1989)を SNP 遺伝子型データにも応用した論文も報告されたが(Kelley et al. 2006)、その使用にも検出力などの観点からの限界があるようである。

その一方で、正の自然選択の影響下にある SNP や遺伝子探しは、その探索法の開発も含めて、花盛りである。特にヒト集団間で自然選択の働き方が異なる SNP や遺伝子について、非常に多くのことが明らかになってきている（Kimura et al. 2007; Sabeti et al. 2007; Voight et al. 2006; Williamson et al. 2007)。異なるグループからの研究成果の間に認められる異同に関しては、Neilsen らの総説（Nielsen et al. 2007）に詳しい内容が記載されているのでそちらも参照されたい。図 3.8 には、正の自然選択が働くと、ゲノムにどのような痕跡が残されるかについて、その一例を模式的に示している。ヒトの進化史のある時点において、個体の適応度を高めるような SNP アレル（図の赤丸）が特定のハプロタイプ（図の中央パネル緑色）上に新生したとする。この新生 SNP アレルに正の自然選択が働くことにより、現在では、赤丸 SNP アレルが集団に固定（集団内のすべてのゲノムに赤丸 SNP アレルが存在すること）している。このように、正の自然選択により集団内のゲノム配列が均一化することを selective sweep（選択的一掃）とよぶが、この現象に伴い、赤丸 SNP 周辺遺伝領域の（1）遺伝的多様度低下、（2）中立モデルからの逸脱、（3）連鎖不平衡の増大などが生じ、ゲノム中に「痕跡」として残されることになる。これらの「痕跡」は、selective sweep がすべてのヒト集団で均一に起きてきた場合、あるいは、ある特定の地理的集団のみに起きてきた場合のいずれにおいても、ゲノム探索時の対象となるのに対して、集団特異的な selective sweep は（4）集団間の対立遺伝子頻度の違い（例えば、図 3.8 では赤丸 SNP アレル頻度の集団間の相違）によっても発見することが出来る。このような「外れ値」探し研究により見いだされた SNP や遺伝子が、どのような病態生理学的機能と関わるかについての遺伝学的、実験的証明が、今後に残された大きな課題である。

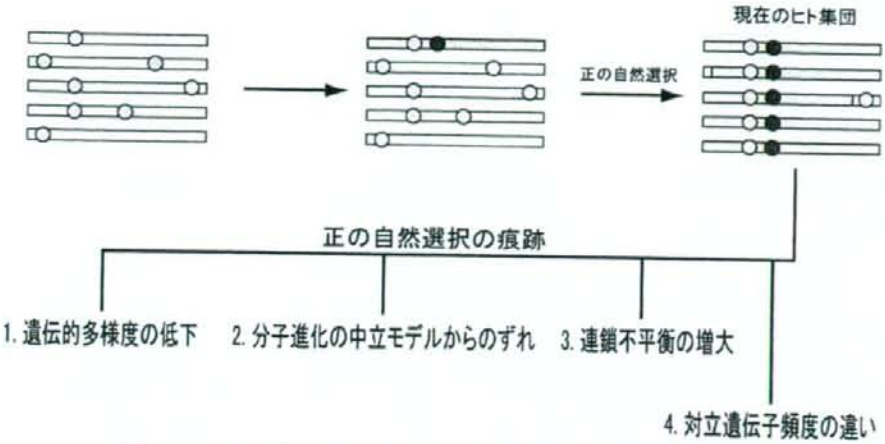

図 3.8　ヒトゲノムに存在する正の自然選択の「痕跡」

特定の SNP アレルに正の自然選択が働いた場合に、ヒトゲノムに残される正の自然選択の「痕跡」について模式的に示している。5 種類の中立的な SNP（水色丸）が観察される 5 本のハプロタイプ（白色）からなる集団を例とする（左パネル）。過去のある時点において、特定ハプロタイプ（中央パネル緑色）に生存上有利な SNP アレル（赤丸）が新たに生じたとする。この新生 SNP アレルに正の自然選択が働くことにより、新生 SNP アレル（赤丸）が集団に固定し、それと同時に、正の自然選択が働く前に存在した 5 種類の中立的 SNP のうち、3 種が消失し、2 種（水色丸）のみが現在観察される（右パネル）。この過程に伴い、（1）遺伝的多様度の低下、（2）分子進化の中立モデルからのずれ、（3）連鎖不平衡の増大が引き起こされることから、これらの指標となる統計量についての経験値分布を得ることにより、統計学的「外れ値」として、正の自然選択の痕跡を見いだすことができる（図 3.7 参照）。なお、正の自然選択が、特定のヒト分集団（地理的集団など）においてのみ働く場合は、（4）対立遺伝子頻度の違い（この例では、赤丸 SNP アレル頻度の集団間の相違）としても、その痕跡を見出すことが可能である。

