このような統計学的仮説検定の出発は、仮説の設定である。通常、結論したい事柄（「調べたSNPと疾患発症との間に関連がある」）を作業仮説（実験仮説ともいう）として調査研究を実施するが、仮説検定のために、作業仮説を対立仮説 H_1 とし、作業仮説を否定する仮説（「調べたSNPと疾患発症との間に関連がない」）を帰無仮説 H_0 と設定する。仮説検定の対象となるのは帰無仮説であり、帰無仮説が統計学的に棄却されれば、対立仮説が支持されることになる。

すなわち、帰無仮説が棄却されて初めて研究者が結論したい事柄を主張することが統計学的に可能となるわけであり、この意味において、帰無仮説（無に帰される仮説）とよばれる。前節で述べた推定の場合と同じく、仮説検定の背景では、特定のパラメーターに従う母集団の存在を仮定しており、母集団からの無作為抽出標本を調べることにより標本統計量（SNP遺伝子型頻度など）を求め、仮説検定に必要な検定統計量とよばれる数量を算出する、という手順となる。疾患関連SNP解析における具体的な統計手法は他章をご参照いただきたいが、ここでは仮説検定結果の解釈を中心に以下にまとめておこう。

患者-対照関連解析（遺伝モデルを仮定しない場合）では、帰無仮説 H_0 は「SNPアレル（あるいは遺伝子型）と疾患とは関連がない（独立である）」、対立仮説 H_1 は「SNPアレル（あるいは遺伝子型）と疾患とは関連がある（独立でない）」とし、SNPアレル（あるいは遺伝子型）と疾患表現型についての分割表（クロス集計表ともいう）に基づき、独立性の検定を行うことになる。

例えば、表10.1に示すような2×2分割表が得られたとする。独立性を検定するための統計量は χ^2 統計量であり、帰無仮説 H_0 のもとでは、

$$\chi_0^2 = \frac{(A_1+A_2+a_1+a_2)(A_1a_2-A_2a_1)^2}{(A_1+A_2)(a_1+a_2)(A_1+a_1)(A_2+a_2)} = \frac{(2N+2M)(A_1a_2-A_2a_1)^2}{(A_1+A_2)(a_1+a_2)(2N)(2M)} \quad \text{（式 8）}$$

で与えられる。

表 10.1　SNP アレル分割表

	患者群	対照群	計
SNP アレル数			
A	A_1	A_2	A_1+A_2
a	a_1	a_2	a_1+a_2
計	$2N$ ($=A_1+a_1$)	$2M$ ($=A_2+a_2$)	$2N+2M$

患者群 N 名、対照群 M 名を分析した場合。

図 10.2 には、帰無仮説 H_0 のもとでの χ_0^2 統計量の確率密度分布（χ^2 分布ともいう）を示しているが、2×2 分割表に基づく場合、χ_0^2 統計量は自由度 1(=(2-1)×(2-1)) の χ^2 分布に従う（SNP 遺伝子型で集計した 2×3 分割表に基づく場合の自由度は 2 (=(2-1)×(3-1)) である)。

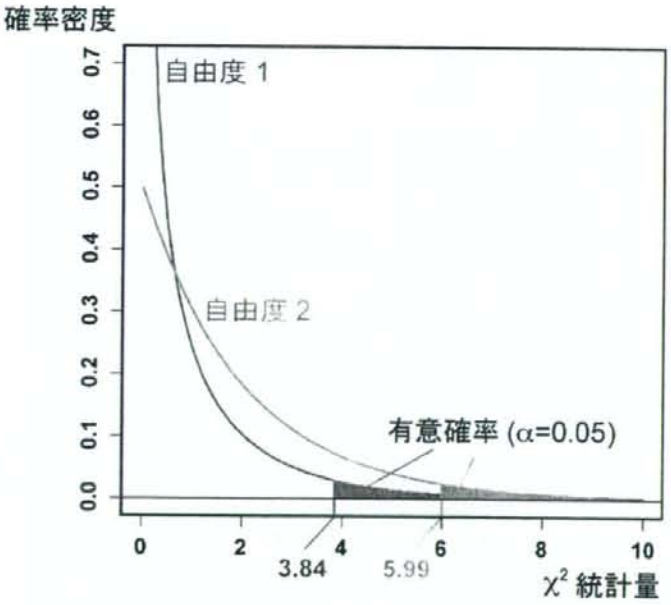

図 10.2　χ^2 統計量の確率密度分布（χ^2 分布）

自由度 1 および 2 における χ^2 分布を、それぞれ緑色、橙色の線で表している。有意確率 α=0.05 (図中の曲線下面積) とする χ_0^2 統計量は、それぞれ 3.84 (自由度 1)、5.99 (自由度 2) である。

一般に、標本より検定統計量を算出し、その値以上の検定統計量が得られる確率のことを有意確率 P とよぶ。有意確率 P がどの程度小さい場合に、帰無仮説 H_0 を棄却（対立仮説 H_1 を採択）するかについての基準を有意水準とよび、通常 α で表す。図 10.2 の χ^2 分布（自由度 1）では、α=0.05 に設定した時の有意確率を緑色（自由度 2 では橙色）で示しているが、χ_0^2 統計量が 3.84（自由度 2 では 5.99）のときに有意確率 P=0.05 となることを意味する。従って、$\chi_0^2 > 3.84$（自由度 1）の場合、有意確率 P<0.05 となり、帰無仮説 H_0 を棄却したうえで、「有意水準 5% のもとで、SNP アレルと疾患とは関連がある」と結論する。一方、$\chi_0^2 \leq 3.84$（自由度 1）では有意確率 $P \geq 0.05$ となり、帰無仮説 H_0 を採択し、この場合の結論は「有意水準 5%のもとで、SNP アレルと疾患とは関連があるとはいえない」となる。

