ゲノム多様性と抗癌剤副作用―関連遺伝子探索のための方法論について―

松浦正明[1]

要旨 ―― 目的. 抗癌剤に対する副作用関連遺伝子探索のための一塩基多型（SNP）を用いたアソシエーション解析を考察し，統計学的検出力の高い方法を検討する．**方法.** 主要なハプロタイプの数を考慮し，SNP 単位での解析手法とハプロタイプ単位での解析手法の比較検討を行う．さらにハプロタイプブロックを考察する事により，ハプロタイプの数とアソシエーション解析における統計学的検出力との関係を調べる．**結果.** 主要なハプロタイプの種類が 2 種類の場合は SNP 単独の解析でも信頼性の高い結果を得られるが，主要なハプロタイプが 3 種類以上の場合は重要な関連遺伝子を見逃す可能性がある．副作用関連遺伝子の探索のためのアソシエーション解析で検出力を高めるためには，ハプロタイプブロックの適切な同定と，ハプロタイプの論理和モデルの利用が重要である．**結論.** 抗癌剤の治療効果や副作用に関連する遺伝子やマーカーの探索のための SNP を用いたアソシエーション解析における注意点と統計的検出力の高い解析法を提示した．（肺癌．2006;46:253-258）

索引用語 ―― 副作用，SNP，ハプロタイプ，治療感受性予測，個別化医療

Genome Diversity And Adverse Effects of Anticancer Drug —On Methodology for Search for Relevant Genes—

Masaaki Matsuura[1]

ABSTRACT ―― Objective. We discuss methodologies using SNPs to search for genes related with adverse effects of anticancer drugs and introduce statistical methods which have high statistical power. ***Method.*** We compared haplotype-based methods and separate-SNP-based methods according to the number of major haplotypes. Furthermore we examined relationships of the number of haplotypes and statistical power of association study by considering haplotype blocks. ***Results.*** Reliable results were obtained when the number of major haplotypes was two. There is a possibility of missing important related genes when the number of haplotypes is three or more. Accurate detection of haplotype blocks and the use of logical union models are important to gain statistical power in association studies to determine genes related with adverse effects. ***Conclusion.*** We identified problems and methodologies with statistically high power of the association study based on SNPs in search for genes or makers of treatment effects or adverse effects of anti-cancer drugs.(*JJLC.* 2006;46:253-258)

KEY WORDS ―― adverse effect, SNP, haplotype, prediction, personalized medicine

はじめに

ゲノム情報の個人差を手がかりとして個々の患者に最適な治療法を提供する「オーダーメイド医療」を実現するために，現在様々な試みがなされている．ゲフィチニブなどの抗癌剤の投与前に，あらかじめ重篤な副作用と関連する遺伝子や治療効果を規定する遺伝子が同定できれば，重篤な副作用の事前の回避や抗癌剤が有効となる患者を特定する事が可能となる．さらに，個々の患者の体質に応じた投薬量の決定や最適薬剤の選定などの次世

[1](財)癌研究会ゲノムセンター情報解析グループ．
別刷請求先：松浦正明，(財)癌研究会ゲノムセンター情報解析グループ，〒135-8550　東京都江東区有明 3-10-6.
[1]Japanese Foundation for Cancer Research, Japan.
Reprints: Masaaki Matsuura, Japanese Foundation for Cancer Research, Japan, Ariake 3-10-6, Koto-ku, Tokyo 135-8550, Japan.

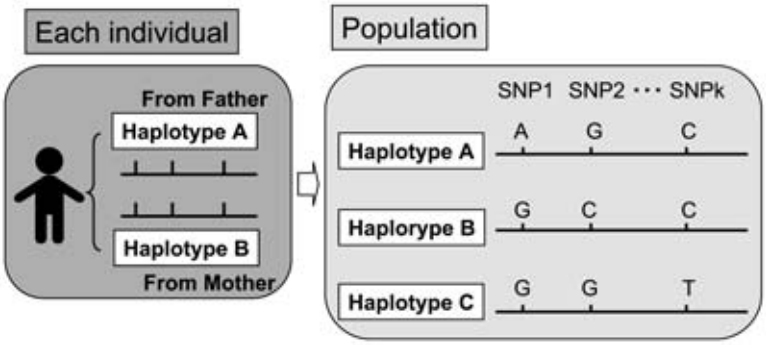

Figure 1. SNPs and Haplotypes.