3.5 SNP と遺伝子発現量多型との関わり

これまでの節では、DNA レベルでの多様性（個人差、集団間差）について紹介してきたが、本節では、近年大きく研究が進展しているヒト遺伝子の発現量多型（Cheung et al. 2003, 2005; Morley et al. 2004）について、SNP や集団遺伝学と関連づけて取り上げたい。

通常、遺伝子発現量は量的形質とみなされることから、遺伝子発現量の個人差が観察された場合、それらは量的多型として取り扱うことが可能である。マイクロアレイとよばれる技術を用いることで、これまでよりも安価に、個々人の数十万以上の SNP 遺伝子型と、特定細胞（白血球細胞など）における全遺伝子の転写物量とを併せて探査することが容易になってきたという背景を受け、eQTL (expression Quantitative Trait Loci)マッピングが精力的に行われるようになってきている。例えば、昨年（2007 年）の Nature Genetics 10 月号には、eQTL マッピングに関連する異なる研究グループからの論文が連続した（Dixon et al. 2007; Göring et al. 2007; Stranger et al. 2007）。この eQTL マッピングでは、図 3.9 に示すように、ある遺伝子の転写物量に関する多様性情報（Expression phenotype：図 3.9 右パネル）が得られたときに、その転写物量多型を最も良く説明する SNP をゲノム全域から相関解析（SNP 遺伝子型と遺伝子転写物量多寡との相関解析：図 3.9 左パネル）により探索するという方法がとられている。これらの研究から、遺伝子の転写物量は、その遺伝子座に位置する SNP により規定されている（シス制御）とは限らず、トランス制御(標的遺伝子座から数 Mbp 以上離れた位置に存在する SNP や、異なる染色体上に位置する SNP による発現量制御のこと）されている可能性の高い遺伝子が数多く存在することが明らかになってきた。このことは、例えば、GWAS から感受性 SNP を同定することが出来たとしても、研究対象とすべき遺伝子を必ずしも絞り込んだことにはならず、SNP 近傍遺伝子のみを対象とすれば良いわけではないことを暗示している。ミシガン大学 Chen らのグループは、彼らの eQTL マッピングの成果を mRNA by SNP Browser (http://www.sph.umich.edu/csg/liang/asthma/) として公開しており、疾患感受性遺伝子の探索にも活用することが可能となっている。一方、遺伝子転写物量の個人差に加えて、転写物量の集団間差に関しても、転写物量多型に対する集団遺伝学アプローチから明らかにされつつある

（Spielman et al. 2007）。

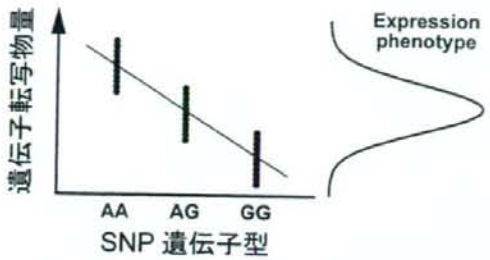

図 3.9　SNP 遺伝子型と遺伝子転写物量多型との相関についての模式図

ある遺伝子の転写物量を調べたとき、図右に示すような転写物量分布（青色）が得られたとする。転写物量多型が特定の SNP により遺伝的に規定されている場合、その SNP の遺伝子型と転写物量との間に有意な相関が認められる（図左）。この例では、G アレル保有数（0、1、2）が多くなると、転写物量が有意に低値を示す。

前節で述べたような自然選択の影響下にある「外れ値」SNP と eQTL 関連 SNP との関わりや、「外れ値」遺伝子の転写物量制御などは、生物学的に興味深い研究課題である。また、疾患関連研究の実践という側面からみると、「困難は分割せよ」という考え方があるように、疾患関連研究に求められる因果説明、すなわち原因（SNP）と結果（疾患表現型）との結びつきの説明において、eQTL マッピングからの成果は「SNP→中間表現型（転写物量多型）→疾患表現型」という流れを生み出すことが期待され、対象疾患の遺伝学的理解を加速する役割を担うことになるであろう。

3.6 おわりに

近年、進化医学（Darwinian 医学ともいう）の視点からヒト疾患を再考しようという試みが、世界中でなされ始めている。従来の「WHAT & HOW（何がどのようにして病気をおこすのか）」という要因のみならず、「WHY（なぜ病気になるのか）」要因までも、医学・医療の場で考慮すべきであるということであろう。個人ごとに疾患の「WHAT & HOW」要因が明らかになっていく個別医療時代において、進化医学的な考え方が臨床医学をどのように変えていくかは不明なところがある。しかしながら、本章で概説したように、集団遺伝学的立場からの研究が、疾患の「WHY」要因、すなわち疾患の進化的起源を理解・説明する上でも重要な役割を担うことはいうまでもない。この解説を通じて、実学としての「集団遺伝学」の必要性を少しでも感じていただければ幸いである。

