8.4 データ解析（DNA コピー数変異解析）

2 倍体であるヒト細胞には、疾患に関連して遺伝子数の増減があっても通常同じ遺伝子は 2 つあると言うのが通説であった。ところが、DNA マイクロアレイの高密度化などにより、ヒトゲノム全域の高解像度解析が可能となったことにより、近年正常な個人であっても遺伝子数に違いがありうることが次々と明らかにされている。DNA マイクロアレイは本来核酸の定量のために開発されており、SNP タイピング用アレイであってもアレルごとに DNA 定量を行い、その結果から SNPs のタイピングを行っているという面では、その事には変わりがない。

ただしコピー数変異（CNV）領域では、ヘテロの場合にプローブシグナル比が 1：1 以外の値をとる、2 倍体を仮定したタイピングでは Hardy-Weinberg 平衡やメンデル遺伝則からのずれが生じるなどの理由から、SNPs タイピング用に開発されたマイクロアレイでは製品化の段階で CNV 領域のプローブが除外されるため、この領域のプローブ密度が低くなる。

とはいえ SNPs タイピング情報と同時にコピー数情報も得られる点は、SNPs タイピング用マイクロアレイの強みであろう。GeneChip Mapping Array の最新バージョンである SNP5.0、SNP6.0 では、SNPs タイピング用プローブ以外にも CNV 解析用プローブも搭載されている。

ただし SNPs タイピング用マイクロアレイの CNV 解析用のソフトウエアは、アレイの製品化から時間的に遅れることが多い。GeneChip Mapping Array の場合、Affymetrix 社が提供している Chromosome Copy Number Analysis Tool (CNAT)以外にも、dChipSNP（http://biosun1.harvard.edu/complab/dchip/）、CNAG (http://www.genome.umin.jp/)、GEMCA（http://www.genome.rcast.u-tokyo.ac.jp/CNV/gemca_details.html）など優れたソフトウエアが利用可能である。

実験上の問題点としては、前にも述べたように特に Mapping 500K array set 以降プローブの高密度化のために、S/N 比の高いデータを得ることが難しくなっている。実験上のノイズの除去はデータ解析時点では難しいため、信頼性の高い CNV

解析を行うためにも S/N 比の高いデータが得られるように実験を行うべきであろう。

データ解析上の問題点としては、Mapping 500K array set の場合、画像データを含めたファイルデータサイズは、1 検体あたり約 2Gb、シグナル強度を記載したファイルのみでも約 200Mb になる。これらを利用して独自に解析を行うのは、通常の PC では困難を伴うが、データ構造の理解には有効であると考えている。

8.5 まとめ

ヒトゲノム関連データの大規模化に伴って、ヒトゲノム解析技術もハイスループット化、DNA マイクロアレイでは高密度化が推し進められ、このことがまたヒトゲノム関連データの大規模化に寄与するという状況になっている。現在進められているシークエンス技術上の開発、改良を見ると、この動きは今後ますます加速するに違いない。

本章で解説した DNA マイクロアレイに関しては、高密度化が核酸定量性を担保できるぎりぎりのところまで進んでおり、実験のクオリティーとデータ解析手法の優劣が解析結果に大きく影響を与えうる状況になっている。SNPs タイピング実験は、その実験の性格から大規模にならざるを得ないため、実験開始時には実験材料、終了後には実験結果のクオリティーに十分な注意が必要であろう。

9. BeadChip を用いたゲノム全域解析による疾患遺伝子解析

9.1 はじめに

2003 年 4 月のヒトゲノム解読完了以降、ゲノム科学分野における技術革新は著しく発展してきた。DNA マイクロアレイの高密度化はとどまるところを知らず、同時に、ゲノム DNA 検体の前処理の工夫や、検出器の精度の向上もあいまって、より効率よく、より正確に、多様なゲノム配列と複雑な構造を解析することが可能になっている。そのなかでも、我々の研究室で実際に採用している、イルミナ社（Illumina, Inc）製の BeadChip system は、国際ハップマッププロジェクトの SNP 情報をもとに開発された全ゲノム遺伝子多型解析法 Infinium II アッセイがベースとなっており、SNP ジェノタイピングのみならず、CNV（コピー数変異）や LOH（ヘテロ接合性欠失）の解析も可能にしている。本章では、この BeadChip を用いた実験方法とデータ解析手法について解説する。