ところで、有意水準α=0.05 というのは、同一状況下で 20 回検定を行うと、そのうちの 1 回は帰無仮説 H_0 を誤って棄却してしまう危険性があることを意味する。この種の誤り、すなわち、「真は H_0 であるのに H_0 を棄却してしまう誤り」のことを第 1 種の過誤（偽陽性）とよび（表 10.2）、この過誤が生起する確率はαである（このことから、有意水準αのことを危険率ともいう）。

表 10.2　検定における 2 つのタイプの過誤と検出力の関係

真実	検定結果	
	帰無仮説 (H_0) を棄却（有意差あり）	帰無仮説 (H_0) を採択（有意差なし）
疾患と関連する (H_1)	検出力 1-β	第 2 種の過誤β
疾患と関連しない (H_0)	第 1 種の過誤α	特異度 1-α

一方、「H_0 が誤っているにも関わらず H_0 を採択してしまう誤り」のことを第 2 種の過誤（偽陰性）とよび、通常βで表す。図 10.3 には、第 1 種の過誤αと第 2 種の過誤βの関わりを模式的に示す。帰無仮説 H_0 および対立仮説 H_1 のもと、図 10.3 に示すような検定統計量の確率密度分布がそれぞれから得られたとする。一見して明らかなように、ある有意水準を設定すると、第 1 種の過誤α、第 2 種の

過誤βのいずれもが一意に定まる。

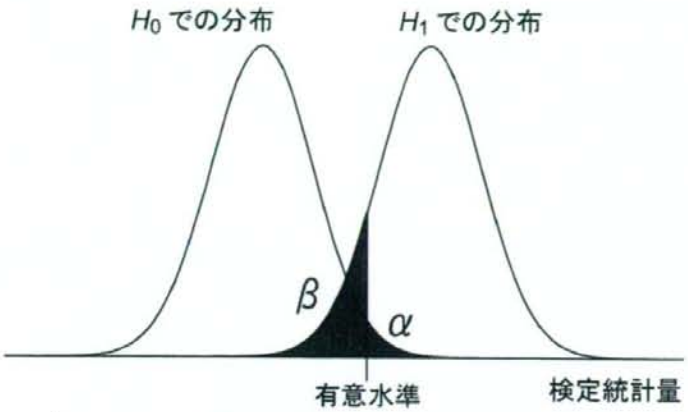

図 10.3　第 1 種の過誤αと第 2 種の過誤βの関係性

この両者の関係はトレードオフであり、有意水準を変更することで片方の過誤を小さくしようとすると、もう片方の過誤が大きくなってしまう（例えば、αを小さくして第 1 種の過誤を減らそうとすると、第 2 種の過誤βが増す）。

それでは、「真が H_1 であるとき、H_0 を正しく棄却し H_1 を採択する確率」は、どのようにして求めることができるであろうか。表 10.2 に示すように、この確率は 1-βで求められ、仮説検定における検出力（検定力）とよぶ。

前述したように、第 1 種の過誤αと第 2 種の過誤βは、有意水準のもとではトレードオフの関係にあるが、標本サイズを大きくすることで、これらの過誤をともに低く調整することが可能となり、ゆえに検出力も高まる。

このような標本サイズへの依存は、検定統計量が標本サイズに影響されることに由来する。図 10.3 の例において、有意水準αは固定し、標本サイズのみを変化させた場合に、第 2 種の過誤βがどのように変化するかについての模式図を図 10.4 に示す。標本サイズを大きくすることにより、帰無仮説 H_0 および対立仮説 H_1 のもとでの検定統計量の標準偏差がいずれも小さくなり、その結果、第 2 種の過誤の確率βが低下する（検出力 1-βは高まる）。

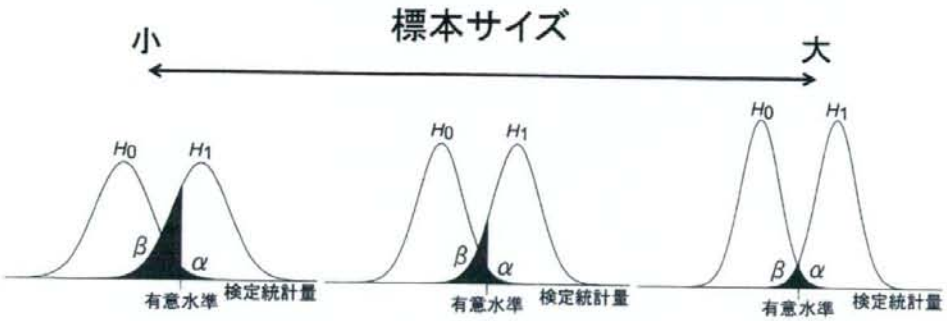

図 10.4　標本サイズと第 2 種の過誤（偽陰性）との関わりについての模式図

有意水準αは固定し、母集団から無作為抽出される標本の大きさのみを変化(小～大)させた時、
第 2 種の過誤β（青色の曲線下面積）が変化する程度を模式的に示している。

従って、疾患関連 SNP 解析において、立証したいと考える作業仮説を支持するために必要な標本サイズをあらかじめ推定しておくことは重要である。患者-対照関連解析から疾患に対してアプローチする場合、標本サイズを決定するためには、第 1 種の過誤α、第 2 種の過誤β（検出力 1-β）、オッズ比（遺伝子型相対リスクなど）、リスクアレル頻度、対象疾患の有病率などを定義する必要がある。実際の検出力の計算には、Genetic Power Calculator (Purcell et al. 2003: http://pngu.mgh.harvard.edu/~purcell/gpc/)、CaTS (Skol et al. 2007: http://www.sph.umich.edu/csg/abecasis/CaTS/index.html)、QUANTO (Gauderman 2002a; 2002b; http://hydra.usc.edu/GxE/)などの WEB サイトやフリーウェアを使用することができ、研究目的に応じて使い分けることになる（2 章に実際の計算例を示しているので、ご参照願いたい）。

