

(19) 世界知的所有権機関
国際事務局(43) 国際公開日
2006年6月8日 (08.06.2006)

PCT

(10) 国際公開番号
WO 2006/059770 A1

- (51) 国際特許分類:
G06F 19/00 (2006.01)
- (21) 国際出願番号: PCT/JP2005/022317
- (22) 国際出願日: 2005年12月5日 (05.12.2005)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願2004-351839 2004年12月3日 (03.12.2004) JP

(71) 出願人 (米国を除く全ての指定国について): 独立行政法人理化学研究所 (RIKEN) [JP/JP]; 〒3510198 埼玉県和光市広沢 2 番 1 号 Saitama (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 鎌谷 直之 (KAMATANI, Naoyuki) [JP/JP]; 〒2300045 神奈川県横浜

市鶴見区末広町一丁目 7 番 2 2 号 独立行政法人理化学研究所横浜研究所内 Kanagawa (JP). 山崎 敏正 (YAMAZAKI, Toshimasa) [JP/JP]; 〒2300045 神奈川県横浜市鶴見区末広町一丁目 7 番 2 2 号 独立行政法人理化学研究所横浜研究所内 Kanagawa (JP). 柳澤 政生 (YANAGISAWA, Masao) [JP/JP]; 〒1698555 東京都新宿区大久保 3 - 4 - 1 学校法人早稲田大学理工学部内 Tokyo (JP).

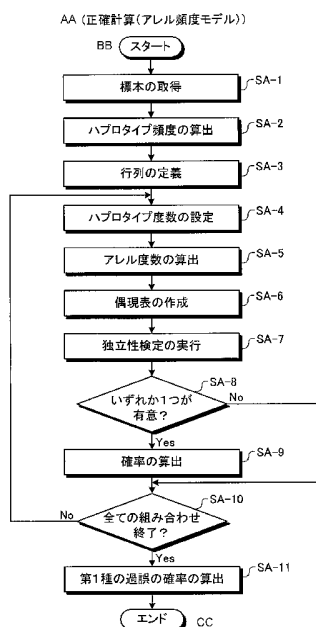
(74) 代理人: 酒井 宏明 (SAKAI, Hiroaki); 〒1006019 東京都千代田区霞が関三丁目 2 番 5 号 霞が関ビルディング 酒井国際特許事務所 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO,

(続葉有)

(54) Title: PROBABILITY CALCULATING METHOD

(54) 発明の名称: 確率算出方法



AA... ACCURACY CALCULATION (ALLELE FREQUENCY MODEL)
BB... START
SA-1... OBTAIN SPECIMEN
SA-2... CALCULATE HAPLOTYPE FREQUENCY
SA-3... DEFINE MATRIX
SA-4... SET HAPLOTYPE FREQUENCY
SA-5... CALCULATE ALLELE FREQUENCY
SA-6... MAKE FORTUITOUS OCCURRENCE TABLE
SA-7... PERFORM INDEPENDENCE TEST
SA-8... ANY ONE SIGNIFICANT?
SA-9... CALCULATE PROBABILITY
SA-10... ALL COMBINATIONS ENDED?
SA-11... CALCULATE PROBABILITY OF TYPE-1 ERROR
CC... END

(57) Abstract: A probability calculating method for directly calculating the probability of type-1 error in consideration of the haplotype frequency thereby to prevent oversight of disease-related SNP. In the method, (1) a specimen is obtained, (2) the haplotype frequency is calculated, (3) a matrix composed of components which are predetermined values corresponding to the alleles is defined, (4) a combination of the haplotype frequencies is given, (5) the allele frequency at each locus is calculated, (6) the fortuitous occurrence table at each locus is made, (7) the independence test at each locus in the fortuitous occurrence table is carried out, (8) the probability corresponding to the combination of the haplotype frequencies is calculated using the joint distribution if any one of the results of the independence tests are significant, steps (4), (5), (6), (7), (8) are repeated for all the combinations of the haplotype frequencies, and the probabilities obtained by the repetition are added, thereby calculating the probability of type-1 error.

(57) 要約: 第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、その結果、疾患関連SNPの見落としを防ぐことができる確率算出方法を提供することを課題とする。本発明にかかる確率算出方法は、(1) 標本を取得し、(2) ハプロタイプ頻度を算出し、(3) 各アレルに対応する所定値を成分とする行列を定義し、(4) ハプロタイプ度数の組合せを与え、(5) 各座位でのアレル度数を算出し、(6) 各座位に対応する偶現表を作成し、(7) 偶現表につき、座位ごとに独立性検定を実行し、(8) 独立性検定の検定結果のいずれか1つが有意であった場合、ハプロタイプ度数の組合せに対応する確率を結合分布を用いて算出する。そして、(4)、(5)、(6)、(7) および (8) を、ハプロタイプ度数の全ての組合せについて繰り返し、繰り返すごとに (8) で算出される確率を加算することで、第1種の過誤の確率を算出する。



RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR,
TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML,
MR, NE, SN, TD, TG).

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR),

添付公開書類:

— 国際調査報告書

2文字コード及び他の略語については、定期発行される各PCTガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

明 細 書

確率算出方法

技術分野

- [0001] 本発明は、SNP(Single Nucleotide Polymorphism)を用いたcase-control相関解析で行う独立性検定における第1種の過誤の確率(有意水準)を算出する確率算出方法に関するものである。

背景技術

- [0002] case-control研究では、ある疾患に罹った者またはある病態にある者(case)とそうでない者(control)とがあらかじめどのような危険因子にさらされたかを調べる。SNPを用いたcase-control相関解析では、当該危険因子として、(1)アレル頻度モデル(A/a)、(2)優性遺伝モデル($AA+Aa/aa$)、(3)劣性遺伝モデル($AA/Aa+aa$)、(4)遺伝子型モデル($AA/Aa/aa$)、を想定する。なお、 A はメジャーアレル(major allele)、 a はマイナーアレル(minor allele)を示す。そして、各モデルについて、 2×2 または 2×3 の偶現表を作成し、独立性検定を行う。
- [0003] しかし、上述したモデル間で共通したデータを使っているので、多重検定の偶然の影響を除いた上で依然として検定の統計的有意差が認められるかが課題となる。
- [0004] そこで、多重検定の影響を除外するための方法として、まず、Bonferroniの補正の方法が知られている。Bonferroniの補正の方法では、独立性検定で得られたp値に独立性検定の回数(上述した4つのモデルの場合では4回)を乗じた値を多重検定全体から得られたp値とみなすことで、多重検定の影響を除外している。この方法は、検定結果を最も保守的(厳格)に解釈するものである。また、多重検定の影響を除外するための方法として、つぎに、非特許文献1に記載の方法が知られている。非特許文献1に記載の方法では、少数個のSNPを対象として各2つのSNP間の連鎖不平衡指標値(Δ)を計算し当該指標値を成分とする行列の固有値分解を行うことで近似的な補正を施している。つまり、非特許文献1に記載の方法は連鎖不平衡に着目した多重検定におけるp値の補正の方法である。
- [0005] 非特許文献1:Dale R. Nyholt, A simple correction for multiple test

ing for single-nucleotide polymorphisms in linkage disequilibrium with each other, Am. J. Hum. Genet., Vol. 74, pp. 765-769

発明の開示

発明が解決しようとする課題

[0006] しかしながら、Bonferroniの補正の方法では、補正されたp値が真のp値より大きい可能性が高いので、疾患関連SNPを見落とす危険性がある、という問題点があった。また、大量のSNPデータを対象とした場合、多重検定全体のp値を一般的に用いられている例えば0.01や0.05などに設定すると、設定したp値を許容するためには各SNPのp値が非常に小さな値、つまり非常に厳しい条件となるので、同様に疾患関連SNPを見落とす危険性がある、という問題点があった。例えば、50000個のSNPデータを対象として、多重検定全体のp値を0.01に設定した場合、各SNPのp値は0.0000002 ($0.01 \div 50000$)と非常に小さな値となる。

[0007] また、非特許文献1に記載の方法では、ペアワイズ(pairwise)な連鎖不平衡に着目しているので、ハプロタイプ頻度が考慮されていない、という問題点があった。さらに、直接的なp値を算出していない、という問題点があった。

[0008] 本発明は上記問題点に鑑みてなされたもので、SNPを用いたcase-control相関解析で行う独立性検定におけるp値、つまり第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、その結果、疾患関連SNPの見落としを防ぐことができる確率算出方法を提供することを目的とする。

課題を解決するための手段

[0009] 上記目的を達成するために、本発明にかかる請求項1に記載の確率算出方法は、SNPを用いたcase-control相関解析で行う独立性検定における第1種の過誤の確率を算出する確率算出方法において、複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する標本取得ステップと、前記標本取得ステップで取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出するハプロタイプ頻度算出ステップと、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた

所定値を設定することで、当該所定値を成分とする行列を定義する行列定義ステップと、第1標本における各ハプロタイプのハプロタイプ度数および第2標本における各ハプロタイプのハプロタイプ度数を設定することで、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せを与えるハプロタイプ度数設定ステップと、前記ハプロタイプ度数設定ステップで設定した第1標本に対応する複数のハプロタイプ度数および前記行列定義ステップで定義した行列に基づいて第1標本における各座位でのアレル度を算出し、設定した第2標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第2標本における各座位でのアレル度を算出するアレル度数算出ステップと、前記アレル度数算出ステップで算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する偶現表作成ステップと、前記偶現表作成ステップで作成した偶現表について、座位ごとに独立性検定を実行する独立性検定実行ステップと、前記独立性検定実行ステップで実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のハプロタイプ度数、第2標本のハプロタイプ度数およびハプロタイプ頻度に基づいて、前記ハプロタイプ度数設定ステップで設定した第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのハプロタイプ度数に関する結合分布を用いて算出する確率算出ステップと、を含み、前記ハプロタイプ度数設定ステップ、前記アレル度数算出ステップ、前記偶現表作成ステップ、前記独立性検定実行ステップおよび前記確率算出ステップを、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との全ての組合せについて繰り返し、繰り返すごとに算出される前記確率を加算することで、最終的に第1種の過誤の確率を算出すること、を特徴とする。

[0010] また、本発明にかかる請求項2に記載の確率算出方法は、SNPを用いたcase-control相関解析で行う独立性検定における第1種の過誤の確率を算出する確率算出方法において、複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する標本取得ステップと、前記標本取得ス

テップで取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出するハプロタイプ頻度算出ステップと、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する行列定義ステップと、第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与えるディプロタイプ形度数設定ステップと、前記ディプロタイプ形度数設定ステップで設定した第1標本に対応する複数のディプロタイプ形度数および前記行列定義ステップで定義した行列に基づいて第1標本における各座位でのアレル度を算出し、設定した第2標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第2標本における各座位でのアレル度を算出するアレル度数算出ステップと、前記アレル度数算出ステップで算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する偶現表作成ステップと、前記偶現表作成ステップで作成した偶現表について、座位ごとに独立性検定を実行する独立性検定実行ステップと、前記独立性検定実行ステップで実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、前記ディプロタイプ形度数設定ステップで設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する確率算出ステップと、を含み、前記ディプロタイプ形度数設定ステップ、前記アレル度数算出ステップ、前記偶現表作成ステップ、前記独立性検定実行ステップおよび前記確率算出ステップを、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せについて繰り返し、繰り返すごとに算出される前記確率を加算することで、最終的に第1種の過誤の確率を算出すること、を特徴とする。

[0011] また、本発明にかかる請求項3に記載の確率算出方法は、SNPを用いたcase-control相関解析で行う独立性検定における第1種の過誤の確率を算出する確率算出方法において、複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する標本取得ステップと、前記標本取得ステップで取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出するハプロタイプ頻度算出ステップと、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する行列定義ステップと、第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与えるディプロタイプ形度数設定ステップと、前記ディプロタイプ形度数設定ステップで設定した第1標本に対応する複数のディプロタイプ形度数および前記行列定義ステップで定義した行列に基づいて第1標本における各座位での遺伝子型度を算出し、設定した第2標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第2標本における各座位での遺伝子型度を算出する遺伝子型度数算出ステップと、前記遺伝子型度数算出ステップで算出した第1標本における各座位での遺伝子型度数および第2標本における各座位での遺伝子型度数に基づいて、各座位に対応する偶現表を作成する偶現表作成ステップと、前記偶現表作成ステップで作成した偶現表について、座位ごとに独立性検定を実行する独立性検定実行ステップと、前記独立性検定実行ステップで実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、前記ディプロタイプ形度数設定ステップで設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する確率算出ステップと、を含み、前記ディプロタイプ形度数設定ステップ、前記遺伝子型度

数算出ステップ、前記偶現表作成ステップ、前記独立性検定実行ステップおよび前記確率算出ステップを、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せについて繰り返し、繰り返すごとに算出される前記確率を加算することで、最終的に第1種の過誤の確率を算出すること、を特徴とする。

- [0012] また、本発明にかかる請求項4に記載の確率算出方法は、SNPを用いたcase-control相関解析で行う独立性検定における第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法であって、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する行列定義ステップと、ハプロタイプ度数の組合せを予め設定した第1標本および第2標本に対して生成するハプロタイプ度数生成ステップと、第1標本または第2標本のいずれかの標本を選択する標本選択ステップと、第1ハプロタイプを選択する第1ハプロタイプ選択ステップと、前記標本選択ステップで選択した標本における前記第1ハプロタイプのハプロタイプ度数が0でない場合、当該第1ハプロタイプとは異なる第2ハプロタイプを選択する第2ハプロタイプ選択ステップと、予め設定した前記第1ハプロタイプのハプロタイプ頻度および前記第2ハプロタイプのハプロタイプ頻度に基づいて、前記第1ハプロタイプのハプロタイプ度数および前記第2ハプロタイプのハプロタイプ度数を更新するハプロタイプ度数更新ステップと、前記ハプロタイプ度数更新ステップで更新した後の第1標本に対応する複数のハプロタイプ度数および前記行列定義ステップで定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、更新した後の第2標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第2標本における各座位でのアレル度数を算出するアレル度数算出ステップと、前記アレル度数算出ステップで算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する偶現表作成ステップと、前記偶現表作成ステップで作成した偶現表について、座位ごとに独立性検定を実行する独立性検定実行ステップと、前記独立性検定実行ステップで実行した独立性検定の検定結果のいずれか1つが有意であった場合、当該場合の累積回数を更新する累積回数更新ステップと、を含み、前記標本選択ステップ、前記第1ハプ

ロタイプ選択ステップ、前記第2ハプロタイプ選択ステップ、前記ハプロタイプ度数更新ステップ、前記アレル度数算出ステップ、前記偶現表作成ステップ、前記独立性検定実行ステップおよび前記累積回数更新ステップを所定回数繰り返し、前記所定回数に対する最終的な累積回数の割合を第1種の過誤の確率として算出すること、を特徴とする。

発明の効果

[0013] 本発明にかかる請求項1に記載の確率算出方法は、(1)複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得し、(2)取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出し、(3)各ハプロタイプの各座位がマイナーアレル(同じ座位で頻度の低い方のアレル)またはメジャーアレル(同じ座位で頻度の高い方のアレル)を有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義し、(4)第1標本における各ハプロタイプのハプロタイプ度数および第2標本における各ハプロタイプのハプロタイプ度数を設定することで、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せを与え、(5)設定した第1標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第1標本における各座位でのアレル度を算出し、設定した第2標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第2標本における各座位でのアレル度を算出し、(6)算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表(なお、偶現表とはcontingency tableのことであり、日本語では分割表と訳すこともある。)を作成し、(7)作成した偶現表について、座位ごとに独立性検定を実行し、(8)実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のハプロタイプ度数、第2標本のハプロタイプ度数およびハプロタイプ頻度に基づいて、設定した第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのハプロタイプ度数に関する結合分布を用いて算出する。そして、(4)、(

5)、(6)、(7)および(8)を、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との全ての組合せについて繰り返し、繰り返すごとに(8)で算出される確率を加算することで、最終的に第1種の過誤の確率を算出する。これにより、第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、その結果、疾患関連SNPの見落としを防ぐことができる、という効果を奏する。

[0014] 本発明にかかる請求項2に記載の確率算出方法は、(1)複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得し、(2)取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出し、(3)各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義し、(4)第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与え、(5)設定した第1標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、設定した第2標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第2標本における各座位でのアレル度数を算出し、(6)算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成し、(7)作成した偶現表について、座位ごとに独立性検定を実行し、(8)実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する。そして、(4)、(5)、(6)、(7)および(8)を、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せについて繰り返し、繰り返すごとに(8)で算出される確率を加算することで、最

最終的に第1種の過誤の確率を算出する。これにより、第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、その結果、疾患関連SNPの見落としを防ぐことができる、という効果を奏する。

[0015] また、本発明にかかる請求項3に記載の確率算出方法は、(1)複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得し、(2)取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出し、(3)各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義し、(4)第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与え、(5)設定した第1標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第1標本における各座位での遺伝子型度を算出し、設定した第2標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第2標本における各座位での遺伝子型度を算出し、(6)算出した第1標本における各座位での遺伝子型度数および第2標本における各座位での遺伝子型度数に基づいて、各座位に対応する偶現表を作成し、(7)作成した偶現表について、座位ごとに独立性検定を実行し、(8)実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する。そして、(4)、(5)、(6)、(7)および(8)を、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せについて繰り返し、繰り返すごとに(8)で算出される確率を加算することで、最終的に第1種の過誤の確率を算出する。これにより、第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、その結果、

疾患関連SNPの見落としを防ぐことができる、という効果を奏する。

- [0016] また、本発明にかかる請求項4に記載の確率算出方法は、(1)各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義し、(2)ハプロタイプ度数の組合せを予め設定した第1標本および第2標本に対して生成し、(3)第1標本または第2標本のいずれかの標本を選択し、(4)第1ハプロタイプを選択し、(5)選択した標本における第1ハプロタイプのハプロタイプ度数が0でない場合、当該第1ハプロタイプとは異なる第2ハプロタイプを選択し、(6)予め設定した第1ハプロタイプのハプロタイプ頻度および第2ハプロタイプのハプロタイプ頻度に基づいて、第1ハプロタイプのハプロタイプ度数および第2ハプロタイプのハプロタイプ度数を更新し、(7)更新した後の第1標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、更新した後の第2標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第2標本における各座位でのアレル度数を算出し、(8)算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成し、(9)作成した偶現表について、座位ごとに独立性検定を実行し、(10)実行した独立性検定の検定結果のいずれか1つが有意であった場合、当該場合の累積回数を更新する。そして、(3)、(4)、(5)、(6)、(7)、(8)、(9)および(10)を所定回数繰り返し、所定回数に対する最終的な累積回数の割合を第1種の過誤の確率として算出する。これにより、第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、その結果、疾患関連SNPの見落としを防ぐことができる、という効果を奏する。また、本確率算出方法では第1種の過誤の確率をMCMC法で近似的に算出しているので、計算時間を短縮することができ、実用性がより高まる、という効果を奏する。