9.2 BeadChip の原理

従来の DNA マイクロアレイは、プローブとなるオリゴヌクレオチドや一本鎖 DNA をスライドグラスなどの基板上で合成する、もしくは隣接する領域と混雑しないようにプローブ溶液を基盤に載せてプローブを貼り付けるものがほとんどである。それに対して、イルミナ社の BeadChip と Array Matrix はまったく新しい技術（Beadarray テクノロジー）によって構成されたマイクロアレイである。また、これらに適用される SNP ジェノタイピング解析法にもイルミナ社独自のアッセイ（「GoldenGate」、「Infinium」、「Infinium II」）が用意されている。その中でも、我々の研究室で実際に採用している BeadChip と Infinium II アッセイについてその原理を詳しく解説したい。

まずは、BeadChip から説明しよう。BeadChip は全ゲノムを対象とした SNP タイピング用となっている。基板には何万ものマイクロウェルが存在し、プローブの結合したビーズが何千種類も整然と配置されている（図 9.1）。一枚の基板上には同一種類のビーズを 30 個以上含んでいるため、その冗長性を活用して、信頼度の高い定量的測定が可能である。各ビーズはすべて ID 番号のようなものがふられているので、各マイクロウェルにどのビーズが固定されているか、識別できるようになっている。これは各ビーズに対する品質管理も兼ねている。ゲノム DNA 検体をハイブリダイゼーションさせることで、相補的配列をもつ DNA 断片がビーズ表面に結合できる。そしてこの基板上で SNP タイピングを行う。ちなみに BeadChip には、アミノ酸変異を伴う nsSNP（non synonymous SNP）を搭載したものと、ハプロタイプを代表したタグ SNP を搭載したものがあり、いずれも国際 HapMap プロジェクトのデータを基に設計されている。もちろんカスタムアレイなどにも対応している。

図 9.1　BeadChip のマイクロウェルの拡大図。
イルミナ社ウェブサイトから抜粋・引用

次に、SNP ジェノタイピングの解析法である「Infinium II」について説明しよう。ID のついたビーズには、SNP 座位から 1 塩基手前までの約 50 塩基長のポリヌクレオチド鎖がプローブとして結合している。このヌクレオチド鎖にサンプル DNA をハイブリダイズしたあと、一塩基伸長反応を行うわけであるが、このときに用いる基質は、4 種類の塩基をそれぞれ別々の波長の蛍光色素で標識したヌクレオチドである（図 9.2）。そして、ビーズごとに発せられる蛍光を検出することで、SNP 座位にどの塩基が対合したかがわかる仕組みになっている。一塩基伸長反応であるため、蛍光強度が、ハイブリした DNA 量に比例することも利点である。この点については、データ解析の項で詳述したい。

イルミナ社のホームページ（http://www.illuminakk.co.jp/index.shtml）上でもこれら原理の説明がまとめられているので参照されたい。

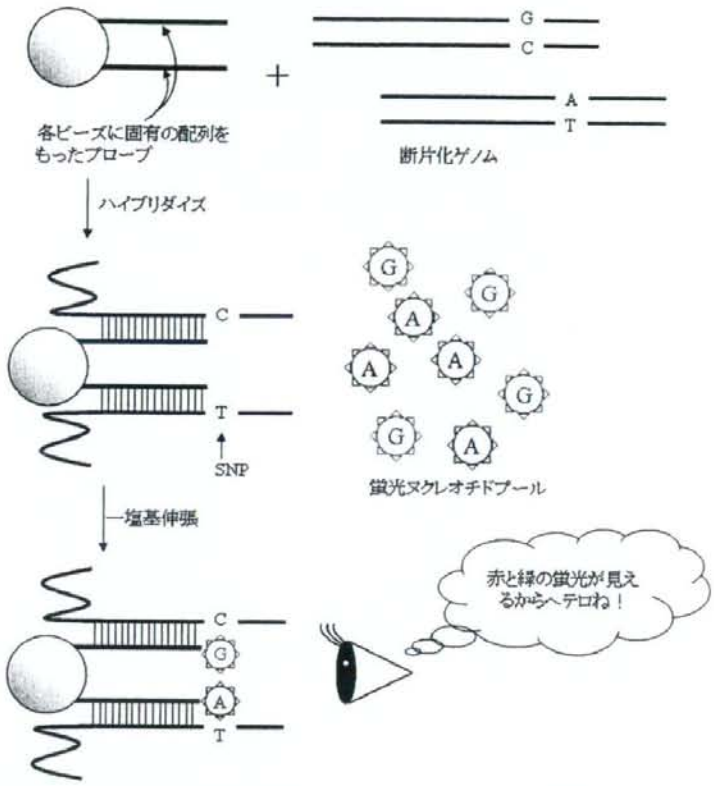

図 9.2　Infinium II アッセイの原理。イルミナ社ウェブサイトより抜粋・改変

9.3 BeadChip 実験

BeadChip を用いた SNP タイピングアレイ実験操作について、簡単に実験工程を解説しよう（図 9.3）。タイピングアレイ実験は、大まかに 3 つの段階に分けることができる。1.ゲノム DNA の増幅、断片化、2.アレイと DNA 検体のハイブリダイゼーション、3.アレイの洗浄と一塩基伸長反応とイメージスキャン、である。