当たり前のことではあるが、明確な目標（感受性 SNP の効果の大きさ）を設定して研究を開始しなければ、見つかるものも見つからず（β）、見つかったとしても偽物（α）である可能性が高まるということである。

10.3 ベイズ推定

前節までで述べたように、従来の統計学は、母集団およびそれを規定する唯一のパラメーターが存在すると仮定し、母集団より抽出した標本からパラメーターを推定する、という推計統計学である。すなわち、観測データが得られたのはパラメーターで規定される母集団が存在するからである、という見方にほかならない。一方、ベイズ統計学とよばれる学問分野では、母集団の存在そのものを仮定しない。もちろん、観測データが得られたのには何らかの原因があるはずなので、ベイズ統計では、観測データからの推定の不確実性を考慮した上で、ベイズ確率とよばれる「原因の確率」を考える。このような確率を推定するためのもとになる「ベイズの定理」を以下に紹介する。

簡単のために、X および Y を 2 つの離散的な確率変数とし、X が原因、Y はその結果である場合を考える。定義より、

$P(\mathrm{X})$： 事象 X が生起する確率（事前確率）

$P(\mathrm{X}|\mathrm{Y})$： 事象 Y が生起したという条件下、事象 X が生起する条件付き確率（事後確率）

であり、ベイズの定理から事後確率 $P(\mathrm{X}|\mathrm{Y})$は式 9 のように表すことができる。

$$P(\mathrm{X}|\mathrm{Y})=\frac{P(\mathrm{Y}|\mathrm{X})P(\mathrm{X})}{P(\mathrm{Y})} \quad （式 9）$$

一般に、原因 X が同一でも結果 Y まで同じとは限らないのであるが、観測データに基づき、「原因 X で条件付けた時の結果 Y の条件付き確率 $P(\mathrm{Y}|\mathrm{X})$」を求めることは十分可能である。また、原因 X や結果 Y の事前確率 $P(\mathrm{X})$, $P(\mathrm{Y})$も、実験・調査研究などにより見積もることができる。従って、ベイズの定理は、観測データなどの結果（式 9 の右辺）から、原因の確率（式 9 の左辺）、すなわち「結果 Y で条件付けた時の原因 X の条件付き確率 $P(\mathrm{X}|\mathrm{Y})$」を推定することができるということを意味している。

近未来の遺伝子診断を例として、具体的に説明してみよう。実際に疾患に罹患している（事象 X）場合、開発された遺伝子診断を行うことにより 99.5%の確率で診断結果は「陽性」となり、一方、疾患に罹患していない（排反事象 $\overline{X}$）場合は、その遺伝子診断により、99%の確度で「陰性」とされる場合を考える。また、疫学調査などから、この疾患の有病率は 5%（事前確率 $P(X)$）であることが情報として与えられているとする。ある人がこの遺伝子診断を受けた結果「陽性」と判定された（事象 Y）場合、この診断結果が真の「陽性」、すなわち、この人が実際に疾患に罹患している確率（事後確率 $P(X|Y)$）を計算してみると、ベイズの定理から、

$$P(X|Y)=\frac{P(Y|X)P(X)}{P(Y|X)P(X)+P(Y|\overline{X})P(\overline{X})} =\frac{0.995\times 0.05}{0.995\times 0.05+(1-0.99)\times(1-0.05)}=0.8397\approx 0.84 \quad \text{（式 10）}$$

となる。従って、遺伝子診断を受ける前は疾患に罹患している確率は 5%であった（有病率からの推定）が、受診後には罹患確率 84%へと推定値が改訂されたことになる(このような診断からも 16%の偽陽性が出現するという意味でもある)。ここで、事後および事前におけるオッズの割合のことをベイズ因子（Bayes factor）とよぶ。上記の例では、事象 X（疾患に罹患しているという事象）に対するベイズ因子は、

$$\frac{0.8397/(1-0.8397)}{0.05/(1-0.05)}=99.5 \quad \text{（式 11）}$$

と求められる。

ベイズ因子は、ベイズ統計学においてモデル選択の際に使用されている数値である（ベイズ情報量基準 BIC や偏差情報量基準 DIC など、モデル選択のための他の基準については他書[Congdon 2007 など]に譲る)。一般化した数式を以下に示した上で、ベイズ因子についてもう少し説明を加えてみよう。データ x を観測した時（例えば、遺伝子診断で陽性；2 種以上のデータ観測を行った場合はデータベ

クトル x）に、2 つのモデル M_0, M_1 を比較し、いずれかを選択するという問題を考える。ベイズ因子 B_{10} は下式により与えられる。

$$B_{10} = \frac{p(x|M_1)}{p(x|M_0)} = \frac{\int p(\theta_1|M_1)p(x|\theta_1,M_1)d\theta_1}{\int p(\theta_0|M_0)p(x|\theta_0,M_0)d\theta_0} \quad \text{（式 12）}$$

ここで、$p(x|M_i)$はモデル M_i の周辺尤度（Marginal likelihood）とよばれ、各モデルを構成するパラメーター θ_0, θ_1（固有分布をもつランダム変数）を用いることで、式 12 の右辺のように表すことができる。モデルごとにパラメーター θ_i の分布範囲で尤度を積分（尤度の平均化）し、モデル M_i の周辺尤度を求め、それらの比をベイズ因子と定義している（尤度比に似ているが、尤度比では 2 つの仮説［帰無仮説 H_0 と対立仮説 H_1］におけるパラメーター θ の最尤推定量の比を求めていることに留意）。実際の周辺尤度の求め方などは本章の範囲を逸脱しているのでここでは触れないが、ご興味を持たれた方は他の論文・成書（Bos 2002; Congdon 2007 など）をご参考にしていただきたい。ベイズ因子 $B_{10}>1$ とは、観測データからはモデル M_1 がモデル M_0 よりも確からしいことを意味し、$B_{10}<1$ の場合はその逆となる。