代個別化医療や，医薬品を開発するゲノム創薬にも重要な役割を果たすものと期待されている．

癌研究会ゲノムセンターでは，癌患者の遺伝学的体質診断に向けた研究として，化学療法剤に対する薬剤感受性に関連する遺伝子の探索・同定を試みている．その方法は一塩基多型（SNP）を用いる候補遺伝子アプローチであり，各患者に対して副作用と関連すると思われる 507 遺伝子，計 3144 個の SNP を調べている．抗癌剤投与後の各患者の副作用の有無と，事前に調べた SNP 情報を結びつける事により，副作用と相関の高い遺伝子を探索し，これらの情報を基に患者毎の副作用予測システムや治療効果予測システム，薬剤選択システムの構築を目指している．

本報告では，抗癌剤の治療効果や副作用に関連する遺伝子やマーカーの探索のための解析手法を検討し，アソシエーション解析において統計学的に検出力の高い手法について検討する．

ゲノム多様性と副作用関連遺伝子探索アプローチ

ヒトの一塩基多型である SNP は，ゲノム中の塩基が 1 つ異なる遺伝子多型であり，1000 個の塩基中 1 個の割合，30 億塩基対に 300 万個も存在すると報告されている．[1] SNP は，病気の易罹患性，薬効や副作用の現れ方などといった個人の体質の違いに影響を及ぼすものと考えられており，さらに表現型の原因変異のマーカーとしても有用である．ヒトの SNP のデータベースについては世界的には dbSNP が有名であり，日本人の SNP に関しては遺伝子コード領域とその近傍の SNPs を主な対象とした JSNP があり，日本人 SNP データベースとして重要な基盤情報として利用されている．また，国際ハップマッププロジェクトでは欧州人，アジア人，アフリカ人についてのハプロタイプ地図作成が目標とされ，合計 60 万 SNP が解析の対象となっている．

多型の種類が豊富なマイクロサテライトと異なり，SNP は通常，メジャーとマイナーの 2 種類のアレルを有す．特に医学的に興味ある SNP は遺伝子領域上（exon）にある塩基の変異がアミノ酸のコードを変える cSNP（coding SNP）である．実際の SNP のタイピングはインベーダー法などにより行われ，[2] インベーダー法から得られたデータを基に各個体の SNP のアレルを統計学的に自動判定する方法も開発されている．[3]

疾患関連遺伝子や薬剤感受性遺伝子を探索するアプローチとして，マーカーとしての SNP を全染色体領域に亘ってほぼ等間隔あるいは遺伝子の領域を中心に配置して調べるゲノムワイドアプローチと，表現型のメカニズムに関連する機能を持つと予想される遺伝子に絞って調べる候補遺伝子アプローチがある．[4] ゲノムワイドアプローチでは未知の遺伝子を探索する事が可能であるが，全ゲノムを網羅するためには数十万個の非常に多数の SNP をタイピングして関連解析を行う必要がある．一方，候補遺伝子アプローチでは，疾患の発症メカニズムに関連する機能を持つと予想される遺伝子を候補遺伝子とし，それらの遺伝子についてアミノ酸置換あるいは遺伝子発現に変化を及ぼすと推測される SNP を検索し患者の表現型との相関を解析する．未知のメカニズムとの関連は探索不可能であるが，コストを控え重要な遺伝子を検討する事ができる．現在，癌研ゲノムセンターでは，癌関連遺伝子として，薬物動態（146 遺伝子），DNA 損傷修復（151），アポトーシス（63），細胞周期制御（31），血管新生（11），炎症（105）の計 507 遺伝子について 3144 個の SNP を調べている．

ゲノムワイドと候補遺伝子のどちらのアプローチにおいても，疫学的手法としてのアソシエーション解析を個々の SNP とハプロタイプに対して行い，研究対象の表現型と患者のジェノタイプとの関連性を統計学的に評価する．ハプロタイプとは，父または母から由来する 1 本の配偶子上で比較的近隣に存在する遺伝子変異の組合せの事であり，従って各個体は父母由来の 2 種類のハプロ

