遺伝統計学分野における解析技術の基礎調査

平成 19 年 3 月 15 日

情報・システム研究機構

目 次

1. 遺伝統計学とは

ヒトの全ゲノム配列の解読が宣言され，個体間のゲノム配列の違いである多型を調べて医療や創薬などの分野に利用しようとする動きが高まりつつある。個体の遺伝子多型情報を用いて形質（疾患の有無や薬剤の効果・副作用の有無など）に関係している遺伝子を探索する作業は形質マッピングとよばれ，現在世界中で盛んに行われている。形質マッピングを行うための有効な手法の１つに遺伝統計学的手法がある。遺伝統計学的手法は，統計学を用いて特定の形質に関係している遺伝子を探索する手法の総称であり，連鎖解析と連鎖不平衡(Linkage Disequilibrium：LD)を利用した解析とに大別される。前者は家系データを対象とするのに対して，後者は主に集団データを対象とする。また，疾患の有無や薬剤の効果・副作用の有無などの質的形質に加えて，血圧やコレステロール値，身長などの量的形質に関係する遺伝子座位（Quantitative Trait Locus：QTL）を対象として遺伝統計学的解析を行うことを QTL 解析と呼ぶ。

1.1. 連鎖解析

連鎖とは，メンデルの独立の法則の例外である。すなわち，同一染色体上の異なる座位の対立遺伝子（アレル）が非独立に親から子へ伝達される現象である。連鎖の程度は，座位間の組み換え割合に依存する。一般に，異なる 2 つの座位が離れているほど組み換えが起こりやすい。2 つの座位間の組み換え割合が 0.5 より小さいとき，2 つの座位は連鎖していると言う。2 つの座位が連鎖していない場合は，組み換え割合が 0.5 となる（自由組み換え）。連鎖解析では，この連鎖を利用して，染色体上の疾患関連遺伝子座位を推定する。この際，家系情報，各個体の表現型，固定されたマーカー座位における遺伝子型などのデータを観察データとして用いる。連鎖解析には，遺伝形式の仮定を必要とするパラメトリック連鎖解析と遺伝形式の仮定を必要としないノンパラメトリック連鎖解析がある。

1.1.1. パラメトリック連鎖解析

パラメトリック連鎖解析では，遺伝形式（常染色体優性遺伝，常染色体劣性遺伝など）や浸透率，対立遺伝子頻度などを仮定して疾患関連遺伝子座位の推定を行う。浸透率とは，特定の遺伝子型をもつ個体がある表現型をとる確率である。また，「疾患関連遺伝子座位とマーカー座位との間に連鎖がない」という帰無仮説の下で連鎖の有無を検定する。実際の解析では，尤度を用いた推定及び検定が行われる。まず，疾患関連遺伝子座位がマーカー座位と連鎖していると仮定したときに観察データが得られる尤度を組み換え割合 θ の関数として求める（$L(\theta)$）。次に，$L(\theta)$ と自由組み換えのときの尤度（$L(\theta = 0.5)$）との比を考え，この尤度比を最大化する θ を求める。更に，このときの尤度比をもとに検定を行う。この尤度比の常用対数を LOD スコアと呼ぶ。通常，LOD スコアが 3.0 以上となる場合に連鎖があると判断され，このときの組み換え割合から疾患関連遺伝子座位を推定する。

$$\text{LOD スコア} = \log_{10} \frac{L(\theta)}{L(\theta = 0.5)} \qquad (1)$$

パラメトリック連鎖解析には大きな家系が必要であり，遺伝病のような単一遺伝子性疾患の解析に有効である。しかし，不完全浸透や表現模写，遺伝的異質性が存在する多因子疾患の場合には遺伝様式の仮定が困難であり，パラメトリック連鎖解析の適用は妥当ではない。不完全浸透は疾患に関わる遺伝子型を持っていても罹患しないことであり，表現模写は疾患に関わる遺伝子型を持っていないのに環境要因等により罹患することである。遺伝的異質性は複数の遺伝子座位が疾患の発症に関わっていることである。このような多因子疾患の場合には，次に述べるノンパラメトリック連鎖解析を用いる。

1.1.2. ノンパラメトリック連鎖解析

ノンパラメトリック連鎖解析では，遺伝形式や浸透率を仮定せずに，家系内における罹患者間で共有される同祖由来の対立遺伝子数に着目する。家系内の 2 人が共通の祖先に由来する対立遺伝子を共有する状態を IBD(identical by descent)という。(よく似た用語として IBS（identical by state）があるが，これは単に同じ対立遺伝子を持っていることを表す。) ノンパラメトリック連鎖解析では，IBD 対立遺伝子数の分布が「疾患関連遺伝子座位と連鎖がない」と仮定した場合に期待される分布から大きくずれる座位を探索することが基本となる。そのため，解析を行うためには少なくとも 2 人以上の罹患者を含む家系を集めなければならない。

代表的なノンパラメトリック連鎖解析の手法として罹患同胞対 (Affected Sib Pair : ASP) 法（Penrose 1935，Fishman et al. 1978）がある。同じ疾患に罹患した同胞対は疾患関連遺伝子を共有している確率が高いため，多数の罹患同胞対に共有される確率の高い染色体領域に疾患関連遺伝子が存在する，という原理に基づいている。同胞における IBD 対立遺伝子の数は 0，1，2 の場合があり，疾患関連遺伝子座位とマーカー座位との間に連鎖がないときの確率はそれぞれ 0.25，0.5，0.25 である。すなわち，IBD 対立遺伝子の数が i となる確率を z_i とし，$z = (z_0, z_1, z_2)$ とすると，$z = (0.25,\ 0.5,\ 0.25)$ と表される。連鎖がある場合には，z の分布はこの分布からずれる。今，N 家系を対象として，w_{ij} を j 番目の家系で 2 人の同胞が i 個の IBD 対立遺伝子を持つ確率とした場合, 全ての罹患同胞対による観察データの尤度は

$$L(z) = \prod_{j=1}^{N} \sum_{i=0}^{2} z_i w_{ij} \qquad (2)$$

と表される。この尤度を最大にする z を求める。ここで，$z_0 + z_1 + z_2 = 1$ より z_2 は z_0 と z_1 を用いて表すことができ，変数は 2 つとなる。この z_0 と z_1 がとり得る範囲は図 1 において黒く示した部分である。この範囲を Holmans' possible triangle（ホールマンの三角形）と呼ぶ（Holmans 1993）。疾患関連遺伝子座位とマーカー座位との間に連鎖がない（$z = (0.25,\ 0.5,\ 0.25)$）という帰無仮説の下での尤度を分母とし，最尤推定値での尤度を

分子とする尤度比を用いて，連鎖の有無を検定する。

連鎖解析では，パラメトリックであるかノンパラメトリックであるかに関わらず，単点解析と多点解析を行うことができる。単点解析は，疾患関連遺伝子座位と 1 つのマーカー座位の間の連鎖を検討する。多点解析は，疾患関連遺伝子座位と複数のマーカー座位を用いて解析を行う。多点解析を行う際には，用いるマーカー座位の並び順が正確でなければならない。

連鎖解析を行うためには家系情報が必要である。したがって，データが得にくいという問題点がある。また，連鎖解析による陽性範囲は広く，fine マッピングには向かない。

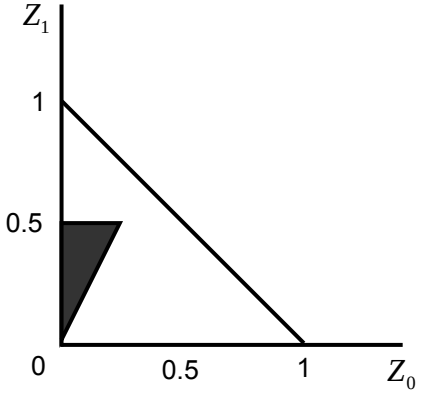

図 1：ホールマンの三角形

1.2. 連鎖不平衡を利用した解析

連鎖不平衡とは，連鎖する 2 つ以上の遺伝子座位における対立遺伝子間に関連が存在することを言う。連鎖不平衡が存在すると，その領域における対立遺伝子の組み合わせ（ハプロタイプ）の頻度が，ハプロタイプを構成する対立遺伝子の頻度の積からずれる。例として，ある 2 つの遺伝子座 X，Y を考える。各遺伝子座位には 2 つの対立遺伝子が存在し，遺伝子座位 X の対立遺伝子は A と a，遺伝子座位 Y の対立遺伝子は B と b であるとする。連鎖平衡にあるとき，ハプロタイプ AB の頻度 P_{AB} は式(3)で表される。

$$P_{AB} = p_A p_B \qquad (3)$$

ここで，p_A，p_B はそれぞれ対立遺伝子 A 及び対立遺伝子 B の頻度を示す。連鎖不平衡が存在する場合は式(3)が成り立たない。連鎖不平衡を利用した代表的な解析として，関連解析，ハプロタイプ解析，伝達不平衡テスト(transmission/disequilibrium test : TDT)（Spielman et al. 1993）が挙げられる。

関連解析は，マーカー座位と疾患関連遺伝子座位の間に存在する関連の有無を検定する手法である。最もよく行われる関連解析は，case-control study である。case-control study では，ある集団から多数の case（患者）と control（対照者）をサンプリングし，着目するマーカー座位における対立遺伝子の頻度を各群で比較する。今，2 つの対立遺伝子 A，a を持つ 1 つの SNP をマーカーとして選択し，この SNP に対して case-control study を行う場

合を考える。case 群の個体数を n，control 群の個体数を m とし，各群における遺伝子型 A/A，A/a，a/a の個体数がそれぞれ表 1 のようになったとする。この時，各群での対立遺伝子 A 及び a の総数は表 2 で示すことができる。この分割表を用いて，「マーカー座位と疾患関連遺伝子座位の間に関連がない」を帰無仮説とした χ^2 検定を行う。case 群における対立遺伝子 A 及び a の個数をそれぞれ n_A， n_a とし，control 群での個数をそれぞれ m_A， m_a とすると，

$$\begin{aligned}\chi^2 &= \frac{(m_A + m_a + n_A + n_a)(m_A n_a - m_a n_A)^2}{(m_A + n_A)(m_a + n_a)(m_A + m_a)(n_A + n_a)} \\ &= \frac{(m+n)(m_A n_a - m_a n_A)^2}{2mn(m_A+n_A)(m_a + n_a)}\end{aligned} \qquad (4)$$

で表される χ^2 統計量は自由度 1 の χ^2 に近似的に従う。

表 1：関連解析における分割表（遺伝子型頻度）

	A/A	A/a	a/a
case 群	a	b	c
control 群	d	e	f

表 2：関連解析における分割表（対立遺伝子頻度）

	A	a
case 群	2a+b	b+2c
control 群	2d+e	e+2f

上記の例では対立遺伝子の頻度に着目したが，対立遺伝子 a に対して優性遺伝であるような疾患を考えた場合には，表 3 に示す分割表を用いて同様に検定を行うことができる。

表 3：関連解析における分割表（優性遺伝）

	AA	Aa+aa
case 群	a	b+c
control 群	d	e+f

関連解析では血縁関係の無い一般集団を対象とするためサンプルを集めやすいが，疾患や作用の発症率・発現率が異なる集団からサンプリングを行った場合，この集団の違いを記述するマーカーを偽陽性として検出してしまう可能性がある。集団間の遺伝的な異質性を集団の構造化と呼ぶ。関連解析を行う際には，予め集団の構造化を解析し，最適な解析

デザインを選択する必要がある。

ハプロタイプ解析は，連鎖解析における多点解析に相当する。単一の遺伝子座位だけでは形質との関連が示されない場合でも，ハプロタイプを用いることにより関連が示される場合もある。しかし，多くの場合において個体のハプロタイプの相を知ることはできず，観察データとして与えられるのは複数の座位における遺伝子型データだけである。現状では，個体のハプロタイプの相を実験的に決めるためには多くの費用と時間が必要となるため，計算機を用いてハプロタイプの相を推定する場合が多い。ハプロタイプ推定の詳細は第 2 章で述べる。

推定されたハプロタイプを用いて関連解析を行う際には，単一マーカーの場合のような分割表による χ^2 検定は適さない。分割表を用いる代わりに，尤度を利用した検定を行う。遺伝子型 G_i に対して可能なディプロタイプ形を D，ディプロタイプ形 D の頻度を $P(D)$ とすると，観察された遺伝子型データ全体の尤度は

$$L = \prod_i \sum_{D \sim G_i} P(D) \qquad (5)$$

となる。case 群のみのデータから求めた最大尤度を L_{case}，control 群のみのデータから求めた最大尤度を L_{control}，両群を合わせたデータから求めた最大尤度を L_{all} とすると，χ^2 統計量は式(6)で表される。

$$\chi^2 = 2\ln\left(\frac{L_{\text{case}} L_{\text{control}}}{L_{\text{all}}}\right) \qquad (6)$$

自由度は，頻度が 0 にならないハプロタイプの数である。サンプルサイズが小さい場合には自由度を正確に推定することが困難なため，並べ替え検定を行う。

TDT は，関連があるとされるマーカー座位と疾患関連遺伝子座位の間に連鎖があるかどうかを検定する手法である。大きな家系を必要とせず，罹患した子 1 人とその両親の遺伝情報を用いて，両親から子に伝達された対立遺伝子の頻度と伝達されなかった対立遺伝子の頻度を比較する。したがって，少なくとも片方の親が対象遺伝子のヘテロ接合でなければならない。親が持つ 2 つの対立遺伝子の一方が子に伝達される確率は 1/2 であるが，疾患と関連するマーカー座位では 1/2 以上の確率で子に伝達されるはずだという考えに基づいて検定を行う。

簡単な例として，図 2 に示す家系について考える。左側の家系図において，四角は男性，丸は女性，黒い四角は罹患した子を表し，その下に 2 種類の対立遺伝子 A と a を持つマーカー座位での遺伝子型を各個体に対して示している。父親の遺伝子型がホモ（A/A）であるため，罹患者である子には A が伝達される。子の遺伝子型はヘテロ（A/a）であり，父親から A が伝達されているのであるから，母親からは a が伝達されたことになる。これらの情報から，子に伝達されなかった対立遺伝子を組み合わせた A/A という遺伝子型を持つ仮想的な内部対照を考えることができる。このように仮想的な内部対照群を生成することにな

るため，集団の構造化による偽陽性をほぼ抑えることができる。

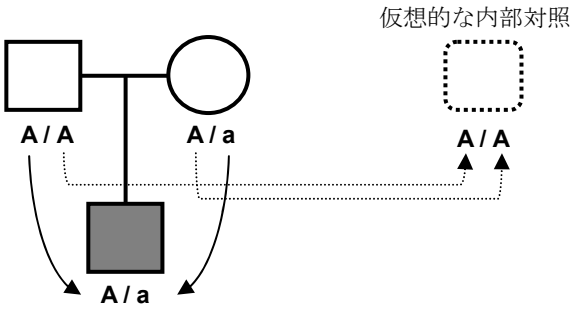

図 2：家系の例

対立遺伝子の頻度を比較する際には，表 4 に示す分割表を考える。図 2 に示す家系では，父親から子に伝達された対立遺伝子，伝達されなかった対立遺伝子ともに A であるから，父親は w の中に数えられる。一方，母親からは a が伝達されて A が伝達されなかったため，y の中に数えられる。n 家系について考えた場合，2n 人の親について同様の作業を行う。得られた分割表において，帰無仮説を「マーカー座位と疾患関連遺伝子座位との間に連鎖がない」として検定を行う。帰無仮説の下では，ある対立遺伝子が伝達された数と伝達されなかった数は等しくなるため，「$w+x=w+y$」及び「$y+z=x+z$」が成り立つことになる。これらは，「$x=y$」が成り立つことに等しい。Mcnemar 検定を用いてこれを検定する。すなわち，

$$\chi_{td}^{\ 2}=\frac{(x-y)^2}{x+y} \qquad (7)$$

が自由度 1 の χ^2 分布に従うことを利用して検定を行う。

表 4：罹患した子に伝達された対立遺伝子と伝達されなかった対立遺伝子

子に伝達された対立遺伝子	子に伝達されなかった対立遺伝子		合計
	A	a	
A	w	x	w + x
a	y	z	y + z
合計	w + y	x + z	2n

このように TDT では罹患者の両親の遺伝情報が必要であるため，遅発性疾患への適用におけるデータの収集が困難である。この問題点に対する方法として，ある疾患に罹患した子と，その疾患に罹患していない兄弟の遺伝情報を用いて連鎖と連鎖不平衡による関連を検定する同胞対 TDT（sibpair-TDT：S-TDT）（Spielman and Ewens 1998）がある。