このように、ベイズ因子が 1 より大きいか小さいかの情報に基づきモデル選択するのであるが、得られた数値をどのように解釈すればよいかについてのガイドラインのひとつ（Jefreys 1961; Congdon 2007）を表 10.3 に示す。先の遺伝子診断に関する例ではベイズ因子が 99.5 と求められたことから、疾患罹患性についての「遺伝子診断」モデルが「有病率」モデルよりも確からしいと解釈することが可能である。

表 10.3　ベイズ因子に基づくモデル選択のガイドライン

ベイズ因子 B_{10}	モデル選択に関するベイズ因子からの解釈
1 より小さい	モデル 0 を支持
1–3	弱いながらもモデル 1 を支持
3–20	モデル 1 を支持
20–150	モデル 1 についての強い根拠
150 以上	モデル 1 を強く支持

Jefreys 1961, Congdon 2007 より改変

これまでに述べたように、ベイズ推定法では従前の科学的知見などを事前確率（あるいは事前分布）として取り込むことができる。逆に言えば、推定される事後確率の有用性は、設定した事前確率（と尤度）の正確性に大きく依存する。また、パラメーター数が多かったり、それらの確率分布型が複雑であったりすると、前述した周辺尤度の解を解析的に求めることが困難となる場合が多い。このような欠点を補うために、現在では、乱数を用いた数値計算法のひとつであるマルコフ連鎖モンテカルロ法を用いて、ベイズ推定からの事後確率（あるいは事後分布）やベイズ因子をシミュレーションにより求めることが主流となっている。

ヒト疾患や形質と関連する SNP を探査する場合を考えると、例えば、SNP を原因、疾患や形質をその結果とすることで、ベイズ推定からのアプローチが疾患関連 SNP 解析へも応用可能であることは容易に想像されることであろう。ベイズ確率やベイズ因子などの適用範囲はそれぞれであるが、実際、多くの研究において、ベイズ推定法を用いた疾患・形質関連多型の探索がすでになされている（Lunn et al. 2006; Marchini et al. 2007; Morris 2006; Servin and Stephens 2007; Sillanpää and Bhattacharjee 2005）。

10.4 おわりに

今後の疾患関連 SNP 解析を展望すると、ゲノム全域での体系的アソシエーション・スタディ（Genome-wide association study）により疾患感受性変異を同定するという流れは継続され、従って、異なる疾患・形質をもつより多くのヒトから、より多くの SNP 遺伝子型が決定されることになるであろう。

また、そのような研究からの知見も大量に蓄積していくことであろう。このことは、豊富な事前情報のもとベイズ確率に基づきヒト疾患や形質に解釈をつけていく、というベイズ統計学的アプローチにより適した研究環境が形成されるであろうことを意味する。一方、そのような状況にあっても、従来の統計学体系（検定、推定など）の基礎的な理解は必須であろう。ベイズ統計学を含めた統計学全般に対する知識をレパートリーとし、それらを使いこなすことで研究が進展することはいうまでもない。

しかしながら、統計の本質は「いかにして良質なデータを収集するか」にあることを忘れてはいけない。正確に決定された SNP 遺伝子型や、研究サンプルに関して正確・詳細に記述された情報（例えば、臨床情報）を収集することが、統計学的な疾患関連遺伝解析の出発点であることを最後に強調しておきたい。

＜用語説明＞

Bonferroni の補正法（Bonferroni correction）

有意水準 α を検定の数 N で除した値 $\frac{\alpha}{N}$ を、多重検定における有意水準 α' とする手法である。従来、多重検定における有意水準の補正法として、最も一般的に用いられてきたが、保守的な有意水準を与え、タイプ II エラー（本来棄却するべき帰無仮説を、誤って採択してしまうこと）を多数生じさせる傾向にあるため、近年はこれに代わる手法が主流となっている。

CGH（Comparative Genome Hybridization）法

スライドガラスに正常染色体を貼り付け、別々の色素で標識した正常 DNA と異常 DNA（例えば腫瘍細胞の DNA）を競合的にハイブリダイゼーションさせることで、コピー数の変化を検出することができる手法のこと。

Cockran-Armitage trend test

アソシエーション・スタディの有意差検定としてアレル頻度でカイ検定をおこなう手法が一般的であった。遺伝子型でのカイ検定の方が生物学的意味を有するので優先すべきという意見もあるが、アレルでの検定は見掛け上遺伝子型の検定の 2 倍の検体数となっているので、低い P 値を得ることができのでよく使われてきた。ただ、Hardy-Weinberg 平衡(HWE)からのずれがあると（例えば、マイナーアレルのホモ接合体が患者に多いため平衡からのずれが生じる）、遺伝子型での検定の方が適切な場合がある。ゲノム全域解析では、2 次スクリーニングのため有意な SNP を順位付けしたい。その際、HWE からの乖離があっても対応できる Cockran-Armitage trend test が適切な検定法である。

Common Disease-Common Variant (CD-CV)仮説／Common Disease-Rare Variant (CD-RV)仮説

高血圧症や糖尿病などのありふれた疾患（common disease）は、遺伝要因と環境要因とが複雑に絡み合って発症にいたる多因子疾患である。また、その遺伝要因も単一でなく、複数の感受性遺伝子が単独で、あるいは相互作用を介して、疾患発症に関わると考えられている。ゲノムの同祖性からの帰結により、ありふれた疾患の発症に関わる疾患感受性アレルの少なくとも一部は、複数の患者間で同祖共有していると考えることは十分妥当である。患者間で共有すると仮定する感受性アレルの集団頻度の違い（比較的高頻度[common variant]、低頻度[rare variant]）が、CD-CV および CD-RV 仮説の主要な相違点である。

copy number polymorphism (CNP：コピー数多型)