図 9.3　BeadChip の実験手順。イルミナ社ウェブサイトより抜粋・改変

9.3.1 ターゲット配列の増幅、断片化

まずはゲノム DNA 検体を増幅する。イルミナ社のプロトコールには明記されていないが、通常の MDA (multiple displacement amplification)のように Phi29 ポリメラーゼを用いた増幅を応用したものと思われる。特定の配列に偏った増幅はないことになっている。少量のゲノム DNA サンプル(750ng)を用いるだけで、100,000SNP 座位のアッセイに十分な量が得られる（0.0075ng (7.5pg)/1SNP タイピング）。次に、酵素反応を用いて増幅 DNA サンプルを切断して断片化する。その後、ハイブリダイゼーション反応に影響を及ぼすような塩やその他夾雑物を除去するためにアルコール沈殿と再溶解を行なう。

このように得られた DNA サンプルは、大量の SNP をタイピングするための十分な量と、正確なタイピングに必要な純度がみたされる。

9.3.2 アレイとターゲット配列のハイブリダイゼーション

BeadChip をキャピラリー・フロースルー・チェンバーにセットし、DNA 検体を添加して一晩インキュベーションする。この間に、増幅され断片化された DNA サンプルが、SNP 座位特異的プローブとハイブリダイズする。

9.3.3 アレイの洗浄とイメージスキャン

ハイブリダイゼーションが終了したら、次に DNA ポリメラーゼ伸張反応で蛍光標識された塩基を取り込み、アリル特異性を決定する。イルミナ社の Beadarray Reader を用いて各ビーズの蛍光強度を測定、SNP アリル自動検出ソフトウェアを用いて解析する。

9.4 データ解析（SNP タイピング）

スキャン画像の各プローブのシグナル強度から SNP の遺伝子型を決定する。先述のように、一つのアレイには複数個の同一ビーズが存在するため、その冗長性のため、より高精度データを得ることができ、また、ハイブリダイゼーションのムラや検出面の汚れなどスキャン画像に乱れがあった場合でもデータの欠落を抑えることができる。

イルミナ社から提供されている「Beadstudio」を用いれば、BeadChip をスキャンして得られた蛍光強度のデータをそのまま使って遺伝子型を決定してくれる。図 9.4 のように 2 色の蛍光強度にしたがってプロットすると 3 つのクラスター（プロット点のかたまり）が現れる。一見すると TaqMan®による SNP タイピングのデータにも似ている。その 3 つのクラスターがホモ接合（遺伝子型 AA もしくは BB の組）とヘテロ接合（遺伝子型が AB）をあらわしている。しかし、TaqMan との決定的な違いは、蛍光強度と SNP 量（相補 DNA 量）との間により強い相関があることである。その利点を活かすため、通常は、図 9.4 のように改変した散布図を用いる(Colella et al. 2007; Peiffer et al. 2006)。どのように改変するのか詳しくみてみよう。まず、原点とプロット点を直線で結び、その直線と横軸からの角度の値 θ をあらたに横軸に設定し、原点からの距離 R を縦軸にする。次に、イルミナ社独自のアルゴリズムを用いて検出した各クラスターの中心点を直線で結ぶ。その線は、ある θ に対する R の期待値を示している。そして、R の実測値と期待値の比（Rs/Re＝R ratio）をとり、さらにその対数 log R ratio を縦軸に設定する。すると、2 種類の蛍光がそれぞれもっている蛍光強度の差が補正された値が得られる。R-θ 図でも遺伝子型を示すクラスターを表示できるが、こうすることによって遺伝子型をより確からしく判別することができるうえに、このあと解説する CNV/LOH 解析も可能にしている。