図面の簡単な説明

- [0017] [図1]図1は、アレル頻度モデルで第1種の過誤の確率を正確に算出する確率算出方法の一例を示すフローチャートである。

[図2]図2は、優性・劣性遺伝モデルで第1種の過誤の確率を正確に算出する確率算

出方法の一例を示すフローチャートである。

[図3]図3は、遺伝子型モデルで第1種の過誤の確率を正確に算出する確率算出方法の一例を示すフローチャートである。

[図4]図4は、正確なハプロタイプ頻度がわからない場合に第1種の過誤の確率を正確に算出する確率算出方法の一例を示すフローチャートである。

[図5]図5は、アレル頻度モデルで第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法の一例を示すフローチャートである。

[図6]図6は、優性・劣性遺伝モデルで第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法の一例を示すフローチャートである。

[図7]図7は、遺伝子型モデルで第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法の一例を示すフローチャートである。

[図8]図8は、第1標本および第2標本に含まれる情報の一例を示す図である。

[図9]図9は、存在し得る全てのハプロタイプに関する情報の一例を示す図である。

[図10]図10は、存在し得る全てのハプロタイプに関するハプロタイプ頻度の一例を示す図である。

[図11]図11は、定義した行列の一例を示す図である。

[図12]図12は、ハプロタイプ度数に関する情報の一例を示す図である。

[図13]図13は、アレル頻度モデルの場合に作成する偶現表の一例を示す図である。

[図14]図14は、ディプロタイプ形番号とハプロタイプ番号との関係を示す図である。

[図15]図15は、ディプロタイプ形度数に関する情報の一例を示す図である。

[図16]図16は、劣性遺伝モデルの場合に作成する偶現表の一例を示す図である。

[図17]図17は、優性遺伝モデルの場合に作成する偶現表の一例を示す図である。

[図18]図18は、遺伝子型モデルの場合に作成する偶現表の一例を示す図である。

[図19]図19は、アレル頻度モデルの場合に作成する偶現表の一例を示す図である。

[図20]図20は、アレル頻度モデルの場合に作成する偶現表の一例を示す図である。

[図21]図21は、アレル頻度モデルの場合に作成する偶現表の一例を示す図である。

[図22]図22は、アレル頻度モデルの場合に作成する偶現表の一例を示す図である。

[図23]図23は、アレル頻度モデルの場合に作成する偶現表の一例を示す図である。

[図24]図24は、アレル頻度モデルの場合に作成する偶現表の一例を示す図である。

[図25]図25は、アレル頻度モデルの場合に作成する偶現表の一例を示す図である。

発明を実施するための最良の形態

[0018] 以下に、本発明にかかる確率算出方法の実施の形態を図面に基づいて詳細に説明する。なお、この実施の形態によりこの発明が限定されるものではない。

[0019] [第1実施形態]

ここでは、第1種の過誤の確率を正確に算出する本発明にかかる確率算出方法について、(1-1)アレル頻度モデル、(1-2)優性・劣性遺伝モデル、(1-3)遺伝子型モデル、(1-4)正確なハプロタイプ頻度がわからない場合、の順に図を参照して詳細に説明する。

[0020] (1-1)アレル頻度モデル

図1は、アレル頻度モデルで第1種の過誤の確率を正確に算出する確率算出方法の一例を示すフローチャートである。

まず、複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する(ステップSA-1)。具体的には、図8に示すように、 L (L は正の整数)個の連鎖したSNPで構成されるハプロタイプ情報を N_1 個含む第1標本および L 個の連鎖したSNPで構成されるハプロタイプ情報を N_2 個含む第2標本を取得する。

[0021] ついで、ステップSA-1で取得した第1標本に含まれるハプロタイプ情報および第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出する(ステップSA-2)。ここで、SNPの場合には2つのアレルが存在するので、存在し得るハプロタイプの総数は、具体的には、図9に示すように 2^L (以下、 M と記載することがある)個である。そして、図10に示すように、存在し得る M 個のハプロタイプのそれぞれに対して、ステップSA-1で取得した第1標本に含まれるハプロタイプ情報および第2標本に含まれるハプロタイプ情報に基づいて、ハプロタイプ頻度 h_j (j はハプロタイプを識別する番号であり、 $1 \leq j \leq M$ を満たす整数)を算出する。なお、ハプロタイプ頻度 h_j は下記数式1を満たす。

[数1]

$$\sum_{j=1}^M h_j = 1 \quad \cdots \quad (\text{数式 1})$$

[0022] ついで、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する(ステップSA-3)。具体的には、ハプロタイプ番号jのハプロタイプにおけるi番目(iは座位を識別する番号であり、 $1 \leq i \leq L$ を満たす整数)の座位がマイナーアレル(a)である場合は1を、メジャーアレル(A)である場合は0を所定値として定める。そして、定義する行列は、図11のように表すことができ、座位番号i($1 \leq i \leq L$ を満たす整数)とハプロタイプ番号jとの組合せごとに、0または1の値を格納する。

[0023] ついで、第1標本における各ハプロタイプのハプロタイプ度数および第2標本における各ハプロタイプのハプロタイプ度数を設定することで、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せを与える(ステップSA-4)。ここで、第1標本におけるハプロタイプ番号jのハプロタイプ度数 X_{1j} および第2標本におけるハプロタイプ番号jのハプロタイプ度数 X_{2j} は、ランダム変数であり、それぞれ下記数式2を満たす。そして、下記数式2を満たすように、第1標本におけるハプロタイプ番号jのハプロタイプ度数 X_{1j} の実現値 x_{1j} を全てのハプロタイプ番号jに対して設定し、第2標本におけるハプロタイプ番号jのハプロタイプ度数 X_{2j} の実現値 x_{2j} を全てのハプロタイプ番号jに対して設定する(図12参照)。これにより、 $(x_{11}, \cdots, x_{1M}, x_{21}, \cdots, x_{2M})$ の組合せが与えられた。

[数2]

$$\sum_{j=1}^M X_{1j} = 2N_1 \quad , \quad \sum_{j=1}^M X_{2j} = 2N_2 \quad \cdots \quad (\text{数式 2})$$

[0024] ついで、ステップSA-4で設定した第1標本に対応する複数のハプロタイプ度数およびステップSA-3で定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、ステップSA-4で設定した第2標本に対応する複数のハプロタイプ度数およびステップSA-3で定義した行列に基づいて第2標本における各座位でのア

レル度数を算出する(ステップSA-5)。ここで、第1標本における座位番号*i*でのマイナーアレルのアレル度数 Y_{1i} および第2標本における座位番号*i*でのマイナーアレルのアレル度数 Y_{2i} は、ランダム変数であり、それぞれ下記数式3で表される。そして、下記数式3に基づいて、第1標本における座位番号*i*でのマイナーアレルのアレル度数 Y_{1i} の実現値 y_{1i} および第2標本における座位番号*i*でのマイナーアレルのアレル度数 Y_{2i} の実現値 y_{2i} を算出する。

[数3]

$$Y_{1i} = \sum_{j=1}^M \alpha_{ij} \times X_{1j} \quad , \quad Y_{2i} = \sum_{j=1}^M \alpha_{ij} \times X_{2j} \quad \cdots \quad (\text{数式 3})$$

数式3において、 α_{ij} はステップSA-3で定義した行列である。

[0025] ついで、ステップSA-5で算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する(ステップSA-6)。具体的には、図13に示す2×2の偶現表を座位数分(L個)だけ作成する。

[0026] ついで、ステップSA-6で作成した偶現表について、座位ごとに独立性検定を実行する(ステップSA-7)。ここで、独立性検定について簡単に説明する。一般に、偶現表に対して、下記数式4で定義される X_n を算出し、算出した X_n の値が、自由度(1, 1)で有意水準 α (例えば0.01や0.05など)の χ (カイ)二乗分布値 $\chi_{1,1}^2$ 以上ならば有意とする。

[数4]

$$X_n = \sum_{i=1}^2 \sum_{j=1}^2 \frac{\left(n_{ij} - \frac{n_{i.} \times n_{.j}}{n} \right)^2}{\frac{n_{i.} \times n_{.j}}{n}} \quad \cdots \quad (\text{数式 4})$$

数式4において、*i*は標本を識別する番号である。ここでは、*i*=1の場合はケース群を示し、*i*=2の場合はコントロール群を示す。*j*はアレルを識別する番号である。こ

では、 $j=1$ の場合はマイナーアレルを示し、 $j=2$ の場合はメジャーアレルを示す。 n_{ij} はアレル度数である。 n_i は標本番号 i におけるマイナーアレル度数(n_{i1})およびメジャーアレル度数(n_{i2})の和である。 n_j はアレル番号 j に対応するケース群のアレル度数(n_{1j})およびコントロール群のアレル度数(n_{2j})の和である。 n はケース群のサンプル数とコントロール群のサンプル数の和である。具体的には、 $n = \sum_{i=1}^2 \sum_{j=1}^2 n_{ij}$ である。

[0027] ついで、ステップSA-7で実行した独立性検定の検定結果のいずれか1つが有意であった場合(ステップSA-8: Yes)、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のハプロタイプ度数、第2標本のハプロタイプ度数およびハプロタイプ頻度に基づいて、ステップSA-4で設定した第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのハプロタイプ度数に関する結合分布を用いて算出する(ステップSA-9)。具体的には、第1標本の標本数 N_1 、第2標本の標本数 N_2 、存在し得るハプロタイプの総数 $M(2^L)$ 、第1標本のハプロタイプ度数の実現値 x_{1j} 、第2標本のハプロタイプ度数の実現値 x_{2j} およびハプロタイプ頻度 h_j に基づいて、ステップSA-4で設定した第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せ($x_{11}, \dots, x_{1M}, x_{21}, \dots, x_{2M}$)に対応する確率 P を、下記数式5に示す第1標本と第2標本とのハプロタイプ度数に関する結合分布を用いて算出する。

[数5]

$$P(x_{11}, x_{12}, \dots, x_{1M}, x_{21}, x_{22}, \dots, x_{2M}) \\ = \frac{(2N_1)!(2N_2)!}{\prod_{j=1}^M \prod_{k=1}^2 x_{kj}!} \prod_{j=1}^M h_j^{\sum_{k=1}^2 x_{kj}} \quad \dots \quad (\text{数式5})$$

数式5において、 k は標本を識別する番号である。ここでは、 $k=1$ の場合は第1標本を示し、 $k=2$ の場合は第2標本を示す。

[0028] ついで、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との全ての組合せが終了したか否かを判定し、全ての組合せが終了した場合(ステップSA-10: Yes)、ステップSA-4～ステップSA-9を繰り返すごとにステップSA-9で算

出された確率を全て加算し、加算した確率を最終的に第1種の過誤の確率とする(ステップSA-11)。一方、全ての組合せが終了していない場合(ステップSA-10:No)、ステップSA-4～ステップSA-9を実行し、全ての組合せが終了するまで繰り返す。

[0029] 以上、説明したように、本発明にかかる確率算出方法は、(1)複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得し、(2)取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出し、(3)各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義し、(4)第1標本における各ハプロタイプのハプロタイプ度数および第2標本における各ハプロタイプのハプロタイプ度数を設定することで、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せを与え、(5)設定した第1標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、設定した第2標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第2標本における各座位でのアレル度数を算出し、(6)算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成し、(7)作成した偶現表について、座位ごとに独立性検定を実行し、(8)実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のハプロタイプ度数、第2標本のハプロタイプ度数およびハプロタイプ頻度に基づいて、設定した第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのハプロタイプ度数に関する結合分布を用いて算出する。そして、(4)、(5)、(6)、(7)および(8)を、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との全ての組合せについて繰り返し、繰り返すごとに(8)で算出される確率を加算することで、最終的に第1種の過誤の確率を算出する。これにより、第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、その結果

、疾患関連SNPの見落としを防ぐことができる。

[0030] ここで、ステップSA-5において、数式3に示す第1標本における座位番号iでのマイナーアレルのアレル度数 Y_{1i} および第2標本における座位番号iでのマイナーアレルのアレル度数 Y_{2i} は、座位番号iでのマイナーアレルのアレル頻度 p_i に関するハプロタイプ頻度 h_j の関数(下記数式6参照)に基づいて定められる。

[数6]

$$p_i = \sum_{j=1}^M \alpha_{ij} \times h_j \quad \cdot \cdot \cdot \quad (\text{数式 6})$$

[0031] また、ステップSA-9において、数式5に示す第1標本と第2標本とのハプロタイプ度数に関する結合分布は、下記数式7に示す第1標本のハプロタイプ度数に関する結合分布と下記数式8に示す第2標本のハプロタイプ度数に関する結合分布との積で与えられる。ここで、case-control association研究などから得られた観察データを用いて検定を行う場合、帰無仮説(H_0)として2つの標本が同一の集団から得られたことを仮定する。この2つの標本(第1標本および第2標本)におけるハプロタイプ番号jのハプロタイプ度数をそれぞれランダム変数 X_{1j} 、 X_{2j} とすると、2つの標本抽出は独立なので、ランダム変数 X_{1j} および X_{2j} はそれぞれ二項分布に従う。よって、 X_{11} 、 X_{12} 、 $\dots$ 、 X_{1M} の結合分布(joint distribution)は多項分布であり、 X_{21} 、 X_{22} 、 $\dots$ 、 X_{2M} の結合分布はそれとは独立な多項分布である。従って、ランダム変数 X_{1j} および X_{2j} の確率は、それぞれ下記数式7および下記数式8で示される。ここで、2つの標本はいずれも、同じハプロタイプ頻度(複数)を持つ集団からのものであると仮定していることに注意する(帰無仮説)。また、下記数式7および下記数式8において、 x_{1j} および x_{2j} は、それぞれランダム変数 X_{1j} および X_{2j} の実現値である。

[数7]

$$\begin{aligned}
 & P(X_{11} = x_{11}, X_{12} = x_{12}, \dots, X_{1M} = x_{1M}) \\
 &= \frac{(2N_1)!}{\prod_{j=1}^M x_{1j}!} \prod_{j=1}^M h_j^{x_{1j}} \quad \cdot \cdot \cdot \quad (\text{数式 7})
 \end{aligned}$$

[数8]

$$\begin{aligned}
 & P(X_{21} = x_{21}, X_{22} = x_{22}, \dots, X_{2M} = x_{2M}) \\
 &= \frac{(2N_2)!}{\prod_{j=1}^M x_{2j}!} \prod_{j=1}^M h_j^{x_{2j}} \quad \cdot \cdot \cdot \quad (\text{数式 8})
 \end{aligned}$$

[0032] また、集団から、第1標本として $2N_1$ 個、第2標本として $2N_2$ 個のハプロタイプコピーを抽出する実験において、ハプロタイプコピーとハプロタイプとの概念の違いは次の通りである。例えば、1つの個体に2つの同じハプロタイプ(ホモ接合体)が存在するならば、この個体は1つのハプロタイプと2つのハプロタイプコピーを保有すると考える。

[0033] また、ステップSA-11で算出した第1種の過誤の確率は、下記数式9で表現することができる。そして、下記数式9で求められる確率は、帰無仮説のもとでの確率であり、ステップSA-7で実行した独立性検定の検定結果のいずれか1つが有意となる確率を意味する。

[数9]

$$\begin{aligned}
& P[Z=1] \\
&= \sum_{x_{11}=0}^{2N_1} \sum_{x_{12}=0}^{2N_1-x_{11}} \cdots \sum_{x_{1M}=0}^{2N_1-\sum_{j=1}^{M-1} x_{1j}} \sum_{x_{21}=0}^{2N_2} \sum_{x_{22}=0}^{2N_2-x_{21}} \cdots \sum_{x_{2M}=0}^{2N_2-\sum_{j=1}^{M-1} x_{2j}} \\
& Z \times \frac{(2N_1)!(2N_2)!}{\prod_{j=1}^M \prod_{k=1}^2 x_{kj}!} \prod_{j=1}^M h_j^{\sum_{k=1}^2 x_{kj}} \\
& \quad \cdot \cdot \cdot \quad (\text{数式 9})
\end{aligned}$$

数式9において、Zは、ステップSA-7で実行した独立性検定の検定結果のいずれか1つが有意であった場合には1を、そうでなかった場合には0をとるように定義したランダム変数である。変数Zは上述したランダム変数 Y_{ki} の関数であり、さらに当該ランダム変数 Y_{ki} は上述したランダム変数 X_{kj} の関数であるので、変数Zは関数 $f(x_{11}, \cdots, x_{1M}, x_{21}, \cdots, x_{2M})$ で表すことができる。

[0034] ここで、アレル頻度モデルで第1種の過誤の確率を算出する確率算出方法において、主要なハプロタイプによる保守的な確率算出方法について説明する。

上述では $M(2^L; L$ はSNP座位数)個全てのハプロタイプについて計算を行ったが、 $M(2^L)$ 個の全てのハプロタイプが存在するとは限らない。その場合、ある程度以上の累積頻度(例えば95%)に達するまで頻度の高いハプロタイプを数え、その個数がsだとすると、残りのハプロタイプ(以下では、「その他ハプロタイプ」と記載する場合がある)を全てs+1番目のハプロタイプとして取り扱うことが適切であると思われる。ただし、その他ハプロタイプとして算出される第1種の過誤の確率の誤差を考慮するべきである。

[0035] 具体的には、主要なハプロタイプによる保守的な確率算出方法は以下の手順で行われる。

(1) 累積頻度の高いハプロタイプから加え、s個のハプロタイプを選択する。ここで、累積度数を p_s とする。その他ハプロタイプの頻度は $1-p_s$ である。また、i番目のハプロタイプ頻度を h_i ($i=1 \sim s$)とし、 h_{s+1} を $1-\sum_{i=1}^s h_i$ とする。