集団に 1%以上の頻度で存在する CNV のこと。アミラーゼ遺伝子群や臭い受容体遺伝子群は CNP の代表的な例である。

False discovery rate（FDR）

帰無仮説が棄却された検定のうち、実際には帰無仮説が真であるものの割合。ある p 値が有意水

準として与えられたとき、FDR は $\frac{\hat{\pi}_0 Np}{i}$ で表される。ここで、N はすべての検定の数を、また i は、与えられた p 値よりも有意性の高い検定の数、そして $\hat{\pi}_0$ は、帰無仮説が真である検定の、すべての検定に対する割合の推定値を表す。

FDR を一定の水準以下に制御する手法としては、Benjamini and Hochberg（1995）、Storey et al.（2003）、および Storey et al.（2004）が提唱されている。ここでは、すべての検定を有意性の低い順に N、N - 1、...、i、...、2、1 と並べ替えた後、各検定での p 値をもとに q 値（各検定での p 値を有意水準としたときの FDR）を計算し、この値がある基準値 q^* より小さいものを有意と見なす。

Fisher's exact test

カイ二乗検定において、度数が 5 以下のセルが存在すると精度を欠く。そこで、分割表の行、列の合計を変えず、すべての組み合わせを検討し、得られたデータが得られる確率を算出する。通常のアソシエーション・スタディでは多くの検体数を用いるので、カイ二乗検定で十分である。

Representational Oligonucleotide Microarray Analysis （ROMA）

2003 年に Rob Lucito らによって開発された、CNV を検出するための手法である。ゲノムを制限酵素で消化したり、PCR で増幅することで、従来の手法よりもゲノムの複雑さの低減、検出感度の向上を実現した。

アソシエーション・スタディ

ある SNP が疾患と関連しているかを検定する手法である。連鎖と関連は本来の日本語ではほぼ同じ意味であろうが、遺伝学において関連とは直接、ある遺伝子変異もしくはそれと連鎖不平衡にある遺伝子変異が病気と関連することをいい、連鎖とはその遺伝子変異が家系内で病気と連鎖して存在するということである。連鎖の場合、その遺伝子変異が原因である必要はない。

アレイ CGH

従来の CGH 法で用いられる染色体標本のかわりに、スライドガラスなどの基板上にプローブ DNA を結合させた、いわゆるマイクロアレイを CGH に応用した手法。

遺伝子型相対リスク（Genotype relative risk）

2 つのアレル（M, m）からなる SNP の場合、3 種類の遺伝子型（MM, Mm, mm）が観察され得るが、各遺伝子型の浸透率（発症率）比のことを、遺伝子型相対リスクとよぶ。例えば、m アレルがリスクアレルである場合、遺伝子型 MM に対する Mm あるいは mm 遺伝子型の遺伝子型相対リスクを計算することにより、この SNP の疾患への関わり方（優性、劣性、相乗遺伝モデルなど）を推量することが可能となる。

遺伝子発現量多型

遺伝子発現量の個人差のこと。遺伝子がコードするタンパク質配列の相違はないものの、比較的多くの遺伝子において、（1）量的な違いが観察されること、（2）その量的多型が遺伝的に決定されていることなどから、ヒトの形質や疾患との関連で近年注目されている。現時点では、大多

数の研究において、タンパク質レベルではなくRNAレベルで量的な違いを観察していることから、遺伝子転写物量多型ともよぶ。

遺伝的異質性／表現模写
疾患表現型は同一でありながら、家系、集団ごとに発症に関わる要因が異なること。多因子疾患においては、遺伝要因が疾患発症に関わることは共通するものの、関与する遺伝因子が異なることを遺伝的異質性とよび、一方、遺伝要因は関与せず、特定の環境要因に暴露されることなどにより疾患発症にいたる場合を表現模写とよぶ。

遺伝的浮動
生物集団におけるアレル頻度（突然変異遺伝子頻度）の継代的な確率的変動のこと。ある世代から次の世代への配偶子を介したアレルの伝達は、遺伝子プールとよばれる配偶子の有限集合からのランダムサンプリング（機械的抽出）によるとみなすことができ、抽出過程後の次世代におけるアレル頻度は確率分布に従い変動する。したがって、仮に生存に有利な突然変異アレルが新たに生じたとしても、この偶然性により集団から失われることがある。

カイ二乗検定
２要因以上が存在する場合、要因を行、列の分割表で示し、それぞれのセルの独立性を検定する手法である。帰無仮説はそれぞれのセルが独立であり、対立仮説はセル間の度数に隔たりがあるである。

帰無仮説
統計学では差があるかどうかを知りたいのだが、実際には差がないということを検定する。それを帰無仮説という。帰無仮説が否定できると有意差があるといい、帰無仮説が否定できないと、有意差がないとなる。

偽陽性
あるアレルが疾患と関連するという結果を得たがそれが間違いであることである。あるテストが陽性であると必ず疾患である、が成り立つと偽陽性はない。ありふれた疾患では、ある遺伝子型が必ず疾患と関連するということはなく、統計的な解析を要し、それに伴う偽陽性が生じる。

組換え（Recombination）
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現象。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、相互に異なる染色体に位置することになる。
組換えの起こる確率（組換え率）は、染色体上の距離と関連しており、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。

グラフィカルモデリング（graphical modeling）
条件付き独立性を基本的概念とし、多変量データの関連構造を表す統計モデルを、ネットワークグラフによって表す手法。大規模なデータへの応用や、多重共線性の問題にも対応が可能であり、