家系を扱う TDT は，通常の case-control study に比べてサンプルの収集が困難であること，また，ゲノムに共通部分が多く検出力が低下することが弱点となる。

連鎖不平衡を利用した解析では陽性領域が狭いため，候補領域を絞った上での fine マッピングに適する。ただし，集団の構造化や多重比較による偽陽性の増加が問題となる。

疾患関連遺伝子座位の推定において，連鎖解析では比較的広範囲の候補領域を抽出し，関連解析ではより細かく領域を絞り込む。

最後に，連鎖解析及び連鎖不平衡を利用した解析に関して，特徴を表 5 にまとめた。

表 5：連鎖解析と連鎖不平衡を利用した解析の比較

	対象データ	対立遺伝子異質性に対する頑健性	集団の構造化による影響	陽性領域	サンプル収集
パラメトリック連鎖解析	大家系	高い	小さい	広い	困難
ノンパラメトリック連鎖解析	多数の小家系	高い	小さい	広い	困難
関連解析	患者と対照者	低い	大きい	狭い	易しい
TDT	多数の小家系	低い	小さい	狭い	中程度

1.3. QTL 解析

ヒトの ABO 式血液型や罹患の有無のように，不連続で質的な違いとして示される形質を「質的形質」と言う。これに対して，血圧やコレステロール値，身長のように，連続した実数あるいは整数で示される形質を「量的形質」と言い，量的形質の原因となる遺伝子座位を QTL（Quantitative Trait Locus）と呼ぶ。量的形質は 1 つの遺伝子で規定されているのではなく，複数の遺伝子の作用によって生じると考えられている。QTL に対して連鎖解析などの遺伝統計学的解析を行うことを QTL 解析と言う。

人為的な純系個体の作成や多くの子孫を得ることが可能な実験動物や農作物と異なり，ヒトの場合は，対立遺伝子がどちらの親由来であるかの決定や環境異質性の除去ができない。また，1 人の親における子の数も限られているため，実験動物などに対する手法とは異なる方法を用いて QTL 解析を行わなければならない。以下では，ヒトを対象とした QTL 解析について述べることとする。

QTL 解析を行うための手法はいくつかあるが，いずれも次に示す原理を基本としている。

i. 類似した形質の値を持つ個体同士は，偶然によって期待されるよりも対象とする形質座位における対立遺伝子を共有しやすい
ii. 組み換えが起こる場合を除いて，染色体はおおよそ親から子へそのまま伝えられ

るため，類似した形質の値を持つ個体同士は，対象とする形質座位で共有する対立遺伝子及び，その周辺座位で共有する対立遺伝子の数が多くなる

iii. 組み換えはあらゆる世代で起こるため，対象とする形質座位の周辺座位では，対立遺伝子の共有されやすさは世代を経るごとに減少する

共有する対立遺伝子数が増加しているマーカー座位があれば，その近くに QTL が存在している，すなわち QTL とマーカー座位が連鎖している可能性が高い。QTL 解析では，家系内で共有する IBD 対立遺伝子の数に着目して QTL とマーカー座位との間における連鎖の検定や QTL の同定を行う。

QTL 解析を行う際には，ある特定の形質モデルを用いる。最も簡単でよく使われるモデルは主要な遺伝子を 1 つ用いたモデルである。形質の値 X は，

$$X = \mu + g + e \qquad (8)$$

と表すことができる。ただし，μ は集団全体の平均値，g はその座位における遺伝子型に起因する平均値からのずれ，e は環境変動を表す。環境変動は通常，平均 0，分散 σ_e^2 の正規分布に従うと仮定される。遺伝子型と環境との間に関連がなければ，全体の形質分散は，遺伝的分散 σ_g^2 を用いて

$$\sigma^2 = \sigma_g^2 + \sigma_e^2 \qquad (9)$$

と分解することができる。遺伝的分散はさらに相加的効果による成分（相加的遺伝分散：σ_a^2）と優性効果による成分（優性遺伝分散：σ_d^2）に分けることができる。相加的効果とは，その座位での対立遺伝子による線形の効果である。優性効果は同じ遺伝子座の対立遺伝子間の相互作用を表す。式(8)に示すモデルには，この他に共分散やその他の主要遺伝子，遺伝子と環境の相互作用などがよく組み込まれる。

QTL 解析を行うための手法の多くは，環境変動に加えて集団全体における形質の分布も正規分布に近似できると仮定している。このような分布の仮定に対する頑健性は，用いる手法によって異なる。代表的な手法である Haseman-Elston 法（Haseman ane Elston 1972）と variance components 法（Amos 1994）については，次章で説明を行う。

2. アルゴリズム

2.1. 連鎖解析

連鎖解析における尤度計算を行うためのアプローチは，Elston-Stewart アルゴリズム（Elston and Stewart 1971），Lander-Green アルゴリズム（Lander and Green 1987），Markov chain Monte Carlo（MCMC）のいずれを利用しているかによって，大きく 3 つに分けることができる。各アルゴリズムに関して，解析に用いる家系内の個体数，マーカー数，欠損データの増加に対する影響を計算量という観点から表 6 にまとめた。

表 6：連鎖解析を行うアルゴリズムの比較

アルゴリズム	各指標の増加に伴う計算量の増加		欠損データによる影響の受けやすさ
	家系内個体数	マーカー数	
Elston-Stewart	線形	指数	大
Lander-Green	指数	線形	中
Markov chain Monte Carlo	線形	線形	小

2.1.1. Elston-Stewart アルゴリズム

Elston-Stewart アルゴリズムでは，再帰的な方法を用いてパラメトリック連鎖解析における尤度計算を行う。

今，i 番目の個体が n 個の座位についての遺伝子型 $g_i = (g_i^1, g_i^2, \cdots, g_i^n)$ をもち，それに対応する表現型が x_i であるとする。このとき，m 人から成る家系全体の尤度 L は，

$$\begin{aligned} L = P(x) &= \sum_g P(x, g) = \sum_g P(x \mid g) P(g) \\ &= \sum_{g_1} \sum_{g_2} \cdots \sum_{g_m} \prod_i P(x_i \mid g_i) P(g_i \mid \cdot) \end{aligned} \qquad (10)$$

として求めることができる。ここで，$x = (x_1, x_2, \cdots, x_m)$，$g = (g_1, g_2, \cdots, g_m)$ とする。また，$P(g_i \mid \cdot)$ は両親の遺伝子型が与えられたときに i 番目の子が遺伝子型 g_i をもつ確率を表す。ただし，創始者の場合には単に遺伝子型 g_i をもつ確率を表す。式(10)を用いて尤度を算出する場合，各個体において可能な遺伝子型の組み合わせ全てについての計算が必要となる。Elston-Stewart アルゴリズムでは，この尤度を以下のように計算することで効率化を図っている。

$$L = \sum P(x_1 \mid g_1) P(g_1 \mid \cdot) \cdots \sum P(x_{m-1} \mid g_{m-1}) P(g_{m-1} \mid \cdot) \sum P(x_m \mid g_m) P(g_m \mid \cdot) \qquad (11)$$

これは，親の遺伝子型が与えられた下での子の遺伝子型は，兄弟間で独立に分離することを基本としている。家系内の各個体の番号は，祖先側から順番に付けなければならない。家系の末端の個体における観察データから順に尤度計算を行い，家系の上方向に畳み込んでいく。これにより，家系内でより上にいる個体の尤度計算の際に考慮すべき遺伝子型の数を絞り込むことができる。Elston-Stewart アルゴリズムの概念図を図 3 に示す。点線で囲んだ核家族部分の尤度計算を子孫側から順に行う。それぞれの核家族部分において，黒く塗りつぶした個体は祖先側と繋がっており，核家族部分の尤度をこの個体の確率に畳み込む。このアルゴリズムは，ループがなく，両親が子に先行するなど個体に順位があるような単純な家系に対して非常に有力である。現在では，ループを含む家系へと拡張が行われている。

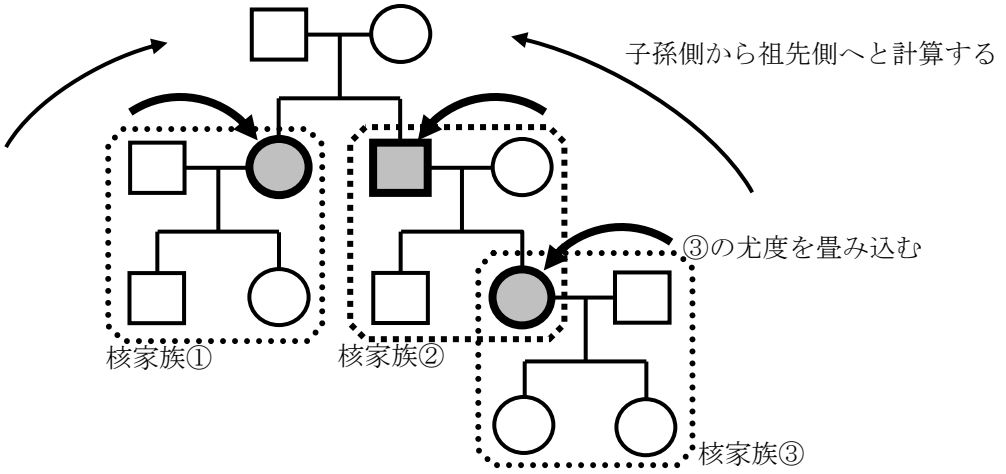

図 3 : Elston-Stewart アルゴリズムの概念図

家系に含まれる個体数の増加に対する計算量の増加は線形であるが，解析に用いるマーカー座位の増加に対しては指数的に計算量が増加する。従って，少数の遺伝子座位に関する大家系データの解析に向いたアルゴリズムである。

Elston-Stewart アルゴリズムを用いたソフトウェアには，LINKAGE package（Lathrop et al. 1984，Lathrop et al. 1985）やその拡張版である FASTLINK（Cottingham et al. 1993, Schaffer et al. 1994），Vitesse（O'Connell and Weeks 1995）などがある。

2.1.2. Lander-Green アルゴリズム

Lander-Green アルゴリズムでは，継承ベクトルという概念を用いて，隠れマルコフモデルにより家系データの尤度を求める。Lander-Green アルゴリズムは，ノンパラメトリック連鎖解析にも適用することができる。

継承ベクトルは家系内の各非創始者の各座位について，父親由来の対立遺伝子と母親由来の対立遺伝子それぞれが祖父由来であれば 0, 祖母由来であれば 1 とするビット列を並べて構成される。非創始者の数を n とすると，継承ベクトルは $2n$ 次元のベクトルとなる。例として，図 4 に示す家系の継承ベクトルを考える。四角は男性，丸は女性を示し，各番号は個体番号とする。各四角と丸の中に示す点は，左側が父親由来の対立遺伝子，右側が母親由来の対立遺伝子を表す。この家系における非創始者は個体番号 4，5，7，8，9，10 の 6 人である。個体番号 4 の個体について考えると，父親由来の対立遺伝子は父方の祖父から継承されており，母親由来の対立遺伝子は母方の祖父から継承されている。したがって，継承ベクトルの個体番号 4 の個体に関する部分は $(1, 1)$ となる。他の非創始者に関しても同様に考えると，個体番号の小さい順に要素を並べた最終的な継承ベクトル v は，$v = (1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0)$ となる。継承ベクトルは，座位毎に定めることができる。ある 2 つの座位間に関して，家系内の全ての継承で組み換えがなければ両座位での継

承ベクトルは一致する。特定の行における継承ベクトルの因子が 0 → 1，または 1 → 0 と変化している場合は，その行に対応する継承において組み換えが起きたことを示す。隣り合う 2 つの座位間の継承ベクトルの関係は，その座位間の組み換え割合に依存する。

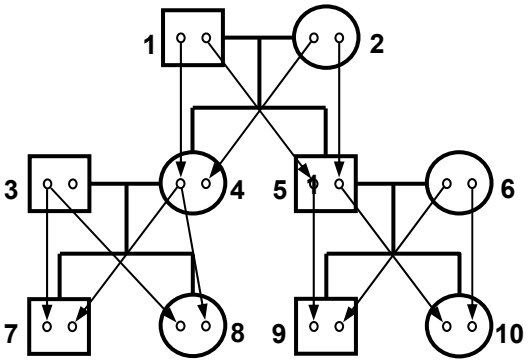

図 4：家系の例

Lander-Green アルゴリズムでは，解析に用いる全てのマーカー座位における継承ベクトルの分布と創始者についてのディプロタイプ形（ハプロタイプの組み合わせ）を用いて，対象とする家系の対立遺伝子の流れを記述する。そして座位間の継承ベクトルの遷移を，組み換え割合の関数である遷移行列に基づくマルコフ過程として考え，隠れマルコフモデルに適用させる。隠れマルコフモデルにおける時刻 t が各マーカー座位に対応し，時刻 t における状態がその座位の継承ベクトルに対応する。記号の出力確率は，各マーカー座位での表現型の浸透度に対応する。また，観察結果は各マーカー座位に与えられた遺伝子型の列である。このような仮定のもと，継承ベクトルを隠れ層とする隠れマルコフモデルを用いて全ての座位の継承ベクトルの事後分布を計算する。

解析に用いるマーカー座位の増加に対する計算量の増加は線形であるが，家系に含まれる個体数の増加に対する計算量の増加は指数的である。これは，Elston-Stewart アルゴリズムと全く反対の特徴となっている。

Lander-Green アルゴリズムを用いた代表的なソフトウェアには，GeneHunter（Kruglyak et al. 1996），Merlin（Abecasis et al. 2002）などがある。

2.1.3. Markov chain Monte Carlo（MCMC）

表 6 に示したように，Elston-Stewart アルゴリズムは家系内の個体数の増加に対する計算量の増加は線形であるが，解析に用いるマーカー座位の増加に対しては指数的に計算量が増加する。また，Lander-Green アルゴリズムは解析に用いるマーカー座位の増加に対する計算量の増加は線形であるが，家系内の個体数の増加に対する計算量の増加は指数的である。したがって，個体数の多い大規模な家系における多数のマーカー座位のデータを解析する場合，上記の 2 つのアルゴリズムで exact な尤度を計算すると，尤度計算時に考慮する（観察データに合致する）場合の数が膨大となり，計算時間が膨大になる。この傾向は欠

損データが多い場合により顕著となる。このような場合は，exact に尤度計算を行なうのではなく，MCMC を用いて尤度を推定する方法が用いられる。

MCMC による家系尤度計算においては，マーカー座位における遺伝子型が与えられたもとでの，家系内における gene flow（すなわち対立遺伝子の伝達）がサンプリングされる。この gene flow は，図 5 に示すような descent graph で表される。descent graph は，マーカー座位数を l，家系内の個体数を n とすると $2ln$ 個のノードからなる。各ノードにはソース（図 5 における矢印の起源）が存在し，ソースをたどることによりその対立遺伝子が父親由来か母親由来かが特定できる。descent graph は Lander-Green アルゴリズムで用いられる継承ベクトルと同様のものである。MCMC によるサンプリングにおいて，各 descent graph がマルコフ連鎖の各状態に相当する。

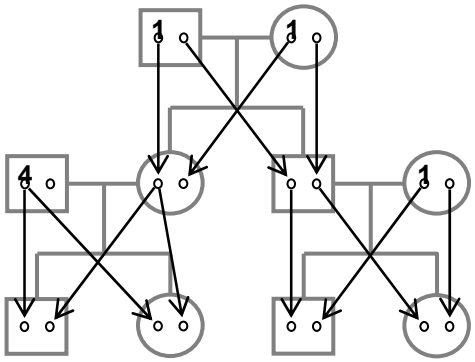

図 5：descent graph の例（Sobel E. & Lange K. 1996 より改変）

Tompson（Tompson 1994a, 1994b）は，MCMC を用いて descent graph をサンプリングするために，descent graph 間を遷移するためのルールを定式化した。最も基本的な遷移ルールは遷移ルール T_0 として定義されるものである。遷移ルール T_0 は，descent graph のあるノードのソースを，父親由来から母親由来へ，または母親由来から父親由来へとスイッチする。図 6 に遷移ルール T_0 の例を示す。図 6 の遷移では，図中に黒丸で示すノードのソースがスイッチされていることがわかる。

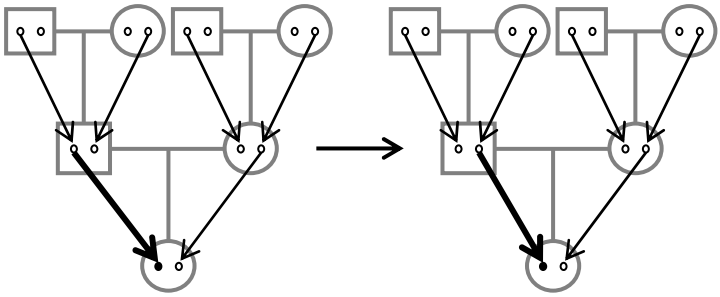

図 6：遷移ルール T_0 の例（Sobel E. & Lange K. 1996 より改変）

そのほかにも複合的な遷移ルールとして遷移ルール T_1 および遷移ルール T_2 が定義されている。遷移ルール T_1 は，家系内のある個体 i のあるマーカー座位 l について，個体 i から子

へと伝わる対立遺伝子を，父親由来であれば母親由来に，母親由来であれば父親由来にスイッチする。図 7 に遷移ルール T_1 の例を示す。図 7 の遷移では，図中に黒丸で示す個体から子へ伝達される対立遺伝子のソースがスイッチされていることがわかる。

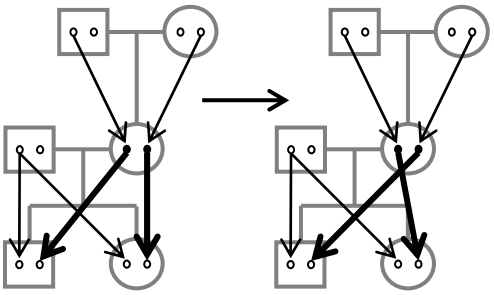

図 7：遷移ルール T_1 の例（Sobel E. & Lange K. 1996 より改変）

遷移ルール T_2 は，遷移ルール T_{2a} および遷移ルール T_{2b} からなる。図 8 に遷移ルール T_2 の例を示す。遷移ルール T_2 では，あるマーカー座位 l について，家系内のある個体 i および個体 j からなる夫婦を選択し，夫婦がもつ 4 つのノードをスイッチすることによって，4 つのノードを基点とする subtree（すなわち descent graph の一部分）をスイッチする。個体 i と個体 j の父親由来の対立遺伝子を示すノード同士，及び母親由来の対立遺伝子を示すノード同士をスイッチする場合が遷移ルール T_{2a}，個体 i の父親由来の対立遺伝子を示すノードと個体 j の母親由来の対立遺伝子を示すノード，及び個体 i の母親由来の対立遺伝子を示すノードと個体 j の父親由来の対立遺伝子を示すノードをスイッチする場合が遷移ルール T_{2b} である。

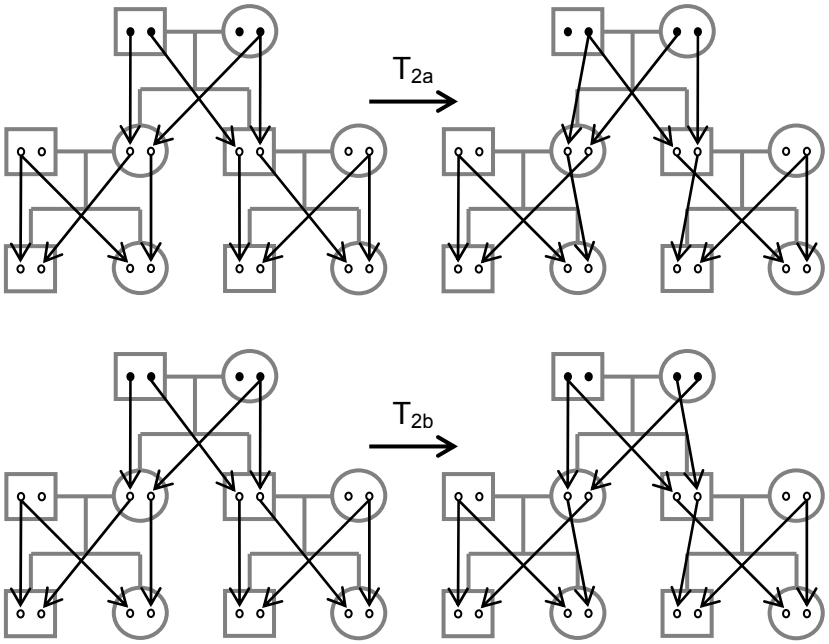

図 8：遷移ルール T_2 の例（Sobel E. & Lange K. 1996 より改変）

MCMC による家系尤度計算においては，これらの遷移ルールを家系内の各個体および各マーカー座位において適用し，マーカー座位における遺伝子型の観察データ M のもとでの descent graph $\hat{G}$ の条件付分布 $\Pr(\hat{G}|M)$をサンプリングして尤度の推定を行なう。

2.2. ハプロタイプ解析

ここでは，ハプロタイプの頻度及び相の推定を行うアルゴリズムについて述べる。

2.2.1. Clark アルゴリズム

Clark アルゴリズム（Clark 1990）では，ハプロタイプの頻度推定は行わず，相の決定を行う。すでに決定されたハプロタイプを用いることによる曖昧な遺伝子型の除去が基本となる。

まず，個体における複数座位の遺伝子型情報について，ホモ接合の対立遺伝子のみを持つ，あるいはヘテロ接合の対立遺伝子を 1 つだけ含む個体を探索する。該当する個体の場合，ハプロタイプの相は唯一に決定される。相が決定されたハプロタイプをまとめて，決定済みハプロタイプの集合とする。次に，曖昧性を含む残りの個体に対して決定済みハプロタイプの集合を調査し，当てはまり得るハプロタイプを探索する。そのようなハプロタイプが発見されれば，対を成すハプロタイプを決定済みハプロタイプの集合に加え，その個体のハプロタイプの相は決定されたものとする。曖昧性を含む個体がなくなるまで，あるいは新しいハプロタイプが発見されなくなるまで以上の操作を続ける。

Clark アルゴリズムは単純で直感的に理解しやすいが，以下に示すような欠点がある。

i. 曖昧性を含まない個体が 1 つ以上必要である
ii. 全ての個体に対してハプロタイプの相を決定できるとは限らない
iii. 調査する決定済みハプロタイプの順番によって最終結果が異なる

Clark アルゴリズムの利用により，非常に高速なハプロタイプの相の決定を行うことができる。しかし，対象とする集団は連鎖不平衡により比較的少数の共通ハプロタイプを共有すると想定しているため，集団がこの条件を満たさない場合には性能が悪くなる。1990 年代初期にはよく用いられていたが，今日ではあまり利用されていない。

Clark アルゴリズムを利用したソフトウェアとして HAPINF（Clark 1990）がある。

2.2.2. EM アルゴリズム

Excoffier と Slatkin によって最初にハプロタイプ頻度推定に適用された手法で，観察データが得られる尤度を最大にするような集団のハプロタイプ頻度を推定する（Excoffier and Slatkin 1995）。集団のハプロタイプ頻度を推定することができれば，その推定値を用いて各遺伝子型に対応するディプロタイプ形の分布を求めることができる。以下では，EM アルゴリズムを用いたハプロタイプ頻度推定の一般的な手順について説明する。

今，n 個の個体からなる集団について，$G=(G_1,\cdots,G_n)$ を各個体の遺伝子型であるとす

る。G に含まれる遺伝子型のうち相異なる遺伝子型を g_i $(i=1,\cdots,n')$ とし，遺伝子型 g_i をとる個体の数を x_i とする。また，集団内で考え得るハプロタイプを $H=(h_1,\cdots,h_m)$，その頻度を $F=(f_1,\cdots,f_m)$ とおく。ハプロタイプが M 個の SNP 座位からなるとき，m の最大値は 2^M である。遺伝子型 g_i を与える相異なるディプロタイプ形 D_{ij} $(j=1,\cdots,m_i)$ が 2 つのハプロタイプ h_k，h_l から成るとき，その確率は

$$\Pr(D_{ij})=\begin{cases} {f_k}^2 & (k=l) \\ 2f_k f_l & (k\neq l) \end{cases} \qquad (12)$$

となる。集団のハプロタイプ頻度 F を直接知ることはできないため，最尤法を用いた推定を行う。ここで，ハプロタイプ頻度全体の尤度は

$$L(F)=\Pr(G|F)\approx\prod_{i=1}^{n'}\Pr(g_i)^{x_i}=\prod_{i=1}^{n'}\left(\sum_{j=1}^{m_i}\Pr(D_{ij})\right)^{x_i} \qquad (13)$$

と表すことができる。尤度を最大にする集団のハプロタイプ頻度を算出すれば，その頻度から各遺伝子型に対応するディプロタイプ形の分布を得ることができる。

最尤ハプロタイプ頻度の推定に，EM アルゴリズムを用いる。まず，対象となるハプロタイプ頻度 F に任意の初期値を与え，$F^{(0)}=({f_1}^{(0)},\cdots,{f_m}^{(0)})$ とする。その後，以下の期待値算出ステップ（E ステップ）と最大化ステップ（M ステップ）を繰り返す。t 番目の反復操作におけるハプロタイプ頻度を $F^{(t)}=({f_1}^{(t)},\cdots,{f_m}^{(t)})$ と表す。E ステップでは，ハプロタイプ h_k，h_l から成るディプロタイプ形に対して，

$$\Pr(D_{ij})^{(t)}=\begin{cases} \left({f_k}^{(t)}\right)^2 & (k=l) \\ 2{f_k}^{(t)}{f_l}^{(t)} & (k\neq l) \end{cases} \qquad (14)$$

$$\Pr(D_{ij}|g_i)^{(t)}=\frac{\Pr(D_{ij})^{(t)}}{\sum_{s=1}^{m_i}\Pr(D_{is})^{(t)}} \qquad (15)$$

に従って，遺伝子型が g_i である個体について期待されるディプロタイプ形の分布を算出する。M ステップでは，E ステップで算出されたディプロタイプ分布を基に個体のディプロタイプ中の各ハプロタイプを数え上げる。

$$f_u^{(t+1)}=\frac{1}{2n}\sum_{i=1}^{n'}\left(x_i\sum_{j=1}^{m_i}\delta_{iju}\Pr(D_{ij}|g_i)^{(t)}\right) \qquad (16)$$

ただし，δ_{iju} はディプロタイプ D_{ij} に含まれるハプロタイプ u の数（0，1，2 のいずれか）である。このように得られた集団のハプロタイプ頻度を新たに $F^{(t+1)}=({f_1}^{(t+1)},\cdots,{f_m}^{(t+1)})$ とおく。以上の操作を，ハプロタイプ頻度の変化が収束するまで繰り返す。収束した時点

でのハプロタイプ頻度を母集団におけるハプロタイプ頻度の最尤推定値とする。

EM アルゴリズムの欠点として，SNP 中のヘテロ接合の数の増加に伴って考え得るハプロタイプ数が指数的に増加すること，局所解が存在するため初期値の設定によって異なる解が得られる可能性があることが挙げられる。また，連鎖不平衡の程度が低い領域では性能が低下する。

EM アルゴリズムを利用したソフトウェアである PL-EM は，計算量増加の問題に対応するために Partition-Ligation（PL）と呼ばれる方法（Niu et al. 2002）を用いている。PL 法では，ハプロタイプを構成する多数のマーカーを隣接する少数のマーカーから成るグループに分割し（Partition），グループ毎に EM アルゴリズムを用いたハプロタイプ推定を行う。推定したハプロタイプを部分ハプロタイプとする。そして，隣接する 2 つのグループで推定された部分ハプロタイプを, EM アルゴリズムを用いて 2 つの連鎖する対立遺伝子の相を決定するように結合していく（Ligation）。Ligation でも EM アルゴリズムを用いる PL-EM では，局所解の問題を避けることができない。この問題に対応するため，特定の閾値以上の頻度を持つ部分ハプロタイプのみで Ligation を行うのではなく，頻度が閾値未満となる部分ハプロタイプも一定の容量を超えない範囲で全体の結果の候補に加える。PL 法において，ハプロタイプを結合していく過程の概念図を図 9 に示す。ただし，最初に行われる分割の後の各グループのマーカー数を M ，i 番目の Ligation 後における各グループのマーカー数を L_i とする。

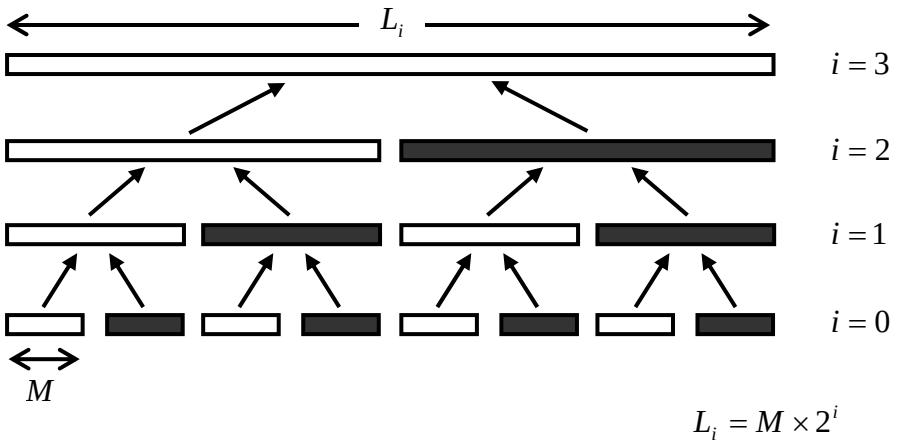

図 9 : PL 法の概念図（Qin Z.S. et al. 2002 より改変）

2.2.3. PHASE アルゴリズム

Stephens らによって最初に提案されたアルゴリズムで，ベイズ法を用いて遺伝子型データが与えられた下での個体のディプロタイプ形の事後分布を推定する（Stephens et al. 2001）。

ベイズ法によるアプローチでは，未知のハプロタイプを任意の値として扱い，事前分布

と尤度を結合することで事後分布を予測する。事前分布は集団サンプルに期待されるハプロタイプパターンであり，尤度は観察データの情報である。ハプロタイプ自身は，得られた事後分布を基に推定する。事後分布を厳密に求めることができないため，近似するための計算手法が用いられる。計算手法には，一般的にマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo：MCMC）が採用される。

事前分布には, これまでに Dirichlet 分布及びコアレセントモデルの近似が適用されてきた。Dirichlet 分布は, 簡単だが集団の進化による遺伝的過程に対しては非現実的な仮定だと言える。一方，コアレセントモデルは，より複雑であるが現実的な仮定に基づいている。両者の違いは，遺伝子型からは一意に決定できない個体のディプロタイプ形を推定する際に生じる。例として，図 10 に示すようなディプロタイプ形未知の個体 A，B を考える。個体 A のように既知のハプロタイプに分解できる場合には，どちらの事前分布を用いても既知のハプロタイプに一致するものが選択されやすいように重み付けされる。個体 B の場合は，既知のハプロタイプと一致するハプロタイプに分解することができない。この場合，コアレセントモデルに基づいた事前分布を用いるアプローチでは，一致しないものの類似するハプロタイプが選択されやすいように重み付けされる。従って，個体 B に対して「33434 / 11444」というディプロタイプ形が選択されやすくなる。一方，Dirichlet 分布を事前分布に用いるアプローチでは，類似するハプロタイプに対して特別な重み付けをしない。実際のデータの特徴をよりうまく捉えている事前分布を用いた場合に，事後分布の推定が良くなる。コアレセントモデルは遺伝的な変異をより忠実にモデルに取り込んでいるため推定精度の向上が期待されるが，シミュレーションの計算量が膨大になること，また，収束判定基準の決め方が難しいことが課題である。

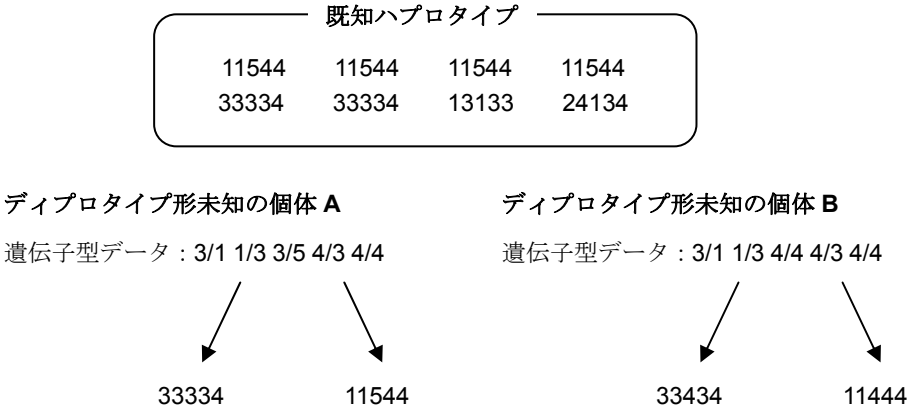

図 10 : 類似ハプロタイプの選択（Stephens M. et al. 2002 より改変）

MCMC を用いて事後分布を推定する際のマルコフ過程は，$D^{(t)}$ を t 番目のサンプリングにおける全ディプロタイプ形の集合とすると，$D^{(0)}, D^{(1)}, \cdots, D^{(n)}$ と表すことができる。遺

伝子型データ $G=(g_1,\cdots,g_n)$（g_i は個体 i に関してハプロタイプを構成する座位の遺伝子型情報の組み合わせを表す）が与えられたときの推定手順を説明する。ここで，$D_{-i}^{(t)}$ を $D^{(t)}$ から g_i に関するディプロタイプ形を除いた集合，$D_i^{(t)}$ を $D^{(t)}$ に含まれる g_i に関するディプロタイプ形のみからなる集合であるとする。MCMC の一種であるギブスサンプリングによる以下のステップを繰り返し，事後分布からの近似解を求める。

i. 全ての個体に対するディプロタイプ形の初期値 $D^{(0)}$ を設定する
ii. 曖昧性を含む遺伝子型 g_i を持つ個体 i を任意に選択する
iii. $D_{-i}^{(t)}$ が正しいと仮定して，条件付分布 $\Pr\left(D_i\middle|G,D_{-i}^{(t)}\right)$ から $D_i^{(t+1)}$ をサンプリングする
iv. $D_j^{(t+1)}=D_j^{(t)}\quad(j=1,\cdots,n:\ j\neq i)$ とする

ステップiiiにおける条件付分布は，多くの場合未知である。そのため，g_i に一致するディプロタイプ形 $D_i=(h_{i1},h_{i2})$ に対して式(17)で示す書き換えを行う。

$$\begin{aligned}\Pr\left(D_i\middle|G,D_{-i}\right)&\propto\Pr\left(D_i\middle|D_{-i}\right)\\&\propto\pi\left(h_{i1}\middle|D_{-i}\right)\pi\left(h_{i2}\middle|D_{-i},h_{i1}\right)\end{aligned}\qquad(17)$$

ここで $\pi(\cdot|D)$ は，前もってサンプリングされたディプロタイプ形の集合 D が与えられたという条件の下でサンプリングされるハプロタイプの分布を示す。この条件付分布も一般に未知である。そこで，未知の $\pi(\cdot|D)$ に対して，式(18)のような近似を行う。

$$\pi\left(h|D\right)=\sum_{\alpha\in E}\sum_{s=0}^{\infty}\frac{r_\alpha}{r}\left(\frac{\theta}{r+\theta}\right)^s\frac{r}{r+\theta}\left(P^s\right)_{\alpha h}\qquad(18)$$

E は一般的な突然変異モデルのためのハプロタイプの集合である。P は突然変異行列であり，ハプロタイプ α が次にサンプリングされるハプロタイプ h に変わる確率を表す。r_α はディプロタイプ形の集合 D に含まれるハプロタイプ α の数，r は D に含まれるハプロタイプの総数である。また，θ は scaled mutation rate である。この式を式(17)に代入することによって，ステップiiiを実行する。

ベイズ法を利用した最初のソフトウェアは PHASE（Stephens et al. 2001）である。その後，PL を取り入れたソフトウェアとして HAPLOTYPER（Niu et al. 2002）が開発された。PHASE はコアレセントモデルを事前分布に用いており，HAPLOTYPER は Dirichlet 分布を用いている。

なお，任意のベクトル $\mathbf{X}=\left(X_1,\ldots,\ X_n\right)$ の密度が以下の式で表せるとき，ベクトル $\mathbf{X}$ は Dirichlet 分布 $\left(\beta_1,\ \ldots\ ,\ \beta_n\right)$ に従うと言う。ただし，$\sum x_i=1,\ x_i\geq 0$ である。

$$f\left(x_1,\ldots,\ x_n\right)=\frac{\Gamma\left(\beta_1+\cdots+\beta_n\right)}{\Gamma\left(\beta_1\right)\times\cdots\Gamma\left(\beta_n\right)}x_1^{\beta_1-1}\times\cdots\times x_n^{\beta_n-1}\qquad(19)$$

2.3. QTL 解析

2.3.1. Haseman-Elston 法

Haseman-Elston 法はヒトの QTL 解析のために開発された最初の統計学的手法で，同胞対で共有される IBD 対立遺伝子の数を利用する。この手法における従属変数は 2 人の同胞の間における量的形質値の差の二乗であり，図 11 に示すように，共有する IBD 対立遺伝子の数（独立変数）が多くなるほどその値は減少すると予想される。

i 番目の同胞対が持つそれぞれの形質値 X_{i1}， X_{i2} を用いて，量的形質の回帰モデルを

$$E\left(Y_i \middle| \pi_{im}\right) = \left(X_{i1} - X_{i2}\right)^2 = \alpha + \beta\pi_{im} + e \qquad (20)$$

と表す。π_{im} は，i 番目の同胞対がマーカー座位 m で共有する IBD 対立遺伝子の数の推定割合である。この回帰モデルを基に，連鎖がないという帰無仮説に対して Student の t 検定を行う。帰無仮説の下では，回帰係数（傾き） β は 0 となり，連鎖がある場合には， β は負の値をとる。優性遺伝分散を考えないときには，式(20)は分散を用いて以下のように書き換えることができる。

$$E\left(Y_i \middle| \pi_{im}\right) = \left(\sigma_e^2 + 2\sigma_g^2\right) - 2\sigma_g^2\pi_{im} \qquad (21)$$

このとき，帰無仮説の下では $\sigma_g^2 = 0$ となる。

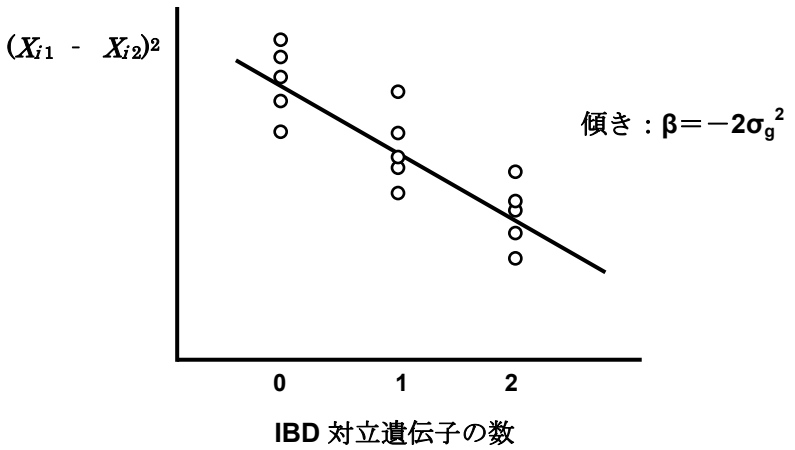

図 11：Haseman-Elston モデルの概念図

Haseman-Elston 法のような回帰モデルでは，従属変数と独立変数との間に線形の関係があると仮定している。形質座位に優性効果がない場合にはこの仮説が成り立つが，優性効果が存在する場合には成り立たない。現在では，この仮説を緩めたノンパラメトリックな手法も提案されている。

回帰モデルは，計算が簡単で効率が良く，リサンプリング手法を用いたパラメータの標準偏差が簡単に推定できるため，広く使われている。しかし，収集されたデータが同胞対

のような血縁者のペアから得られたものでない場合には，QTL 解析への適用は好ましくない。

2.3.2. variance components 法

ヒトの QTL 解析に初めて variance components 法を用いたのは Goldgar（1990）だが，現在よく使われている一般的な枠組みと方法論は Amos（1994）によるものである。基本的な variance components 法では，家系メンバーの形質値は多変量正規分布に従うと仮定する。また，各個体の形質値は主となる遺伝子の効果，ランダムポリジーンの効果，環境の効果，共分散によって決定される。血縁者ペアの形質値間の共分散は，マーカー座位で共有する IBD 対立遺伝子の数によって決まり，

$$\mathrm{Cov}(X_i, X_j \mid \pi_{ij}) = f(\theta, \pi_{ij})\sigma_a^2 + g(\theta, z_2)\sigma_d^2 + \Phi_{ij}\sigma_G^2 \qquad (22)$$

と表される。ただし，σ_G^2 はポリジーンに起因する分散，Φ_{ij} は 2 者間の関係を示す係数，θ はマーカー座位と形質座位との間の組み換え割合を表す。π_{ij} は，血縁者ペアがマーカー座位で共有する IBD 対立遺伝子の数である。また，関数 $f(\theta, \pi_{ij})$ と $g(\theta, z_2)$ は用いる血縁者ペアの種類（同胞，親子など）に依存する。

ここで，n_r 人のメンバーから成る R 個の家系について μ_r は形質値の平均のベクトル，$\mathbf{V}_r$ は分散共分散行列であるとする。共分散行列の各要素は，式(22)で与えられる。形質値がおおよそ正規分布に従うと考えると，全体の対数尤度は全ての形質値の尤度値を結合して

$$\log(L) = c - \frac{1}{2}\sum_{r=1}^{R}\log(\det(\mathbf{V}_r)) - \frac{1}{2}\sum_{r=1}^{R}(\mathbf{X}_r - \mu_r)'\mathbf{V}_r^{-1}(\mathbf{X}_r - \mu_r) \qquad (23)$$

と表すことができる。ただし，c は定数項を表す。連鎖がないという帰無仮説の下では，$\sigma_a^2 = 0$ 及び $\sigma_d^2 = 0$ が成り立つ。式(23)に示す対数尤度を最大にするようにパラメータ（各分散及び組み換え割合）の推定を行い，その時の尤度と分散を 0 とした時の尤度を比較する尤度比検定によって，連鎖の有無を検定する。

Amos によるモデルを基本として，より複雑なモデルに適応させるための拡張が行われている。例えば，Towne ら（1997）は遺伝子と環境間の相互作用をモデルに導入し，Stern ら（1996）と Mitchell ら（1997）は複数の遺伝子間の相互作用を導入している。

variance components 法は，Genehunter や Merlin，SOLAR（Almasy and Blangero 1998）に実装されている。

モデルにおける正規分布の仮定がほぼ成り立つときには，variance components 法は Haseman-Elston 法より検出力が高い。また，Haseman-Elston 法と異なり，大家系や複雑な家系への適用も可能である。しかし，正規性の仮定が成り立たない場合には，偽陽性が増加するという問題点がある。正規性が成り立たない主な要因には，集団における形質値の分布が非正規分布である可能性があること，元の集団が正規分布に従っていても選択的サンプリングによって正規性から逸脱し得ることなどがある。

Haseman-Elston 法と variance components 法の比較を表 7 にまとめた。

表 7：Haseman-Elston 法と variance components 法の比較

	Haseman-Elston	variance components
形質値の分布に対する仮定	必要なし	多変量正規分布を仮定
非正規分布に対する頑健性	高い	低い
大家系への適用	×	○
連鎖の検定	Student の t 検定	尤度比検定

3. ソフトウェア

現在，学術機関及び非営利組織を対象に，多くのソフトウェアが無料でダウンロードできる。本章では，一般によく用いられるソフトウェアに関して，その概要を説明する。

なお，無料でダウンロード可能な資料及びソフトウェアの範囲内で調査を行った。

3.1. linkage 形式

まず，連鎖解析や連鎖不平衡解析での入力として一般的に用いられる linkage 形式の説明を行う。linkage 形式による入力ファイルには, 家系ファイルとデータファイルが含まれる。家系ファイルはテキストエディタを用いて ASCII フォーマットで作成する。各行に一個体の情報を記載する。個体の情報には以下に示す情報を含み，これらを空白区切りで順に記載する。

i. 家系 ID
ii. 個体 ID
iii. 個体の父親の ID（家系内に父親が不在の場合は 0）
iv. 個体の母親の ID（家系内に母親が不在の場合は 0）
v. 性別情報（男性：1，女性：0）
vi. 座位 1 での表現型
vii. 座位 2 での表現型　　以下，同様に必要な座位数の情報を記載

リストの 6 番目以降に示す表現型は，座位のタイプによって 4 種の異なる記載方法がある。罹患状態を表す場合は，罹患していれば 2，罹患していなければ 1，不明であれば 0 となる。対立遺伝子番号の場合は，その座位で個体がもつ対立遺伝子を記す。該当座位に 2 つの対立遺伝子が存在すれば，2 つの番号を記すことになる。2 値因子の場合，i 番目の因子の有無を示す 0 及び 1 の列を記す。量的形質の場合，各個体の測定値を記載する。

例として，図 12 の左側に示す家系について考える。四角または丸の横に記した番号は個体 ID であり，下に記した番号は各個体の遺伝子型である。個体 ID が 3 の個体を考えた場合，父親の ID は 1，母親の ID は 2 となり，また男性であるから性別情報は 1 となる。罹患

しているため，罹患状態を表す記号は 2 となり，対立遺伝子情報は「1 1」となる。家系 ID を 1 とすると，この家系全体に対する家系ファイルは図 12 の右側に示す形式で与えられる。

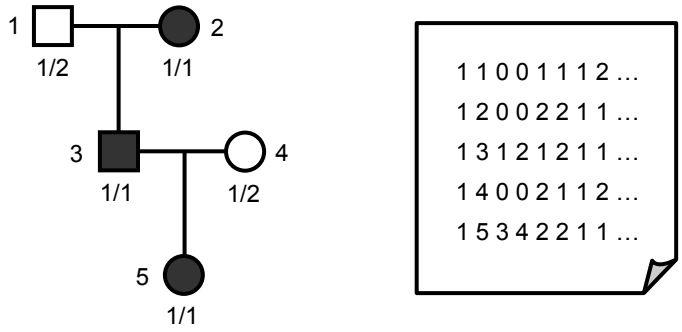

図 12：一般的な家系ファイル（linkage 形式）

家系ファイルでは，6 列目以降に示す座位のタイプは解析ごとに異なる。従って，各座位のタイプを明記するためのデータファイルが必要となる。データファイルは一般に，PREPLINK というプログラムを用いて作成される。テキストエディタを用いて直接作成することもできる。データファイルには，以下の情報を記す。

i. 座位とその順序に関する一般的な情報（座位数，突然変異率，染色体上での順序など）
ii. 座位のタイプ（量的変数，罹患状態，2 値因子，対立遺伝子状態）
iii. 組み換えに関する情報（組み換え率，性別による違いの有無，マップ関数による干渉の有無）
iv. 使用するプログラムに必要な情報（プログラムによって異なる）

データファイルの例を図 13 に示す。

```
3 0 0 1          << no loci,  risk locus,  sexlinked(if 1),  program code
0 0.0 0.0 0      << mut locus,  male mut rate,  female mut rate,  haplotype freq(if 1)
1 2 3            << order of loci

1 2              << binary factors,  # alleles
0.99 0.01        << gene freqs
2                << number of binary factors
0 1              << allelic codes
1 1
                    .
                    .
                    .
0 0              << sex difference (if 1),  interference (if 1)
0.1              << recombination values
1 1
```

図 13：データファイルの例（linkage 形式）

3.2. LINKAGE package

【概要】

任意数の座位における連鎖解析及び遺伝的リスクの計算を行ういくつかのプログラムを統合したシステムである。これらのプログラムは，解析プログラムとサポートプログラムに大別される。前者は連鎖解析の中心となるプログラムから成り，組み換え率の最尤推定や LOD スコアの計算，遺伝リスクの解析を行う。後者は利用者に使いやすいインタフェースを提供すると共に，データチェックや解析結果からのレポート作成を行う。解析プログラムの中で中心となるのは ILINK，MLINK，LINKMAP であり，それぞれヒトの家系に対して単点解析あるいは多点解析を行う。

ILINK は任意数のマーカー座位と疾患関連遺伝子座位との間の組み換え割合の最尤推定を行う。2 座位での解析の場合には，最大 LOD スコアの決定も行う。また，浸透率，遺伝子頻度，他のパラメータの推定も行うことができる。MLINK も主に単点解析を行うプログラムであり，２ヶ所以上の座位での LOD スコアの計算及び遺伝リスクの計算を行う。座位間の組み換え割合を段階的に変化させ，各組み換え割合での尤度を計算する。LINKMAP は多点解析を行うプログラムであり，固定された多座位の地図に対して，ある１つの座位の位置スコアを計算する。

ILINK，MLINK，LINKMAP はいずれも一般的な家系に対する解析を行うプログラムであるが，LINKAGE package にはこの他に CEPH（Centre d’Etude du Polymorphisme Humain）の基準家系で共優性マーカー分析を行うための特殊バージョンがある。

【入力】

linkage 形式による入力ファイルを用いる。

【アルゴリズム】

Elston-Stewart アルゴリズムを用いて尤度を計算するが，オリジナルの Elston-Stewart アルゴリズムでは家系の末端から上方向に畳み込むのに対して，上方向からの畳み込みも行うことができる。

【出力】

結果はテキストファイルに出力される。ILINK を実行した場合には，尤度を最大にする組み換え割合の値及び LOD スコアが出力される。MLINK の実行では，自由組み換えでの対数尤度値と LOD スコア，指定した各組み換え割合での対数尤度値と LOD スコアが出力される。LINKMAP の実行では，観察データの対数尤度が出力される。

【参考文献】

Lathrop G.M., Lalouel J.M., Julier C. and Ott J. (1984) Strategies for multilocus linkage analysis in humans. *Proc. Nat. Acad. Sci. USA* **81**:3443-3446

Lathrop G.M., Lalouel J.M., Julier C. and Ott J. (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am. J. Hum. Genet.* **37**:482-498

Lathrop G.M. and Ott J. (1990) Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. *Am. J. Hum. Genet.* **47**:A188

Ott J. (1999) *Analysis of Human Genetic Linkage*, 3rd edition. Johns Hopkins University Press, Baltimore

Terwilliger J.D. and Ott J. (1994) *Handbook for Human Genetic Linkage.* Johns Hopkins University Press, Baltimore

【利用形態】

自由にダウンロードして利用することができる。

【ダウンロード先】

http://www.genemapping.cn/LinkageWin.zip　（Windows 用）

【動作環境】

FreePascal で書かれており，MS-DOS/PC-DOS，Windows，Linux，Unix など様々なプラットフォームで利用できる。

3.3. FASTLINK

【概要】

LINKAGE package の拡張版であり，より高速である。Linkage package は FreePascal で書かれているが，FASTLINK は C 言語で書かれている。入出力は LINKAGE package と同様である。

現在では，LINKAGE package ではなく FASTLINK を使用することが推奨されている。

【参考文献】

Lathrop G.M., Lalouel J.-M., Julier C. and Ott J. (1984) Strategies for Multilocus Analysis in Humans. *PNAS* **81**:3443-3446.

Lathrop G.M. and Lalouel J.-M. (1984) Easy Calculations of LOD Scores and Genetic Risks on Small Computers. *Am. J. Hum. Genet.* **36**:460-465.

Lathrop G.M., Lalouel J.-M. and White R.L. (1986) Construction of Human Genetic Linkage

Maps: Likelihood Calculations for Multilocus Analysis. *Genet. Epidemiol.* **3**:39-52.
Cottingham R.W. Jr., Idury R.M. and Schaffer A.A. (1993)Faster Sequential Genetic Linkage Computations. *Am. J. Hum. Genet.* **53**:252-263.
Schaffer A.A., Gupta S.K., Shriram K. and Cottingham R.W. Jr. (1994) Avoiding Recomputation in Linkage Analysis. *Hum. Hered.* **44**:225-237.
Schaffer A.A. (1996) Faster Linkage Analysis Computations for Pedigrees with Loops or Unused Alleles Hum. Hered. **46**:226-235

【利用形態】
自由にダウンロードして利用することができる。

【ダウンロード先】
ソースコード及び各種 README ファイル：
ftp://fastlink.nih.gov/pub/fastlink/fastlink.tar.Z
Windows 用（実行ファイル）：
ftp://fastlink.nih.gov/pub/fastlink/windows/
MacOS X 用（実行ファイル）：
ftp://fastlink.nih.gov/pub/fastlink/mac/FASTLINK_executables.zip

【動作環境】
MS-DOS，VMS，UNIX 及び Linux で実行可能である。

3.4. Vitesse

【概要】
Elston-Stewart アルゴリズムを用いて家系内の尤度計算を行う。LINKAGE package に含まれる LINKMAP 及び MLINK の機能を持つ。正確な尤度計算に必要となる遺伝子型の数を減少させるために，set-recording，fuzzy inheritance と呼ばれる 2 つの方法を新しく組み入れ，計算の高速化を行っている。これにより，多数の高次元多型に対する正確な多点尤度を高速に計算できる。ただし，ループのない家系を扱う。

【入力】
linkage 形式の入力ファイルを用いる。

【出力】
結果は，画面及びファイルに出力される。出力内容は，LINKAGE や FASTLINK によるものと同様である。

【参考文献】

O'Connell J.R. and Weeks D.E. (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nat. Genet.* **11**:402-408

O'Connell J.R. (2001) Rapid Multipoint Linkage Analysis via Inheritance Vectors in the Elston-Stewart Algorithm. *Hum. Hered.* **51**:226-40.

【利用形態】

プログラムをダウンロードするためには，下記ダウンロード先のサイトにて氏名，メールアドレス，所属，国籍を登録しなければならない。

【ダウンロード先】

http://watson.hgen.pitt.edu/register/soft_doc.html

【動作環境】

DOS，UNIX 系の OS（Solaris，SunOS，OSF/1.3，HP-UX，AIX，IRIX）で実行可能である。

3.5. GeneHunter

【概要】

統合された方法でパラメトリック連鎖解析及びノンパラメトリック連鎖解析を行う。罹患同胞対解析は Mapmaker/sibs に実装されている方法を同胞のみでなく罹患した家系メンバーへと拡張したものである。多数のマーカー座位に対する高速な多点解析を行うことができるが，小家系に限られる。TDT 解析及び variance components 法による QTL 解析も行うことができる。ハプロタイプ解析も実行できるが，連鎖平衡を仮定していることに注意が必要だとする指摘もある（Schaid D.J. et al. 2002，Becker T. and Knapp M. 2003）。

【入力】

linkage 形式の入力ファイルを用いる。データファイルの読み込みには load コマンドを用いて「load markers <filename>」と入力し，家系ファイルの読み込みには scan コマンドを用いて「scan pedigrees <filename>」と入力する。全家系データを総合した結果を出力するためには「total stat」コマンドを用いる。各スコアのグラフを postscript 形式で出力したい場合には，「total stat」を実行する前に「ps on」コマンドを入力する必要がある。その他，細かな設定などに必要なコマンドについては「help <commandname>」を実行するかマニュアルを参照されたい。

【アルゴリズム】

Lander-Green アルゴリズムに基づく方法によって継承ベクトルの事後分布を求める。第 2 章で述べたように，オリジナルの Lander-Green アルゴリズムでの継承ベクトルは，非創始者の数を n とすると $2n$ 次元のベクトルである。しかし，創始者由来の配偶子伝達には相の情報は含まれていないため，継承ベクトルの要素には加えない。従って，創始者数を f とすると継承ベクトルの次元は $2n-f$ となる。求めた継承ベクトルが観察データである個体の表現型にどれだけ合っているかを評価するための採点関数に基づいて，パラメトリック連鎖解析の LOD スコアやノンパラメトリック解析の NPL スコアなどを計算する。採点関数は継承ベクトル w と個体の表現型全体の集合 Φ の関数 $S(w,\Phi)$ となる。

点 x 上におけるパラメトリック連鎖解析での採点関数は，

$$\overline{S}(x,\Phi)=\frac{\sum_{w\in V}P(\Phi|w)P_{\text{complete}}(w)}{\sum_{w\in V}P(\Phi|w)P_{\text{uniform}}(w)} \qquad (24)$$

と表される。ここで，V は可能な継承ベクトル全体の集合である。$P_{\text{complete}}(w)$ は全ての座位でのデータが与えられた上での継承ベクトル w の事後確率，$P_{\text{uniform}}(w)$ は一様分布の下での継承ベクトル w の確率である。この採点関数の常用対数を取れば LOD スコアとなる。

ノンパラメトリック連鎖解析での採点関数には，S_{pairs} と S_{all} の 2 つがある（Whittemore and Halpern 1994）。S_{pairs} は，家系内の罹患者に関する全てのペアで対立遺伝子を比較し，IBD 対立遺伝子の数を数え上げたものである。S_{all} は，

$$S_{\text{all}}(v)=2^{-a}\sum_{h}\left[\prod_{i=1}^{2f}b_i(h)!\right] \qquad (25)$$

と表される。ここで，h は各罹患者から 1 つの対立遺伝子を選択することによってできる集合であり，$b_i(h)$ は i 番目の創始者の対立遺伝子が集合 h 内にいくつあるかを表す。また，a は罹患者の数である。これら S_{pairs} と S_{all} に対して，以下のように正規化したスコアを定義する。

$$Z(v)=\frac{S(v)-\mu}{\sigma} \qquad (26)$$

μ 及び σ は，それぞれ P_{uniform} の下での $S(v)$ の平均値と標準偏差である。家系が複数ある場合には，

$$Z=\sum_{i=1}^{m}\gamma_i Z_i \qquad (27)$$

とする。ただし，m は家系数，Z_i は各家系での正規化スコア，γ_i は重み付け因子とする。この Z の値を NPL スコアと定義する。

【出力】

「scan pedigrees」コマンドにより家系ファイルが読み込まれると，家系ごとにパラメトリック連鎖解析及びノンパラメトリック連鎖解析が行われ, 染色体上の位置, LOD スコア, NPL スコア，p-value 及び情報量が出力される。その後「total stat」コマンドを実行すると，全家系を合わせた計算結果が出力される。postscript output オプションが on になっていれば，LOD スコア，NPL スコア及び情報量の PS ファイルが作成される。

家系ファイルの読み込み後に「estimate」コマンドを実行すると，罹患同胞対解析が行われる。出力はテキストファイルと PS ファイルで得られる。テキストファイルには位置，観察データの尤度が最大になる時の z_0，z_1，z_2 の値及び対数尤度の値が出力される。ただし，z_i は IBD 対立遺伝子の数が i となる確率を表す。また, 出力される PS ファイルは 2 種あり，それぞれ位置と対数尤度の関係，最尤の IBD 共有割合が領域に渡ってどのように変化するかが示される。

Haseman-Elston アルゴリズムによる QTL 解析では，位置，回帰モデルの傾き β，LOD スコア，t 値がテキストファイルに出力される。

variance components 法による QTL 解析では，各位置での LOD スコア及び平均値，分散成分，共分散回帰係数の推定値がテキストファイルに出力される。また，帰無モデルでの推定値も出力される。PS ファイルを出力するようにオプション指定することにより，LOD スコアの PS ファイルも得ることができる。

TDT 解析ではまず「tdt」コマンドを実行する。その後，「perm1」または「perm2」を実行すると並べ替え検定が行われる。結果は観察データより高い最大値を持つデータセットの数と，閾値 $(0.01,\ 0.001)$ より高い結果を持つデータセットの数が表示される。

【参考文献】

Kruglyak L., Daly M.J., Reeve-Daly M.P. and Lander E.S. (1996) Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach. *Am. J. Hum. Genet.* **58**:1347-1363

Kruglyak L. and Lander E.S. (1998) Faster Multipoint Linkage Analysis Using Fourier Transforms. *J. Comput. Biol.* **5**:1-7

【利用形態】

自由にダウンロードして利用することができる。

【ダウンロード先】

http://www.broad.mit.edu/ftp/distribution/software/gh2.1/gh-v2_1_r2_tar.tar.gz

【動作環境】

C 言語でかかれており，Sun 及び Linux X86 用には実行可能ファイル（gh.sun 及び gh.linux）が提供されている。その他の UNIX マシンでは Makefile を数箇所書き換えた上でコンパイルする必要がある。

3.6. EH

【概要】

異なるマーカー座位間やマーカー座位と疾患関連遺伝子座位との間の連鎖不平衡の検定及びハプロタイプ頻度の推定を行うプログラムである。連鎖不平衡が存在する場合において，EM アルゴリズムを用いた最尤ハプロタイプ頻度の推定を行う。また，観察データから算出した対立遺伝子頻度を基に連鎖不平衡がない場合のハプロタイプ頻度も示す。

個体数が少なく対立遺伝子数が多いときには統計量の検定に用いられている χ^2 近似が信頼できないため，より高性能なプログラムの使用が推奨されている。

【入力】

入力ファイルには各マーカーにおける対立遺伝子数及び各遺伝子型の観察数を記載する。図 14 に入力ファイルの例を示す。ここでは，解析に用いるマーカーは 3 つ，各マーカーが持つ対立遺伝子数は順に 2 つ，2 つ，3 つであるとする。入力ファイルの 1 行目には，この対立遺伝子数 2，2，3 を順に記す。図の左側に示す表のように各マーカー座位の遺伝子型全ての組み合わせを考え，解析に用いる個体を該当するセルに数え上げる。入力ファイルにはセルの数値を書き込む。

		座位 3					
座位 1	座位 2	1/1	1/2	2/2	1/3	2/3	3/3
	1/1	0	0	0	10	3	5
1/1	1/2	1	11	2	2	3	3
	2/2	5	5	4	3	0	15
	1/1	7	2	2	0	6	5
1/2	1/2						
	2/2						
	1/1						
2/2	1/2						
	1/3						

入力ファイル

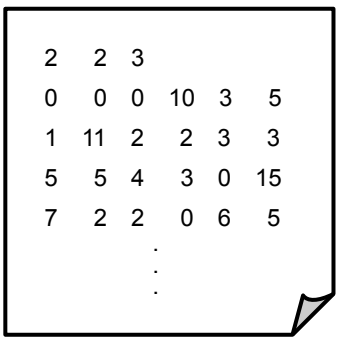

図 14 : 入力ファイルの例 (EH)

EH の入力ファイル形式では，1 つの座位における対立遺伝子数が多くなる場合，入力ファイルの行数や列数がサンプルサイズに比べて非常に大きくなることがある。メモリを効率良く使用するように入力ファイル形式を変更したソフトウェアとして EHPLUS がある。EHPLUS では観察データ中の遺伝子型の各組み合わせに対して ID を付け，入力ファイルの各行には ID 及び対応する組み合わせのカウント数を記載する。観察データに存在しない組み合わせ，すなわち EH の入力形式でカウント数が 0 となる組み合わせは入力しない。

【出力】

観察データから算出した対立遺伝子頻度，ハプロタイプ頻度の推定値及び対応する対数尤度，χ^2 値を出力する。また，EM アルゴリズムによる反復操作の回数も出力する。

【参考文献】

Xie X, Ott J (1993) Testing linkage disequilibrium between a disease gene and marker loci. *Am. J. Hum. Genet.* **53**:1107 (abstract)

【ダウンロード先】

http://www.genemapping.cn/eh.zip

【動作環境】

Free Pascal で書かれており，Windows や Linux など様々なプラットフォームで利用できる。

3.7. Merlin

【概要】

一般的な家系解析全般に対応しており，パラメトリック連鎖解析，ノンパラメトリック連鎖解析，QTL 解析（回帰ベース），IBD 推定，ハプロタイプ解析，エラー検出，シミュレーションを行うことができる。ほとんどの解析において，マーカー間の連鎖不平衡を入れることが可能である。

【入力】

コマンドによる入力を行う。入力ファイルには，一般的な linkage 形式，あるいは QTDT 形式を用いる。QTDT 形式における家系ファイルは linkage 形式と同等である。linkage 形式でのデータファイルに相当する入力は，座位のタイプを示すデータファイルとその座位の染色体上での位置情報を記したマップファイルに分けられる。データファイルとマップファイルの入力例を図 15 に示す。データファイルの 1 列目に示す記号は，M はマーカー，A は罹患状態，T は量的形質，C は共変量を表す。この他，解析によって入力する情報の追

加が必要となる場合もある。各解析で必要となる追加情報については，Web 上のチュートリアルを参照されたい。

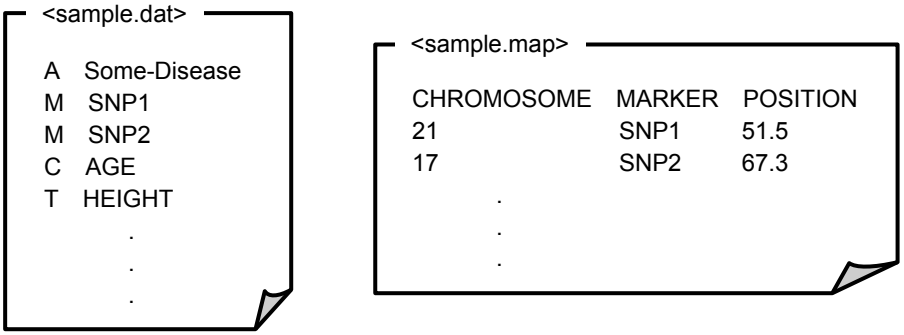

図 15：データファイルとマップファイル（Merlin）

【アルゴリズム】

家系解析において，家系内の gene flow パターンを表す二分木として sparse inheritance trees を用いることによって，高速化を行っている。sparse inheritance trees では，図 16(a) に示すように考え得る gene flow のパターン全てに対応するノードを生成するのではなく，尤度が 0 となるノードを省き，また，同胞と結果が等しくなるノードに対する計算を行わない（図 16(b)）。QTL 解析には，Sham ら（2002）によって柔軟性を持つように拡張された Haseman-Elston 法を用いている。

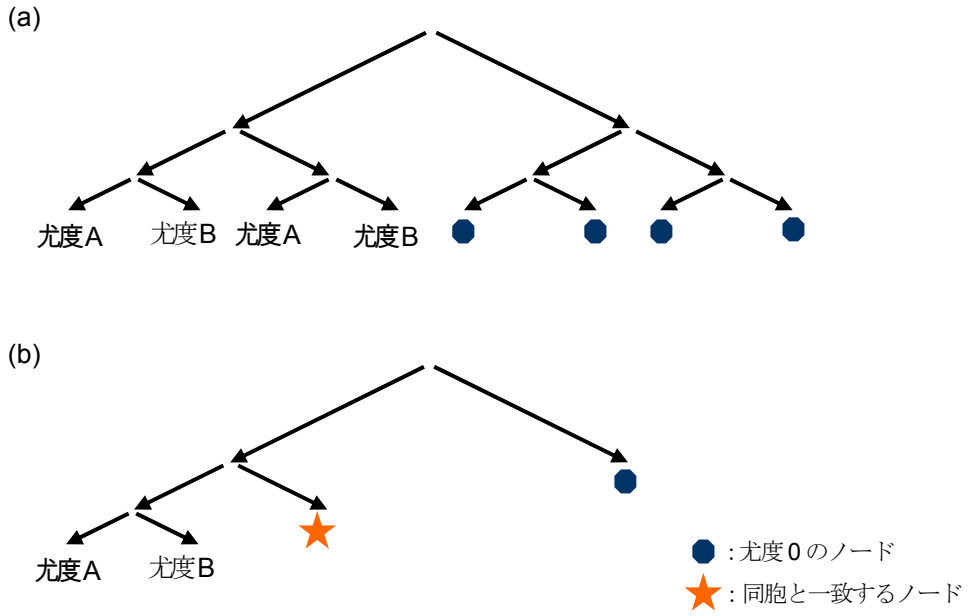

図 16：木構造による gene flow パターンの表示（(a)：通常，(b)sparse inheritance tree）

【出力】

出力は，実行する解析によって異なる。パラメトリック連鎖解析では，実行時に指定した各位置での LOD スコアの推定値，連鎖した家系の割合及びそれに対応する最大 heterogeneity LOD スコアを出力する。オプション指定によってグラフの表示も可能である。ノンパラメトリック連鎖解析では，染色体上の位置，Z スコア，Kong&Cox のデルタ値，K&C の LOD スコア，K&C の p-value を出力する（Kong and Cox 1997）。回帰による QTL 解析では，染色体上の位置，推定された座位特有の遺伝力とその標準偏差，LOD スコア及び p-value が出力される。また，ハプロタイプ解析では，各個体の推定ディプロタイプ形が示される。

【参考文献】

Abecasis G.R., Cherny S.S., Cookson W.O. and Cardon L.R. (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**:97-101

【利用形態】

営利目的であるか非営利目的であるかに関わらず，フリーで使用することができるが，氏名，メールアドレス及び所属の登録が必要である。

【ダウンロード先】

GNU/LINUX systems 用：

http://www.sph.umich.edu/csg/abecasis/Merlin/download/Linux-merlin.tar.gz

Sun Workstations 用：

http://www.sph.umich.edu/csg/abecasis/Merlin/download/SunOS-merlin.tar.gz

Windows Workstations 用：

http://www.sph.umich.edu/csg/abecasis/Merlin/download/Windows-merlin.zip

MacOS X G5 Workstations 用：

http://www.sph.umich.edu/csg/abecasis/Merlin/download/Darwin-merlin.tar.gz

【動作環境】

UNIX，Linux，Windows，MacOS で利用可能である。

3.8. SimWalk2

【概要】

パラメトリック連鎖解析，ノンパラメトリック連鎖解析，IBD 解析，ハプロタイプ解析，ミスタイプ解析を行うことができる。

【入力】

マップデータファイル，座位データファイル，家系ファイル，浸透率ファイル，コントロールファイルの 5 つが入力ファイルとして必要である。Mega2 を利用すれば linkage 形式のデータから Simwalk2 での解析に必要な入力ファイルに変換することができる。ただし，コントロールファイルにおける解析の種類はユーザが指定しなければならない。

【アルゴリズム】

MCMC 及びシミュレーテッドアニーリング法を用いて各種解析を行う。

【出力】

出力はテキストファイルで得られる。パラメトリック連鎖解析では location スコアを出力し，ノンパラメトリック連鎖解析では 5 つの NPL 統計量（BLOCKS, MAX-TREE, ENTROPY, NPL_PAIR，NPL_ALL）の p-value を出力する。BLOCKS は劣性形質で最も検出力が良くなる傾向にあり，MAX-TREE は優性形質，残りの統計量は相加的な形質で最も検出力が良くなる。ハプロタイプ解析では，各個体の最尤ディプロタイプ形を出力する。IBD 解析では，IBD 対立遺伝子数が 0，1，2 となる場合のそれぞれの確率を出力する。ミスタイプ解析では，観察された各遺伝子型データに対してミスタイプである確率を出力する。

【参考文献】

Sobel E. and Lange K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am. J. Hum. Genet.* **58**:1323-1337

Sobel E., Sengul H. and Weeks D.E. (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum. Hered.* **52**:121-131

Sobel E., Papp J.C. and Lange K. (2002) Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.* **70**:496-508

【利用形態】

SimWalk2，Mega2 共にダウンロードするためには登録が必要である。

【ダウンロード先】

SimWalk2 をダウンロードするための登録ページ：

http://www.genetics.ucla.edu/software/registration

Mega2 をダウンロードするための登録ページ：

http://watson.hgen.pitt.edu/register/cgi-bin/new_register.CGI

【動作環境】

Fortran で書かれており，Linux，Mac OSX，Solaris，Windows 及び OSF/Tru64 で実行することができる。location スコアオプションを有効にするためには，Mendel のバージョン 3.35 が必要である。実行ファイルは”simwalk2snp”と”simwalk2”の 2 種類あり，用いるデータが SNP のみの場合には, ”simwalk2snp”を使用することでより多数のマーカーを扱うことができる。

3.9. Mapmaker/sibs

【概要】

罹患同胞対法による解析を行うための標準的ソフトである。

【入力】

linkage 形式の入力ファイルを用いる。

【出力】

以下に示す 3 つのファイルが出力として得られる。

i. mls.out：座位ごとの LOD 値及びそれを与える z_0，z_1，z_2 の最尤推定値
ii. share.ps：座位を横軸に，z_0，z_1，z_2 各々の最尤推定値を縦軸にとったグラフ
iii. mls.ps：座位を横軸に，LOD 値を縦軸にとったグラフ

【参考文献】

Kruglyak L. and Lander E.S. (1995) Complete Multipoint Sib Pair Analysis of Qualitative and Quantitative Traits. *Am. J. Hum. Genet.* **57**:439-454

【利用形態】

フリーで利用することができる。

【ダウンロード先】

ftp://ftp-genome.wi.mit.edu/distribution/software/sibs/sibs-2.1.tar.Z

【動作環境】

C 言語で書かれており，UNIX 系 OS で実行可能である。SunOS や DEC Alpha での使用には実行ファイルが提供されているが，その他 UNIX 系 OS での使用には，ソースコードのコンパイルを行う必要がある。

3.10. PL-EM

【概要】

EM アルゴリズムによるハプロタイプ推定に PL 法を取り入れたソフトウェアである。PL 法におけるグループ毎の部分ハプロタイプの推定では，複数の部分ハプロタイプを候補として残し，EM アルゴリズムを用いて隣接ハプロタイプの結合を行う。局所解に陥る可能性が低くなるように，推定頻度の低い部分ハプロタイプも候補として残している。欠損データも扱うことができる。

【入力】

PL-EM の入力ファイル形式は HAPLOTYPER と同様である。

入力ファイルには，用いるマーカー（SNP）の情報を個体ごとに記載する。マーカーの情報とは遺伝子型のことを言い，ヘテロ接合体のときは 0，野生型ホモ接合体のときは 1，変異型ホモ接合体のときは 2 となる。入力ファイルの例を図 17 に示す。各行は各個体に対応している。1 行目に示す個体の場合，1 列目の記号は「1」であるから，対応するマーカーにおいて野生型ホモ接合体であることが分かる。

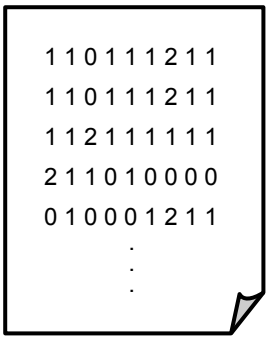

図 17：入力ファイルの例（PL-EM）

【出力】

集団の推定ハプロタイプ頻度及び個体毎の推定ディプロタイプ形を出力する。また，集団のハプロタイプの標準偏差も出力する。

【参考文献】

Qin Z.S., Niu T. and Liu J.S. (2002) Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **71**:1242-1247

【利用形態】

学術機関及び非営利機関の場合はフリーでダウンロードすることができる。ダウンロー

ドするためには商用目的に使用しないなどの同意書に同意し，Web 上で氏名，所属，メールアドレス，使用する OS を記入する必要がある。営利目的の機関の場合は技術ライセンス及び商標ライセンスのためハーバード大学の事務所と連絡をとらなければならない。

【ダウンロード先】

http://www.people.fas.harvard.edu/~junliu/plem/click.html

【動作環境】

UNIX 系の OS 及び Linux で実行可能である。

3.11. Haploview

【概要】

ハプロタイプ解析を行うための総合的なツールを提供する。複数の処理を共通のインタフェースで提供することにより，ユーザにとって利用しやすいソフトウェアとなっている。Haploview には，以下の機能が含まれている。

i. LD 解析及びハプロタイプブロック解析
ii. 集団のハプロタイプ頻度の推定
iii. SNP 関連解析及びハプロタイプ関連解析（TDT または case-control study）
iv. 関連の優位性を検定するための並べ替え検定
v. Tagger による tag SNP の選択

2007 年 2 月現在の最新バージョンは Haploview version 3.32 である。

【入力】

linkage 形式を含む 3 つの形式による入力が可能である。入力の際には，Haploview では 3 つ以上の対立遺伝子を持つマーカーを扱うことができない点に注意しなければならない。ここでは，linkage 形式を除く 2 つの入力形式（phased haplotypes 形式及び HapMap data 形式）について説明する。

```
Family1   Person1   0   1   1   1
Family1   Person1   0   1   1   1
Family1   Person2   2   3   h   1
Family1   Person2   2   3   h   1
```

図 18 : 入力ファイルの例（Haploview : phased haplotypes）

phased haplotypes 形式では，全ての相または一部の相が決定したハプロタイプを入力する。各行は家系 ID，個体 ID 及び遺伝子型データから成り，各個体につき 2 行（1 本の染色

体で 1 行）ずつ入力する。図 18 は，2 人の個体についての入力例を示す。相の決まっていないヘテロ接合の対立遺伝子である場合は，該当する座位の遺伝子型情報には「h」を入力する。また，欠損データの場合は 0 を入力する。

HapMap Data 形式では，HapMap プロジェクトの GBrowse を用いて取り込んだファイルが入力ファイルとなる。HapMap プロジェクトの Web サイトにおいて，”Report & Analysis”の”Download SNP genotype data”を選択すればよい。

linkage 形式または phased haplotypes 形式による入力を行う際には，マーカー情報を記載した入力ファイルが別に必要となる。1 つのマーカーにつき 1 行であり，各行にはマーカーの名前と位置を入力する。

【アルゴリズム】

ハプロタイプの相の推定及び集団ハプロタイプ頻度の推定には，EM アルゴリズムを用いている。特に 10 個を超えるマーカーから成るブロックの場合には，PL 法を適用している。

【出力】

結果は 5 つのタブを持つ GUI で提供される。ただし，”Association”タブは linkage 形式による入力を行った上，オプションで関連解析を選択した場合のみ表示される。図 19 に出力例を示す。この例では関連解析を行っていないため，4 つのタブが表示されている。

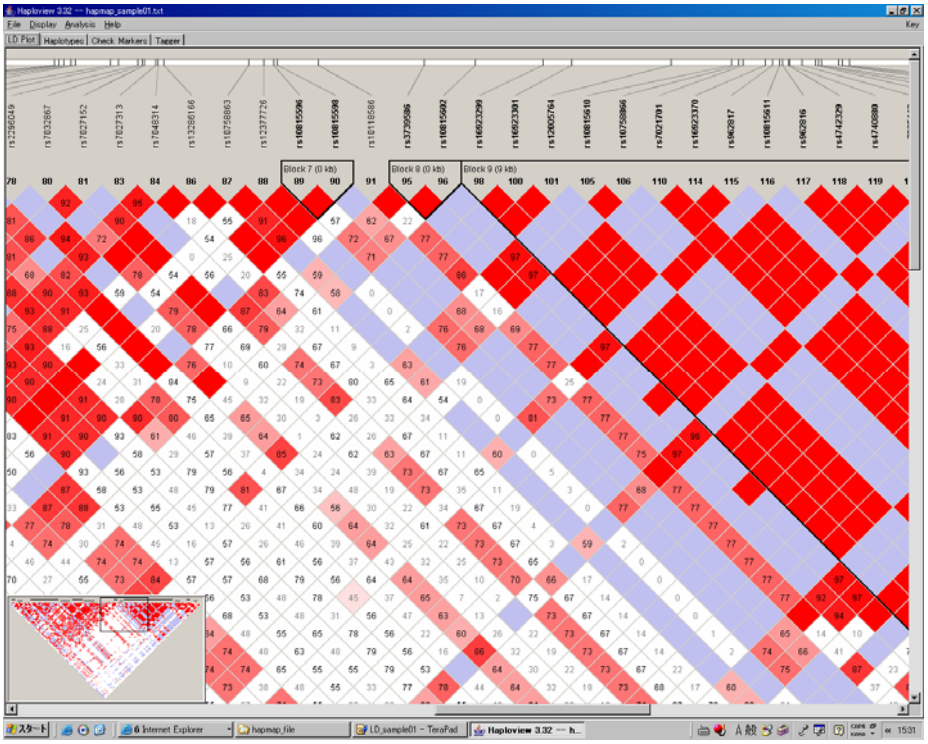

図 19：出力例（Haploview）

“Check Markers”タブでは，各マーカーにおけるジェノタイピングの質を示す数的指標を表示する。HWE（Hardy-Weinberg equilibrium）を満たしているか，また，メンデルの遺伝法則からの逸脱はないか確認することができる。各マーカーにおける MAF（minor allele frequency）や非欠損データの割合なども表示されており，このタブ上で，ハプロタイプ解析に利用する SNP の取捨選択を行うことも可能である。

“LD plot”タブでは, マーカー間の連鎖不平衡の強さ及び全体のブロック構造をカラーで表示する。連鎖不平衡の尺度は D’や r^2 値などいくつか用意されており，自由に選択することが可能である。ここで表示された図は，Export オプションで PNG 形式の画像として保存することができる。また，詳細な数値情報をテキスト形式で保存することもできる。

“Haplotypes”タブでは，ブロック内の集団のハプロタイプ及びその推定頻度，また，2 ブロック間の LD の尺度である Hedrick の他座位 D’を表示する。これらの数値情報は，Export オプションによりテキスト形式で保存することが可能である。

“Association”タブでは，関連解析の結果として，χ^2 値及び p-value を表示する。TDT 及び case-control study のいずれを行うかはデータの入力時に選択する。また，TDT 関連の優位性を評価する並べ替え検定を実行することができる。結果はテキスト形式で保存することができる。

“Tagger”タブでは，最適化された tag SNP のセットを表示する。

【参考文献】

Barrett J.C., Fry B., Maller J. and Daly M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**:263-265

【利用形態】

営利，非営利に関係なくフリーでダウンロードすることができる。

【ダウンロード先】

JRE のダウンロード： http://www.java.com/

Windows 用：

http://www.broad.mit.edu/mpg/haploview/downloads/hapinstall.exe

Mac OSX 及び UNIX 用：

http://www.broad.mit.edu/mpg/haploview/downloads/Haploview.jar

【動作環境】

Java で書かれており，Windows，Mac OSX 及び UNIX で実行できる。実行するためには Java JRE 1.3 以降のバージョンを事前にインストールしておく必要である（Java JRE 1.4 以降のバージョンの利用を推奨)。

3.12. PHASE

【概要】

集団の遺伝子型データからハプロタイプ及び個体のディプロタイプ形の推定を行う。組み換え率の推定及び組み換えホットスポットの同定も行うことができる。遺伝子型データには SNP, マイクロサテライト, その他複数の対立遺伝子を持つ座位を扱うことができる。また，case 群と control 群との間におけるハプロタイプ頻度の違いの検定も可能である。検定には，「case 群と control 群は共通のハプロタイプ頻度を持つ集団からランダムに取り出された」という事象を帰無仮説とする並べ替え検定が用いられる。

2005 年 4 月の時点で，バージョン 2.2.1 が提供されている。

【入力】

実行するためには，PHASE 固有の形式に従った入力ファイルが必要である。入力ファイルの 1 行目には解析する個体数を入力し, 2 行目には解析に用いるマーカー座位の数を入力する。3 行目には各座位の位置を記載する。行頭には大文字の P を記載し，各座位の順序は染色体に沿った物理的順序と同じでなければならない。4 行目には，各座位のタイプを入力する（S : SNP または対立遺伝子が 2 つの座位，M : マイクロサテライトまたは 3 つ以上の対立遺伝子を持つ座位）。5 行目以降には，各個体の情報を入力する。3 行で 1 個体を表し，最初の行には ID，2 番目及び 3 番目の行には遺伝子型を記載する。各座位について一方の対立遺伝子を 2 番目の行に，もう一方の対立遺伝子を 3 番目の行に入力する。

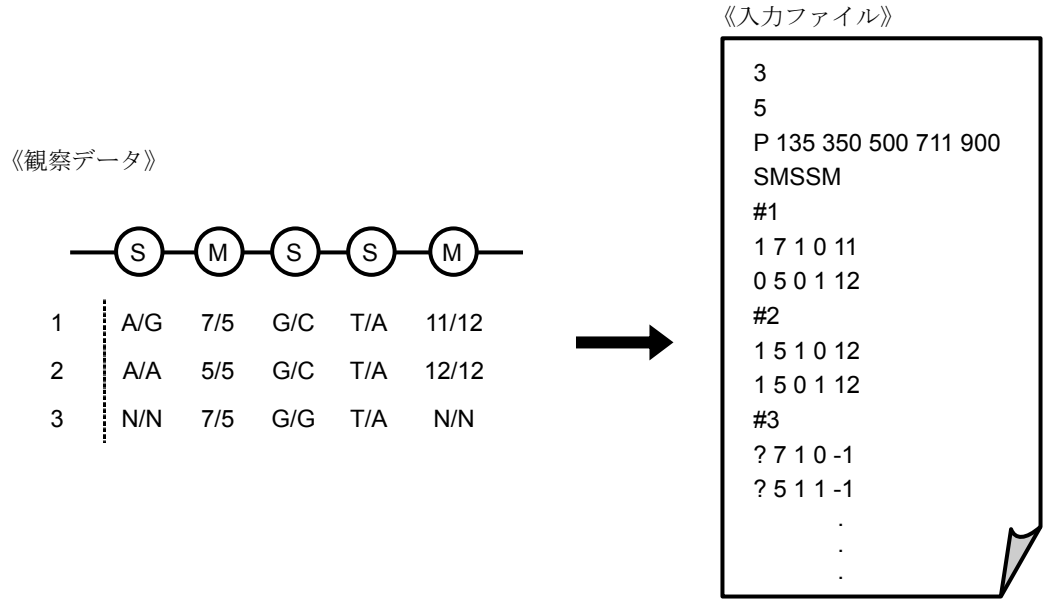

図 20：入力ファイルの例（PHASE）

図 20 に観察データの例及びそれに対応する入力ファイルの例を示す。観察データの上部における S は座位のタイプが SNP であることを表し, M はマイクロサテライトまたは複数

の対立遺伝子を持つ座位であることを表す。この例では個体数 3，座位数 5 であることがそれぞれ 1 行目と 2 行目に表されている。各座位の位置（3 行目）は必ずしも必要ではない。個体 1 の遺伝子型データは，「#1」の後に示す 2 行で与えられる。入力ファイル作成の際には，欠損データの入力方法に対して注意が必要である。個体 3 の観察データでは，SNP タイプである 1 番目の座位及び，マイクロサテライトタイプである 5 番目の座位が欠損データとなっている。このような場合，SNP タイプの座位における欠損データの入力には「?」，マイクロサテライトタイプの座位における欠損データの入力には「-1」を用いなければならない。

【アルゴリズム】

オリジナルバージョンは 2001 年に Stephens らによって提案された手法を用いており，コアレセントモデルに基づいた事前分布を用いたベイズ法によって，個体のディプロタイプ形の事後分布を推定する。その際の計算には MCMC を利用している（Pseudo Gibbs Sampling（PGS）と呼ばれる）。2003 年には Stephens と Donnelly によって PL 法の変形が組み込まれ，また，距離が広くなるに従って連鎖不平衡が弱くなるモデルが導入された。これによって，ハプロタイプ推定のスピードと精度が向上した。更に，2005 年には Stephens と Scheet によって”blocklike”な連鎖不平衡パターン及び”nonblocklike”な連鎖不平衡を扱うことのできる新しい組み換えモデルが導入された。

【出力】

PHASE では，結果の概要に関する出力ファイル及び，より詳細な結果を示す複数の出力ファイルが作成される。

概要に関する出力ファイルには，最良な結果を与えるハプロタイプのリスト及び対応する各ハプロタイプの頻度，各個体に対する最良な推定ディプロタイプ形のリストが出力される。その他，サンプル集団のハプロタイプ推定頻度と標準偏差，各個体のディプロタイプ形の分布，組み換えパラメータの推定値などを出力するファイルがある。組み換えパラメータの推定を行うためには，入力ファイルに各座位の位置を入力しておかなければならない。case 群と control 群との間における違いの有無を検定する並べ替え検定の p-value を出力するためには，PHASE を実行する際のコマンドに「-c」を追加しなければならない。また，入力ファイルに各個体が case 群に属するか control 群に属するかを記載しなければならない。

【参考文献】

Stephens M., Smith N. and Donnelly P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**:978-989

Stephens M. and Donnelly P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**:1162-1169

Li N. and Stephens M. (2003) Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* **165**:2213-2233

Crawford et al (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.* **36**:700-706

Stephens M. and Scheet P. (2005) Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. *Am. J. Hum. Genet.* **76**:449-462

【利用形態】

非営利目的の使用はフリーであるが，ダウンロードには，氏名，メールアドレス，住所，所属の登録が必要である。また，営利目的の使用にはライセンスが必要である。

【ダウンロード先】

非営利目的使用でのダウンロードを行うための登録ページ：

http://www.uwopendoor.org/LicenseSoftware.asp?softwareid=10

ライセンス取得のためのフォーマット：

http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/PHASE_commercial.pdf

【動作環境】

C 言語で書かれており，Linux 及び Windows で実行可能なファイルが提供されている。ただし，バージョン 2.1 に関しては大規模データに対する Windows での実行は不安定との報告もあり，Linux での利用が推奨されている。

3.13. fastPHASE

【概要】

欠損した遺伝子型の推定，相の決まっていない SNP データからのハプロタイプ再構成を行う。

【入力】

入力ファイルの形式は PHASE と同様である。ただし，2006 年 11 月現在のバージョン 1.1 では SNP しか扱うことができないため, 入力ファイル 4 行目の各座位のタイプには「S」のみを用いる。基本となる実行コマンドは,

./fastPHASE –T10 –oMyresults sample.inp

である。この場合，入力ファイル sample.inp に対して EM アルゴリズムを 10 回行い，接頭

辞が Myresults となるファイルを結果として出力する。

【アルゴリズム】

ハプロタイプのクラスターモデルに基づいた手法を用いる。各クラスターは局所的に共通するハプロタイプあるいは対立遺伝子の組み合わせを表しているものとし，クラスターの構成要素は HMM に従ってゲノムに沿って連続的に変化する。これは，ゲノムを連鎖不平衡の強い部分（ブロック）で分割したブロックベースのクラスターモデルをより柔軟性のあるモデルにしたものだと言える。

EM アルゴリズムを用いてモデルのパラメータを推定し，得られたパラメータの下で欠損した遺伝子型の推定及びディプロタイプ形の推定を行う。

【出力】

各個体につき 2 行で推定されたハプロタイプまたは遺伝子型をファイルに出力する。また，いくつかのパラメータによる結果の要約もサマリーとして出力する。また，switch error を最小とする推定結果を「<接頭辞>_switch.out」ファイルに出力し，individual error を最小とする推定結果を「<接頭辞>_indiv.out」ファイルに出力する。ここで，switch error はヘテロ接合座位の中で相を誤って推定した座位の割合，individual error は遺伝子型データからハプロタイプを決定できない個体の割合を示す。

【参考文献】

Scheet P. and Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**:629-644

【利用形態】

非営利目的の使用はフリーであるが，ダウンロードには，氏名，メールアドレス，住所，所属の登録が必要である。また，営利目的の使用にはライセンスが必要である。

【ダウンロード先】

非営利目的使用でのダウンロードを行うための登録サイト：

http://www.uwopendoor.org/LicenseSoftware.asp?softwareid=3

ライセンス取得のためのフォーマット：

http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/fastPHASE_commercial.pdf

【動作環境】

C 言語で書かれており，Linux，Solaris，Darwin 及び Windows で実行可能なファイルが提供されている。Solaris，Darwin，Windows については実行ファイルのダウンロードが可能である。

3.14. HAPLOTYPER

【概要】

Dirichlet 分布を事前分布としたベイズ法を用いてハプロタイプの推定を行う。

【入力】

入力ファイルには，用いるマーカー（SNP）の情報を個体ごとに記載する。マーカーの情報とは遺伝子型のことを言い，ヘテロ接合体のときは 0，野生型ホモ接合体のときは 1，変異型ホモ接合体のときは 2 となる。入力ファイルの例を図 21 に示す。各行は各個体に対応している。1 行目に示す個体の場合，1 列目の記号は「1」であるから，対応するマーカーにおいて野生型ホモ接合体であることが分かる。

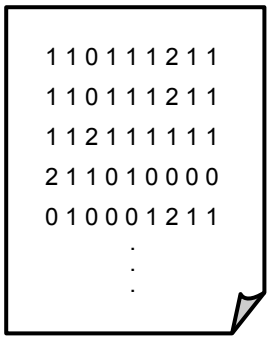

図 21 : 入力ファイルの例（HAPLOTYPER）

【アルゴリズム】

PL 法及びギブスサンプリング法を使用している。また，局所解を回避するために Prior annealing を用いている。

【出力】

出力は 2 つの部分より成る。出力ファイルの前半部分には，各個体の推定ディプロタイプ形及びその事後確率が記される。後半部分には，推定された各個体のディプロタイプ形から算出された全体のハプロタイプ頻度が記される。

【参考文献】

Niu T., Qin Z.S., Xu X. and Liu J. (2002) Bayesian Haplotype Inference for Multiple Linked Single Nucleotide Polymorphisms. *Am. J. Hum. Genet.* **70**:157-169

【利用形態】

学術機関及び非営利機関の場合はフリーでダウンロードすることができる。ダウンロードするためには商用目的に使用しないなどの同意書に同意し，Web 上で氏名，所属，メールアドレス，使用する OS を記入する必要がある。営利目的の機関の場合は技術ライセンス及び商標ライセンスのためハーバード大学の事務所と連絡をとらなければならない。

【ダウンロード先】

http://www.people.fas.harvard.edu/~junliu/Haplo/click.html

【動作環境】

Unix，Linux で利用できる。

3.15. SOLAR

【概要】

variance components 法による解析を行うソフトウェアパッケージであり，連鎖解析（単点解析，多点解析），QTL 解析，共変量のスクリーニングに対応している。20 までの量的あるいは離散的な形質を含むモデルに対する多変量解析を行うことができる。パッケージの中には Linkage package 及び FASTLINK も含まれている。

2007 年 2 月末現在の最新バージョンは 4.0.7 である。

【入力】

入力ファイルには家系ファイル，表現型ファイル，マーカーファイル，マップファイル，頻度ファイルがある。マップファイルと頻度ファイルを除いて，入力ファイルには PEDSYS ファイルまたはコンマ区切りのファイルを用いる。ここではコンマ区切りのファイルについて説明する。コンマ区切りのファイルでは，フィールド名をコンマ区切りでリスト表示したヘッダーが必要となる。例えば，

ID,AGE,EF,Q1,Q2,Q3,Q4,Q5

のように記載する。空白を用いてはならない点に注意が必要である。

家系ファイルは各個体につき 1 レコードのデータから成り，それぞれのレコードには個体 ID，父親の ID，母親の ID，性別が含まれる。1 つのデータセットの中に同じ個体 ID が存在する場合には，家系 ID も必要となる。親の ID が空白または 0 となっている場合は「親不明」として扱う。SOLAR では，両親とも不明（創始者）であるか，両親とも既知でなければならない。片親のみが分かっている場合には，既知の親を空白にするか，あるいはダミーの親を採用する。性別は M/F，m/f，1/2 のいずれかで表す。

表現型ファイルも各個体につき 1 レコードのデータから成る。それぞれのレコードには

個体 ID と 1 つ以上の表現値が含まれる。年齢のような量的共変量も含むことができる。家系ファイルと同様に，1 つのデータセットの中に同じ個体 ID が存在する場合には家系 ID が必要である。表現型ファイルにおける欠損データは空白で表される。連続する 2 つの整数のみからなるフィールドは離散変数であると判断され，それ以外は量的データとみなされる。従って，3 つ以上のクラスを持つ離散変数を形質として扱うことはできない。共変量として用いる場合には，複数の 2 値変数に分解しなければならない。

マーカーファイルも各個体につき 1 レコードのデータから成り，マーカー座位における遺伝子型情報が入力される。標準的な遺伝子型のコード方法の多くを認識できるが，「/」を用いて対立遺伝子を分ける方法が推奨される。欠損データは空白，「0/0」または「-/-」で表す。

マップファイルには，マーカーの染色体上での位置が染色体ごとに入力される。単位は cM である。1 行目には染色体番号及び，マーカー間の距離を組み換え割合に変換する際に使用するマッピング関数の名前を入力し，2 行目以降にマーカー名とそのマーカーの染色体上での位置を空白区切りで入力する。2007 年 2 月末現在，マッピング関数として利用できるのは Kosambi 関数と Haldane 関数であり，デフォルトでは Kosambi 関数が使用される。

頻度ファイルには一連のマーカー座位に対して対立遺伝子頻度を入力する。1 行に 1 マーカーのデータを入力し，各行にはマーカー名，1 番目の対立遺伝子名，1 番目の対立遺伝子頻度，2 番目の対立遺伝子名，2 番目の対立遺伝子頻度（以下，必要数の対立遺伝子について同様に記載する）を空白区切りで表す。

解析を行う際には，最初に家系ファイルの読み込みが必要である。家系ファイル名を sample.ped とすると，

```
load pedigree sample.ped
```

というコマンドを入力する。特定の作業ディレクトリ内では 1 つの家系ファイルを用いた解析しか行うことができない。一度家系ファイルを読み込めば，同じディレクトリで解析を行う際に再読み込みを行う必要はない。

量的遺伝解析を行うには，家系ファイルの読み込みの後，次のコマンドによって表現型ファイルを読み込む。

```
load phenotypes sample.phen
```

家系ファイルと同様に，ファイルに変更がない限り，同じ作業ディレクトリでの解析には再読み込みの必要はない。表現型ファイルの読み込みに続いて形質とする表現型を選択し，共変量を設定する。共変量における相互作用は「*」を用いて表す。

```
trait q4                          ;# 形質の選択
covariate sex age age*sex         ;# 共変量の設定
```

その後，polygenic または polygenic –screen コマンドを入力すれば標準的な量的遺伝解析が実行され，尤度を最大とするパラメータが求められる。「-screen」オプションは各共変量の有意性の有無を調べるために用いる。オプションの有無に関わらず，polygenic コマンド

を実行すると連鎖解析に必要な null0 というモデルが生成される。

単点解析を行うためには IBD 行列が必要である。load コマンドでマーカーファイルを読み込んだ後，ibd コマンドを実行すると IBD ファイルが圧縮ファイルとして保存される。遺伝子型データに欠損のある個体が存在する場合には，マーカーファイルを読み込む前に頻度ファイルの読み込みを行う。trait コマンドで形質を選択した後，twopoint コマンドを実行すると，単点解析を実行することができる。

多点解析には多点 IBD ファイルが必要である。多点解析と同様にマーカーファイルを読み込んだ後 mibd コマンドを実行する。その後，chromosome コマンドで対象とする染色体を選択し，interval コマンド及び finemap コマンドで必要な情報を与えた後，multipoint コマンドで多点解析を実行する。

その他詳細な説明は，ユーザガイドを参照されたい。

【アルゴリズム】

variance components 法を利用した家系解析を行う。

【出力】

量的遺伝解析を行うと, 対数尤度値が得られる。「-screen」オプションを用いた場合には，各共変量の p 値と共に有意性の有無が表示され，最終的なモデルから除去された共変量が示される。単点解析の結果の概要は twopoint.out，多点解析の結果の概要は multipoint.out という名のファイルに出力される。多点解析では，全ての LOD スコアが出力され，LOD スコアが最大となる連鎖モデルが保存される。

【参考文献】

Almasy L. and Blangero J. (1998) Multipoint quantitative trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**:1198-1211

【利用形態】

フリーでダウンロードできるが，氏名，メールアドレス，所属及び使用している OS の登録が必要である。

【ダウンロード先】

ダウンロードするための登録ページ：

http://www.sfbr.org/solar/solarreg.html

【動作環境】

Linux，Solaris，Mac OSX で実行できる。

3.16. SAGE

【概要】

連鎖解析から関連解析まで, 様々な解析を行うためのプログラムが用意されている。2007 年 2 月現在で利用できる機能を以下に示す。

i. 用いるデータに関する統計量の算出及び質の評価
ii. 対立遺伝子頻度の推定
iii. 遺伝力及び家系内相関関係の推定
iv. 発病年齢を変数として含めた遺伝的伝達及び浸透率関数の混合モデルの推定
v. 血縁者ペアにおける IBD 対立遺伝子の共有確率の推定
vi. パラメトリック連鎖解析
vii. ノンパラメトリック連鎖解析
viii. TDT 解析
ix. 形質と対立遺伝子間の関連解析

入出力は行う解析によって異なる。詳細についてはユーザマニュアルまたは SAGE のホームページを参照されたい。

【参考文献】

特にない。

【利用形態】

非営利目的での使用はフリーであるが，開発元と連絡をとらなければならない。営利目的での使用にはライセンスが必要である。

【ダウンロード先】

非営利目的で使用する際の連絡先：

sage@darwin.cwru.edu

ライセンス取得のための登録ページ：

http://darwin.cwru.edu/sage/?q=node/29

ユーザマニュアル（バージョン 5.3.1）：

http://darwin.cwru.edu/files/SAGE_UserDoc_v5.3.1.pdf

【動作環境】

C++で書かれており，Linux，Mac OSX，Solaris，Windows で実行できる。事前に J2SE JRE をインストールしておかなければならない。Windows 2000 での使用には Service Pack 4，Windows XP での使用には Service Pack 2 が必要である。

3.17. 商用ソフトの紹介

代表的な商用ソフトとして，Helix Tree®と SAS/Genetics™ を挙げることができる。

Helix Tree®は Golden Helix 社と GlaxoSmithKline 社によって開発された統計解析ソフトウェアである。分かりやすくインタラクティブなユーザインタフェースを提供している。主な機能を以下に挙げる。

i. Hardy-Weinberg 平衡解析
ii. EM アルゴリズムによるハプロタイプ頻度推定
iii. 連鎖不平衡解析
iv. 多数の遺伝子と環境因子および臨床データとの相関解析
v. スクリプトによるバッチ処理可能

入力ファイルには，Affymetrix 社のチップデータ(CHP ファイル）や Excel データなどの様々なデータ形式のファイルを用いることができる。Windows，Linux，Mac OSX での利用が可能である。7 日間の無料試用期間があり，オンラインセミナーを受講すると更に 14 日間，試用期間が延長される。

SAS/Genetics™ は SAS Institute 社によって開発されたアプリケーションで, 統計パッケージ SAS 上で動く。ユーザフレンドリーであり，遺伝統計の分野で用いられる基本的な解析のほとんどが網羅されている。7 つのプロシージャーと 1 つのマクロで構成される。以下に主な機能を示す。

i. 個々のマーカーの特性を示す統計量の算出
ii. Hardy-Weinberg 平衡解析
iii. EM アルゴリズムによるハプロタイプ頻度推定
iv. 連鎖不平衡解析
v. ハプロタイプまたはマーカーと 2 値変数（疾患状態など）との間の関連解析
vi. 家系データによる関連解析：TDT，S-TDT，SDT など
vii. tag SNP の選択
viii. 家系内の近交係数及び共分散係数の算出

Windows，Linux，Solaris，Mac で利用することができる。利用するためには，Base SAS のライセンスが必要である。

参考文献

鎌谷直之(編) (2001) ポストゲノム時代の遺伝統計学. 羊土社

斎藤聡, 鎌谷直之 (2002) ゲノム研究における遺伝統計学の主な手法. *Molecular Medicine* (臨時増刊号 癌ゲノム学) **39**: 226-235

J･オット(著), 五條堀孝(監訳), 安田徳一(訳) (2002) ヒトゲノムの連鎖分析 - 疾患遺伝子の探索. 講談社

Abecasis G.R., Cherny S.S., Cookson W.O. and Cardon L.R. (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**:97-101

Almasy L. and Blangero J. (1998) Multipoint quantitative trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**:1198-1211

Amos C.I. (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54**:535-543

Becker T. and Knapp M. (2003) Comment on "The Impact of genotyping error on haplotype reconstruction and frequency estimation". *Eur. J. Hum. Genet.* **11**:637 (Letter to the Editor)

Clark A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7**:111-122

Cottingham R.W. Jr., Idury R.M. and Schaffer A.A. (1993)Faster Sequential Genetic Linkage Computations. *Am. J. Hum. Genet.* **53**:252-263

Elston R.C. and Stewart J. (1971) A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**:523–542

Excoffier L. and Slatkin M. (1995) Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**:921-927

Fishman P.M., Suarez B., Hodge S.E. and Reich T. (1978) A robust method for the detection of linkage in familial diseases. *Am. J. Hum. Genet.* **30**:308–321

Goldgar D.E. (1990) Multipoint analysis of human quantitative genetic variation. *Am. J. Hum. Genet.* **47**:957-967

Haseman J.K and Elston R.C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**:3-19

Holmans P. (1993) Asymptotic properties of affected-sib-pair linkage analysis. *Am. J. Hum. Genet.* **52**:362-374

Kong A. and Cox N.J. (1997) Allele-sharing models: LOD scores and accurate linkage tests. *Am. J.Hum. Genet.* **61**:1179-1188

Kruglyak L., Daly M.J., Reeve-Daly M.P. and Lander E.S. (1996) Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach. *Am. J. Hum. Genet.* **58**:1347-1363

Lander E.S. and Green P. (1987) Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84**:2363–2367

Lathrop G.M., Lalouel J.M., Julier C. and Ott J. (1984) Strategies for multilocus linkage analysis in humans. *Proc. Nat. Acad. Sci. USA* **81**:3443-3446

Lathrop G.M., Lalouel J.M., Julier C. and Ott J. (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am. J. Hum. Genet.* **37**:482-498

Mitchell B.D., Ghosh S., Schneider J.L., Birznicks G. and Blangero J. (1997) Power of variance component linkage analysis to detect epistasis. *Genet. Epidemiol.* **14:**1017–1022.

Niu T., Qin Z.S., Xu X. and Liu J.S. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet.* **70**:157-169

O'Connell J.R. and Weeks D.E. (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nat. Genet.* **11**:402-408

Ott J. (1999) Analysis of Human Genetic Linkage, 3rd ed. Johns Hopkins University Press, Baltimore.

Penrose L.S. (1953) The general purpose sib-pair linkage test. *Ann. Eugen.* **18**:120−124

Qin Z.S., Niu T. and Liu J.S. (2002) Partition-Ligation Expectation-Maximisation algorithm for haplotype inference with single nucleotide polymorphisms. *Am. J. Hum. Genet.* (letter), **71**:1242-1247

Schaffer A.A., Gupta S.K., Shriram K. and Cottingham R.W. Jr. (1994) Avoiding Recomputation in Linkage Analysis. *Hum. Hered.* **44**:225-237

Schaid D.J., McDonnell S.K., Wang L., Cunningham J.M. and Thibodeau S.N. (2002) Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am. J. Hum. Genet.* **71**:992–995 (Letter to the Editor)

Sham P.C. Purcell S. Cherny S.S. and Abecasis G.R. (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am. J. Hum. Genet.* **71**:238-253

Sobel E. and Lange K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am. J. Hum. Genet.* **58**:1323-1337

Spielman R.S., McGinnis R.E. and Ewens W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**:506–516

Spielman R.S. and Evens W.J. (1998) A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62**:450-458

Stephens M., Smith N.J. and Donnelly P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Gen.* **68**:978-989

Stern M.P., Duggirala R., Mitchell B.D., Reinhart J.L., Shivakumar S. et al. (1996) Evidence for linkage of regions on chromosomes 6 and 11 to plasma glucose concentrations in Mexican Americans. *Genome Res.* **6**:724−734

Thompson E.A. (1994a) Monte Carlo estimation of multilocus autozygosity probabilities. In: Sall J, Lehman A (eds) Proceedings of the 1994 Interface Conference. Interface Foundation of North America, Fairfax Station, VA, pp 498–506.

Thompson E.A. (1994b) Monte Carlo likelihood in genetic mapping. *Stat. Sci.* **9**:355-366.

Towne B., Siervogel R.M. and Blangero J. (1997) Effects of genotype-by-sex interaction on quantitative trait linkage analysis. *Genet. Epidemiol.***14**:1053−1058

Whittemore A.S. and Halpern J. (1994) A class tests for linkage using affected pedigree members. *Biometrics.* **50**:118-127