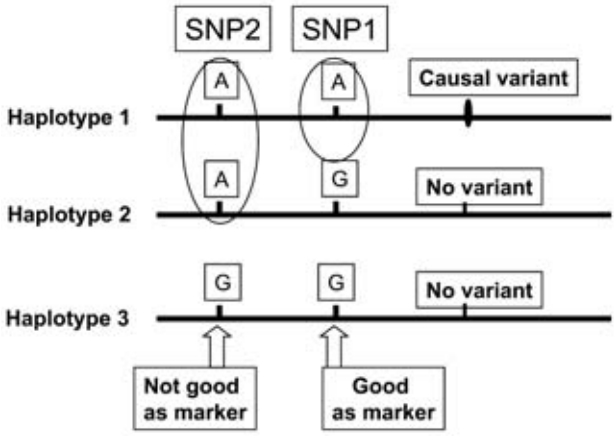

Figure 2. Example of SNPs for inadequate markers.
In case of three haplotypes or more, SNP1 may not be able to identify the causal variant. We can identify the causal variant using haplotype 1.

タイプを有し，その組み合わせをディプロタイプと呼ぶ（Figure 1）．ハプロタイプ内に k 個の SNP がある場合，理論的には 2^k 個のタイプが存在するが，実際には主要なハプロタイプの種類は種々の民族においても 2 から 6 種程度と報告されている．次節でハプロタイプを用いた解析の重要性を示す．

副作用関連遺伝子探索法における問題点

1. 単独 SNP とハプロタイプに基づく解析の比較

原因と考えられる要因としての遺伝変異と臨床結果との関連性を統計学的に評価するにはアソシエーション解析が一般的に用いられている．SNP ごとのアソシエーション解析では，特定のジェノタイプに対応して特定の表現型が出現しているかどうかを調べる．関連が無ければ，どのジェノタイプでも表現型の出現頻度はほぼ同じ程度となる．従ってジェノタイプ間での副作用出現頻度に差がない事を帰無仮説として独立性の検定を行う．1 つの SNP に対しては，患者が取り得るジェノタイプは 2 通りのホモ接合体と 1 通りのヘテロ接合体の計 3 種類である．例えば，G と A を持つ SNP の場合，GG または AA のホモと GA のヘテロである．また副作用などの表現型を表すカテゴリーは 2 値の場合が多く用いられ，従って 1 つの SNP に対して 2×3 の分割表が作製される．GG か AA のどちらか片方のホモと GA のヘテロデータを合わせた表現型頻度と，もう片方のホモの表現型頻度とを比較すれば，SNP が優性か劣性に働いていることが調べられる．さらに患者数の 2 倍の数となるアレルを基に 2×2 の分割表も構成できる．いずれの場合も独立性の検定を行い，有意確率 p 値を算出して評価を行なう．またハーディーワインベルグ平衡の検定[4] はアレル型測定の誤りのチェックにも用いられている．

SNP のゲノム上の位置が近い場合，2 つの SNP の分割表が非常に似た結果を示す場合がある．これは連鎖不平衡によるもので，2 つの SNP の間で組み替えの起こった頻度が低いことを示す．組み替えの強さは種々の連鎖不平衡係数で測られ，この係数が 1 の時には 2 つの分割表は完全に一致する．従って，連鎖不平衡の強い領域の SNP を多数調べても有用な情報はあまり得られない．

ここでは，SNP 毎の解析だけでは原因変異の探索において重要な SNP を見逃す可能性を示す．Figure 2 に原因変異を特定しやすい例と特定しにくい例を示す．ここでは主要なハプロタイプが 3 種類であるとし，その 1 本のハプロタイプ 1 の上に副作用と関連する原因変異があると仮定し，その他 2 本のハプロタイプ 2 と 3 には変異がないものとする．さらに SNP1 と SNP2 は原因変異の近傍に存在して連鎖不平衡が強いものと仮定する．SNP1 ではアレルの A が原因変異と同じハプロタイプ 1 上に存在し，さらにアレルの G は原因変異が無いハプロタイプに存在している．従って，原因変異の有無と SNP のアレルの型とが 1 対 1 に対応し，SNP1 は原因変異を探索するマーカーとして適切に働く事がわかる．しかしながら SNP2 では，アレルの A では原因変異と同じハプロタイプ 1 の上に存在しているが，ハプロタイプ 2 のアレル A では原因が存在しない．従って，同じアレル A が 2 種類の結果と対応しており，1 対 1 の関係が満たされず，SNP2 で原因変異を同定する事は困難である事がわかる．各 SNP が原因変異のマーカーとして適切かどうかはデータを収集する前には不明である．確実に原因変異を同定するには，SNP1 と SNP2 を組み合わせて両者がアレル A を持つ場合のみ，すなわちハプロタイプ 1 であれば原因変異に対応することを利用する．なお，ハプロタイプの種類が 4 種以上の場合に単独 SNP だけで原因変異を特定できるのは，片方のアレルだけが原因変異ありの場合に対応し，もう片方のアレルの全てが原因変異のないハプロタイプ上に乗っている場合だけである．なお，主要なハプロタイプの種類が 2 種の場合には，単独 SNP を用いてもアレルの種類が 2 通りのため，原因変異を持つハプロタイプと原因変異を持たないハプロタイプと 1 対 1 の対応関係が保持でき，原因変異の探索が容易となる．