今後最も期待されるアプローチの一つである。

欠失変異（deletion variants）
1 kb 以下の塩基配列が欠失するゲノム構造変異。

ゲノム全域アソシエーション・スタディ
国際 HapMap 計画等によりゲノムを網羅する SNP データベースができている。かつ 30−50 万 SNP を同時にタイピングするプラットフォームができている。膨大なタイピングにより、多くの多因子疾患やヒト形質の感受性遺伝子が同定されるようになり、その有効性ゆえに疾患遺伝子解析法の主流となっている。

検出力
仮説検定における帰無仮説が偽であるときに、正しく帰無仮説を棄却する確率のこと。定義より、（検出力）＝1−（第 2 種の過誤の確率）で求められる。

コピー数変異（copy number variation、CNV）
1 キロベースから数メガベースに及ぶゲノム領域が重複もしくは欠失することで、その領域に含まれる遺伝子のコピー数が増加もしくは減少するゲノム構造変異のこと。

糸球体腎炎（glomerulonephritis）
糸球体の炎症が原因で、蛋白尿や血尿といった症状がでる病気の総称が糸球体腎炎である。原発性は腎臓のみに障害がある場合、続発性は膠原病や紫斑病などの病気に伴う場合の 2 グループに大別され、原発性はさらに急性と慢性にわけることができる。

自然選択
生物集団内における突然変異アレル頻度に影響を与える主要因のひとつであり、自然淘汰ともよばれる。自然選択は、大別すると、正の自然選択と負の自然選択とに分けられ、C. Darwin が「種の起原」において進化の原動力として提唱した「Natural selection」は、主として、正の自然選択のことである。新しく生じた突然変異アレルが既存のものよりも相対的に有利である場合には正の自然選択が、相対有害な場合には負の自然選択がそれぞれ働き得る。

疾患感受性遺伝子
本章では、多因子疾患の発症に関わる遺伝子のことを疾患感受性遺伝子とよぶ。

シャルコ・マリー・トゥース病（Charcot-Marie-Tooth disease）
遺伝性末梢神経障害のひとつで、下腿・足をはじめとした四肢遠位筋の萎縮や筋力の低下を主な特徴とする疾患である。欧米では多い病気で、2500 人に一人の割合で罹患し、平均 12 歳で発症する。遺伝形式としては常染色体優性遺伝、常染色体劣性遺伝、X 染色体連鎖優性遺伝、と様々なタイプがあり、孤発例も多くみられる。症状によっていくつかのタイプにわけることができる。CMT1 は節性脱髄が生じるために末梢神経伝導速度が著しく低下する。また、末梢神経の生体検査でタマネギ様バルブがみられるのも特長です。変異している遺伝子領域の違いで

CMT1A:17P11.2-P12 (PMP22)と CMT1B:1q21.2-q23 (MPZ), CMTX1:Xq13.1 (Cx32)がある。それに対して、CMT2 は軸索変性を中心とする神経障害で、神経伝導速度は正常もしくは若干低い程度だがが、複合筋活動電位に低下が見られる。今のところ、有効な治療法はなく、対症療法を処置する。遺伝子検査もおこなわれている。

浸透率
ある遺伝子型をもつ個体が、特定の表現型を呈する確率のこと。例えば、多因子疾患においては、ある遺伝子型をもつ集団のなかで、疾患表現型を呈する個体の割合から浸透率（penetrance）を推定することができる。

第 1 種の過誤
仮説検定における帰無仮説が真であるのにも関わらず、帰無仮説を棄却する誤りのこと。Type I error や偽陽性ともよばれる。第 1 種の過誤が生じる確率は、仮説検定の有意確率に一致する。

第 2 種の過誤
仮説検定における帰無仮説が偽であるにも関わらず、帰無仮説を採択する誤りのこと。Type II error や偽陰性ともよばれる。

多重検定の補正
有意差検定においてたくさんのテストをおこなうとどこかに小さい P 値を得る。当然、テストの回数で補正する必要があろう。

中立突然変異
生物の生存に対して有利にも不利にもならない突然変異のこと。中立突然変異には自然選択が働くことはなく、遺伝的浮動により生物集団内のアレル頻度が変動する。なお、「中立的」突然変異には、厳密に「中立な変異」のみならず、集団内でのふるまいが中立突然変異と同様である「中立に近い変異」も含まれることに留意されたい。

ニューラルネットワーク（**neural network**）
脳内のニューロン（神経細胞）が、他のニューロンから樹状突起を介して信号を受け取り、シナプスを通じて、その信号をまた別のニューロンへ伝えるモデルをコンピューター上で再現し、それによって最適解を得るための手法。多くの場合、多層パーセプトロンを想定し、バックプロパゲーションを用いて解を得るアプローチを指すが、遺伝子間相互作用解析においては、最適化に遺伝的プログラミングを用いる手法も提唱されている。

パーミュテーションテスト（並べ替え検定、**permutation test**）
ノンパラメトリック検定法の一つであり、データセットのランダムな並べ替えと検定統計量の計算を多数繰り返し、得られた検定統計量の分布から、元のデータセットにおける検定統計量の有意性を判定する手法である。
仮説間の独立性を最も適切に反映した有意水準を与える反面、多くの計算量を必要とするため、大規模なデータの解析には適さない。

「外れ値」SNP
ある SNP のもつ遺伝学的特徴が、全 SNP の平均的な特徴から大きく乖離した場合、本章ではその SNP のことを「外れ値」SNP とよぶ。「外れ値」SNP 探しでは、すべての SNP が分子進化的に中立であることを仮定しており、したがって、「外れ値」SNP は中立進化から逸脱した SNP であることが期待される。類似の用語である「外れ値」遺伝子とは、分子レベルでの中立モデルに適合しない遺伝子である可能性が高いものを指し示す。