(2) 1～s番目のハプロタイプについて、i番目のハプロタイプのk番目の座位がマイナーアレルであれば1、メジャーアレルであれば0となる変数 α_{ik} ($i=1\sim s$, $k=1\sim L$)を定義する。 $\{\alpha_{ik}\}$ は $s\times L$ の行列となる。

(3) $f(x_{11}, x_{12}, \dots, x_{1s}, x_{21}, x_{22}, \dots, x_{2s})$ なる関数を以下のように定義する。ただし、 x_{ji} はi($i=1\sim s$)番目のハプロタイプの第j($j=1, 2$)標本におけるハプロタイプコピーの数に対応するランダム変数 X_{ji} の実現値である。

まず、任意のk($k=1\sim L$)について、図19に示す偶現表を作成する。ここで、図19において、 b_{jk} は、第j標本の第s+1番目のハプロタイプのハプロタイプコピー x_{js+1} 個のうち、k番座位についてマイナーアレルである割合である。これは、結果の第s+1番目のハプロタイプの x_{js+1} 個の実際のハプロタイプの構成により変化する値であり、ランダム変数 B_{jk} の実現値と考えられる。

ついで、作成した偶現表について、座位ごとに独立性検定を実行する。

ついで、独立性検定の結果のいずれかが有意になった場合は1、それ以外の場合は0となる関数fを定義する。

[0036] ここで、下記数式10に示す標本オッズ比は、 $b_{1k}=1$ および $b_{2k}=0$ の時に最大で、 $b_{1k}=0$ および $b_{2k}=1$ の時に最小である。なぜなら、下記数式10の標本オッズ比の分子は b_{1k} について単調増加、分母は b_{2k} について単調減少だからである。

[数10]

$$\frac{\left(\sum_{i=1}^s \alpha_{ik} x_{1i} + b_{1k} x_{1s+1} \right) \left(2N_2 - \sum_{i=1}^s \alpha_{ik} x_{2i} - b_{2k} x_{2s+1} \right)}{\left(\sum_{i=1}^s \alpha_{ik} x_{2i} + b_{2k} x_{2s+1} \right) \left(2N_1 - \sum_{i=1}^s \alpha_{ik} x_{1i} - b_{1k} x_{1s+1} \right)} \dots \text{(数式 10)}$$

[0037] ここで、表1に示す偶現表があるとする。

[表1]

(表 1)

a	m - a
b	n - b

但し、表1において、 a, b は非負の整数で、 $m \geq a, n \geq b$ とする。

[0038] このとき、Pearsonの χ^2 統計量は下記数式11のようになる。また、 T を a で偏微分すると下記数式12となり、 b で偏微分すると下記数式13となる。

[数11]

$$T = \frac{(m+n)(bn-am)^2}{mn(a+b)(m+n-(a+b))} \quad \dots \quad (\text{数式 1 1})$$

[数12]

$$\frac{\partial T}{\partial a} = \frac{(m+n)^2(an-bm)((a+2b)(n-b)+b(m-a))}{mn(a+b)^2(a+b-(m+n))^2} \quad \dots \quad (\text{数式 1 2})$$

[数13]

$$\frac{\partial T}{\partial b} = \frac{(m+n)^2(bm-an)((2a+b)(m-a)+a(n-b))}{mn(a+b)^2(a+b-(m+n))^2} \quad \dots \quad (\text{数式 1 3})$$

[0039] 数式12は、 $a < bm/n$ で負、 $a > bm/n$ で正である。なお、数式12および数式13の $m-a$ および $n-b$ は同時に0になることはないものとする。従って、 T は $a < bm/n$ で単調減少、 $a \geq bm/n$ で単調増加である。 b が固定された時、 $a_0 \leq a \leq a_1$ の範囲で T の最大値を考えると、 b に関わらず $a=a_0$ または $a=a_1$ で最大値をとる。同様にして、数式13は、 $b < an/m$ で負、 $b > an/m$ で正である。すなわち、 T は $b < an/m$ で単調減少、 $b > an/m$ で単調増加である。従って、 T は a が固定された時、 $b_0 \leq b \leq b_1$ の範囲では $b=b_0$ または $b=b_1$ で最大値をとる。

[0040] こうして、数式11は、 $a_0 \leq a \leq a_1$ および $b_0 \leq b \leq b_1$ の範囲では、 $(a=a_0, b=b_0)$ 、 $(a=a_0, b=b_1)$ 、 $(a=a_1, b=b_0)$ 、 $(a=a_1, b=b_1)$ のいずれかの点で最大値をとる。なぜなら、それ以外の点が T に最大値を与えるとすると、それは a が a_0 でも a_1 でもないか、 b が b_0 でも b_1 でもないことを意味するが、前者の場合は a を a_0 または a_1 にすることにより更に大きい T を得ることに矛盾し、後者の場合は b を b_0 または b_1 にすることにより更に大きい T を得ることに矛盾するからである。ただし、 $a_0 \leq a \leq a_1$ および $b_0 \leq b \leq b_1$ で、 a

$=b=0$ でなく、 $a=m$, $b=n$ でないとする。従って、図19の偶現表では、 $(b_{1k}=0, b_{2k}=0)$ 、 $(b_{1k}=0, b_{2k}=1)$ 、 $(b_{1k}=1, b_{2k}=0)$ 、 $(b_{1k}=1, b_{2k}=1)$ のいずれかの時に最大の χ^2 統計量を示す。

[0041] 改めて主要なハプロタイプによる保守的な確率算出方法を具体的に述べると、 $s+1$ 番目のハプロタイプの k 番座位のアレルが第1標本、第2標本のそれぞれにおいて、全てマイナーアレルまたは全てメジャーアレルとして検定を行い、いずれかの検定で有意となれば k 番座位について有意と判定する方法である。第1標本、第2標本のそれぞれについて、 $s+1$ 番ハプロタイプの全ての k 番座位をマイナーアレルまたはメジャーアレルであると考えることにより、図20から図23に示す4種類の偶現表ができる。この内、1つでも有意と判定される場合、図20から図23に示す4つの偶現表のいずれかが有意と判定される。すなわち、図20から図23に示す4つの偶現表のいずれかが有意であれば、関数 $f=1$ とする。

[0042] つぎに、アレル頻度モデルで第1種の過誤の確率を算出する確率算出方法において、上述したその他ハプロタイプのそれぞれの座位のアレル数のかわりに、頻度に応じてサンプルして第1種の過誤の確率を算出する確率算出方法について説明する。

上述した保守的な確率計算方法において、 p_s (主要ハプロタイプの累積頻度)は出来る限り大きくすべきであるが、そうすると主要ハプロタイプの頻度の推定値の信頼性は低くなるであろう。ここで述べる確率算出方法は、第 k 座位に対して、第1標本、第2標本のそれぞれについて、その他ハプロタイプのマイナーアレル数の代わりに、頻度によりサンプルした数を用いる方法である。全てのハプロタイプの頻度が既知であるとすれば、主要ハプロタイプの第 k 座位でのマイナーアレル累積頻度は $\sum_{i=1}^s \alpha_{ik} h_i$ であり、その他ハプロタイプの第 k 座位でのマイナーアレル累積頻度は $\sum_{i=s+1}^L \alpha_{ik} h_i$ である。しかし、一般に、 $s+1, \dots, L$ 座位でのハプロタイプ頻度の信頼性は低い可能性がある。一方、一般に、集団における第 k 座位でのマイナーアレル頻度は既知であることが多い。この集団における k 座位でのマイナーアレル頻度を q_k とすると、その他ハプロタイプにおける第 k 座位のマイナーアレル頻度は、 $q_k - \sum_{i=1}^s \alpha_{ik} h_i$ と考えられる。第 j 標本のその他ハプロタイプのコピー数は X_{js+1} である。ゆえに、 $B(X_{js+1}, q_k - \sum_{i=1}^s \alpha_{ik} h_i)$ の二項分布に従い、標本ごと、座位ごとに第 k 座位のマイナーアレル数を

サンプルする。その実現値を第 k 座位、第 j 標本についてランダム変数 Z_{jk} とすると、図24に示す偶現表を得る。そして、全ての座位について図24に示す偶現表を作成し、作成した偶現表について、座位ごとに独立性検定を行った後、そのいずれかが有意であれば関数 $f=1$ とし、それ以外は0とする。

[0043] これまで述べてきた確率算出方法では、各SNP座位での有意水準(第1種の過誤の確率)が与えられた時、全体の第1種の過誤の確率を算出した。そこで、ここでは、全体の有意水準(例えば $P=0.05$)が与えられた時に、望ましい有意水準を各SNP座位での有意水準を算出する確率算出方法について説明する。具体的には、以下の手順で行う。

(1) 各SNPでの検定のための有意水準を、例えば $P_{\text{locus}}=0.001, 0.002, 0.003, \dots, 0.010$ のように少しずつ増やしていき、上述した確率算出方法または後述する近似的に確率を計算する方法により全体の第1種の過誤の確率 P_{all} を算出する。

(2) P_{locus} と P_{all} の関係により、望ましい P_{all} の値(例えば0.05)に対応する P_{locus} を計算する。なお、計算方法は直線回帰や曲線回帰などの方法を用いればよい。

(3) 計算した各SNPの検定における望ましい有意水準 P_{locus} を用いて各SNPの検定を行う。これにより、連鎖不平衡を考慮した多数のSNP座位での検定が可能になる。

[0044] 以上、アレル頻度モデルで第1種の過誤の確率を算出する確率算出方法に関する説明を終了する。

[0045] (1-2) 優性・劣性遺伝モデル

図2は、優性・劣性遺伝モデルで第1種の過誤の確率を正確に算出する確率算出方法の一例を示すフローチャートである。

まず、複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する(ステップSB-1)。なお、第1標本および第2標本の取得に関する説明は、上述した(1-1)アレル頻度モデルと同様である。

[0046] ついで、ステップSB-1で取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出する(ステップSB-2)。なお、ハプロタイプ頻度の算出に

関する説明は、上述した(1-1)アレル頻度モデルと同様である。

[0047] ついで、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する(ステップSB-3)。なお、行列の定義に関する説明は、上述した(1-1)アレル頻度モデルと同様である。

[0048] ついで、第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与える(ステップSB-4)。ここで、第1標本におけるディプロタイプ形番号 jj' (j' はハプロタイプ番号であり、 $1 \leq j' \leq M$ を満たす整数、ディプロタイプ形番号については図14参照)のディプロタイプ形度数 $X_{1jj'}$ および第2標本におけるディプロタイプ形番号 jj' のディプロタイプ形度数 $X_{2jj'}$ は、ランダム変数であり、それぞれ下記数式14を満たす。そして、下記数式14を満たすように、第1標本におけるディプロタイプ形番号 jj' のディプロタイプ形度数の実現値 $x_{1jj'}$ を全てのディプロタイプ形番号 jj' に対して設定し、第2標本におけるディプロタイプ形番号 jj' のディプロタイプ形度数の実現値 $x_{2jj'}$ を全てのディプロタイプ形番号 jj' に対して設定する(図15参照)。これにより、 $(x_{111}, \dots, x_{1MM}, x_{211}, \dots, x_{2MM})$ の組合せが与えられた。

[数14]

$$\sum_{j=1}^M \sum_{j'=1}^M X_{1jj'} = N_1 \quad , \quad \sum_{j=1}^M \sum_{j'=1}^M X_{2jj'} = N_2 \quad \cdots \quad (\text{数式 14})$$

[0049] ついで、ステップSB-4で設定した第1標本に対応する複数のディプロタイプ形度数およびステップSB-3で定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、ステップSB-4で設定した第2標本に対応する複数のディプロタイプ形度数およびステップSB-3で定義した行列に基づいて第2標本における各座位でのアレル度数を算出する(ステップSB-5)。ここで、第1標本における座位 i でのマイナーアレルのアレル度数 Y_{1i} および第2標本における座位 i でのマイナーアレルのアレル度数 Y_{2i} は、ランダム変数であり、それぞれ下記数式15で表される。そして、下

記数式15に基づいて、第1標本における座位iでのマイナーアレルのアレル度数の実現値 y_{1i} および第2標本における座位iでのマイナーアレルのアレル度数の実現値 y_{2i} を算出する。

[数15]

$$Y_{1i} = \sum_{j=1}^M \sum_{j'=1}^M \alpha_{ij} \alpha_{ij'} \times X_{1jj'} \quad , \quad Y_{2i} = \sum_{j=1}^M \sum_{j'=1}^M \alpha_{ij} \alpha_{ij'} \times X_{2jj'} \quad \cdots \quad (\text{数式 15})$$

数式15において、 α_{ij} および $\alpha_{ij'}$ はステップSB-3で定義した行列である。

[0050] ここで、劣性遺伝モデルの場合では、数式15を用いてマイナーアレル(aa+Aa)のアレル度数を算出するが、優性遺伝モデルの場合では下記数式16を用いてマイナーアレル(aa)のアレル度数を算出する。

[数16]

$$Y_{1i} = \sum_{j=1}^M \sum_{j'=1}^M (1 - \alpha_{ij})(1 - \alpha_{ij'}) \times X_{1jj'} \quad , \quad Y_{2i} = \sum_{j=1}^M \sum_{j'=1}^M (1 - \alpha_{ij})(1 - \alpha_{ij'}) \times X_{2jj'} \quad \cdots \quad (\text{数式 16})$$

[0051] ついで、ステップSB-5で算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する(ステップSB-6)。具体的には、劣性遺伝モデルの場合では数式15を用いて算出したアレル度数に関する図16に示す2×2の偶現表を座位数分(L個)だけ作成し、優性遺伝モデルの場合では数式16を用いて算出したアレル度数に関する図17に示す2×2の偶現表を座位数分(L個)だけ作成する。

[0052] ついで、ステップSB-6で作成した偶現表について、座位ごとに独立性検定を実行する(ステップSB-7)。

[0053] ついで、ステップSB-7で実行した独立性検定の検定結果のいずれか1つが有意であった場合(ステップSB-8:Yes)、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、ステップSB-4で設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応す

る確率を、予め設定した第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する(ステップSB-9)。具体的には、第1標本の標本数 N_1 、第2標本の標本数 N_2 、存在し得るハプロタイプの総数 $M(2^L)$ 、第1標本のディプロタイプ形度数の実現値 x_{1ij} 、第2標本のディプロタイプ形度数の実現値 x_{2ij} およびハプロタイプ頻度 h_j に基づいて、ステップSB-4で設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せ($x_{111}, \dots, x_{1MM}, x_{211}, \dots, x_{2MM}$)に対応する確率 P を、下記数式17に示す第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する。

[数17]

$$P(x_{111}, x_{112}, \dots, x_{1MM}, x_{211}, x_{212}, \dots, x_{2MM}) \\ = \frac{\prod_{k=1}^2 N_k!}{\prod_{j=1}^M \prod_{j'=1}^M \prod_{k=1}^2 x_{kjj'}!} \prod_{j=1}^M \prod_{j'=1}^M (h_j h_{j'})^{\sum_{k=1}^2 x_{kjj'}} \dots \quad (\text{数式 17})$$

数式17において、 k は標本を識別する番号である。ここでは、 $k=1$ の場合は第1標本を示し、 $k=2$ の場合は第2標本を示す。

[0054] ついで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せが終了したか否かを判定し、全ての組合せが終了した場合(ステップSB-10: Yes)、ステップSB-4～ステップSB-9を繰り返すごとにステップSB-9で算出される確率を加算し、加算した確率を最終的に第1種の過誤の確率とする(ステップSB-11)。一方、全ての組合せが終了していない場合(ステップSB-10: No)、ステップSB-4～ステップSB-9を実行し、全ての組合せが終了するまで繰り返す。

[0055] 以上、説明したように、本発明にかかる確率算出方法は、(1)複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得し、(2)取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出し、(3)各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを

有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義し、(4)第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与え、(5)設定した第1標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第1標本における各座位でのアレル度を算出し、設定した第2標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第2標本における各座位でのアレル度を算出し、(6)算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成し、(7)作成した偶現表について、座位ごとに独立性検定を実行し、(8)実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する。そして、(4)、(5)、(6)、(7)および(8)を、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せについて繰り返し、繰り返すごとに(8)で算出される確率を加算することで、最終的に第1種の過誤の確率を算出する。これにより、第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、疾患関連SNPの見落としを防ぐことができる。

- [0056] また、ステップSB-9において、数式17に示す第1標本と第2標本とのディプロタイプ形度数に関する結合分布は、全ての k, j, j' について $X_{kjj'}$ の結合分布、すなわちテンソル $\{X_{kjj'}\}$ の分布は多項分布であるので、下記数式18に示す第1標本のディプロタイプ形度数に関する結合分布と下記数式19に示す第2標本のディプロタイプ形度数に関する結合分布との積で与えられる。ここで、優性遺伝モデルまたは劣性遺伝モデルによる検定について第1の過誤の確率を計算するためにはディプロタイプ形について考える必要がある。なお、優性遺伝モデルによる検定とはメジャーアレルを

保有する個体とそうでない個体の比率を2群で比較する検定法をいい、劣性遺伝モデルによる検定とはマイナーアレルを保有する個体とそうでない個体の比率を2群で比較する検定法をいう。2つのハプロタイプコピーの順列を順列ディプロタイプ形ということにし、ハプロタイプ番号 j のハプロタイプとハプロタイプ番号 j' のハプロタイプとによる(この順番の)順列ディプロタイプ形を $D_{jj'}$ とする。一般に、 $D_{jj'} = D_{j'j}$ は成り立たず、 $j=j'$ であってもよい。ハプロタイプの総数を M とすれば、順列ディプロタイプ形の総数は M^2 である。ハプロタイプに関するHardy-Weinberg平衡が成り立っているとすると、集団内の $D_{jj'}$ の頻度は $h_j h_{j'}$ である。今、第1標本として N_1 人、第2標本として N_2 人をランダムに抽出すると、第 k 標本について $D_{jj'}$ を持つ人の総数は二項分布 $B(N_k, h_j h_{j'})$ に従うランダム変数である。第 k 標本について、このランダム変数を $X_{kjj'}$ とする。特定の k について、全ての $X_{kjj'}$ の結合分布、すなわち行列 $\{X_{1jj'}\}$ および $\{X_{2jj'}\}$ の分布は、多項分布であり、その確率は、第1標本および第2標本のそれぞれについて、下記数式18および下記数式19で与えられる。

