試験を受け得る。より短時間差で行われた２回の連続する試験の間の変化は最小であるが、なお、疾患の進行の速度に関する臨床情報を提供し得る。図７は、図６で与えられる例の疾患の進行の間の遺伝子コピー数（ＧＣＮ）での変異を示す。δＧｓの数は３であり、２と１はそれぞれ種々の段階を示す。例えば、Ｔｊａｄｅｎらの「Ａｐｐｌｉｅｄ Ｍｙｃｏｌｏｇｙ ａｎｄ Ｂｉｏｔｅｃｈｎｏｌｏｇｙ：Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ、６、２００６」の技術が上記増加データを分析するために適用され得る。例えば、同じ疾患を患う種々の患者の上記増加データが、上記疾患の発症から等しい時間例で利用可能であれば、ｋ－平均方法を用いて上記疾患の進行の速度に基づく種々のクラスにクラスタ化し得る。新たな患者の増加データが表される場合には、上記ｋ－平均（又は重心）と比較され、進行速度が推定され得る。これにより上記患者に対する適切な治療を選択することの助けとなる。それぞれのクラスタを用いて、患者のカテゴリを関連付けができ、例えば：「薬物療法に反応性」であると関連付けされる場合は、このクラスタは、「薬物療法に反応しない」クラスタに対してより初期のクラスタ（健康状態）に近く、即ちδＧｓの値が「健康」クラスタでの行列よりもさらに高いことになる。