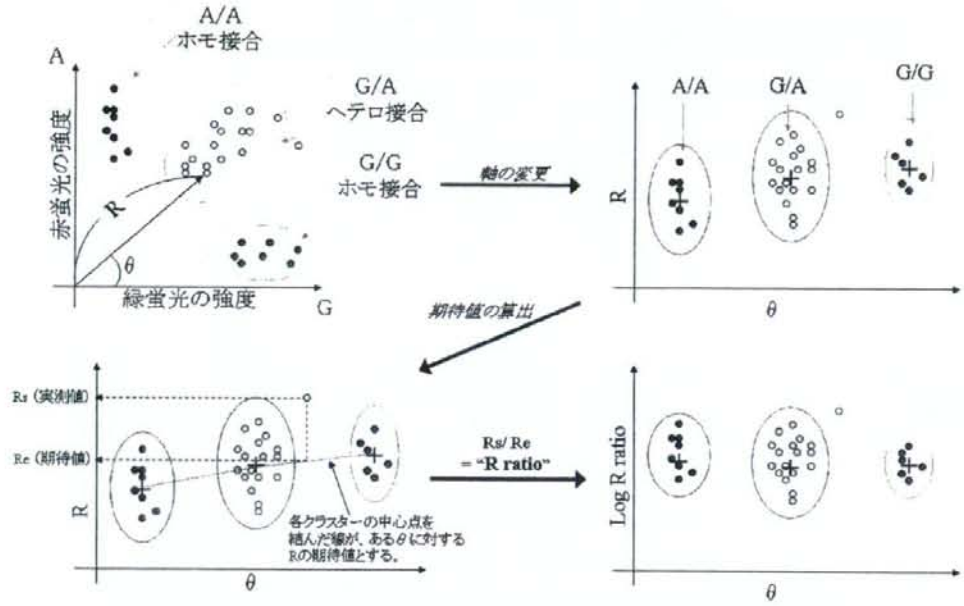

図 9.4　Beadchip のデータ解析。　SNP タイピングの散布図。

ただし、蛍光強度に大きなばらつきがあってクラスターを認識できず、遺伝子型を判別できない場合もある。また無理にクラスター分離しても正確なタイピングができないこととなり、Hardy-Weinberg 平衡やメンデル遺伝則からのずれが観察される。

SNP タイピングの結果は、SNP ID とともに各波長の蛍光強度、遺伝子型頻度、Hardy-Weinberg 平衡とのずれ、などがリストになってあらわれる。同時に散布図も表示されるのでクラスター分離が妥当かどうかの判定も目で見て確認でき、さらに図中のプロット点がどの DNA サンプルかも表示できるため、非常に使い勝手がよい。もちろん、それらのデータを抽出して、他のアプリケーションで解析することも可能である。

9.5 データ解析（遺伝子コピー数変異／ヘテロ接合欠失）

「Beadstudio」では CNV / LOH を解析することもできる。ここで、CNV と LOH を解析する意義を簡単に説明しておこう。LOH（Loss of heterozygosity：ヘテロ接合欠失）は、腫瘍マーカーとして研究されてきたゲノム変異の一つである。腫瘍組織のゲノムでは、片側の遺伝子座が欠失していることがあり、ヘテロ接合性が失われ見かけ上ホモ接合体となる。これを LOH とよんでいる。LOH が生じた領域にがん抑制遺伝子が存在し、その遺伝子に異常が生じるとがん化に至る。実際、病型特異的な領域として、大腸がんでは *APC* 周辺の 5q21 で、そして病型共通の領域として *TP53*（p53）周辺の 17p で観察されることが知られている（Read et al. 1997）。一方、CNV は Copy Number Variation、すなわち遺伝子コピー数変異のことであり、文字通り、あるゲノム領域が増幅もしくは欠失し、遺伝子コピー数が増加もしくは減少する変異である。2004 年、Sebat らによって CNV がゲノム全域に多数存在することがわかり、さらに最近になって、遺伝子コピー数にしたがって遺伝子発現量が変化することも明らかになった(Sebat et al. 2004; Stranger et al. 2007)。このことから、CNV によって引き起こされる疾患の存在が予想され、多くの研究者の注目を集めている。詳細は 7 章を参照されたい。

前項で「BeadChip のデータを一見すると TaqMan のプロットデータに似ている」と書いたが、TaqMan と BeadChip との大きな違いは蛍光強度と DNA 量の相関にある。BeadChip のビーズには等量のプローブが結合していて、しかも、一塩基伸長を行っているので、反応が完全にすすんでいれば、蛍光強度はハイブリする対立遺伝子数 DNA 量と比例する、すなわち定量的に対立遺伝子量を測定することに等しい。Beadstudio は、SNP タイピングとして使ったデータをそのまま CNV 検出に利用している。つまり、対立遺伝子量を測定することにより、遺伝子コピー数もわかる、という具合である。実際にどのように散布図が描かれるかを図 9.5、9.6 に示した。2 倍体ゲノムの検体で、ヘテロ接合であった場合は対立遺伝子 A と B が 1：1 で含まれる。しかし、CNV が生じて、対立遺伝子 A のコピー数が一つ増えた場合、ゲノムに含まれる対立遺伝子は A：B=2：1 となる。すると、A を検出する蛍光強度が増強され、通常の AB クラスターからはずれてしまう。

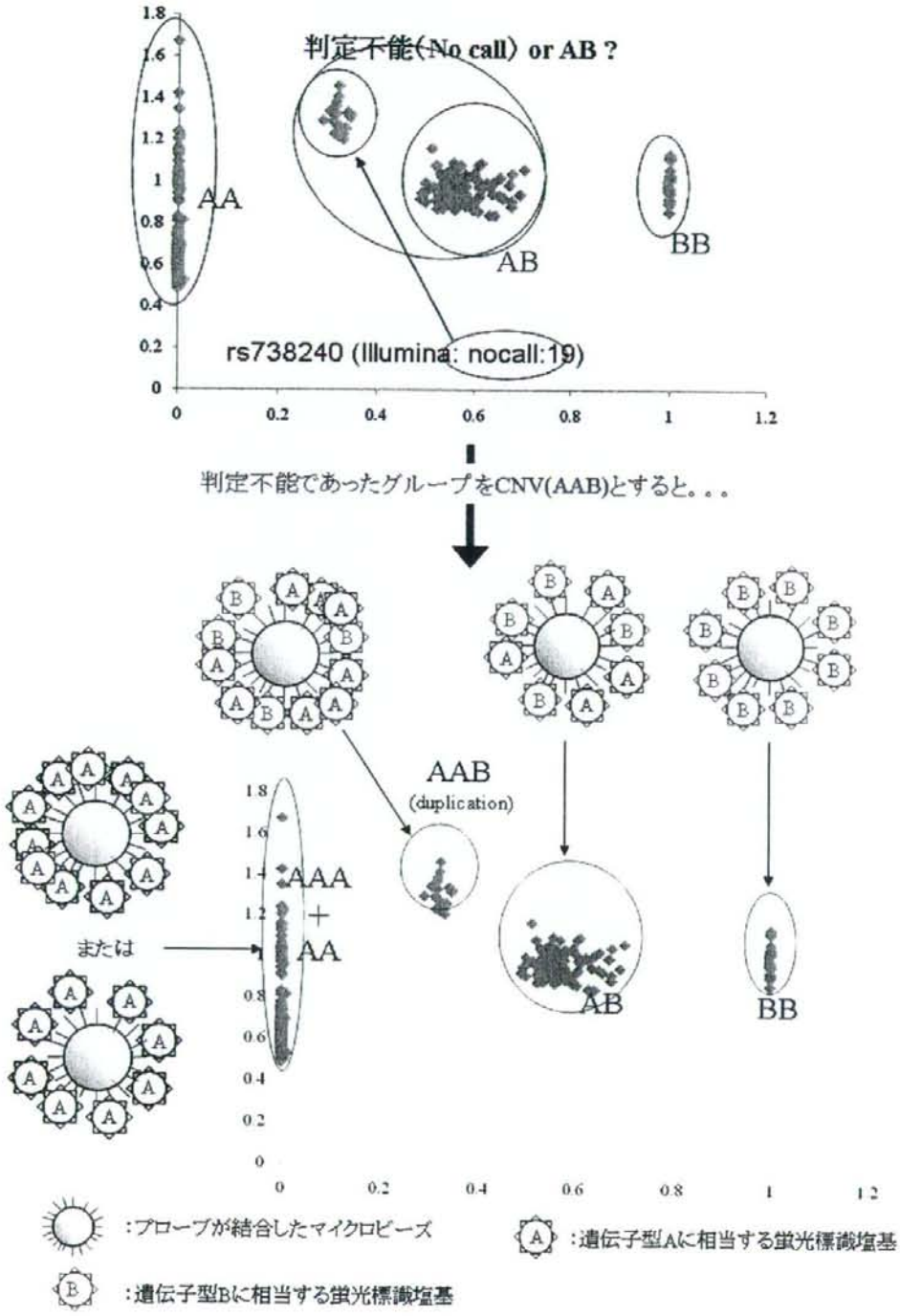

図 9.5　BeadChip による CNV 検出の例。　増幅の場合

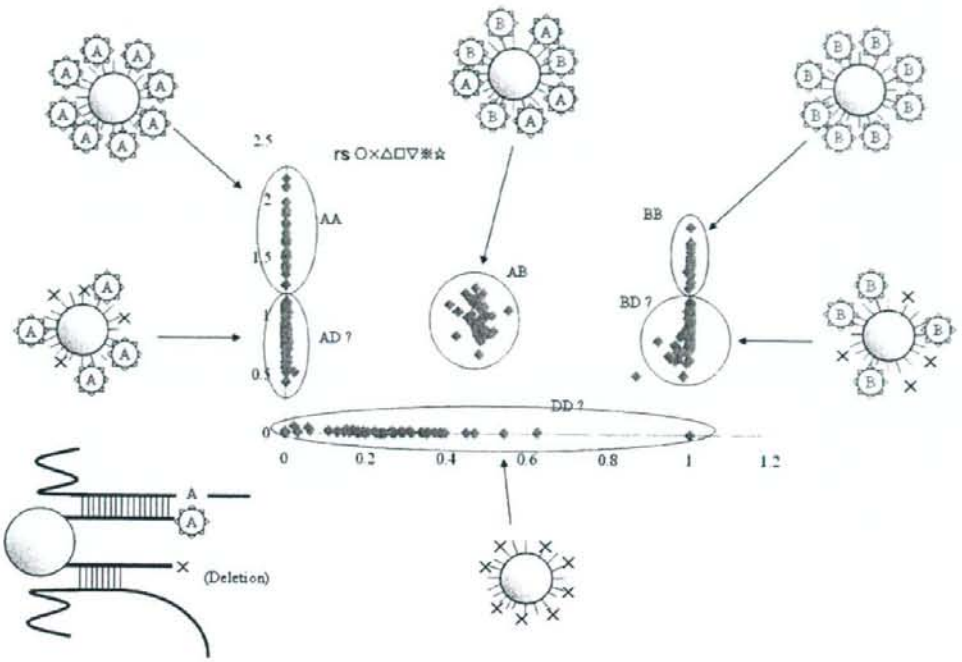

図 9.6　BeadChip による CNV 検出の例。　欠失の場合

こういった散布図が描かれたときには「判定不能（no call）」とみなされるわけだが、単に遺伝子型を決定できないということではなく、CNV の存在が疑われる場合もある。このときには個人ごとにまとめて、横軸に SNP（染色体短腕テロメアから長腕テロメアの順）、縦軸に蛍光強度を示す log R ratio または B アレル頻度（B アレルの個数を意味し、0、1、2 のいずれかの値をとるはずだが、ここでは蛍光強度比から得られる値なので整数とは限らない。）と設定した図のほうが見やすい。Beadstudio で表示可能である。その代表的なものを図 9.7 に示した。こうすると一目瞭然で、染色体のある領域が増幅、欠失を生じているのがよくわかる。さらに、ここでの B アレル頻度とはつまるところヘテロ接合度を示している。0 か 2 に近い値が連続していて 1 付近の値が存在しない領域ではまさしく LOH が生じていることがわかる。ただしこの LOH はシグナル強度に影響を与えていない。すなわち染色体領域が失われたのではなく、ホモ接合体が連続しているためであり、neutral LOH といわれる。通常、近親性を示すと考えていい。両親がいとこ婚の場合、このような neutral LOH が高頻度に観察される。

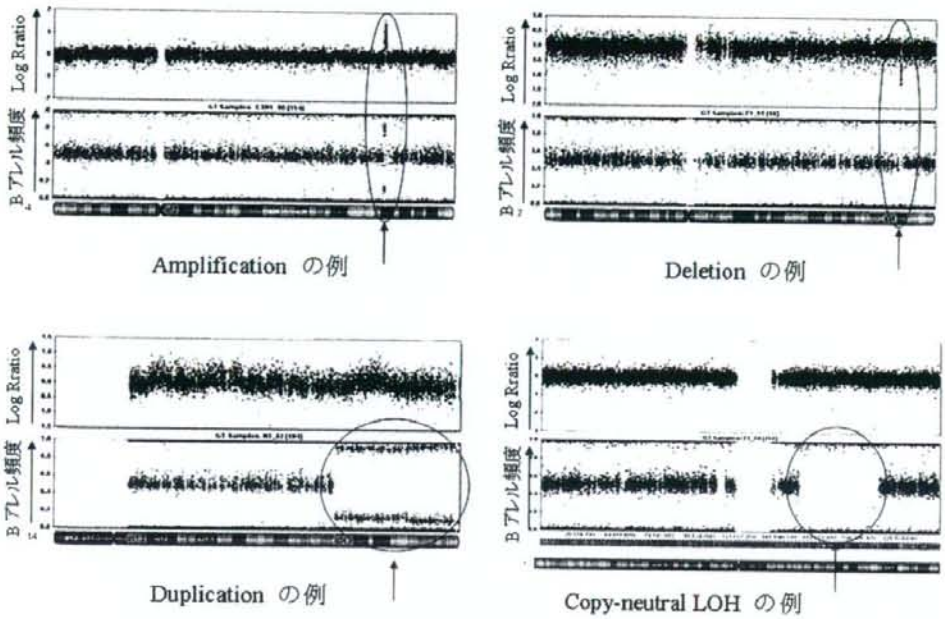

図 9.7　Beadstudio による CNV/LOH データ解析の例

このように、SNP タイピングと CNV/LOH 解析も行うことができ、一粒で二度おいしい利便性の高いプラットフォームといえるだろう。ただし、気をつけねばならないことが一つある。CNV が注目される以前、クラスター分けができなかったり、ハーディー・ワインバーグ平衡やメンデル遺伝則からはずれたりするような SNP 領域のプローブは、SNP タイピングに適していないとみなされ、デザインの段階で除外されていることが多い。つまり、網羅的 SNP ジェノタイピング用として特化された高密度アレイであっても、CNV 領域周辺のプローブ密度は低くなっている。特にコピー数が増幅している領域はその傾向が強い。そのため、最近では CNV 検出用のプローブも新たに搭載した製品も発売されている。

9.6 まとめ

SNP 情報は、そのゲノム科学的意義、医学的意義から、ゲノム変異のなかでも早い段階からデータの蓄積がすすめられ、そのデータベースは猛烈なスピードで大規模化している。ヒトゲノム解析技術のハイスループット化によってこの勢いは維持されるだろう。SNP は、ヒトゲノム 30 億塩基の中で 1 塩基変異の検出という高解像度が最初から要求され、様々な検出法が考え出されてきた。また、多因子疾患などのありふれた病気に関連している例は枚挙にいとまがないので、他のゲノム変異や多型と比べて、網羅的かつ大量処理に適した手法が次々と生みだされてきた。

最近注目されているコピー数変異（CNV）も多因子疾患の関連遺伝子の解明や個人差医療への応用を見越してデータが蓄積されている最中である。CNV は千～10 万塩基に及ぶ様々なサイズで生じている変異なので、顕微鏡の対物レンズをくるくる切り替えながら観察するかのように遠目に引いてみたり近づいてみたりして、余すところなく見逃すことなく検出する必要がある。そのため、これまで CNV を解析しようとする多くの研究室では、BAC ライブラリーや cDNA、オリゴヌクレオチドなどを用いたカスタムメイドのマイクロアレイを用意していた。しかし、幸いにも、すでに実用化されている網羅的 SNP タイピング用マイクロアレイが、CNV の解析へ応用できるようになってきた。さらに最近では、CNV 検出用のプローブも搭載したアレイも製品化されている。これからは、SNP のみならず、CNV についての情報も求められ、そしてデータベースも急速に大規模になっていくに違いない。そして、疾患などへの関連も続々と明らかにされ、個人差医療への応用、実用化へと発展していくことが期待される。

10. これだけは知っておきたい SNP 解析のための統計学

10.1 はじめに

この本章では、他章で詳しく触れることができなかった統計学（本章では従来の統計学とよぶ）の基本体系である検定および推定について、あらためて紹介する。また、従来の統計学とは異なる立場にあるベイズ統計学についても、その基本概念を簡単に解説する。従来の統計学やベイズ統計学に関する知識の整理のために、お役立ていただければ幸いである。

10.2 推定と仮説検定

10.2.1 推定

従来の統計学は、無作為抽出された標本を調査・観測することにより、抽出のもととなる母集団を規定するパラメーター（母数ともよぶ）を推定する学問であるといえる（図 10.1）。すなわち、時間や研究費などの制限から全体調査が不可能な場合（全体調査が無意味であるかどうかはここでは論じない）、標本調査することにより母集団の統計学的特徴などを推測することになる。例えば、日本の「オーダーメイド医療実現化プロジェクト（http://www.biobankjp.org/）」では、5 年間（平成 15-19 年度）で約 40 疾患、30 万人からのゲノム DNA を収集し、SNP と疾患との関わりなどを調べることになっている。

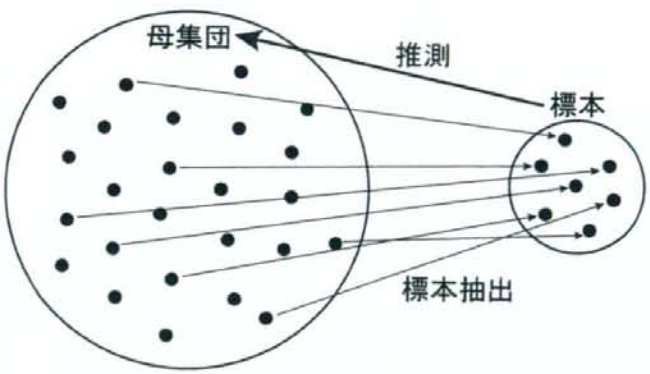

図 10.1　従来の統計学における母集団と標本

標本から求められる統計量としては平均値や分散などがよく知られるが、例えば、疾患関連 SNP 解析においては、患者群での SNP 遺伝子型頻度は標本統計量のひとつであり、その背景にある仮想母集団における遺伝子型頻度の点推定値である。