[数18]

$$P(X_{111} = x_{111}, X_{112} = x_{112}, \dots, X_{1MM} = x_{1MM}) \\ = \frac{N_1!}{\prod_{j=1}^M \prod_{j'=1}^M x_{1jj'}!} \prod_{j=1}^M \prod_{j'=1}^M (h_j h_{j'})^{x_{1jj'}} \quad \dots \quad (\text{数式 18})$$

[数19]

$$P(X_{211} = x_{211}, X_{212} = x_{212}, \dots, X_{2MM} = x_{2MM}) \\ = \frac{N_2!}{\prod_{j=1}^M \prod_{j'=1}^M x_{2jj'}!} \prod_{j=1}^M \prod_{j'=1}^M (h_j h_{j'})^{x_{2jj'}} \quad \dots \quad (\text{数式 19})$$

[0057] また、ステップSB-11で算出した第1種の過誤の確率は、下記数式20で表現することができる。そして、下記数式20で求められる確率は、ステップSB-7で実行した独立性検定の検定結果のいずれか1つが有意となる確率を意味する。

[数20]

$$\begin{aligned}
& P[Z = 1] \\
&= \sum_{x_{111}=0}^{s_{111}} \sum_{x_{112}=0}^{s_{112}} \cdots \sum_{x_{11M}=0}^{s_{11M}} \\
&\quad \sum_{x_{121}=0}^{s_{121}} \sum_{x_{122}=0}^{s_{122}} \cdots \sum_{x_{12M}=0}^{s_{12M}} \\
&\quad \cdots \\
&\quad \sum_{x_{1M1}=0}^{s_{1M1}} \sum_{x_{1M2}=0}^{s_{1M2}} \cdots \sum_{x_{1MM}=0}^{s_{1MM}} \\
&\quad \sum_{x_{211}=0}^{s_{211}} \sum_{x_{212}=0}^{s_{212}} \cdots \sum_{x_{21M}=0}^{s_{21M}} \\
&\quad \sum_{x_{221}=0}^{s_{221}} \sum_{x_{222}=0}^{s_{222}} \cdots \sum_{x_{22M}=0}^{s_{22M}} \\
&\quad \cdots \\
&\quad \sum_{x_{2M1}=0}^{s_{2M1}} \sum_{x_{2M2}=0}^{s_{2M2}} \cdots \sum_{x_{2MM-1}=0}^{s_{2MM-1}} \\
&Z \times \frac{\prod_{k=1}^2 N_k!}{\prod_{j=1}^M \prod_{j'=1}^M \prod_{k=1}^2 x_{kjj'}!} \prod_{j=1}^M \prod_{j'=1}^M (h_j h_{j'})^{\sum_{k=1}^2 x_{kjj'}}
\end{aligned}$$

．．．（数式 20）

数式20において、Zは、ステップSB-7で実行した独立性検定の検定結果のいずれか1つが有意であった場合は1、そうでなかった場合は0と定義したランダム変数である。すなわち、変数Zは上述したランダム変数 Y_{ki} の関数であり、さらに当該ランダム変数 Y_{ki} は上述したランダム変数 $X_{kjj'}$ の関数であるので、変数Zは関数 $f(x_{111}, \cdots, x_{11M}, x_{211}, \cdots, x_{2MM})$ と表すことができる。また、数式20において、 $s_{kjj'}$ ($k=1, 2; j, j'=1 \sim M$)は、下記数式21で定義される。

[数21]

$$\begin{aligned}
s_{k11} &= N_k, s_{k12} = s_{k11} - x_{k11}, s_{k13} = s_{k12} - x_{k12}, \dots, s_{k1M} = s_{k1M-1} - x_{k1M-1} \\
s_{k21} &= s_{k1M} - x_{k1M}, s_{k22} = s_{k21} - x_{k21}, \dots, s_{k2M} = s_{k2M-1} - x_{k2M-1} \\
&\dots \\
&\dots \\
s_{kM1} &= s_{kM-1M} - x_{kM-1M} \\
s_{kM2} &= s_{kM1} - x_{kM1} \\
&\dots \\
s_{kMM-1} &= s_{kMM-2} - x_{kMM-2} \\
&\dots \quad (\text{数式 2 1})
\end{aligned}$$

[0058] このように、優性遺伝モデルや劣性遺伝モデルで検定を行う場合、ハプロタイプコピーではなく、ディプロタイプ形コピーの数を数える必要がある。しかし、ディプロタイプ形コピーの数からハプロタイプコピーの数を数えることは可能である。このことは重要であり、ランダム変数 X_{kij} の実現値 x_{kij} を用いて、優性遺伝モデルや劣性遺伝モデルでの検定だけでなく、アレル頻度モデルの検定を行うことが可能だからである。そして、連鎖した多数のSNP座位について、優性遺伝モデル、劣性遺伝モデルおよびアレル頻度モデルで検定を行い、そのいずれかが有意の場合は、有意とした場合の全体の第1種の過誤の確率を正確に計算することができる。具体的には、順列ディプロタイプ形コピーの数であるランダム変数 X_{kij} の実現値 x_{kij} を用いて、図25に示す偶現表を作成すればよい。そして、全ての座位について独立性検定を行い、いずれかが有意の場合は、上述した関数 $f(x_{111}, \dots, x_{1MM}, x_{211}, \dots, x_{2MM})$ を1とし、それ以外の場合は0とすればよい。この検定で算出した第1の過誤の確率は、上述した $(1-1)$ アレル頻度モデルで算出した第1の過誤の確率と一致する。

[0059] さらに、優性遺伝モデル、劣性遺伝モデルおよびアレル頻度モデルの3つの検定を行い、いずれかで有意の場合に有意と判定する方法による第1種の過誤の確率も、正確に計算することが可能である。具体的には、図13、図16および図17の偶現表について各座位のそれぞれで独立性検定を行い、そのいずれかで有意の場合は関数 f を1とし、それ以外では0とすればよい。

[0060] 以上、優性・劣性遺伝モデルで第1種の過誤の確率を算出する確率算出方法に関

する説明を終了する。

[0061] (1-3) 遺伝子型モデル

図3は、遺伝子型モデルで第1種の過誤の確率を正確に算出する確率算出方法の一例を示すフローチャートである。

まず、複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する(ステップSC-1)。なお、第1標本および第2標本の取得に関する説明は、上述した(1-1)アレル頻度モデル、(1-2)優性・劣性遺伝モデルと同様である。

[0062] ついで、ステップSC-1で取得した第1標本に含まれるハプロタイプ情報および第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出する(ステップSC-2)。なお、ハプロタイプ頻度の算出に関する説明は、上述した(1-1)アレル頻度モデル、(1-2)優性・劣性遺伝モデルと同様である。

[0063] ついで、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する(ステップSC-3)。なお、行列の定義に関する説明は、上述した(1-1)アレル頻度モデル、(1-2)優性・劣性遺伝モデルと同様である。

[0064] ついで、第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与える(ステップSC-4)。なお、ディプロタイプ形度数の設定に関する説明は、上述した(1-2)優性・劣性遺伝モデルと同様である。

[0065] ついで、ステップSC-4で設定した第1標本に対応する複数のディプロタイプ形度数およびステップSC-3で定義した行列に基づいて第1標本における各座位での遺伝子型度を算出し、ステップSC-4で設定した第2標本に対応する複数のディプロタイプ形度数およびステップSC-3で定義した行列に基づいて第2標本における各座位での遺伝子型度を算出する(ステップSC-5)。ここで、第1標本における座位 i でのメジャーホモ(AA)の遺伝子型度数 Y_{ii} および第2標本における座位 i での

メジャーホモ(AA)の遺伝子型度数 Y_{2i} 、第1標本における座位iでのマイナーホモ(a a)の遺伝子型度数 Y'_{li} および第2標本における座位iでのマイナーホモ(aa)の遺伝子型度数 Y'_{2i} ならびに第1標本における座位iでのヘテロ(Aa)の遺伝子型度数 Y''_{li} および第2標本における座位iでのヘテロ(Aa)の遺伝子型度数 Y''_{2i} は、それぞれランダム変数であり、それぞれ下記数式22で表される。そして、下記数式22に基づいて、遺伝子型度数 Y_{li} の実現値 y_{li} および遺伝子型度数 Y_{2i} の実現値 y_{2i} 、遺伝子型度数 Y'_{li} の実現値 y'_{li} および遺伝子型度数 Y'_{2i} の実現値 y'_{2i} ならびに遺伝子型度数 Y''_{li} の実現値 y''_{li} および遺伝子型度数 Y''_{2i} の実現値 y''_{2i} を算出する。

[数22]

$$\begin{aligned}
 Y_{li} &= \sum_{j=1}^M \sum_{j'=1}^M \alpha_{ij} \alpha_{ij'} \times X_{1jj'} \quad , \quad Y_{2i} = \sum_{j=1}^M \sum_{j'=1}^M \alpha_{ij} \alpha_{ij'} \times X_{2jj'} \\
 Y'_{li} &= \sum_{j=1}^M \sum_{j'=1}^M (1 - \alpha_{ij})(1 - \alpha_{ij'}) \times X_{1jj'} \quad , \quad Y'_{2i} = \sum_{j=1}^M \sum_{j'=1}^M (1 - \alpha_{ij})(1 - \alpha_{ij'}) \times X_{2jj'} \\
 Y''_{li} &= N_1 - (Y_{li} + Y'_{li}) \quad , \quad Y''_{2i} = N_2 - (Y_{2i} + Y'_{2i}) \\
 &\quad \quad \quad \cdot \cdot \cdot \quad (\text{数式 2 2})
 \end{aligned}$$

[0066] ついで、ステップSC-5で算出した第1標本における各座位での遺伝子型度数および第2標本における各座位での遺伝子型度数に基づいて、各座位に対応する偶現表を作成する(ステップSC-6)。具体的には、数式22を用いて算出した遺伝子型度数に関する図18に示す2×3の偶現表を座位数分(L個)だけ作成する。

[0067] ついで、ステップSC-6で作成した偶現表について、座位ごとに独立性検定を実行する(ステップSC-7)。

[0068] ついで、ステップSC-7で実行した独立性検定の検定結果のいずれか1つが有意であった場合(ステップSC-8:Yes)、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、ステップSC-4で設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する(ステップSC-9)。なお、確率の算出に関する説明は、上述

した(1-2)優性・劣性遺伝モデルと同様である。

[0069] ついで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せが終了したか否かを判定し、全ての組合せが終了した場合(ステップSC-10:Yes)、ステップSC-4～ステップSC-9を繰り返すごとにステップSC-9で算出される確率を加算し、加算した確率を最終的に第1種の過誤の確率とする(ステップSC-11)。一方、全ての組合せが終了していない場合(ステップSC-10:No)、ステップSC-4～ステップSC-9を実行し、全ての組合せが終了するまで繰り返す。

[0070] 以上、説明したように、本発明にかかる確率算出方法は、(1)複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得し、(2)取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出し、(3)各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義し、(4)第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与え、(5)設定した第1標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第1標本における各座位での遺伝子型度を算出し、設定した第2標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第2標本における各座位での遺伝子型度を算出し、(6)算出した第1標本における各座位での遺伝子型度数および第2標本における各座位での遺伝子型度数に基づいて、各座位に対応する偶現表を作成し、(7)作成した偶現表について、座位ごとに独立性検定を実行し、(8)実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応する確率を、予め設定した第1標本と第2標本との

ディプロタイプ形度数に関する結合分布を用いて算出する。そして、(4)、(5)、(6)、(7)および(8)を、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せについて繰り返し、繰り返すごとに(8)で算出される確率を加算することで、最終的に第1種の過誤の確率を算出する。これにより、第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、その結果、疾患関連SNPの見落としを防ぐことができる。

[0071] 以上、遺伝子型モデルで第1種の過誤の確率を算出する確率算出方法に関する説明を終了する。

[0072] (1-4) 正確なハプロタイプ頻度がわからない場合

図4は、正確なハプロタイプ頻度がわからない場合に第1種の過誤の確率を正確に算出する確率算出方法の一例を示すフローチャートである。

まず、複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する(ステップSD-1)。なお、第1標本および第2標本の取得に関する説明は、上述した(1-1)アレル頻度モデル、(1-2)優性・劣性遺伝モデル、(1-3)遺伝子型モデルと同様である。

[0073] ついで、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する(ステップSD-2)。なお、行列の定義に関する説明は、上述した(1-1)アレル頻度モデル、(1-2)優性・劣性遺伝モデル、(1-3)遺伝子型モデルと同様である。

[0074] ついで、第1標本における各ハプロタイプのハプロタイプ度数および第2標本における各ハプロタイプのハプロタイプ度数を設定することで、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せを与える(ステップSD-3)。ここで、第1標本におけるハプロタイプ番号 j のハプロタイプ度数 X_{1j} および第2標本におけるハプロタイプ番号 j のハプロタイプ度数 X_{2j} は、ランダム変数であり、それぞれ下記数式23を満たす。そして、下記数式23を満たすように、第1標本におけるハプロタイプ番号 j のハプロタイプ度数 X_{1j} の実現値 x_{1j} を全てのハプロタイプ番号 j に対して設定し、第2標本におけるハプロタイプ番号 j のハプロタイプ度数 X_{2j} の実現値 x_{2j} を全てのハプロ

タイプ番号jに対して設定する。これにより、 $(x_{11}, \dots, x_{1M}, x_{21}, \dots, x_{2M})$ の組合せが与えられた。

[数23]

$$\sum_{k=1}^2 X_{kj} = X_{.j} \quad , \quad \sum_{j=1}^M X_{1j} = N_1 \quad , \quad \sum_{j=1}^M X_{2j} = N_2 \quad \cdots \quad (\text{数式 2 3})$$

[0075] ついで、ステップSD-3で設定した第1標本に対応する複数のハプロタイプ度数およびステップSD-2で定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、ステップSD-3で設定した第2標本に対応する複数のハプロタイプ度数およびステップSD-2で定義した行列に基づいて第2標本における各座位でのアレル度数を算出する(ステップSD-4)。なお、アレル度数の算出に関する説明は、上述した(1-1)アレル頻度モデルと同様である。

[0076] ついで、ステップSD-4で算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する(ステップSD-5)。

[0077] ついで、ステップSD-5で作成した偶現表について、座位ごとに独立性検定を実行する(ステップSD-6)。

[0078] ついで、ステップSD-6で実行した独立性検定の検定結果のいずれか1つが有意であった場合(ステップSD-7:Yes)、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のハプロタイプ度数および第2標本のハプロタイプ度数に基づいて、ステップSD-3で設定した第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのハプロタイプ度数に関する結合分布を用いて算出する(ステップSD-8)。具体的には、第1標本の標本数 N_1 、第2標本の標本数 N_2 、存在し得るハプロタイプの総数 $M(2^l)$ 、第1標本のハプロタイプ度数の実現値 x_{1j} および第2標本のハプロタイプ度数の実現値 x_{2j} に基づいて、ステップSD-3で設定した第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せ $(x_{11}, \dots, x_{1M}, x_{21}, \dots, x_{2M})$ に対応する確率Pを、下記数式24に示す第1標本と第2標本とのハプロタイプ度数に関する結合分布を用いて算出する。なお、下記数式24に示す第1標本と第2標本と

のハプロタイプ度数に関する結合分布は、超幾何分布である。

[数24]

$$P\left(\{X_{kj}\} = \{x_{kj}\} \mid \sum_{k=1}^2 x_{.j}, j = 1, 2, \dots, M\right) \\ = \frac{N_1! N_2! \prod_{j=1}^M x_{.j}}{(N_1 + N_2)! \prod_{j=1}^M \prod_{k=1}^2 x_{kj}!} \cdot \cdot \cdot \quad (\text{数式 24})$$

数式24において、kは標本を識別する番号である。ここでは、k=1の場合は第1標本を示し、k=2の場合は第2標本を示す。

[0079] ついで、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との全ての組合せが終了したか否かを判定し、全ての組合せが終了した場合(ステップSD-9:Yes)、ステップSD-3～ステップSD-8を繰り返すごとにステップSD-8で算出された確率を全て加算し、加算した確率を最終的に第1種の過誤の確率とする(ステップSD-10)。一方、全ての組合せが終了していない場合(ステップSD-9:No)、ステップSD-3～ステップSD-8を実行して、全ての組合せが終了するまで繰り返す。

[0080] 以上、説明したように、本発明にかかる確率算出方法は、(1)複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得し、(2)各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義し、(3)第1標本における各ハプロタイプのハプロタイプ度数および第2標本における各ハプロタイプのハプロタイプ度数を設定することで、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せを与え、(4)設定した第1標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、設定した第2標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第2標本における各座位でのアレル度数を算出し、(5)

算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成し、(6)作成した偶現表について、座位ごとに独立性検定を実行し、(7)実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のハプロタイプ度数および第2標本のハプロタイプ度数に基づいて、設定した第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのハプロタイプ度数に関する結合分布を用いて算出する。そして、(3)、(4)、(5)、(6)および(7)を、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との全ての組合せについて繰り返し、繰り返すごとに(7)で算出される確率を加算することで、最終的に第1種の過誤の確率を算出する。これにより、第1種の過誤の確率を、直接的に算出することができ、その結果、疾患関連SNPの見落としを防ぐことができる。

[0081] また、ステップSD-10で算出した第1種の過誤の確率は、下記数式25で表現することができる。そして、下記数式25で求められる確率は、ステップSD-6で実行した独立性検定の検定結果のいずれか1つが有意となる確率を意味する。

[数25]

$$\begin{aligned}
 & P \left[Z = 1 \mid \sum_{k=1}^2 X_{kj} = x_{.j}, j = 1, 2, \dots, M \right] \\
 &= \sum_{x_{11}=0}^{x_1} \sum_{x_{12}=0}^{\min(x_2, N_1-x_{11})} \dots \sum_{x_{1M-1}=0}^{\min(x_M, N_1-\sum_{j=1}^{M-2} x_{1j})} Z \times \frac{N_1! N_2! \prod_{j=1}^M x_{.j}}{(N_1 + N_2)! \prod_{j=1}^M \prod_{k=1}^2 x_{kj}!} \\
 & \dots \text{ (数式 2 5) }
 \end{aligned}$$

数式25において、Zは、ステップSD-6で実行した独立性検定の検定結果のいずれか1つが有意であった場合には1を、そうでなかった場合には0をとるように定義したランダム変数である。変数Zは上述したランダム変数 Y_{ki} の関数であり、さらに当該ランダム変数 Y_{ki} は上述したランダム変数 X_{kj} の関数である。また、ランダム変数 X_{kj} につ

いては数式23の制限がある。よって、変数 Z は関数 $f(x_{11}, \dots, x_{1M})$ で表すことができる。なお、その他の X_{kj} はその他の変数の関数として表される。

[0082] 以上、正確なハプロタイプ頻度がわからない場合に第1種の過誤の確率を正確に算出する確率算出方法に関する説明を終了する。

[0083] [第2実施形態]

ここでは、第1種の過誤の確率をMCMC法(マルコフ連鎖モンテカルロ法)を用いて近似的に算出する本発明にかかる確率算出方法について、(2-1)アレル頻度モデル、(2-2)優性・劣性遺伝モデル、(2-3)遺伝子型モデル、(2-4)正確なハプロタイプ頻度がわからない場合、の順に図を参照して詳細に説明する。

[0084] (2-1)アレル頻度モデル

図5は、アレル頻度モデルで第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法の一例を示すフローチャートである。なお、全てのハプロタイプ頻度が既知であることを前提とする。

まず、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する(ステップSE-1)。なお、行列の定義に関する説明は、上述した第1実施形態における(1-1)アレル頻度モデルと同様である。

[0085] ついで、ハプロタイプ度数の組合せを予め設定した第1標本および第2標本に対して生成する(ステップSE-2)。具体的には、第1標本におけるハプロタイプ番号 j のハプロタイプ度数 X_{1j} の実現値 x_{1j} および第2標本におけるハプロタイプ番号 j のハプロタイプ度数 X_{2j} の実現値 x_{2j} を、上述した数式2を満たすように生成する。これにより、 $(x_{11}, \dots, x_{1M}, x_{21}, \dots, x_{2M})$ の組合せを生成する。

[0086] ついで、第1標本または第2標本のいずれかの標本を選択する(ステップSE-3)。具体的には、標本番号 k ($k=1, 2$)の標本を等確率で選択する。

[0087] ついで、第1ハプロタイプを選択する(ステップSE-4)。具体的には、ハプロタイプ番号 j ($1 \leq j \leq M (=2^L)$; L は座位数)を選択する。

[0088] ついで、ステップSE-3で選択した標本における第1ハプロタイプのハプロタイプ度数が0でない場合(SE-5: Yes)、当該第1ハプロタイプとは異なる第2ハプロタイプ

を選択する(ステップSE-6)。具体的には、選択した標本番号 k の標本におけるハプロタイプ番号 j のハプロタイプ度数 $x_{kj} \neq 0$ (または $x_{kj} > 0$)の場合、ステップSE-3で選択したハプロタイプ番号 j とは異なるハプロタイプ番号 j' ($1 \leq j' \leq M$)を選択する。なお、ハプロタイプ度数 $x_{kj} = 0$ の場合、ハプロタイプ度数 x_{kj} の値を維持する。

[0089] ついで、予め設定した第1ハプロタイプのハプロタイプ頻度および第2ハプロタイプのハプロタイプ頻度に基づいて、第1ハプロタイプのハプロタイプ度数および第2ハプロタイプのハプロタイプ度数を更新する(ステップSE-7)。具体的には、まず、下記数式26で定義される c の値を算出する。なお、下記数式26において、 h_j および $h_{j'}$ は、それぞれハプロタイプ番号 j のハプロタイプ頻度およびハプロタイプ番号 j' のハプロタイプ頻度であり、その値は予め設定されている。

[数26]

$$c = \frac{h_{j'} x_{kj}}{h_j (x_{kj'} + 1)} \quad \cdot \cdot \cdot \quad (\text{数式 26})$$

ついで、 $c \geq 1$ の場合、 x_{kj} の値から1を減算し、 $x_{kj'}$ の値に1を加算する。また、 $c < 1$ の場合、 c の確率で x_{kj} の値から1を減算し、 $x_{kj'}$ の値に1を加算し、 $1 - c$ の確率で x_{kj} および $x_{kj'}$ の値を維持する。

[0090] ついで、ステップSE-7で更新した後の第1標本に対応する複数のハプロタイプ度数およびステップSE-1で定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、ステップSE-7で更新した後の第2標本に対応する複数のハプロタイプ度数およびステップSE-1で定義した行列に基づいて第2標本における各座位でのアレル度数を算出する(ステップSE-8)。なお、アレル度数の算出に関する説明は、上述した第1実施形態の(1-1)アレル頻度モデルと同様である。

[0091] ついで、ステップSE-8で算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する(ステップSE-9)。

[0092] ついで、ステップSE-9で作成した偶現表について、座位ごとに独立性検定を実行する(ステップSE-10)。

- [0093] ついで、ステップSE-10で実行した独立性検定の検定結果のいずれか1つが有意であった場合(ステップSE-11:Yes)、当該場合の累積回数を更新する(ステップSE-12)。具体的には、累積回数に1を加算する。
- [0094] ついで、所定回数が終了したか否かを判定し、所定回数が終了した場合(ステップSE-13:Yes)、所定回数に対する最終的な累積回数の割合(累積回数÷所定回数)を第1種の過誤の確率として算出する(ステップSE-14)。一方、所定回数が終了していない場合(ステップSE-13:No)、ステップSE-3～ステップSE-12を実行し、所定回数が終了するまで繰り返す。
- [0095] 以上、説明したように、本発明にかかる確率算出方法は、Metropolis-Hastings法により、標本ごとに期待される多項分布に従うハプロタイプのサンプルを生成し、それぞれのサンプルについて各座位で検定を行い、そのいずれかの座位で有意となる割合を求める。なお、Markov-chainの状態は、ハプロタイプ度数の実現値による $2 \times M$ (存在し得るハプロタイプの総数)の偶現表により代表される。
- [0096] ここで、主要ハプロタイプを用いた計算では、主要ハプロタイプに加え、その他ハプロタイプを設ける。従って、マルコフ連鎖(Markov-chain)の状態は x_{kj} ($k=1, 2; j=1, 2, \dots, s$)の2標本の主要ハプロタイプコピーの数とその他ハプロタイプコピーの数 x_{ks+1} により代表される。従って、 x_{kj} ($k=1, 2; j=1, 2, \dots, s+1$)の全てのハプロタイプを用いた場合、 x_{kj} ($k=1, 2; j=1, 2, \dots, M(=2^L)$)と同様にMCMC法を実行すればよい。なお、検定については、上述したように保守的な検定法またはその他ハプロタイプコピー数をサンプルすることにより行う。第1種の過誤の確率は、同様に有意と判定されたステップの割合を算出すればよい。
- [0097] また、多項分布に従うMonte Carloサンプリングにより第1標本、第2標本を独立に生成し、上述した関数fの計算を行い、これが1となる割合を第1種の過誤の確率とする方法が考えられる。
- [0098] 以上、アレル頻度モデルで第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法に関する説明を終了する。
- [0099] (2-2) 優性・劣性遺伝モデル
- 図6は、優性・劣性遺伝モデルで第1種の過誤の確率をMCMC法を用いて近似的

に算出する確率算出方法の一例を示すフローチャートである。なお、全てのハプロタイプ頻度が既知であることを前提とする。

まず、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する(ステップSF-1)。なお、行列の定義に関する説明は、上述した第1実施形態における(1-2)優性・劣性遺伝モデルと同様である。

[0100] ついで、ディプロタイプ形度数の組合せを予め設定した第1標本および第2標本に対して生成する(ステップSF-2)。具体的には、第1標本におけるディプロタイプ形番号 jj' のディプロタイプ形度数の実現値 $x_{1jj'}$ を全てのディプロタイプ形番号 jj' に対して設定し、第2標本におけるディプロタイプ形番号 jj' のディプロタイプ形度数の実現値 $x_{2jj'}$ を、上述した数式14を満たすように生成する。これにより、 $(x_{111}, \dots, x_{1MM}, x_{211}, \dots, x_{2MM})$ の組合せを生成する。

[0101] ついで、第1標本または第2標本のいずれかの標本を選択する(ステップSF-3)。具体的には、標本番号 k ($k=1, 2$)の標本を等確率で選択する。

[0102] ついで、第1ディプロタイプ形を選択する(ステップSF-4)。具体的には、ディプロタイプ形番号 j_1j_1' ($1 \leq$ 順序付きの自然数 j_1 および $j_1' \leq M (=2^L)$; L は座位数)を選択する。

[0103] ついで、ステップSF-3で選択した標本における第1ディプロタイプ形のディプロタイプ形度数が0でない場合(SF-5: Yes)、当該第1ディプロタイプ形とは異なる第2ディプロタイプ形を選択する(ステップSF-6)。具体的には、選択した標本番号 k の標本におけるディプロタイプ形番号 j_1j_1' のディプロタイプ形度数 $x_{k j_1 j_1'} \neq 0$ (または $x_{k j_1 j_1'} > 0$) の場合、ステップSF-3で選択したディプロタイプ形番号 j_1j_1' とは異なるディプロタイプ形番号 j_2j_2' ($1 \leq$ 順序付きの自然数 j_2 および $j_2' \leq M (=2^L)$; L は座位数)を選択する。なお、ディプロタイプ形度数 $x_{k j_1 j_1'} = 0$ の場合、ディプロタイプ形度数 $x_{k j_1 j_1'}$ の値を維持する。

[0104] ついで、予め設定した第1ディプロタイプ形を構成する各ハプロタイプのハプロタイプ頻度および第2ディプロタイプ形を構成する各ハプロタイプのハプロタイプ頻度に基づいて、第1ディプロタイプ形のディプロタイプ形度数および第2ディプロタイプ形

のディプロタイプ形度を更新する(ステップSF-7)。具体的には、まず、下記数式27で定義される c の値を算出する。なお、下記数式27において、 h_{j1} 、 $h_{j1'}$ 、 h_{j2} および $h_{j2'}$ は、それぞれハプロタイプ番号 j_1 のハプロタイプ頻度、ハプロタイプ番号 j_1' のハプロタイプ頻度、ハプロタイプ番号 j_2 のハプロタイプ頻度およびハプロタイプ番号 j_2' のハプロタイプ頻度であり、その値は予め設定されている。

[数27]

$$c = \frac{h_{j_2} h_{j_2'} x_{kj_1 j_1'}}{h_{j_1} h_{j_1'} (x_{kj_2 j_2'} + 1)} \quad \dots \quad (\text{数式 27})$$

ついで、 $c \geq 1$ の場合、 $x_{kj_1 j_1'}$ の値から1を減算し、 $x_{kj_2 j_2'}$ の値に1を加算する。また、 $c < 1$ の場合、 c の確率で $x_{kj_1 j_1'}$ の値から1を減算し、 $x_{kj_2 j_2'}$ の値に1を加算し、 $1 - c$ の確率で $x_{kj_1 j_1'}$ および $x_{kj_2 j_2'}$ の値を維持する。

[0105] ついで、ステップSF-7で更新した後の第1標本に対応する複数のディプロタイプ形度数およびステップSF-1で定義した行列に基づいて第1標本における各座位でのアレル度を算出し、ステップSF-7で更新した後の第2標本に対応する複数のディプロタイプ形度数およびステップSF-1で定義した行列に基づいて第2標本における各座位でのアレル度を算出する(ステップSF-8)。なお、アレル度の算出に関する説明は、上述した第1実施形態の(1-2)優性・劣性遺伝モデルと同様である。

[0106] ついで、ステップSF-8で算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する(ステップSF-9)。

[0107] ついで、ステップSF-9で作成した偶現表について、座位ごとに独立性検定を実行する(ステップSF-10)。

[0108] ついで、ステップSF-10で実行した独立性検定の検定結果のいずれか1つが有意であった場合(ステップSF-11:Yes)、当該場合の累積回数を更新する(ステップSF-12)。具体的には、累積回数に1を加算する。

[0109] ついで、所定回数が終了したか否かを判定し、所定回数が終了した場合(ステップ

SF-13:Yes)、所定回数に対する最終的な累積回数の割合(累積回数÷所定回数)を第1種の過誤の確率として算出する(ステップSF-14)。一方、所定回数が終了していない場合(ステップSF-13:No)、ステップSF-3～ステップSF-12を実行し、所定回数が終了するまで繰り返す。

[0110] 以上、説明したように、本発明にかかる確率算出方法は、Metropolis-Hastings法により、標本ごとに期待される多項分布に従うディプロタイプ形のサンプルを生成し、それぞれのサンプルについて各座位で検定を行い、そのいずれかの座位で有意となる割合を求める。

[0111] 以上、優性・劣性遺伝モデルで第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法に関する説明を終了する。

[0112] (2-3) 遺伝子型モデル

図7は、遺伝子型モデルで第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法の一例を示すフローチャートである。なお、全てのハプロタイプ頻度が既知であることを前提とする。

まず、各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する(ステップSG-1)。なお、行列の定義に関する説明は、上述した第1実施形態における(1-3)遺伝子型モデルと同様である。

[0113] ついで、ディプロタイプ形度数の組合せを予め設定した第1標本および第2標本に対して生成する(ステップSG-2)。なお、ディプロタイプ形度数の生成に関する説明は、上述した(2-2)優性・劣性遺伝モデルと同様である。

[0114] ついで、第1標本または第2標本のいずれかの標本を選択する(ステップSG-3)。具体的には、標本番号 k ($k=1, 2$)の標本を等確率で選択する。

[0115] ついで、第1ディプロタイプ形を選択する(ステップSG-4)。具体的には、ディプロタイプ形番号 $j_1 j_1'$ ($1 \leq \text{順序付きの自然数 } j_1 \text{ および } j_1' \leq M (=2^L)$; L は座位数)を選択する。

[0116] ついで、ステップSG-3で選択した標本における第1ディプロタイプ形のディプロタイプ形度数が0でない場合(SG-5:Yes)、当該第1ディプロタイプ形とは異なる第2

ディプロタイプ形を選択する(ステップSG-6)。なお、第2ディプロタイプ形の選択に関する説明は、上述した(2-2)優性・劣性遺伝モデルと同様である。

- [0117] ついで、予め設定した第1ディプロタイプ形を構成する各ハプロタイプのハプロタイプ頻度および第2ディプロタイプ形を構成する各ハプロタイプのハプロタイプ頻度に基づいて、第1ディプロタイプ形のディプロタイプ形度数および第2ディプロタイプ形のディプロタイプ形度を更新する(ステップSG-7)。なお、ディプロタイプ形度の更新に関する説明は、上述した(2-2)優性・劣性遺伝モデルと同様である。
- [0118] ついで、ステップSG-7で更新した後の第1標本に対応する複数のディプロタイプ形度数およびステップSG-1で定義した行列に基づいて第1標本における各座位での遺伝子型度を算出し、ステップSG-7で更新した後の第2標本に対応する複数のディプロタイプ形度数およびステップSG-1で定義した行列に基づいて第2標本における各座位での遺伝子型度を算出する(ステップSG-8)。なお、遺伝子型度の算出に関する説明は、上述した第1実施形態の(1-3)遺伝子型モデルと同様である。
- [0119] ついで、ステップSG-8で算出した第1標本における各座位での遺伝子型度数および第2標本における各座位での遺伝子型度数に基づいて、各座位に対応する偶現表を作成する(ステップSG-9)。
- [0120] ついで、ステップSG-9で作成した偶現表について、座位ごとに独立性検定を実行する(ステップSG-10)。
- [0121] ついで、ステップSG-10で実行した独立性検定の検定結果のいずれか1つが有意であった場合(ステップSG-11:Yes)、当該場合の累積回数を更新する(ステップSG-12)。具体的には、累積回数に1を加算する。
- [0122] ついで、所定回数が終了したか否かを判定し、所定回数が終了した場合(ステップSG-13:Yes)、所定回数に対する最終的な累積回数の割合(累積回数÷所定回数)を第1種の過誤の確率として算出する(ステップSG-14)。一方、所定回数が終了していない場合(ステップSG-13:No)、ステップSG-3～ステップSG-12を実行し、所定回数が終了するまで繰り返す。
- [0123] 以上、説明したように、本発明にかかる確率算出方法は、Metropolis-Hastings

法により、標本ごとに期待される多項分布に従うディプロタイプ形のサンプルを生成し、それぞれのサンプルについて各座位で検定を行い、そのいずれかの座位で有意となる割合を求める。

[0124] 以上、遺伝子型モデルで第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法に関する説明を終了する。

[0125] (2-4) 正確なハプロタイプ頻度がわからない場合

正確なハプロタイプ頻度がわからない場合に第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法では、上述した数式24の超幾何分布に従うハプロタイプのサンプルを生成し、それぞれのサンプルについて各座位での検定を行い、そのいずれかの座位で有意となる割合、つまり上述した関数fが1となる割合を求める。

[0126] ここで、一般に、真のハプロタイプ頻度がわからないことも多い。そこで、ハプロタイプ頻度がわからなくても、標本内のディプロタイプ形がわかっている場合には、以下のように計算することが可能である。

[0127] 順列ハプロタイプコピー数は、ハプロタイプ頻度が与えられれば多項分布に従う。ここで、標本空間の確実な出来事の下での確率ではなく、特定の観察データの下での確率を考える。例えば、下記数式28または下記数式29のように制限を加え、それを満足する結果の集合という条件の下での確率を考える。

[数28]

$$\sum_{k=1}^2 X_{kij} = x_{.ij} \quad \cdot \cdot \cdot \quad (\text{数式 2 8})$$

[数29]

$$Y_{kij} = \begin{cases} X_{kij} + X_{kji} (i \neq j) \\ X_{kij} (i = j) \end{cases}$$

$$\sum_{k=1}^2 Y_{kij} = y_{.ij} \quad \dots \quad (\text{数式 29})$$

[0128] ここで、数式29において、 Y_{kij} はk番目の標本における順列ディプロタイプ形コピー数ではなく、組合せディプロタイプ形コピー数をランダム変数で表したものである。一般に、順列ディプロタイプ形を知ることは困難であり、 Y_{kij} を用いる方が妥当であろう。また、一般に上述した優性遺伝モデル、劣性遺伝モデル、アレル頻度モデルおよび遺伝子型モデルでの検定は、全て組合せディプロタイプ形コピー数が分かれば可能である。

[0129] 上述した条件の下では、ランダム変数テンソル $\{X_{kij}\}$ ($k=1, 2; i=1, 2, \dots, M(=2^L); j=1, 2, \dots, M(=2^L)$)または $\{Y_{kij}\}$ ($k=1, 2; i=1, 2, \dots, M(=2^L); 1 \leq j \leq i$)は超幾何分布に従う。その確率はそれぞれ下記数式30および下記数式31で示される。

[数30]

$$P\left(\{X_{kij}\} = \{x_{kij}\} \mid \sum_{k=1}^2 X_{kij} = x_{.ij}\right) = \frac{N_1! N_2! \prod_{i=1}^M \prod_{j=1}^M x_{.ij}!}{(N_1 + N_2)! \prod_{i=1}^M \prod_{j=1}^M \prod_{k=1}^2 x_{kij}!} \dots \quad (\text{数式 30})$$

[数31]

$$P\left(\{Y_{kij}\} = \{y_{kij}\} \mid \sum_{k=1}^2 Y_{kij} = y_{.ij}\right) = \frac{N_1! N_2! \prod_{i=1}^M \prod_{j=1}^M y_{.ij}!}{(N_1 + N_2)! \prod_{i=1}^M \prod_{j=1}^M \prod_{k=1}^2 y_{kij}!} \dots \quad (\text{数式 31})$$

[0130] なお、これらの超幾何分布に従って第1種の過誤の確率を正確に算出する確率算出方法は、先に述べたように、多項分布に従って第1種の過誤の確率を算出する確

率算出方法と同等である。

[0131] つぎに、冗長な計算を避けるため、第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法について説明する。

[0132] 実際の手順は以下のとおりである。

(A-1) 観察データ $\{Y_{kij}\}$ ($k=1, 2; i=1, 2, \dots, M(=2^L); 1 \leq j \leq i$) が与えられている。

(A-2) MCMC法により数式31に示す超幾何分布に従うサンプルを生成する。但し、 $\{y_{kij}\}$ は下記数式32を満たす。ここで、MCMC法では、第1標本の1つの組合せディプロタイプ形と第2標本の1つの組合せディプロタイプ形をランダムに選択し、 y_{kij} に1つ加える、または減じるなどの方法で結果的に超幾何分布になるようにすればよい。

[数32]

$$y_{.ij} = \sum_{k=1}^2 y_{kij} \quad \cdot \cdot \cdot \quad (\text{数式 3 2})$$

(A-3) 生成されたサンプルについて、各座位のそれぞれで、優性遺伝モデル、劣性遺伝モデル、アレル頻度モデルおよび遺伝子型モデルで検定を行う。

(A-4) 検定結果のいずれかで有意である場合に $f(y_{kij})=1$ となり、それ以外で0となる関数 f を定義する。

(A-5) MCMC法で生成したサンプルのうち、 $f=1$ となる割合を経験的な第1種の過誤の確率とする。

[0133] さらに詳細に述べると、実際の手順は以下のとおりである。

(B-1) 任意の整数値である y_{kij} ($k=1, 2; i=1, 2, \dots, M(=2^L); 1 \leq j \leq i; L$ は座位数) が、所定の制約のもとに与えられる。ここで、マルコフ連鎖(Markov-chain)の状態空間は、当該所定の制約のもと、互いに異なる値をとる複数の要素 y_{kij} からなる集合 $\{y_{kij}\}$ である。なお、所定の制約とは、“任意の i および j ($i=1, 2, \dots, M(=2^L); 1 \leq j \leq i; L$ は座位数) に対して $\sum_{k=1}^2 y_{kij} = y_{.ij}$ を満たす”および“任意の k ($k=1, 2$) に対して $\sum_{i=1}^M \sum_{j=1}^i y_{kij} = n_k$ を満たす”というものである。ただし、 $y_{.ij}$ は、予め観測さ

れた(与えられた)固定の非負整数値である。また、 n_k は、標本番号 k の標本に含まれる標本数である。

(B-2) 標本番号 k ($k=1, 2$)を等確率で選択し、選択した標本番号 k の値に応じて標本番号 k' ($k'=1, 2$)をさらに選択する。具体的には、標本番号 k として1を選択した場合には、標本番号 k' として2を選択し、標本番号 k として2を選択した場合には標本番号 k' として1を選択する。

(B-3) 2つの順序つき整数である (u, v) を等確率で選択することにより、ディプロタイプ形番号 uv を決定する。なお、 u は「 $1 \leq u \leq M (=2^L)$; L は座位数」を満たし、 v は「 $1 \leq v \leq u$ 」を満たす。

(B-4) 選択した標本番号 k の標本におけるディプロタイプ形番号 uv のディプロタイプ形度数 y_{kuv} が「 $y_{kuv} > 0$ 」を満たす場合には、選択した (u, v) とは異なる2つの順序つき整数である (w, s) をさらに選択することによりディプロタイプ形番号 ws をさらに決定する。ここで、 w は「 $1 \leq w \leq M (=2^L)$; L は座位数」を満たし、 s は「 $1 \leq s \leq w$ 」を満たす。一方、ディプロタイプ形度数 y_{kuv} が「 $y_{kuv} = 0$ 」を満たす場合には、ディプロタイプ形度数 y_{kuv} の値を維持して、後述する(B-6)へ進み、そして上述した(B-2)へ戻る。

さらに、選択した標本番号 k' の標本におけるディプロタイプ形番号 ws のディプロタイプ形度数 $y_{k'ws}$ が「 $y_{k'ws} = 0$ 」を満たす場合には、ディプロタイプ形度数 $y_{k'ws}$ の値を維持して、後述する(B-6)へ進み、そして上述した(B-2)へ戻る。

(B-5) 新たなディプロタイプ形度数(y_{kuv}^* 、 y_{kws}^* 、 $y_{k'uv}^*$ 、 $y_{k'ws}^*$)を算出する。具体的には、まず、下記数式33で定義される c の値を算出する。なお、下記数式33において、 y_{kuv}^* は「 $y_{kuv}^* = y_{kuv} - 1$ 」を満たし、 y_{kws}^* は「 $y_{kws}^* = y_{kws} + 1$ 」を満たし、 $y_{k'uv}^*$ は「 $y_{k'uv}^* = y_{k'uv} + 1$ 」を満たし、 $y_{k'ws}^*$ は「 $y_{k'ws}^* = y_{k'ws} - 1$ 」を満たす。

[数33]

$$c = \frac{y_{kuv}! y_{kws}! y_{k'uv}! y_{k'ws}!}{y_{kuv}^*! y_{kws}^*! y_{k'uv}^*! y_{k'ws}^*!} = \frac{y_{kuv} y_{k'ws}}{(y_{k'uv} + 1)(y_{kws} + 1)} \dots \quad (\text{数式 3 3})$$

ついで、 $c \geq 1$ の場合、新たなディプロタイプ形度数として、 y_{kuv} 、 y_{kws} 、 $y_{k'uv}$ および $y_{k'ws}$ の代わりに y_{kuv}^* 、 y_{kws}^* 、 $y_{k'uv}^*$ および $y_{k'ws}^*$ を用い、後述する(B-6)へ進み、そして上述した(B-2)へ戻る。一方、 $c < 1$ の場合、 c の確率で、新たなディプロタイプ形度数として、 y_{kuv} 、 y_{kws} 、 $y_{k'uv}$ および $y_{k'ws}$ の代わりに y_{kuv}^* 、 y_{kws}^* 、 $y_{k'uv}^*$ および $y_{k'ws}^*$ を用い、 $1-c$ の確率で y_{kuv} 、 y_{kws} 、 $y_{k'uv}$ および $y_{k'ws}$ の値を維持し、後述する(B-6)へ進み、そして上述した(B-2)へ戻る。

(B-6)劣性遺伝モデル、優性遺伝モデル、アレル頻度モデル、遺伝子型モデルの各モデルで、3つの偶現表のどれかを用いて、表現型と遺伝子型との間の独立性検定を座位ごとに実行する。ここで、劣性遺伝モデルの場合には図16の偶現表を用い、優性遺伝モデルの場合には図17の偶現表を用いる。ただし、図16において、数式15の x_{1ij} および x_{2ij} の代わりにそれぞれ、 y_{1ij} および y_{2ij} を用いる。また、図17において、数式16の x_{1ij} および x_{2ij} の代わりにそれぞれ、 y_{1ij} および y_{2ij} を用いる。また、遺伝子型モデルの場合には図18の偶現表を用い、アレル頻度モデルの場合には図25の偶現表を用いる。ただし、図18において、数式22の x_{1ij} および x_{2ij} の代わりにそれぞれ、 y_{1ij} および y_{2ij} を用いる。また、図25において、 x_{1ij} および x_{2ij} の代わりにそれぞれ、 y_{1ij} および y_{2ij} を用いる。

(B-7)実行した独立性検定の検定結果のいずれか1つが有意であった場合、実行した全体(全ての座位について)の独立性検定の検定結果が有意であると判定する。

(B-8)これまでのMCMC法による処理を所定回数分繰り返し実行し、(B-7)で有意であると判定された回数(有意判定回数)の割合(有意判定回数÷所定回数)を、全体(全ての座位について)の独立性検定における経験的な第1種の過誤の確率として算出する。

ここで、上述した手順では3つのモデル(優性遺伝モデル、劣性遺伝モデル、遺伝子型モデル)のどれかを用いた独立性検定について述べたが、本発明にかかる確率算出方法は、3つのモデル全てを用いた独立性検定に容易に拡張することができる。したがって、上述した(B-6)で述べたように、独立性検定は3つの異なる偶現表を用いて各座位について行われ、いずれかのモデル・いずれかの座位での独立性検

定の検定結果が有意であるなら、全体(全ての座位について)の独立性検定の検定結果を有意であるとする。

- [0134] 以上、正確なハプロタイプ頻度がわからない場合に第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法に関する説明を終了する。

産業上の利用可能性

- [0135] 以上のように、本発明にかかる確率算出方法は、第1種の過誤の確率を、ハプロタイプ頻度を考慮して直接的に算出することができ、その結果、疾患関連SNPの見落としを防ぐことができる。そのため、本発明にかかる確率算出方法は、医療や創薬などの分野において極めて有用である。

請求の範囲

- [1] SNPを用いたcase-control相関解析で行う独立性検定における第1種の過誤の確率を算出する確率算出方法において、
- 複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する標本取得ステップと、
- 前記標本取得ステップで取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出するハプロタイプ頻度算出ステップと、
- 各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する行列定義ステップと、
- 第1標本における各ハプロタイプのハプロタイプ度数および第2標本における各ハプロタイプのハプロタイプ度数を設定することで、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せを与えるハプロタイプ度数設定ステップと、
- 前記ハプロタイプ度数設定ステップで設定した第1標本に対応する複数のハプロタイプ度数および前記行列定義ステップで定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、設定した第2標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第2標本における各座位でのアレル度数を算出するアレル度数算出ステップと、
- 前記アレル度数算出ステップで算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する偶現表作成ステップと、
- 前記偶現表作成ステップで作成した偶現表について、座位ごとに独立性検定を実行する独立性検定実行ステップと、
- 前記独立性検定実行ステップで実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のハプロタイプ度数、第2標本のハプロタイプ度数およびハプロタイプ頻度に基づいて、前記ハプロタイプ度数設定ステップで設定した第1標本の各

ハプロタイプ度数と第2標本の各ハプロタイプ度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのハプロタイプ度数に関する結合分布を用いて算出する確率算出ステップと、

を含み、

前記ハプロタイプ度数設定ステップ、前記アレル度数算出ステップ、前記偶現表作成ステップ、前記独立性検定実行ステップおよび前記確率算出ステップを、第1標本の各ハプロタイプ度数と第2標本の各ハプロタイプ度数との全ての組合せについて繰り返し、繰り返すごとに算出される前記確率を加算することで、最終的に第1種の過誤の確率を算出すること、

を特徴とする確率算出方法。

- [2] SNPを用いたcase-control相関解析で行う独立性検定における第1種の過誤の確率を算出する確率算出方法において、

複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する標本取得ステップと、

前記標本取得ステップで取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出するハプロタイプ頻度算出ステップと、

各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する行列定義ステップと、

第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与えるディプロタイプ形度数設定ステップと、

前記ディプロタイプ形度数設定ステップで設定した第1標本に対応する複数のディプロタイプ形度数および前記行列定義ステップで定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、設定した第2標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第2標本における各座位でのアレ

ル度数を算出するアレル度数算出ステップと、

前記アレル度数算出ステップで算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する偶現表作成ステップと、

前記偶現表作成ステップで作成した偶現表について、座位ごとに独立性検定を実行する独立性検定実行ステップと、

前記独立性検定実行ステップで実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、前記ディプロタイプ形度数設定ステップで設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する確率算出ステップと、

を含み、

前記ディプロタイプ形度数設定ステップ、前記アレル度数算出ステップ、前記偶現表作成ステップ、前記独立性検定実行ステップおよび前記確率算出ステップを、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せについて繰り返し、繰り返すごとに算出される前記確率を加算することで、最終的に第1種の過誤の確率を算出すること、

を特徴とする確率算出方法。

- [3] SNPを用いたcase-control相関解析で行う独立性検定における第1種の過誤の確率を算出する確率算出方法において、

複数のSNPで構成されるハプロタイプに関するハプロタイプ情報を複数含む第1標本および第2標本を取得する標本取得ステップと、

前記標本取得ステップで取得した第1標本に含まれるハプロタイプ情報および取得した第2標本に含まれるハプロタイプ情報に基づいて、存在し得るハプロタイプごとにハプロタイプ頻度を算出するハプロタイプ頻度算出ステップと、

各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従っ

て予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する行列定義ステップと、

第1標本における各ディプロタイプ形のディプロタイプ形度数および第2標本における各ディプロタイプ形のディプロタイプ形度数を設定することで、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せを与えるディプロタイプ形度数設定ステップと、

前記ディプロタイプ形度数設定ステップで設定した第1標本に対応する複数のディプロタイプ形度数および前記行列定義ステップで定義した行列に基づいて第1標本における各座位での遺伝子型度を算出し、設定した第2標本に対応する複数のディプロタイプ形度数および定義した行列に基づいて第2標本における各座位での遺伝子型度を算出する遺伝子型度数算出ステップと、

前記遺伝子型度数算出ステップで算出した第1標本における各座位での遺伝子型度数および第2標本における各座位での遺伝子型度数に基づいて、各座位に対応する偶現表を作成する偶現表作成ステップと、

前記偶現表作成ステップで作成した偶現表について、座位ごとに独立性検定を実行する独立性検定実行ステップと、

前記独立性検定実行ステップで実行した独立性検定の検定結果のいずれか1つが有意であった場合、第1標本の標本数、第2標本の標本数、存在し得るハプロタイプの総数、第1標本のディプロタイプ形度数、第2標本のディプロタイプ形度数およびハプロタイプ頻度に基づいて、前記ディプロタイプ形度数設定ステップで設定した第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との組合せに対応する確率を、予め設定した第1標本と第2標本とのディプロタイプ形度数に関する結合分布を用いて算出する確率算出ステップと、

を含み、

前記ディプロタイプ形度数設定ステップ、前記遺伝子型度数算出ステップ、前記偶現表作成ステップ、前記独立性検定実行ステップおよび前記確率算出ステップを、第1標本の各ディプロタイプ形度数と第2標本の各ディプロタイプ形度数との全ての組合せについて繰り返し、繰り返すごとに算出される前記確率を加算することで、最

最終的に第1種の過誤の確率を算出すること、
を特徴とする確率算出方法。

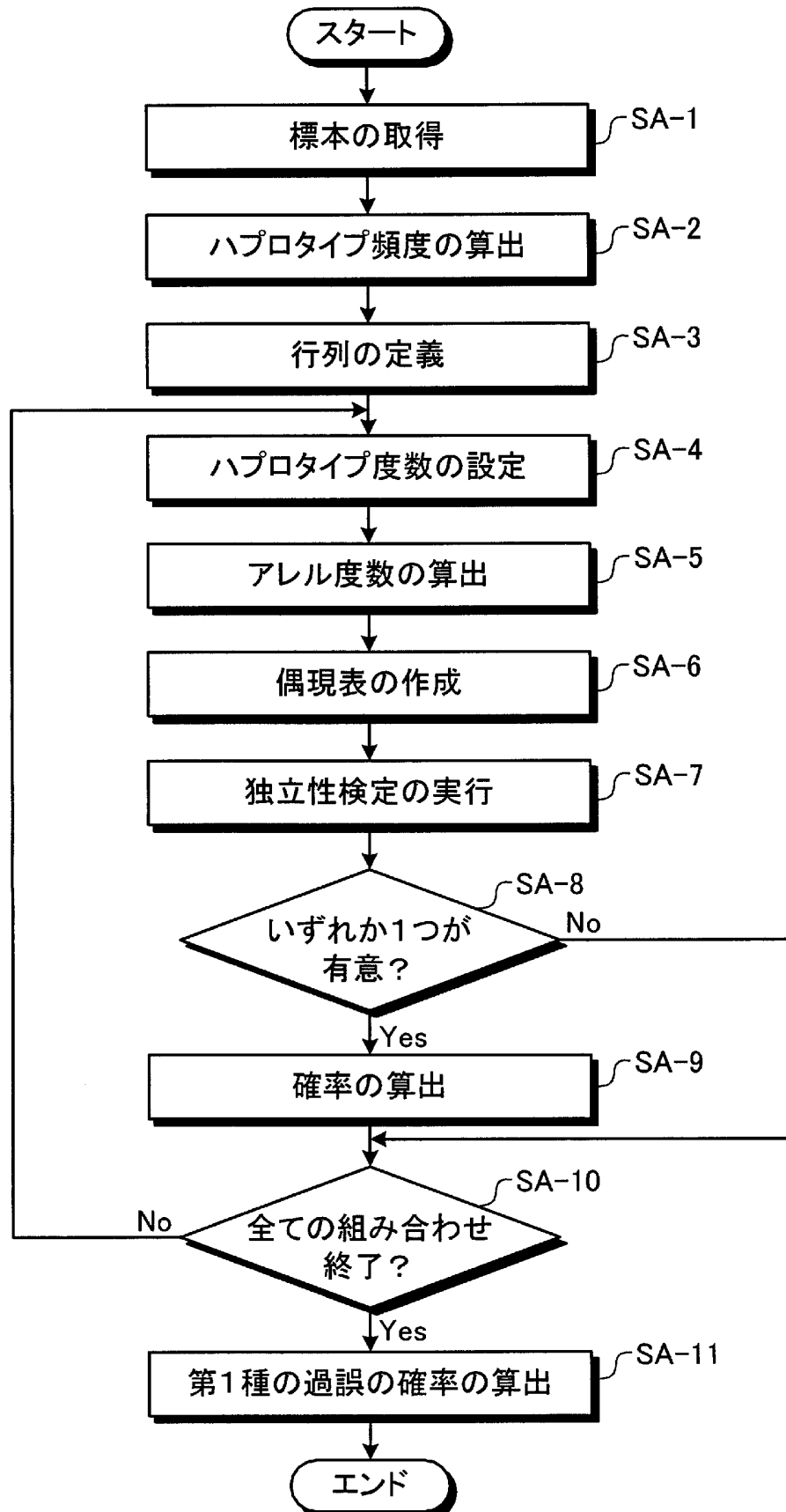
- [4] SNPを用いたcase-control相関解析で行う独立性検定における第1種の過誤の確率をMCMC法を用いて近似的に算出する確率算出方法であって、
各ハプロタイプの各座位がマイナーアレルまたはメジャーアレルを有するかに従って予め定めた所定値を設定することで、当該所定値を成分とする行列を定義する行列定義ステップと、
ハプロタイプ度数の組合せを予め設定した第1標本および第2標本に対して生成するハプロタイプ度数生成ステップと、
第1標本または第2標本のいずれかの標本を選択する標本選択ステップと、
第1ハプロタイプを選択する第1ハプロタイプ選択ステップと、
前記標本選択ステップで選択した標本における前記第1ハプロタイプのハプロタイプ度数が0でない場合、当該第1ハプロタイプとは異なる第2ハプロタイプを選択する第2ハプロタイプ選択ステップと、
予め設定した前記第1ハプロタイプのハプロタイプ頻度および前記第2ハプロタイプのハプロタイプ頻度に基づいて、前記第1ハプロタイプのハプロタイプ度数および前記第2ハプロタイプのハプロタイプ度数を更新するハプロタイプ度数更新ステップと、
前記ハプロタイプ度数更新ステップで更新した後の第1標本に対応する複数のハプロタイプ度数および前記行列定義ステップで定義した行列に基づいて第1標本における各座位でのアレル度数を算出し、更新した後の第2標本に対応する複数のハプロタイプ度数および定義した行列に基づいて第2標本における各座位でのアレル度数を算出するアレル度数算出ステップと、
前記アレル度数算出ステップで算出した第1標本における各座位でのアレル度数および第2標本における各座位でのアレル度数に基づいて、各座位に対応する偶現表を作成する偶現表作成ステップと、
前記偶現表作成ステップで作成した偶現表について、座位ごとに独立性検定を実行する独立性検定実行ステップと、
前記独立性検定実行ステップで実行した独立性検定の検定結果のいずれか1つ

が有意であった場合、当該場合の累積回数を更新する累積回数更新ステップと、
を含み、

前記標本選択ステップ、前記第1ハプロタイプ選択ステップ、前記第2ハプロタイプ
選択ステップ、前記ハプロタイプ度数更新ステップ、前記アレル度数算出ステップ、
前記偶現表作成ステップ、前記独立性検定実行ステップおよび前記累積回数更新
ステップを所定回数繰り返し、前記所定回数に対する最終的な累積回数の割合を第
1種の過誤の確率として算出すること、
を特徴とする確率算出方法。

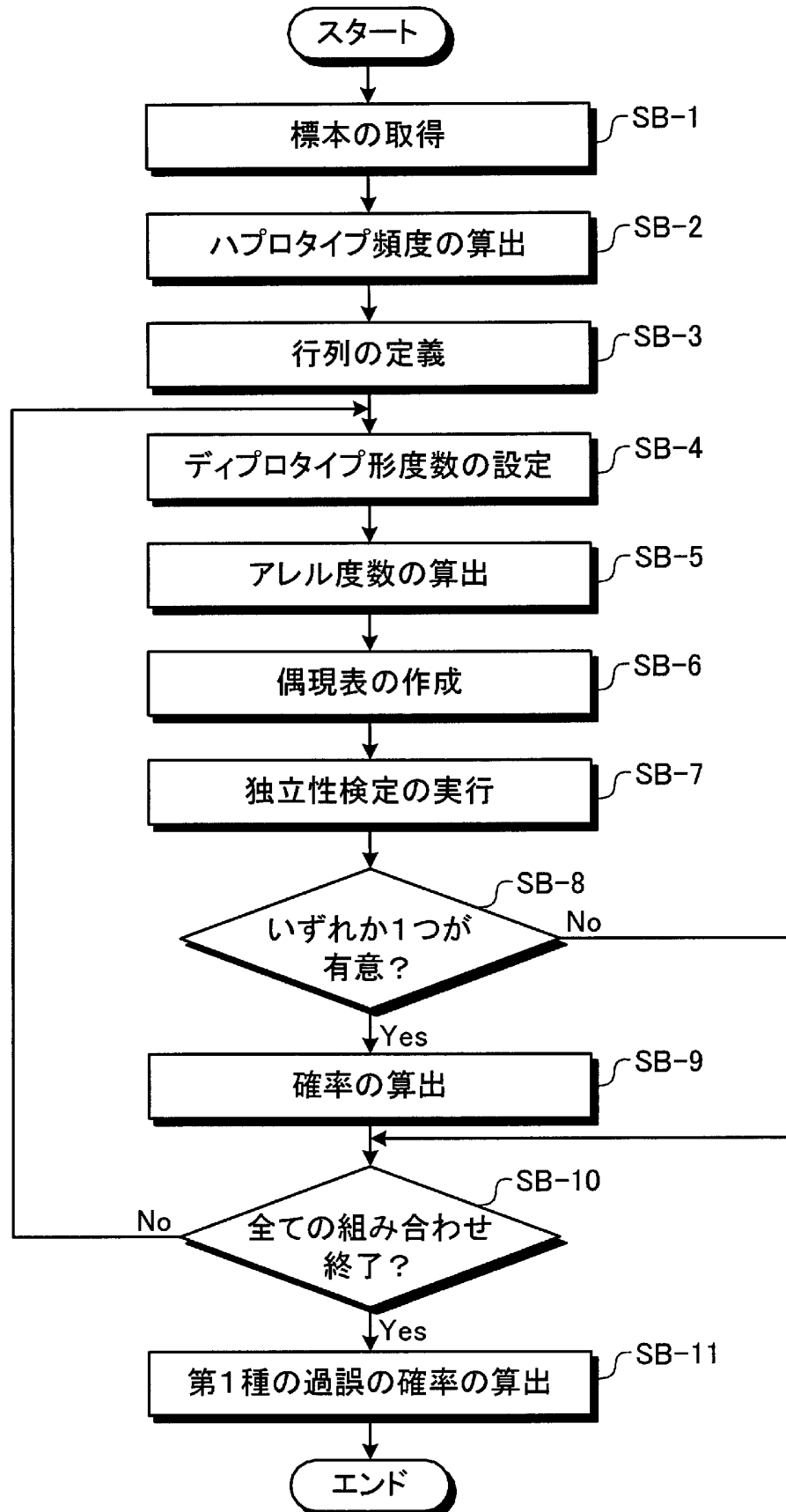
[図1]

(正確計算(アレル頻度モデル))



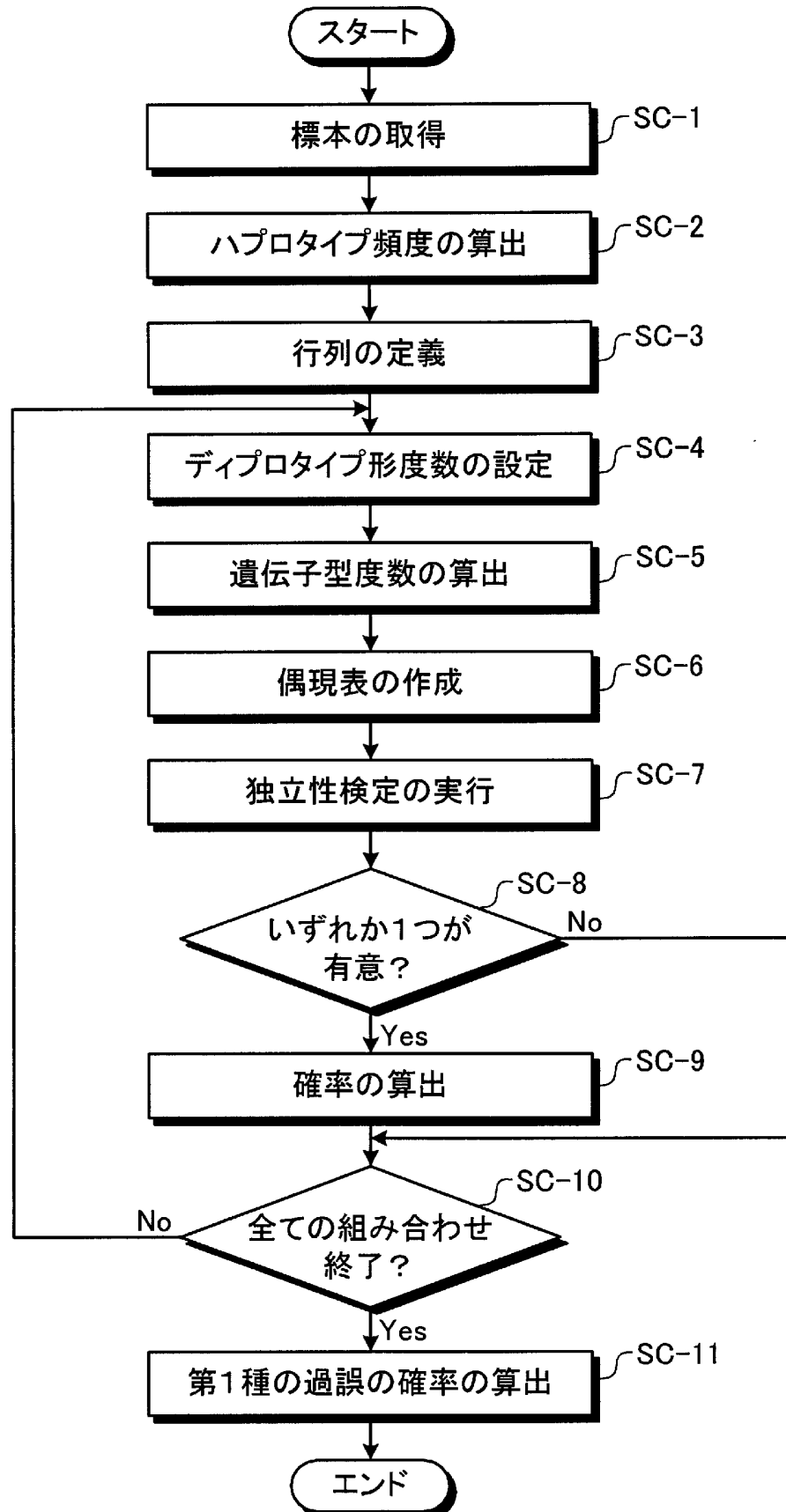
[図2]

(正確計算(優性・劣性遺伝モデル))



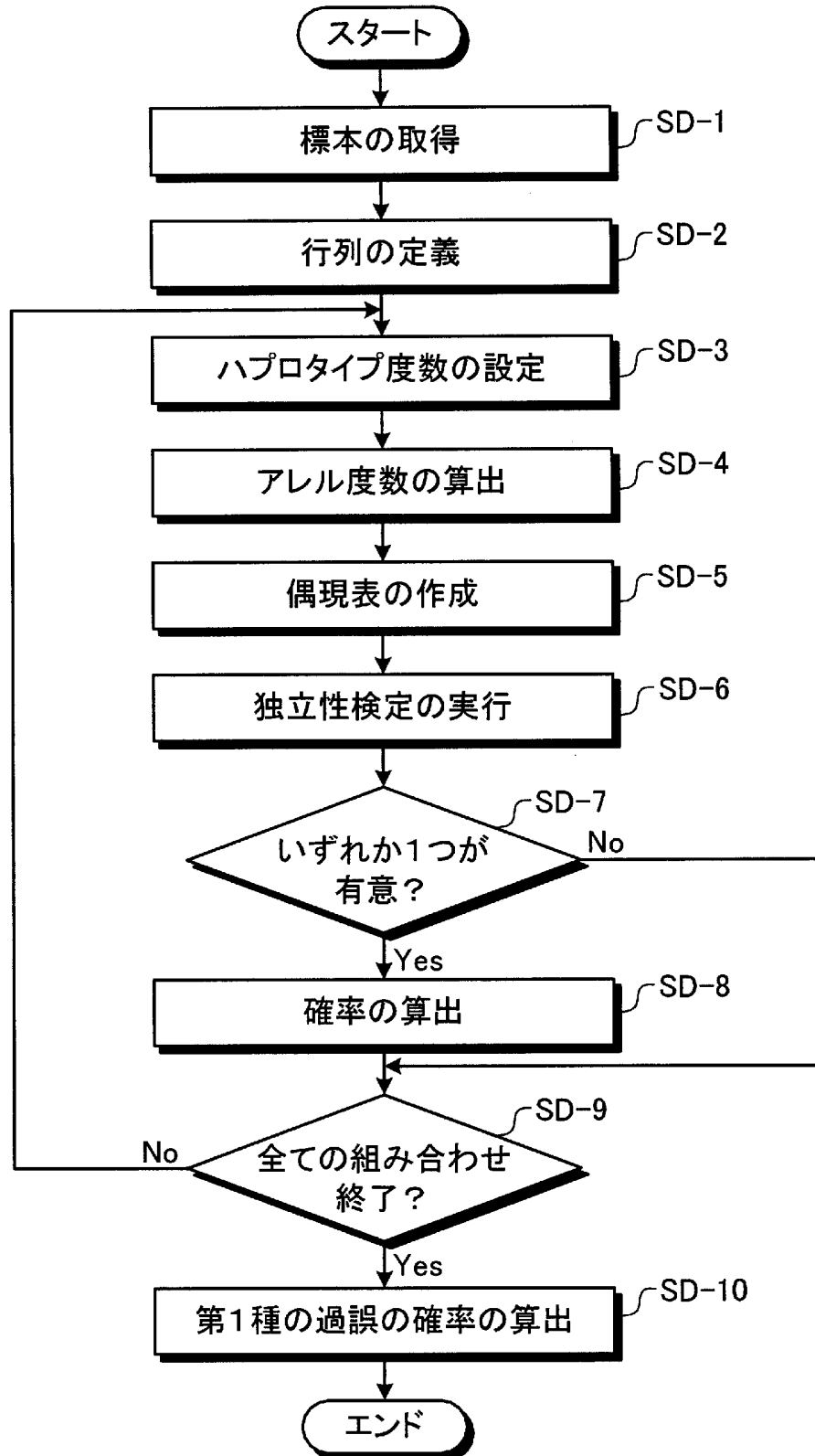
[図3]

(正確計算(遺伝子型モデル))

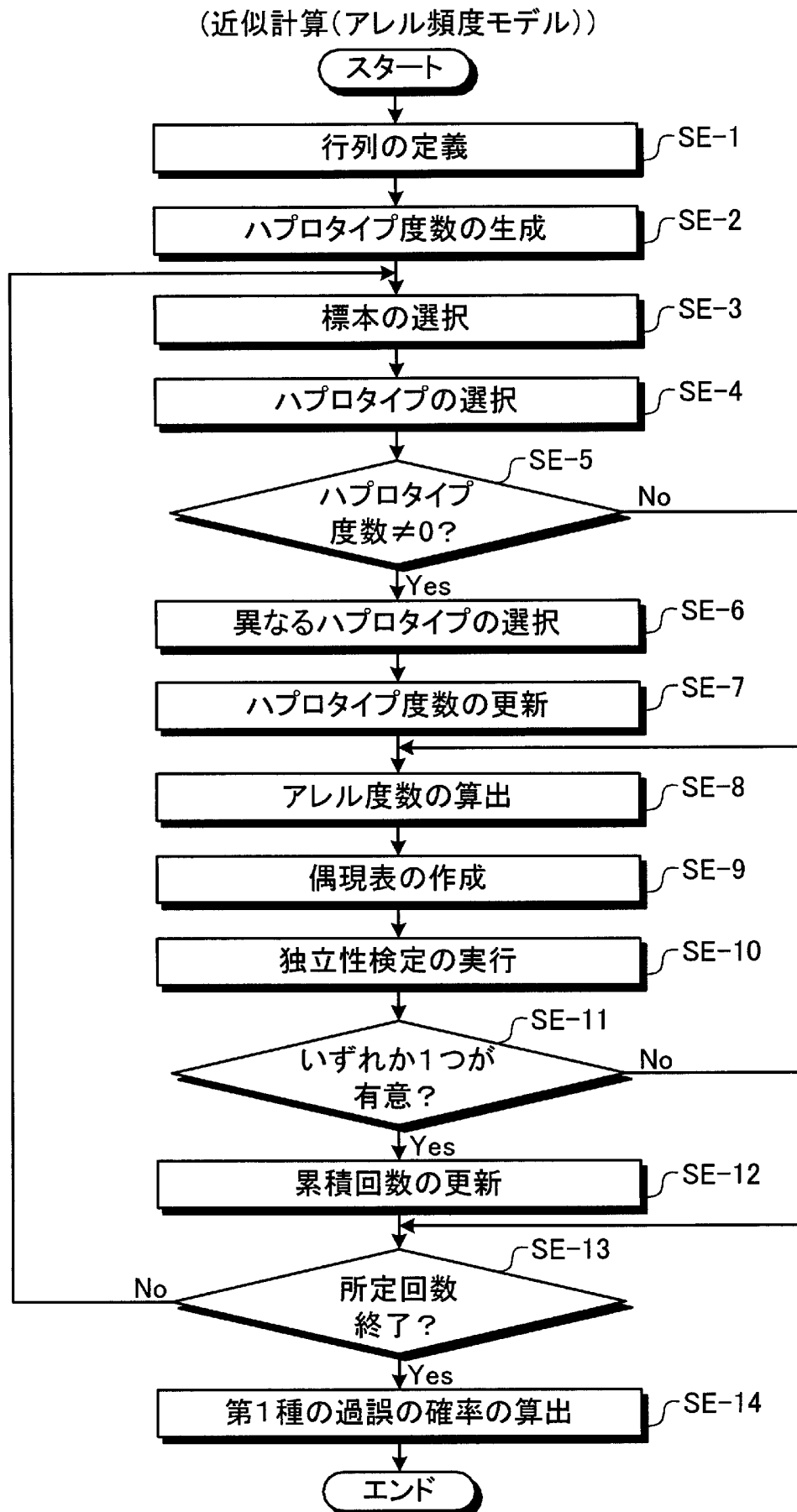


[図4]