4. SNP による連鎖解析

4.1 はじめに

連鎖解析は、疾患遺伝子同定の基本的手法として挙げられ、アルツハイマー病、本態性高血圧、II 型糖尿病などの多因子疾患においても、その効力を発揮してきた。従来の連鎖解析で用いられる遺伝マーカーは、多型性に富むマイクロサテライトが主流であったが、近年、国際 HapMap 計画が完了するなど、SNP データベースの充実が著しく、それを受けて、SNP を用いた連鎖解析が行なわれるようになった。膨大な数の SNP を同時に用いることにより、多型性の低さを補って余りある、豊富な情報に基づいた信頼度の高い解析が可能となっている。ここでは、連鎖と組換えの定義といった基本的な事柄から、SNP を用いた連鎖解析の世界的な実情に至るまでの詳細を述べる。

4.2 連鎖と組換え

メンデルの独立の法則によれば、ある遺伝子座における二つのアレルのどちらが親から子に遺伝するかは、他の遺伝子座に関係なく決定される。しかし、同一の染色体上に、しかも相互に近い領域に位置する二つの遺伝子座の場合、この法則は当てはまらない。すなわち、一方の遺伝子座 A において、子に伝達されるアレルが決定すれば、もう一方の遺伝子座 B においても、A において伝達されるアレルと同一の相同染色体に位置するアレルが、一つのハプロタイプとして同時に伝達されることになる。この場合、この二つの遺伝子座は連鎖しているという。しかし、この非独立性は完全ではなく、同一の相同染色体に位置する二つのアレルが、常に同時に伝達されるとは限らない。これは減数分裂の際、対合した二つの相同染色体の一部が、ある一定の確率で入れ替わり（乗換え）、相同染色体とは異なる遺伝子の組み合わせである配偶子が形成されるためである。この現象は組換えとよばれ、これによって完全な連鎖状態が崩れ、結果として、もともと一つの染色体上に位置していた二つのアレルが、別の染色体に位置することになる。

組換えの例を図 4.1 に示す。ショウジョウバエの翅形（正常翅ー優性、痕跡翅ー劣性）と体色（灰色ー優性、黒色ー劣性）を支配する遺伝子は同一染色体上の近傍に位置し、連鎖の関係にある。ここで、正常翅と灰色、痕跡翅と黒色をそれぞれ表現型とする二つの純系を交配し、得られたヘテロ接合体と変異型の戻し交配を行なうと、多くは、正常翅である個体の体色は灰色となり、痕跡翅である個体の体色は黒色となる。しかし組換えにより、一部の個体は正常翅で黒色、または痕跡翅で灰色となっている。このように、完全連鎖の下では起こり得ない表現型の組み合わせが、組換えによって少数ながら生じる。

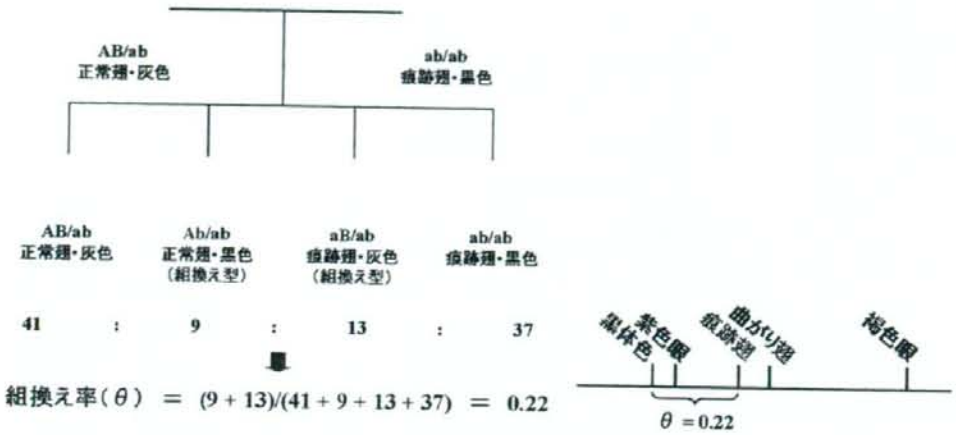

図 4.1　組換えと連鎖解析

任意の二形質における表現型の観察数から、両者を支配する遺伝子座の相対的な位置関係が推定可能である。連鎖解析は組換えの概念を応用した遺伝子探索法であり、マイクロサテライトなどの遺伝子多型を用いて大規模に行なわれる。

組換えの生起確率（組換え率、θ で表す）が取り得る値の範囲は、0（完全連鎖）から 0.5（完全独立）までであり、上記のような系統間交雑集団の場合は、戻し交配で得られた全個体の数に対する、組換え型（純系では見られない表現型の組み合わせを示す個体）の数の割合となる。図 4.1 に示した例では、$\theta = \dfrac{9 + 13}{41 + 9 + 13 + 37} = 0.22$、すなわち、親から子へ遺伝子が伝えられる過程で、およそ 5 回に 1 回の割合で組換えが起こっていることを示す。

4.3 連鎖解析とは

組換え率は、一般に染色体上の距離と相関する。すなわち、同一の染色体上に位置する二つの遺伝子座間の距離が大きいほど、乗換えが起こりやすく、結果として組換え率も高くなる。この関係を利用し、DNA 多型などをマーカー（目印）として、様々な形質を支配する遺伝子座の位置を推定するアプローチを連鎖解析とよぶ。連鎖解析法の大半は、統計学における尤度比検定を基本としたものである。すなわち、「着目する表現型（を支配する遺伝子）とあるマーカーが、組換え率 θ で連鎖している」という仮説に基づいて、θ などをパラメーターとする尤度関数 L_1 を構築し、L_1 を最大化する各パラメーターの最尤推定値（観察されるデータを説明するのに、最も「それらしい」値）を求める。この最大化された L_1 と、帰無仮説（表現型とマーカーは独立である、すなわち、$\theta = 0.5$）での尤度 L_0 の対数尤度比の 2 倍（$2*ln(\frac{L_1}{L_0})$）は、近似的に χ^2 分布にしたがうため、これを利用した有意性検定が行なわれる。

連鎖解析法には大きく分けて、パラメトリック連鎖解析法とノンパラメトリック連鎖解析法の二種類が存在する。前者は、疾患の遺伝形式（常／性染色体遺伝、優／劣性など）をあらかじめ仮定して解析を行なう手法であるため、主に単一遺伝性疾患を対象として、ある程度の規模の家系内に複数の罹患者が存在するようなデータを解析する場合に有効である。一方で後者は、特定の遺伝形式を仮定せず、罹患者間で共有されるアレルの数に着目した手法であるため、複雑な遺伝形式を示す多因子疾患の解析に有効であり、また大規模な家系のデータは必ずしも必要でなく、小規模なデータであっても効力を発揮する。

ヒトに関しては、実験動物などのように、解析用の集団を人為的に作成することは不可能であり、一般集団からのデータの収集も、個人情報保護の問題などから、やはり容易ではない。また生活習慣病など、成人期以降に発症する疾患の多くにおいて、複数の世代にわたるデータの収集はほぼ不可能であるため、利用可能なデータの数は必然的に限られたものとなる。さらに、このような疾患の多くは多因子性であり、その発症リスクには、複数の遺伝子の他、環境因子なども影響を及ぼすため、遺伝形式の適切な仮定は事実上不可能である。

以上の理由から、ヒトの疾患を対象とした研究においては、ノンパラメトリック連鎖解析法が主流であり、その中でも、二名以上の同胞（兄弟姉妹）の記録を多数収集して行なう罹患同胞対解析が最も広く用いられている。

罹患同胞対解析の概略を以下に述べる。ゲノム上のある領域について、任意のペアが持つアレルが、ともに共通の祖先個体が有していた一つのアレルの複製である状態を、identical by descent（IBD）であるという。ヒトは二倍体であるから、IBD であるアレルの数は、0（双方のアレルとも無関係）、1（一方のアレルのみ IBD）、および 2（ともに IBD）のいずれかの値を示し、例えば、血縁関係がない場合は 0、また親子間では常に 1、などとなる。同胞（兄弟姉妹）間では、IBD であるアレルの数が 0、1、および 2 となる確率の期待値が、それぞれ 0.25、0.5、および 0.25（この確率を各々z_0、z_1、および z_2 とする）となる。しかし、ある疾患を発症している同胞のペアの場合、その疾患に関与する遺伝子座については、共通したアレルを受け継いでいる可能性が高いため、z_0 は理論上の値よりも低くなり、同時に z_1 や z_2 は高い値を示すと考えられる。この考え方に基づき、罹患同胞対解析ではまず、調査対象となる疾患を発症している二名以上の同胞の記録を多数収集し、ゲノム上に散在するマーカーにおけるこれら同胞対、およびその両親の遺伝子型から、ゲノム上の任意の点について、尤度関数 L_1 を最大化する z_0、z_1、および z_2 を求める。L_1 は以下に示す式（式 1）で表される。

$$L_1 = \prod_{i=1}^{N} (w_{i0}z_0 + w_{i1}z_1 + w_{i2}z_2) \qquad \text{（式 1）}$$

ここで、N は家系の数、w_{i0}、w_{i1}、および w_{i2} は、家系 i において、同胞対間の IBD が各々0、1、および 2 となる確率である。次に、遺伝子座と疾患は連鎖していない（$z_0 = 0.25$、$z_1 = 0.5$、および $z_2 = 0.25$）とする帰無仮説における尤度 L_0 を計算し、その対数尤度比（$\log_{10}(\frac{L_1}{L_0})$）である LOD スコアーが有意に高い値を示せば、その点に疾患原因遺伝子が存在する可能性が示唆される。

図 4.2 の例で、マーカーA においては、二人の子の遺伝子型がまったく同じで

あり、しかもアレル 2 は父親しか持っていないため、二人の子はともにアレル 2 を父親から、アレル 1 はともに母親から受け継いでいることが分かる。また、両親とも遺伝子型はヘテロ（それぞれ 1/2、1/3）であるから、二人の子が受け継いだアレルは、ともに一つのアレルの複製である、すなわち IBD であるアレルの数は 2 であることが確実である。したがって、$w_{i0} = w_{i1} = 0$、$w_{i2} = 1$ を式（式 1）に代入すると、尤度 $L_1 = z_2 = 1 - z_0 - z_1$ となり、L_1 は $z_0 = z_1 = 0$、$z_2 = 1$ のとき最大（$L_1 = 1$）となる。一方、帰無仮説における尤度 $L_0 = 1 - 0.25 - 0.5 = 0.25$ であるから、LOD スコアーは、$\log_{10}(\frac{L_1}{L_0}) = \log_{10}(\frac{1}{0.25}) \fallingdotseq 0.60$ となる。

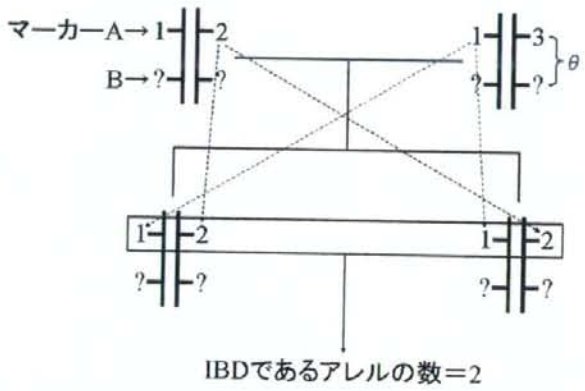

図 4.2　罹患同胞対解析におけるマーカーでの尤度の計算

マーカーA においては、二人の子の遺伝子型はまったく同じであり、しかもアレル 2 は父親しか持っていないため、二人の子はともにアレル 2 を父親から、アレル 1 はともに母親から受け継いでいることが分かる。また、両親とも遺伝子型はヘテロであるから、二人の子が受け継いだアレルは、ともに一つのアレルの複製であり、IBD であるアレルの数は 2 であることが分かる。したがって、$w_{i0} = w_{i1} = 0$、および $w_{i2} = 1$ を本文中の式（式 1）に代入すれば、尤度 $L_1 = z_2 = 1 - z_0 - z_1$ となる。

これは、マーカーA と疾患が連鎖していると考えることが、連鎖していないと考えることと比較して、4 倍「それらしい」ことを示している。統計学的には有意ではないが、真にマーカーA と疾患が連鎖していれば、より多くの記録を収集し、解析を行なうことで、有意な結果が得られる。

マーカー情報が得られていない点での尤度は、近隣に位置するマーカーからの情報と、そのマーカーとの組換え率 θ から計算される。同じく図 4.2 で、この A

から $\theta = 0.01$ の距離に位置する点 B での尤度は、表 4.1 から、$L_1 = \Sigma\, p(v_B \mid v_A) l\,(v_A, v_B, \theta) = \{(1 - 0.01)^4 + 2\,(1 - 0.01)^2\,(0.01)^2 + (0.01)^4\}\,(1 - z_0 - z_1) + \{4\,(1 - 0.01)^3\,(0.01) + 4\,(1 - 0.01)\,(0.01)^3\}\,z_1 + 4\,(1 - 0.01)^2\,(0.01)^2\,z_0$ で、L_1 が最大となるのは、やはり $z_0 = z_1 = 0$、$z_2 = 1$ のとき（$L_1 = 0.96$）であり、LOD スコアーは $\log_{10}(\frac{0.96}{0.25}) \fallingdotseq 0.58$ と、組換えの可能性を考慮した分、値が低くなっている。

表 4.1　罹患同胞対解析におけるマーカー間での尤度の計算

v_A	v_B	$p\,(v_B \mid v_A)$	$l\,(v_A, v_B, \theta)$
[0 0 0 0]	[0 0 0 0]	$(1 - \theta)^4$	$1 - z_0 - z_1\,(= z_2)$
	[0 0 0 1]	$(1 - \theta)^3\,\theta$	z_1
	[0 0 1 0]	$(1 - \theta)^3\,\theta$	z_1
	[0 0 1 1]	$(1 - \theta)^2\,\theta^2$	z_0
	[0 1 0 0]	$(1 - \theta)^3\,\theta$	z_1
	[0 1 0 1]	$(1 - \theta)^2\,\theta^2$	$1 - z_0 - z_1$
	[0 1 1 0]	$(1 - \theta)^2\,\theta^2$	z_0
	[0 1 1 1]	$(1 - \theta)\,\theta^3$	z_1
	[1 0 0 0]	$(1 - \theta)^3\,\theta$	z_1
	[1 0 0 1]	$(1 - \theta)^2\,\theta^2$	z_0
	[1 0 1 0]	$(1 - \theta)^2\,\theta^2$	$1 - z_0 - z_1$
	[1 0 1 1]	$(1 - \theta)\,\theta^3$	z_1
	[1 0 0 0]	$(1 - \theta)^2\,\theta^2$	z_0
	[1 0 0 1]	$(1 - \theta)\,\theta^3$	z_1
	[1 0 1 0]	$(1 - \theta)\,\theta^3$	z_1
	[1 0 1 1]	θ^4	$1 - z_0 - z_1$

マーカー情報が得られていない点での尤度は、近隣に位置するマーカーからの情報と、そのマーカーとの組換え率から計算される。図 4.1 の例において、A における継承ベクトル（親から子へいずれのアレルが伝達されたかを、二値変数（0／1）で表すベクトル）を[0 0 0 0]とすると、そこから θ の距離に位置する点 B での継承ベクトルの確率と尤度は、以下の通りである。子は両親からアレルを一つずつ受け継ぐので、継承ベクトルにおける要素の数は、子の数の二倍（図 4.2 の例では $2 \times 2 = 4$）である。

なお、継承ベクトルの要素は二値変数であり、その数は 4 であるから、B において考えられる継承ベクトルの数は $2^4 = 16$ となる。これらはすべて相互に背反であるから、全体の尤度は、継承ベクトルの事後確率（$p(v_B \mid v_A)$）と尤度（$l\,(v_A, v_B, \theta)$）の積をすべて足し合わせた値（$L_1 = \Sigma\, p(v_B \mid v_A) l\,(v_A, v_B, \theta)$）となる。

ここでは単一のマーカーの情報のみを用いた場合（単点解析）について述べた

が、実際には図 4.3 に示すように、複数のマーカーからの情報を同時に用いた、より複雑な計算（多点解析）を行ない、θ を横軸、また LOD スコアーを縦軸として、θ の値を少しずつ変化させながら、原因遺伝子が存在する可能性が最も高い点を探っていく。

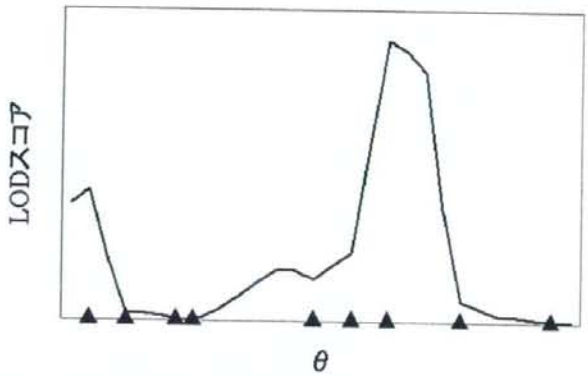

図 4.3　連鎖解析における LOD スコアーのプロット

組換え率 θ を横軸（黒い三角はマーカーを示す）、LOD スコアーを縦軸として、θ の値を少しずつ変化させながら、値が最大となる点（＝原因遺伝子が存在する可能性が最も高い点）を探っていく。

現在、ヒトの遺伝性疾患に関する連鎖解析で主に用いられているソフトウェアは、GENEHUNTER（Kruglyak et al.1996）、Allegro（Gudbjartsson et al. 2000）および MERLIN（Abecasis et al. 2002）などである。これらは、上記の罹患同胞対解析の他、パラメトリック連鎖解析も可能である。この他、研究の目的、対象とする生物種、形質、あるいはデータの構造や規模などにより、実に様々な手法が開発、提唱されており、またそれらを実行するソフトウェアも数多く公開されている。

4.4 SNP と連鎖解析

従来、遺伝解析にマーカーとして用いられる DNA 多型は、主にマイクロサテライト（2〜5 塩基を一単位とする反復配列）であった。マイクロサテライトは、多型性に富む（＝反復数のバリエーションが数多く存在する）ため、ヘテロ接合度が高く、マーカーとして用いた場合、いずれのアレルが親から子へ伝達されたかについて、正確な情報を得やすい。その反面、出現頻度が低く（$\frac{1}{\text{数十万}}$bp）、遺伝子多型そのものが、ある形質における表現型を決定する遺伝子と見なされるケースは比較的まれである。一方、SNP は基本的に正常型/変異型の二種類のみであるため、ヘテロ接合度が低く、アレルの伝達に関する一座位あたりの情報量はマイクロサテライトに劣るが、その出現頻度はマイクロサテライトと比較して非常に高く（$\frac{1}{\text{数百}}$bp）、また疾患の直接的な原因となり得るミスセンス変異（アミノ酸の置換を伴う変異）であるケースもしばしばである。したがって、従来の遺伝解析のプロセスとしては、両者の特長を相補的に活かす形で、まずマイクロサテライトをマーカーとした連鎖解析を行ない、候補となる領域を大まかに絞り込んだ上で、SNP を用いた関連解析により、原因となる多型を詳細に検討していくという流れが一般的であった。

しかし近年、データベースの拡大、整備やゲノム解析技術の飛躍的な進展により、SNP の利用環境が急速に改善されるに及んで、数十万単位の SNP を解析に利用することが可能となった。Kruglyak（1997）は、700〜900 の SNP で構成される連鎖地図であれば、300〜400 セットのマイクロサテライトと同等の情報量を有するとしている。また、近年では Evans と Cardon（2004）も、シミュレーション実験により、マイクロサテライト、SNP 双方について、現在利用可能かつ最も高密度なパネルを用いた場合の情報量を算出し、両親の遺伝子型情報の有無に関わらず、マイクロサテライトよりも SNP を用いた場合の方が、より多くの情報量を有することを示唆した。すなわち、現在の SNP の利用環境は、一座位あたりの情報量の少なさを十分に補うばかりか、全体としては、マイクロサテライトによって提供される情報量を上回るほど充実したものに進化を遂げていると言える。

図 4.4 に、SNP をマーカーとしたある一家系（構成員＝両親、および罹患同胞対である二人の子）の連鎖解析の例を示す。SNP A においては、両親、および二人の子ともすべて遺伝子型がヘテロであるため、二人の子がそれぞれ両親から受け継いでいるアレルやその数は確定できない。ただし、一方のアレルのみを共有している可能性はなく、両親からそれぞれまったく同じようにアレルを受け継いでいる（共有するアレルの数＝2）か、一方のタイプのアレルをそれぞれ別の片親から受け継いでいる（共有するアレルの数＝0）かのいずれかであることまでは判断できる。一方、A に隣接し、ほぼ完全連鎖の関係にある SNP B では、父親の遺伝子型がホモであるため、二人の子が父親からどちらのアレルを受け継いでいるかは分からない。しかし、少なくとも母親由来のアレル 2 は二人に受け継がれ、共有されていることが分かる。

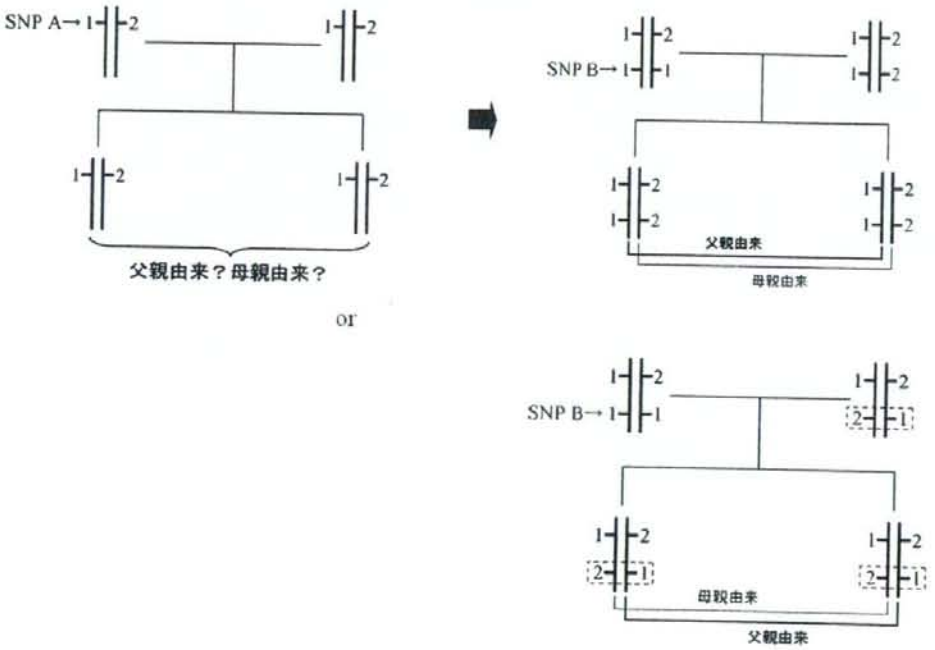

図 4.4. SNP をマーカーとした連鎖解析

左の図から、SNP A においては、罹患同胞対である二人の子が一方のアレルのみを共有している可能性はなく、両親からそれぞれまったく同じアレルを受け継いだ（IBD であるアレルの数＝2）か、それぞれのアレルを別の片親から受け継いだ（IBD であるアレルの数＝0）かのいずれかであることまでしか判断できない。そこへ右の図のように、A のすぐ近くに位置する SNP B の情報を加えることで、両親からそれぞれまったく同じアレルを受け継いだと結論付けられ、

さらに、そのことが疾患発症の一因である可能性も示唆される。なお、下の図のように、母親のディプロタイプが逆である場合も同様の結果となる。

ところで、このきょうだいがAにおいて共有するアレルの数は0か2であるから、Bにおいて母親由来のアレルを共有している以上、父親からも同じアレルを受け継いでいると考えられる（ただし、完全連鎖の関係にない場合はこの限りではない）。すなわち、この領域においては、このきょうだいは両親からそれぞれまったく同じアレルを受け継いでいると結論付けられ、同時に、そのことが疾患の発症に至った一因である可能性も示唆される。なお、母親のディプロタイプ（二本のハプロタイプとしての遺伝子型）が1-2/2-1である場合も、Aにおいて両親から受け継がれるアレルのタイプが逆となる以外は、同様の結果となる。

上記はほんの一例であり、さらに多くのSNPの情報を同時に用いることで、より信頼度の高い連鎖解析が可能となる。

4.5 SNPLINK－高密度 SNP 連鎖解析用 Perl スクリプト

前節で述べたような SNP の高密度化の流れを受けて、連鎖解析の段階から、マイクロサテライトではなく、SNP を利用した研究も実際に報告されてきている。Webb ら（2005）は、Allegro および MERLIN を用いて、高密度 SNP による連鎖解析を実行するフリーの Perl スクリプト、SNPLINK を公開している。Allegro、MERLIN とも、少なくとも 1,000 前後の SNP を一度に計算に取り込むことが可能であるが、強い連鎖不平衡にある SNP を解析に用いた場合、偽陽性が多く検出される傾向にある。SNPLINK では、全 SNP を用いた連鎖解析を行なった後、連鎖不平衡の度合を示す指標 D'、r^2 をもとに、相互に独立な SNP のみを抽出し、再度解析を行なうオプションが選択可能である。

ダウンロードは、イギリス Institute of Cancer Research のサイト（http://www.icr.ac.uk/cancgen/molgen/MolPopGen_Bioinformatics.htm）で可能である。ダウンロードされた圧縮ファイルに含まれるファイルは以下の 8 点である。

- Perl スクリプト本体（snplink.pl）
- モジュール 2 点（snplinkfunctions.pm－ノンパラメトリック連鎖解析用、snplinkfunctionspar.pm－パラメトリック連鎖解析用）
- マニュアル（Installation and user guide.doc）
- データファイルのサンプル（example-chr1.pre－家系情報、表現型、およびマーカー遺伝子型のファイル、example-chr1.dat－遺伝子座の位置情報、遺伝形式、アレル頻度のファイルである。なお両者とも、「《任意の文字列》-chr《染色体番号》」というファイル名でなければならない）
- 解析に関するオプション選択のための入力ファイル 2 点（example1.in、example2.in）

SNPLINK の利用に際しては、Allegro、MERLIN、および Perl の他、結果を postscript ファイルとして出力するため、R のインストールも必要である。また現在のところ、UNIX 環境下でのみ使用可能である。これらのソフトウェアをコンパイルした後、入力ファイルを開いて、解析に関するいくつかのオプションを選

択する。

以下は入力ファイル example1.in の例である。

name|example	・・・	データファイル名（上記の「任意の文字列」に相当する部分）
analysis|nonparametric	・・・	ノンパラメトリック/パラメトリック連鎖解析のいずれを用いるかを選択（なお、ノンパラメトリック連鎖解析には MERLIN が、パラメトリック連鎖解析には Allegro が各々用いられる）
start|1	・・・	解析対象の染色体のうち、番号が最も小さなものを入力
chromosome|1	・・・	解析対象の染色体のうち、番号が最も大きなものを入力
mkstart1|2.88	・・・	染色体の先端から、最も近い地点に位置する SNP までの距離（Mb）
xinclude|no	・・・	X 染色体の解析を行なうかを選択
LDremoval|yes	・・・	連鎖不平衡にある SNP をデータから除去するかを選択
measure|both	・・・	連鎖不平衡のパラメーターを、D'/r^2 より選択（both は双方とも用いることを意味する）
cutoffDprime|0.7	・・・	連鎖不平衡にあると判断する D' の基準値
cutoffrsq|0.4	・・・	連鎖不平衡にあると判断する r^2 の基準値

D'および r^2 は、EM アルゴリズムにより推定される。また、「cutoffDprime|0.7」および「cutoffrsq|0.4」は、John ら（2004）や Schaid ら（2004）の報告に基づいた値であるが、実際に用いる際には、より厳密な形で設定するべきであろう。

適宜これらを設定し、

```
> perl snplink.pl example1.in
```

と入力すれば、すべての計算が自動的に行なわれ、結果が出力される。

現在までに、慢性リンパ急性白血病（Sellick et al. 2005）や、統合失調症（Arinami et al. 2005）、関節リウマチ（Amos et al. 2006）などの感受性遺伝子探索に関する連鎖解析に、このスクリプトが用いられている。

4.6 おわりに

以上述べてきたように、近年のゲノム解析技術の目覚ましい進展により、大規模なSNPの利用が可能となったことに伴い、より信頼性の高い結果が得られることが今後期待される。例えば、以前用いたデータを、高密度SNPをマーカーとして再度解析することで、従来は検出し得なかった重要な連鎖が見出されることも十分考えられる。ただ一方で、タイピングエラーやデータの欠損は、解析結果に重大な影響を及ぼす可能性があり、その点については慎重に検討を進めることが必要である。