ハプロ不全（haploinsufficiency）
姉妹染色体の一方に変異が生じて機能不全に陥り、発現するタンパク質量が不足するために優性遺伝すること。

非定型溶血性尿毒症症候群
溶血性貧血や血小板の減少、急性腎不全を伴う症候群で、発症前に下痢を伴う場合を定型、下痢を伴わない場合を非定型とみなす。具体的には、定型—は、腸管出血性大腸菌感染症患者の 1〜10%で発症し、下痢や発熱の症状があった 4〜10 日に症状があらわれてくる。近年、病原性大腸菌 O-157:H7 株の集団感染などで幼小児・高齢者に多く見られるようになった。治療により完治するものの、一部の患者は死亡したり合併症をきたしたり重篤な病態も示す。それに対して、非定型—は、年長から成人に多く見られ、遺伝的背景があることもわかっていて、再発もみられる。

分子進化の中立説
分子レベルでの進化的変化についての観察データを説明するために、木村資生(KIMURA, Motoo)により 1968 年に提唱された「中立突然変異浮動仮説」に基づく学説のこと。生物進化の過程で生じる遺伝子突然変異の大部分は表現型に効果を及ぼさない中立的な変異であり、分子レベルでの進化的変化の多くは、それらの変異が生物種内に偶然的に浮動し、時に固定化することにより起こる、とする説である。この説では、少数ではあるが生存に有利な突然変異が分子レベルでの進化に寄与することを認めている（明らかに有害な変異は遺伝的浮動、あるいは負の自然選択により生物集団から取り除かれる：図 3.2 参照）のに対して、進化集団遺伝学研究における「分子進化の中立モデル」では、すべての突然変異が進化的に中立であることを仮定していることには注意が必要である。

分節重複 （segmental duplication）
1 kb 以上のまとまったゲノム配列が増幅し、その配列間で 90%以上の相同性があること。非相同染色体間で起こるもの（interchromosomal）と、染色体内で起こるもの（intrachromosomal）がある。

ベイズ推定
ベイズの定理に基づく推論の総称のこと。ベイズの定理は、「未知パラメーターの事後分布（あるいは事後確率）が、そのパラメーターの事前分布（あるいは事前確率）と尤度の積に比例する」ということを意味し、確率の公理から導出される。

ヘテロ接合性欠失 (loss of heterozygosity、LOH)
通常、二倍体のゲノムでは、母方及び父方由来の二種類の染色体を持つため、対立遺伝子の配列

は異なる（ヘテロ接合性を有する）。しかし、近親婚のように遺伝的に近い男女の間に生まれた子は、対立遺伝子の配列が全く同一である領域を占める場合がある。この場合をヘテロ接合性欠失と呼ぶ。マイクロサテライトや SNP の遺伝子型を網羅的に調べることで、連続したホモ接合領域として観察される。また、癌細胞ではがん遺伝子またはがん原遺伝子を含む領域で LOH が生じている例が多い。

前向きコホート研究
研究対象とする疾患が発症していない集団を対象として、ある予測因子（疾患感受性 SNP アレル・遺伝子型の保有など）がその疾患の発生にどのような影響を与えるかについて追跡研究を行うこと。一方、患者-対照関連解析は「後ろ向き研究」の一種である。

尤度
一般に、実際の観測データが、未知のある確率分布（ある仮説）に従い生成されたと考えることの「尤もらしさ Likelihood」を測る尺度のこと。従来の統計学では、尤度はパラメーターθの関数として表現され、尤度関数ともよばれる。

連鎖（Linkage）
ある遺伝子座におけるアレルが、別の遺伝子座におけるアレルとともに、一つのハプロタイプとして親から子へ伝達される現象。アレルの伝達は、一般に他の遺伝子座に関係なく決定される（メンデルの独立の法則）が、同一の染色体、しかも互いに近接した遺伝子座の場合、それらは連鎖の関係にあるケースが多い。

連鎖解析（Linkage analysis）
組換え率と染色体上の距離の関係を利用し、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。
連鎖解析の多くは、統計学における尤度比検定を基本としたものであり、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、疾患の遺伝形式をあらかじめ仮定して解析を行なう手法であり、少数の大規模な家系データから、単一遺伝性疾患の原因遺伝子座を探索する場合に有効である。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り、多数の小規模な家系データから、複雑な遺伝形式を示す多因子疾患の原因遺伝子座を探索する場合に有効である。
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。

連鎖不平衡（Linkage disequilibrium）／タグ SNP
同一染色体上で近接する 2 つの以上の変異が、ランダムではなく連関して存在している状態のこと。連関しているアレルの組み合わせのことをハプロタイプとよぶ。複数の SNP が強い連鎖不平衡にある場合、それらの遺伝学的特徴は同一であると見なすことができるため、それらのうちのひとつの SNP を代表として、疾患との関わりを調べればよいことになる。この代表 SNP のことを「タグ SNP（マーカーSNP）」とよび、感受性遺伝子マッピングの初期段階では、タグ SNP タイピングによる効率的な探査がなされる。

<参考文献>

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97-101

Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, et al. (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439: 851-855

Akagawa H, Narita A, Yamada H, Tajima A, Krischek B, Kasuya H, Hori T, Kubota M, Saeki N, Hata A, Mizutani T, Inoue I (2007) Systematic screening of lysyl oxidase-like (LOXL) family genes demonstrates that *LOXL2* is a susceptibility gene to intracranial aneurysms. Hum Genet 121: 377-387