(正確計算(ハプロタイプ頻度がわからない場合))

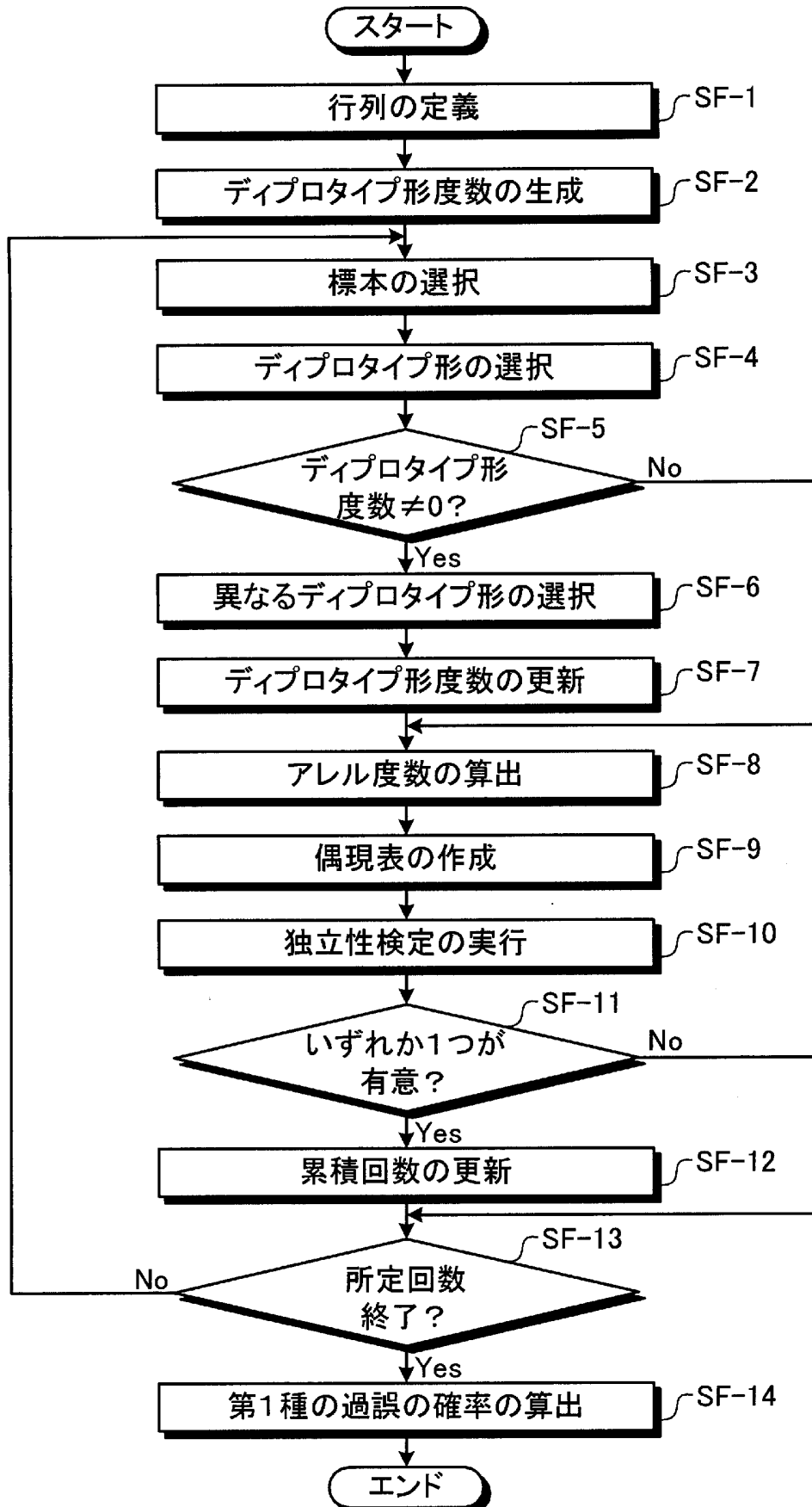


[図5]

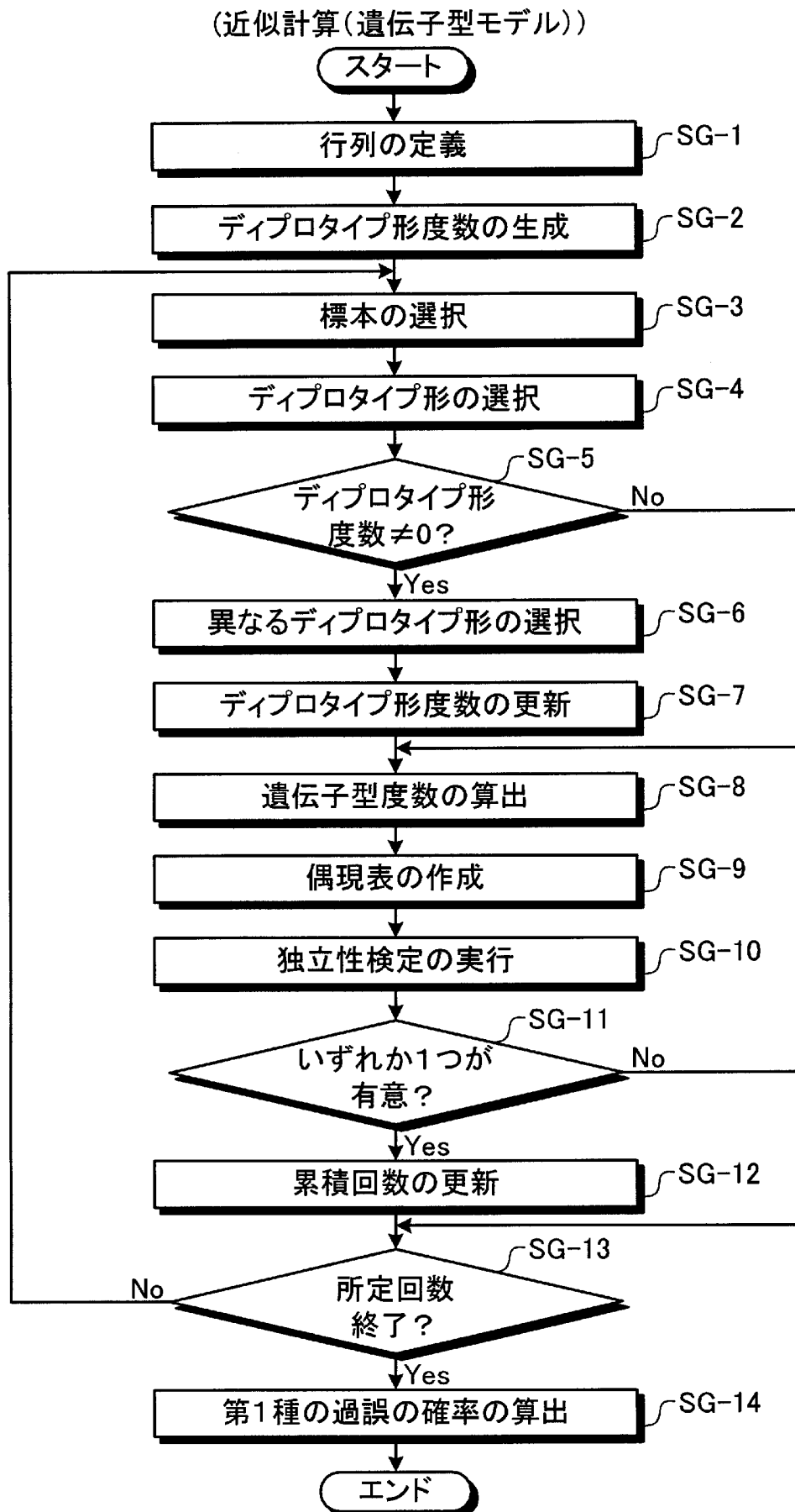


[図6]

(近似計算(優性・劣性遺伝モデル))



[図7]



[図8]

第1標本		第2標本			
サンプル番号	ハプロタイプ情報				
	座位1	座位2	...	座位L	
	1	A	G	...	T
	⋮	⋮	⋮	⋮	⋮
	N ₁				

[図9]

ハプロタイプ 番号	座位1	座位2	...	座位L
1			...	
⋮	⋮	⋮	⋮	⋮
$M(=2^L)$			...	

[図10]

ハプロタイプ番号	ハプロタイプ頻度
1	h_1
$\vdots$	$\vdots$
M	h_M

[図11]

座位番号	ハプロタイプ番号			
	1	2	...	M
1			...	
2			...	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
L			...	

[図12]

ハプロタイプ番号	ハプロタイプ度数	
	第1標本	第2標本
1	X_{11}	X_{21}
$\vdots$	$\vdots$	$\vdots$
M	X_{1M}	X_{2M}

[図13]

	マイナーアレル 度数(a)	メジャーアレル 度数(A)
第1標本	y_{1i}	$2N_1 - y_{1i}$
第2標本	y_{2i}	$2N_2 - y_{2i}$

[図14]

ディプロタイプ形番号	
ハプロタイプ番号	ハプロタイプ番号
1	1
⋮	⋮
M	M

[図15]

ディプロタイプ形 番号	ディプロタイプ形度数	
	第1標本	第2標本
11	X_{111}	X_{211}
12	X_{112}	X_{212}
⋮	⋮	⋮
MM	X_{1MM}	X_{2MM}

[図16]

	マイナーアレル 度数(aa+Aa)	メジャーアレル 度数(AA)
第1標本	y_{1i}	$N_1 - y_{1i}$
第2標本	y_{2i}	$N_2 - y_{2i}$

[図17]

	マイナーアレル 度数(aa)	メジャーアレル 度数(AA+Aa)
第1標本	y_{1i}	$N_1 - y_{1i}$
第2標本	y_{2i}	$N_2 - y_{2i}$

[図18]

	マイナーホモ度数 (aa)	ヘテロ度数 (Aa)	メジャーホモ度数 (AA)
第1標本	y_{1i}	$N_1 - (y_{1i} + y'_{1i})$	y'_{1i}
第2標本	y_{2i}	$N_2 - (y_{2i} + y'_{2i})$	y'_{2i}

[図19]

	マイナーアレル度数	メジャーアレル度数
第1標本	$\sum_{i=1}^s \alpha_{ik} x_{1i} + b_{1k} x_{1s+1}$	$2N_1 - \sum_{i=1}^s \alpha_{ik} x_{1i} - b_{1k} x_{1s+1}$
第2標本	$\sum_{i=1}^s \alpha_{ik} x_{2i} + b_{2k} x_{2s+1}$	$2N_2 - \sum_{i=1}^s \alpha_{ik} x_{2i} - b_{2k} x_{2s+1}$

[図20]

	マイナーアレル度数	メジャーアレル度数
第1標本	$\sum_{i=1}^s \alpha_{ik} x_{li} + x_{ls+1}$	$2N_1 - \sum_{i=1}^s \alpha_{ik} x_{li} - x_{ls+1}$
第2標本	$\sum_{i=1}^s \alpha_{ik} x_{2i}$	$2N_2 - \sum_{i=1}^s \alpha_{ik} x_{2i}$

[図21]

	マイナーアレル度数	メジャーアレル度数
第1標本	$\sum_{i=1}^s \alpha_{ik} x_{li}$	$2N_1 - \sum_{i=1}^s \alpha_{ik} x_{li}$
第2標本	$\sum_{i=1}^s \alpha_{ik} x_{2i} + x_{2s+1}$	$2N_2 - \sum_{i=1}^s \alpha_{ik} x_{2i} - x_{2s+1}$

[図22]

	マイナーアレル度数	メジャーアレル度数
第1標本	$\sum_{i=1}^s \alpha_{ik} X_{li} + X_{1s+1}$	$2N_1 - \sum_{i=1}^s \alpha_{ik} X_{li} - X_{1s+1}$
第2標本	$\sum_{i=1}^s \alpha_{ik} X_{2i} + X_{2s+1}$	$2N_2 - \sum_{i=1}^s \alpha_{ik} X_{2i} - X_{2s+1}$

[図23]

	マイナーアレル度数	メジャーアレル度数
第1標本	$\sum_{i=1}^s \alpha_{ik} X_{li}$	$2N_1 - \sum_{i=1}^s \alpha_{ik} X_{li}$
第2標本	$\sum_{i=1}^s \alpha_{ik} X_{2i}$	$2N_2 - \sum_{i=1}^s \alpha_{ik} X_{2i}$

[図24]

	マイナーアレル度数	メジャーアレル度数
第1標本	$\sum_{i=1}^s \alpha_{ik} x_{li} + z_{lk}$	$2N_1 - \sum_{i=1}^s \alpha_{ik} x_{li} - z_{lk}$
第2標本	$\sum_{i=1}^s \alpha_{ik} x_{2i} + z_{2k}$	$2N_2 - \sum_{i=1}^s \alpha_{ik} x_{2i} - z_{2k}$

[図25]

	マイナーアレル度数	メジャーアレル度数
第1標本	$\sum_{j=1}^M \sum_{j'=1}^M (\alpha_{ij} + \alpha_{ij'}) x_{1jj'}$	$2N_2 - \sum_{j=1}^M \sum_{j'=1}^M (\alpha_{ij} + \alpha_{ij'}) x_{2jj'}$
第2標本	$\sum_{j=1}^M \sum_{j'=1}^M (\alpha_{ij} + \alpha_{ij'}) x_{2jj'}$	$2N_1 - \sum_{j=1}^M \sum_{j'=1}^M (\alpha_{ij} + \alpha_{ij'}) x_{1jj'}$

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2005/022317

A. CLASSIFICATION OF SUBJECT MATTER G06F19/00 (2006.01)		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06F19/00 (2006.01)		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2006 Kokai Jitsuyo Shinan Koho 1971-2006 Toroku Jitsuyo Shinan Koho 1994-2006		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) WPI, PubMed, JSTPlus (JOIS)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2004-192018 A (Japan Biological Informatics Consortium), 08 July, 2004 (08.07.04), Full text; all drawings (Family: none)	1-4
A	Naoyuki KAMATANI et al., "Haplotype o Kiso ni shita Human Genome Takei to Hyogengata no Kanren no Kaiseki Shuho", Japanese Federation of Statistical Science Associations, 06 September, 2004 (06.09.04), Vol.2004, pages 188 to 189	1-4
A	Makoto TOMITA et al., "Rensa Fuheiko Kaiseki ni okeru Tazai no Kankei", Japanese Federation of Statistical Science Associations, 05 September, 2003 (05.09.03), Vol.2003, pages 227 to 228	1-4
☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 22 February, 2006 (22.02.06)		Date of mailing of the international search report 07 March, 2006 (07.03.06)
Name and mailing address of the ISA/ Japanese Patent Office		Authorized officer
Facsimile No.		Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2005/022317

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2004-229511 A (NTT Data Corp.), 19 August, 2004 (19.08.04), Full text; all drawings (Family: none)	1-4
A	NYHOLT D. R., A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other, Am.J.Hum.Genet., 2004, Vol. 74, pp. 765-769	1-4
A	XIONG M, et al., THE HAPLOTYPE LINKAGE DISEQUILIBRIUM TEST FOR GENOME-WIDE SCREENS: ITS POWER AND STUDY DESIGN, Pacific Symposium on Biocomputing, 2000, pp.672-681	1-4

A. 発明の属する分野の分類 (国際特許分類 (IPC)) Int.Cl. G06F19/00 (2006.01)			
B. 調査を行った分野 調査を行った最小限資料 (国際特許分類 (IPC)) Int.Cl. G06F19/00 (2006.01)			
最小限資料以外の資料で調査を行った分野に含まれるもの 日本国実用新案公報 1922-1996年 日本国公開実用新案公報 1971-2006年 日本国実用新案登録公報 1996-2006年 日本国登録実用新案公報 1994-2006年			
国際調査で使用した電子データベース (データベースの名称、調査に使用した用語) WPI, PubMed, JSTPlus (JOIS)			
C. 関連すると認められる文献			
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号	
A	JP 2004-192018 A (社団法人バイオ産業情報化コンソーシアム) 2004.07.08, 全文, 全図 (ファミリーなし)	1-4	
A	鎌谷直之, 外, ハプロタイプを基礎にしたヒトゲノム多型と表現型の 関連の解析手法, 統計関連学会連合大会講演報告集, 2004.09.06, vol. 2004, pp. 188-189	1-4	
☒ C欄の続きにも文献が列挙されている。 ☐ パテントファミリーに関する別紙を参照。			
* 引用文献のカテゴリー 「A」特に関連のある文献ではなく、一般的技術水準を示すもの 「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの 「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す) 「O」口頭による開示、使用、展示等に言及する文献 「P」国際出願日前で、かつ優先権の主張の基礎となる出願		の日の後に公表された文献 「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの 「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの 「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの 「&」同一パテントファミリー文献	
国際調査を完了した日 22.02.2006		国際調査報告の発送日 07.03.2006	
国際調査機関の名称及びあて先 日本国特許庁 (ISA/J P) 郵便番号100-8915 東京都千代田区霞が関三丁目4番3号		特許庁審査官 (権限のある職員) 岩間 直純 電話番号 03-3581-1101 内線 3562	5 L 9287

C (続き). 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	富田誠, 外, 連鎖不平衡解析における多座位の関係, 統計関連学会 連合大会講演報告集, 2003.09.05, vol. 2003, pp. 227-228	1-4
A	JP 2004-229511 A (株式会社エヌ・ティ・ティ・データ) 2004.08.19, 全文, 全図 (ファミリーなし)	1-4
A	NYHOLT D. R., A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other, Am. J. Hum. Genet. 2004, vol. 74, pp. 765-769	1-4
A	XIONG M, et al., THE HAPLOTYPE LINKAGE DISEQUILIBRIUM TEST FOR GENOME-WIDE SCREENS: ITS POWER AND STUDY DESIGN, Pacific Symposium on Biocomputing, 2000, pp.672-681	1-4