それでは、一般に、どのような性質を持つ標本統計量がよい推定値を与えるのであろうか。推計統計学の確立者である R. A. Fisher は、一致性、有効性、十分性を同時に満たす推定量を最適推定量とした。たとえば、パラメーターθをもつ母集

団から、n 個の独立標本 $D_1, D_2, \ldots, D_n$ を抽出し、各標本において統計量 d_i (i=1, 2, ..., n)を観測したとする。ここで、標本の関数である推定量 $\hat{\theta}_n$ を考える。

$$\hat{\theta}_n = \hat{\theta}(d_1, d_2, \ldots, d_n) \quad \text{（式 1）}$$

一致性とは、標本数 n を大きくしたときに、式 1 で表す推定量 $\hat{\theta}_n$ がパラメーター θ に近づくことである。他方、パラメーター θ の推定量 $\hat{\theta}_n$ のなかで、$\hat{\theta}_n$ の分散が最小である推定量のことを最小分散推定量とよび、有効性のひとつの指標とされる。十分性とは、パラメーター θ に関する情報はすべて推定量 $\hat{\theta}_n$ のみに含まれることである。すなわち、大標本においてパラメーターに一致し、ばらつきが小さく、母集団を規定するのに必要なすべての情報を含む推定量が最適ということである。この最適推定量は、一般に、最尤法により最尤推定量として求めることができる。別の言葉に置き換えると、観測された標本統計量に対し「尤もらしさが最大」となるパラメーターを最適推定量とすることができる、ということである。このために、「尤もらしさ」の指標として尤度（Likelihood）という概念を導入し、パラメーターの尤度 $L(\theta|d)$ は式 2 のように表される。

$$L_n(\theta) = \prod_{i=1}^{n} f(d_i \mid \theta) \quad \text{（式 2）}$$

ここで、$f(d_i|\theta)$ はパラメーター θ をもつ母集団より抽出された独立標本 D_i における統計量の確率密度（あるいは確率）であり、それらの積を尤度関数としている。最尤法とは、最尤パラメーターを見つける手段のことであり、多くの場合、式 2 の両辺で対数をとり、対数尤度（式 3）を最大化するパラメーターを数学的に求めることになる。

$$\log L_n(\theta) = \prod_{i=1}^{n} \log f(d_i \mid \theta) \quad \text{（式 3）}$$

例えば、ある SNP の 3 種の遺伝子型 MM, Mm, mm のいずれかをもつ個体からなる母集団を考えてみよう。遺伝子型の出現確率は Hardy-Weinberg 平衡に従うと

し、パラメーターθ $(0 < \theta < 1)$を以下のように定義する。

$$f(\mathrm{MM}\,|\,\theta)=\theta^2,\quad f(\mathrm{Mm}\,|\,\theta)=2\theta(1-\theta),\quad f(\mathrm{mm}\,|\,\theta)=(1-\theta)^2 \qquad (式\ 4)$$

ここで、無作為抽出した 40 個体を調査したところ（1 個体からなる標本を 40 標本調べたという意味）、遺伝子型 MM, Mm, mm である個体数の合計が、それぞれ 10, 20, 10 であったとする。このときのパラメーター尤度 $L(\theta)$は、

$$\begin{aligned} L_{40}(\theta) &= \{f(\mathrm{MM}\,|\,\theta)\}^{10}\{f(\mathrm{Mm}\,|\,\theta)\}^{20}\{f(\mathrm{mm}\,|\,\theta)\}^{10} \\ &= (\theta^2)^{10}\{2\theta(1-\theta)\}^{20}\{(1-\theta)^2\}^{10} \\ &= 2^{20}\theta^{40}(1-\theta)^{40} \end{aligned} \qquad (式\ 5)$$

となり、その対数尤度は、

$$\log L_{40}(\theta)=\log 2^{20}+\log\theta^{40}+\log(1-\theta)^{40} \qquad (式\ 6)$$

で与えられる。最尤推定量を求めるための対数尤度方程式は、上式をθで微分し、右辺を 0 とおけばよく、

$$\frac{\partial}{\partial\theta}\log L_{40}(\theta)=\frac{40}{\theta}-\frac{40}{1-\theta}=0 \quad (0<\theta<1) \qquad (式\ 7)$$

となり、これより最尤推定量 $\hat{\theta}=0.5$ を得る。この値は、40 個体をひとつの集団とみなした時のアレル M の頻度の標本統計量にほかならない。

10.2.2 仮説検定

よくご存知のように、従来の統計学のような推計統計学の目的のひとつに仮説検定がある。例えば、患者-対照関連解析などの疾患関連 SNP 解析では、調べる SNP が疾患発症に関わるか否かを統計学的に検討する。