別な角度からの注意点として，主要なハプロタイプの定義が問題となる．例えば累積ハプロタイプ頻度を 90% までとした場合，10% のマイナーなハプロタイプが無視されるので，副作用頻度が 20～30% の小さい場合は問題となる．また，原因変異があっても副作用が出現するかどうかの浸透率の問題があるが，浸透率が非常に低い場

合はどのような方法を用いても統計学的には判定が困難である．

2. ハプロタイプブロック同定の重要性

本節では，解析すべきハプロタイプの領域の範囲とアソシエーション解析における統計学的検出力の関係について検討する．ここでの検出力とは，統計学的検定問題において真に差のある場合に差があると正しく判定する確率である．

ハプロタイプはハプロタイプブロックと呼ばれる領域が複数繋がって構成されており，隣り合うハプロタイプブロックは組み換えが起こりやすい境界領域によって分けられる．[5,6] この組み替えが起こりやすい箇所をホットスポットと呼ぶ．1つの遺伝子は単一あるいは複数のハプロタイプブロックで構成されており，遺伝子毎に表現型との関連を解析する場合は，ハプロタイプブロックの同定が重要となる．

まず仮想的な例を考察する．1つの遺伝子がAとBの2つのブロックから構成され，Aのブロックには2種類のタイプが，Bのブロックには3種類のタイプが存在すると仮定する．原因変異はAのブロックの片方のハプロタイプに存在すると仮定する．いまブロックの判定（またはホットスポットの検出）に失敗し，これら2つのブロックAとBを合わせて1つのブロックCと判定したとしよう．この場合，Cのハプロタイプの種類は2つのブロックの組み合わせにより，2×3＝6通りになる．ハプロタイプ頻度により，実際は6通りより少ない可能性もあるが，AとBを別々に調べるよりも調べるべきハプロタイプの数が多くなってしまい，統計学的検出力が低下する．しかもAのブロックの片方に原因変異が存在しているため，Cの6通りのハプロタイプの内の3通りの型に原因変異が存在してしまい，単独のハプロタイプの寄与度が小さくなり，これらを個別に検出する事が困難となる．

ハプロタイプブロックを同定する方法として，Gabrielらが提案した2つのSNP間の連鎖不平衡係数の信頼上限，信頼下限を用いて組み換えの生起を判定する方法がある．[7] この方法は現在一般的に用いられているが，多数の組み合わせの結果を総合的に判定するため，個々の判定結果と矛盾する場合が多くなる．新しい方法として，Anderson and Novembreはハプロタイプブロック構造をモデル化する方法を提案している．[8] しかしながら，この方法で判定されたブロックでは，実際のアソシエーション解析で得られたSNP毎の解析結果と矛盾する事が多い．そこで我々は先祖ハプロタイプの概念を取り入れた新たな手法を提案した．[9]

ハプロタイプを大きく取りすぎると，本来分離すべきホットスポットを含んだハプロタイプが含まれるため，メジャーなハプロタイプの数が増え，従って検定の数が増えてしまい検出力が低下する．ハプロタイプブロックを小さくとりすぎると原因となるハプロタイプを分離できず検出力が低下する．このようにハプロタイプブロックを正しく同定する事が検出力を上げるために重要となる．

この新手法を用いた我々の最新の研究ではゲフィチニブの副作用と相関の高い遺伝子が幾つか確認されている．詳細については別途報告される予定である．

ハプロタイプを用いたアソシエーション解析

ハプロタイプ情報を用いて実際に副作用と関連する遺伝子を探索するには，前節で示した方法を用いて，各々のSNPのタイピング結果から各遺伝子内のハプロタイプブロックを同定する必要がある．以下では，1つのブロックが同定されたものとして，そのブロック内でのアソシエーション解析の方法を記述する．

（1）ハプロタイプ別のリスクの解析を行う前準備として，対象集団中の各患者のディプロタイプを推定する事が必要となる．実際のディプロタイプの推定には，SNPAlyze[10] や Haplotyper[11] のプログラムを用いる事ができる．ここで一部の患者についてはディプロタイプの推定が困難な場合があるが，推定した各々のハプロタイプの信頼度が低い場合は解析の対象から外すか，あるいは最も確率の高いハプロタイプを割り当てて解析しておき，その後，解析の対象から外した場合の結果と比較検討して対処する．

（2）次に，ディプロタイプ推定において対象者中に1例あるいは数例に割り付けられた稀なハプロタイプの処理を行う．分割表を用いた解析では稀なハプロタイプを1つのカテゴリーとしても解析に問題はないが，ロジスティックモデルのような多変量モデルを用いると分散共分散行列が算出できない場合がある．この問題を避けるためには，このデータを削除するか，あるいは，稀なハプロタイプをメジャーな先祖ハプロタイプに置き換えると統計学的検出力が向上するという最近報告された考え[12] に基づき，稀なハプロタイプをメジャーな先祖ハプロタイプに置き換える．ただし，この方法は置換するメジャーなハプロタイプと稀なハプロタイプ間で原因変異の有無が同じであれば生物遺伝学的に意味を持つが，そうでない場合は解析結果にバイアスを与えるので，置換する前の解析結果と置換後の解析結果を確認する必要がある．

（3）ブロック内に原因変異が存在している場合，各ハプロタイプは原因変異を有する場合とそうでない場合に分類される．従って，ロジスティックモデルにおける変数xのパラメトリゼーションにおいては，複数のハプロタ

イプを原因変異の有無に対応する2群にまとめる方が，解析すべき次元が少なくなり統計的検出力が向上する．一般の Common disease common variant の仮定[13] の下では，ある1つの特定のハプロタイプの型に原因変異が存在していると考えられている．そこで全てのハプロタイプをモデル化し，その中から有意なハプロタイプが存在しないか調べる．ハプロタイプが複数ある場合は，有意なハプロタイプだけモデル内に残し，他はモデルから外して再解析を行いモデルの適合度を検討する．統計学的に有意でなければ，そのブロックは副作用と相関が認められないと判定して次のブロックを調べる．複数のハプロタイプが最終的に有意性を示す場合，ハプロタイプの論理和を取り1つの変数にまとめ再解析し，適合度で最終モデルを判定する．適合度が落ちれば rare variant 仮説が成立する可能性や，あるいはブロックがさらにサブブロックに分離する可能性も考えられる．

上記の手順によって解析を行えば，連鎖不平衡の強いブロック内でハプロタイプ数を減少させ，解析すべき次元の縮小化が図られているために統計学的検出力が向上し，実際の原因変異がある場合に探索が容易となる．

副作用の有無など2値的な表現型（$y=0,1$）で $y=1$ となる確率は，ロジスティック回帰モデル

$$P(y=1)=1/\{1+\exp(-b_0-b_1x_1-b_2x_2-\cdots-b_kx_k)\}$$

で与えられる．b_i は未知パラメータであり（$0 \leqq i \leqq k$），x_i はハプロタイプ i に対応する説明変数である．説明変数のコード化やモデリングは種々可能である．最も単純なモデルは，推定した個体がハプロタイプの i 番目と j 番目を持てば $x_i=1$，$x_j=1$ とし，他は全て $x_h=0$（$h \neq i,\ j$）とコード化する．型 i をホモで持つ場合は $x_i=1$ とし，他の変数は0とする．このモデルでは特定のハプロタイプのホモの効果とヘテロ単独の効果が等しいと仮定されたモデルになる．別のモデルとして，型 i をホモで持つ場合は $x_i=2$，他の x は0とコード化する事もできる．このモデルでは，型 i の単一ハプロタイプ当たりの表現型への効果が，他のある特定の型と比較して相対リスクが $\exp(b_i)$ 倍だけ変化するように仮定されている．ただし，このモデルの場合，各患者の説明変数の和が常に2となり説明変数行列の階数が落ちるため，1つのハプロタイプを除外してモデルを構成する必要がある．

関連した研究として最近 Stram ら[14] は推定したハプロタイプの不確定性を考慮したハプロタイプ別リスクのオッズ比をロジスティックモデルで推定する方法を開発し，Multiethnic Cohort Study の乳癌症例の *CYP17* の研究に応用している．一般にロジスティックモデルを用いる事の長所は，性や年齢などの交絡要因を調整してリスク評価を行う事ができ，得られたモデルで予測モデルを構築でき有用性が高い．

さいごに—臨床応用に向けて—

本稿では副作用に関連する遺伝子を探索する統計学的な方法を提示したが，1つの遺伝子内でも有意となるブロックとそうでないブロックに分けられるため，探索した遺伝子をマーカーとして臨床応用する場合は，有意となったブロックの SNP に対して独立なデータを収集しバリデーションを行う必要がある．新規の患者に対する治療応答性予測を行うには十分に検証されたモデルを用いる必要がある．なお予測システムの構築には，同定されたブロックに対応するタグ SNP を利用して効率的な診断キットが作成可能となる．複数の遺伝子が副作用に関連している場合は，探索した遺伝子を組み合わせて予測システムを構築する事も可能である．

本報告では，今後の個別化医療に向けての具体的な副作用関連遺伝子の同定における問題点と統計学的検出力の高い解析法の1例を示した．単独 SNP の解析のみで結論を出されている報告を見かけるが，ハプロタイプに基づく解析を行えば統計学的有意性を検出できる可能性があるため再解析をお勧めする．

謝辞：本原稿は NEDO の支援を受けた研究を基とした．また研究の一部は文部科学省科学研究費基盤研究（B）課題番号16300090を用いて行われた．これまで本原稿に関連したコメントをいただいた癌研ゲノムセンターの宮田敏，牛嶋大，榎本友美，磯村実，星川裕，長崎光一，下地尚の各研究員，野田哲生博士，三木義男博士，ならびに統計数理研究所の江口真透教授，藤沢洋徳助教授に感謝致します．

REFERENCES

1. Brookes AJ. Review: The essence of SNPs. *Gene*. 1999; 234:177-186.
2. 中村祐輔．SNP 解析・マイクロアレーによる創薬とオーダーメイド医療．東京：羊土社；2000:30-82.
3. Fujisawa H, Eguchi S, Ushijima M, et al. Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics*. 2004;20:718-726.
4. 鎌谷直之．ポストゲノム時代の遺伝統計学．東京：羊土社；2001:268-269.
5. Daly MJ, Rioux JD, Schaffner SF. et al. High-resolution haplotype structure in the human genome. *Nat Genet*. 2001;29:229-232.
6. Rebbeck TR, Ambrosone CB, Bell DA. et al. SNPs, haplotypes, and cancer: applications in molecular epidemiology. *Cancer Epidemil Biomarkers Prev*. 2004;13:681-687.
7. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science*. 2002; 296:2225-2229.
8. Anderson EC, Novembre J. Finding haplotype block boundaries by using the minimum-description-length

principle. *Am J Hum Genet.* 2003;73:336-354.
9. Fujisawa H, Isomura M, Eguchi S, et al. Identifying haplotype block structure by using ancestor-derived model and minimum-description-length principle (Submitted).
10. Ohmori H, Makita Y, Funamizu M, et al. Haplotype analysis of the human collection placenta 1 (hCL-P1) gene. *J Hum Genet.* 2003;48:82-85.
11. Niu T, Qin, ZS, Xu X, et al. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet.* 2002;70:157-169.
12. Tzeng JY. Evolutionary-based grouping of haplotypes in association analysis. *Genet Epidemiol.* 2005;28:220-231.
13. Taylor JG, Choi E, Foster CB, et al. Using genetic variation to study human disease. *Trends Mol Med.* 2001;7:507-512.
14. Stram DO, Pearce CL, Bretsky P, et al. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity.* 2003;55:179-190.