Akagawa H, Tajima A, Sakamoto Y, Krischek B, Yoneyama T, Kasuya H, Onda H, Hori T, Kubota M, Machida T, Saeki N, Hata A, Hashiguchi K, Kimura E, Kim CJ, Yang TK, Lee JY, Kimm K, Inoue I (2006) A haplotype spanning two genes, *ELN* and *LIMK1*, decreases their transcripts and confers susceptibility to intracranial aneurysms. Hum Mol Genet 15: 1722-1734

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2: e286

Ambrosius WT, Lange EM, Langefeld CD (2004) Power for genetic association studies with random allele frequencies and genotype distributions. Am J Hum Genet 74: 683-693

Amos CI, Chen WV, Lee A, Li W, Kern M, Lundsten R, Batliwalla F, Wener M, Remmers E, Kastner DA, Criswell LA, Seldin MF, Gregersen PK (2006) High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. Genes Immun 7: 277-286

Arinami T, Ohtsuki T, Ishiguro H, Ujike H, Tanaka Y, Morita Y, Mineta M, Takeichi M, Yamada S, Imamura A, Ohara K, Shibuya H, Ohara K, Suzuki Y, Muratake T, Kaneko N, Someya T, Inada T, Yoshikawa T, Toyota T, Yamada K, Kojima T, Takahashi S, Osamu O, Shinkai T, Nakamura M, Fukuzako H, Hashiguchi T, Niwa SI, Ueno T, Tachikawa H, Hori T, Asada T, Nanko S, Kunugi H, Hashimoto R, Ozaki N, Iwata N, Harano M, Arai H, Ohnuma T, Kusumi I, Koyama T, Yoneda H, Fukumaki Y, Shibata H, Kaneko S, Higuchi H, Yasui-Furukori N, Numachi Y, Itokawa M, Okazaki Y; Japanese Schizophrenia Sib-Pair Linkage Group (2005) Genomewide high-density SNP linkage analysis of 236 Japanese families supports the existence of schizophrenia susceptibility loci on chromosomes 1p, 14q, and 20p. Am J Hum Genet 77: 937-944

Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X (2003) Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. Hum Mol Genet 12: 2201-2208

Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, Shaffer LG, Jurka J, Morrow BE (2003) Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. Genome Res 13: 2519-2532

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE (2004) Hotspots of mammalian chromosomal evolution. Genome Biol 5: R23

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers E W, Li PW,

Eichler EE (2002) Recent segmental duplications in the human genome. Science 297: 1003-1007

Bailey JA, Liu G, Eichler EE (2003) An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 73: 823-834

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57: 289-300

Benjamini Y, Hochberg Y (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. J Educ Behav Stat 25: 60-83

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29: 1165-1188

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74: 1111-1120

Biswas S, Akey JM (2006) Genomic insights into positive selection. Trends Genet 22: 437-446

Bos CS (2002) A comparison of marginal likelihood computation methods. In: Härdle W, Ronz B (eds) COMPSTAT 2002: Proceedings in Computational Statistics, pp 111-117

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33: 228-237

Breiman L (2001) Random forests. Mach Learn 45: 5-23

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui LC, Scherer SW (2003) Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol 4: R25

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 33: 422-425

Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, Kim UJ, Kuo WL, Olivier M, et al. (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature 409: 953-958

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365-1369

Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS (2004) Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. Diabetologia 47:549-554

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138: 963-971

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35: 2013-2025

Concato J, Feinstein AR, Holford TR (1993) The risk of determining risk with multivariable models. Ann Intern Med 118:201-210

Congdon P (2007) Bayesian statistical modeling (2nd edn). Wiley, Chichester

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38: 1251-1260

Cook NR, Zee RY, Ridker PM (2004) Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. Stat Med 23:1439-1453

Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, Pooley KA, Scollen S, Baynes C, Ponder BA, Chanock S, Lissowska J, Brinton L, Peplonska B, Southey MC, Hopper JL, McCredie MR, Giles GG, Fletcher O, Johnson N, dos Santos Silva I, Gibson L, Bojesen SE, Nordestgaard BG, Axelsson CK, Torres D, Hamann U, Justenhoven C, Brauch H, Chang-Claude J, Kropp S, Risch A, Wang-Gohrke S, Schürmann P, Bogdanova N, Dörk T, Fagerholm R, Aaltonen K, Blomqvist C, Nevanlinna H, Seal S, Renwick A, Stratton MR, Rahman N, Sangrajrang S, Hughes D, Odefrey F, Brennan P, Spurdle AB, Chenevix-Trench G; Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer, Beesley J, Mannermaa A, Hartikainen J, Kataja V, Kosma VM, Couch FJ, Olson JE, Goode EL, Broeks A, Schmidt MK, Hogervorst FB, Van't Veer LJ, Kang D, Yoo KY, Noh DY, Ahn SH, Wedrén S, Hall P, Low YL, Liu J, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Sigurdson AJ, Stredrick DL, Alexander BH, Struewing JP, Pharoah PD, Easton DF; Breast Cancer Association Consortium (2007) A common coding variant in CASP8 is associated with breast cancer risk. Nat Genet 39:688

Culverhouse R, Klein T, Shannon W (2004) Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 27: 141-152

Culverhouse R, Suarez BK, Lin J, Reich T (2002) A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 70: 461-471

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, Palmer SM, et al. (1994) A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371: 130-136

Dhami P, Coffey AJ, Abbs S, Vermeesch JR, Dumanski JP, Woodward KJ, Andrews RM, Langford C, Vetrie D (2005) Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. Am J Hum Genet 76: 750-762

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO (2007) A genome-wide association study of global gene expression. Nat Genet 39: 1202-1207

Eichler EE (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet 17: 661-669

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Pääbo S (2002) Molecular evolution of *FOXP2*, a gene involved in speech and language. Nature 418: 869-872

Evans DM, Cardon LR (2004) Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. Am J Hum Genet 75: 687-692

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide.