傳統之基因定位法

● 數量性狀基因座定位的資料型式

回交族群

個 體	性狀值	標 識
1	179.80	1 1 1 1 1 1 1 1 1 …
2	207.80	1 1 1 1 0 0 0 0 0 …
3	240.80	1 1 1 1 0 0 1 1 0 …
4	212.17	0 0 0 0 0 0 1 1 1 …
5	202.17	1 1 1 0 0 0 0 0 0 …
6	197.17	0 0 0 0 0 0 0 0 1 …
⋮	⋮	⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮

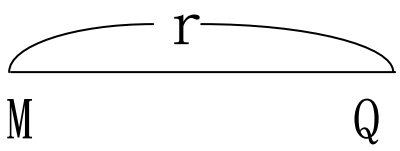
F_2 族群

個 體	性狀值	標 識
1	85.5	0 1 1 1 1 1 1 2 2 …
2	89.0	2 1 2 2 2 2 2 2 2 …
3	94.5	2 2 2 2 2 2 1 1 2 …
4	84.5	1 1 2 2 2 2 2 2 1 …
5	86.5	2 1 0 0 0 1 1 1 0 …
6	86.5	0 1 1 1 1 2 2 2 1 …
⋮	⋮	⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮

- Sax(1923)首先利用豆子的斑點和色素形態這些外表可辨識的標識來研究控制種子大小的QTL。他得到有一個以上的QTL與這些標識連鎖的結論。Thoday(1961)也利用了一些標識研究果蠅的一些性狀。Sax和Thoday的方法主要是比較不同標識基因型的個體間之數量性狀平均值是否有差異。若有差異存在，則顯示控制此數量性狀的QTL與標識間有連鎖的關係。

- Sax和Thoday的方法一次利用一個標識進行不同標識基因間平均值差異的比較。這類方法顯然無法滿足數量性狀是受多個QTL控制的事實，也無法估計QTL 的位置。

● 單一標識t檢定(single-marker t-test)



以回交族群為例：

標 識		QTL		
基因型	比例	基因型	比例	基因型值
MM	$\frac{1}{2}$	QQ	$\frac{1-r}{2}$	$\mu+\frac{1}{2}a$
		Qq	$\frac{r}{2}$	$\mu-\frac{1}{2}a$
Mm	$\frac{1}{2}$	QQ	$\frac{r}{2}$	$\mu+\frac{1}{2}a$
		Qq	$\frac{1-r}{2}$	$\mu-\frac{1}{2}a$

$$\mu_{MM}=\left[\frac{1-r}{2}\left(\mu+\frac{1}{2}a\right)+\frac{r}{2}\left(\mu-\frac{1}{2}a\right)\right]\Big/\frac{1}{2}=\frac{1-2r}{2}a+\mu$$

$$\mu_{Mm}=\left[\frac{r}{2}\left(\mu+\frac{1}{2}a\right)+\frac{1-r}{2}\left(\mu-\frac{1}{2}a\right)\right]\Big/\frac{1}{2}=-\frac{1-2r}{2}a+\mu$$

$$\mu_{MM}-\mu_{Mm}=(1-2r)a$$

假設檢定：

H_0： $r = 0.5$ （Q 與 M 連鎖）

H_1： $r < 0.5$

$$t = \frac{\hat{\mu}_{MM} - \hat{\mu}_{Mm}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

既然a必不為 0，顯著的t值表示 Q 與 M 連鎖。

- 單一標識迴歸法(single-marker regression)

在回交族群中，任一標識x_i的期望值和變方(variance)為

$$E(x_i)=\frac{1}{2}\times\frac{1}{2}+\left(-\frac{1}{2}\right)\times\frac{1}{2}=0$$

$$V(x_i)=\left(\frac{1}{2}\right)^2\times\frac{1}{2}+\left(-\frac{1}{2}\right)^2\times\frac{1}{2}=\frac{1}{4}$$

任二標識（基因）x_i與x_j間之共變異數(covariance)為

$$COV(x_i,x_j)=\frac{1-r}{2}\left(\frac{1}{2}\times\frac{1}{2}\right)+\frac{r}{2}\left(\frac{1}{2}\times-\frac{1}{2}\right)$$

$$+\frac{r}{2}\left(-\frac{1}{2}\times\frac{1}{2}\right)+\frac{1-r}{2}\left(-\frac{1}{2}\times-\frac{1}{2}\right)$$

$$=\frac{1-2r}{4}$$

數量性狀y_i對某個標識x_i的迴歸模式為

$$y_i = \mu + a_m x_i + \varepsilon_i$$

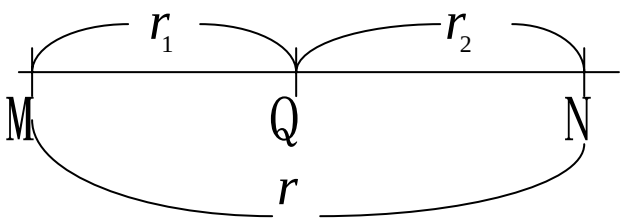
$$a_m = \frac{\mathrm{cov}(y_i, x_i)}{V(x_i)}$$

$$= \frac{\mathrm{cov}(\mu + a x_i^* + \varepsilon_i, x_i)}{V(x_i)} = (1-2r)a$$

上式之x_i^*代表控制y_i的 QTL Q。

- 在期望上，單一標識分析法以（$1-2r$）的比例低估 QTL 的作用a，其缺點有：
  （一）QTL 的作用a會被a_m低估。
  （二）QTL 與標識間的重組比例r無法估得。
  （三）無法區分連鎖的 QTL，而可能使估計偏差。(Lander & Botstein 1989; Jansen 1994; Zeng 1994）。

區間定位法(Interval mapping, Lander & Botstein 1989)



- 大量生產分子遺傳標識的技術逐漸成熟。

- 一次利用一個標識區間（一對相鄰的標識）對 QTL 進行定位。

- IM 利用基因重組的遺傳特性在區間內的每個位置都建立一假設的基因 Q，然後檢驗 Q 與數量性狀 y 是否有顯著的關係。若關係顯著，則表示 Q 為一 QTL。反之，則不是。

- 條件機率（Conditional probability）

回交族群中的任一個標識有兩種基因型。當同時考慮兩個標識（M 和 N），則共同四種可能的基因型，分別是 $\frac{MN}{MN}, \frac{MN}{Mn}, \frac{MN}{mN}$ 和 $\frac{MN}{mn}$ 。

假設有一 QTL Q 位於 M 和 N 所形成的區間內，則 Q 的兩種可能基因型（QQ 和 Qq）在四種標識基因型內的機率分布如下：

標識基因型	QTL 基因型	
	QQ	Qq
MN/MN	$\frac{(1-r_1)(1-r_2)}{1-r}$	$\frac{r_1 r_2}{1-r}$
MN/Mn	$\frac{(1-r_1)r_2}{r}$	$\frac{r_1(1-r_2)}{r}$
MN/mN	$\frac{r_1(1-r_2)}{r}$	$\frac{(1-r_1)r_2}{r}$
MN/mn	$\frac{r_1 r_2}{1-r}$	$\frac{(1-r_1)(1-r_2)}{1-r}$

例如 $prob\left(\frac{MQN}{MqN}\middle|\frac{MN}{MN}\right) = prob(MqN|MN) = \frac{prob(MqN, MN)}{prob(MN)}$

$$= \frac{prob(MqN)}{prob(MN)} = \frac{r_1 r_2}{1-r}$$

- QTL 基因型在四種標識基因型內的機率會因區間的大小和 QTL 在區間內的位置的改變而改變。

- 若標識區間夠小（標識夠密），則區間內發生雙重組(double recombination)事件的機會會很小，而可忽略之。QTL 基因型的條件機率分布可簡化為

標識基因型	QTL 基因型	
	QQ	Qq
MN/MN	1	0
MN/Mn	$1-p$	p
MN/mN	p	$1-p$
MN/mn	0	1

$$p=\frac{r_1}{r}$$

- IM 在所假設的基因 Q 所在位置的統計模式為

$$y_i = \mu + a x_i^* + \varepsilon_i \quad , \quad i = 1,2,\cdots\cdots,n$$

x_i^*代表假設的基因 Q，a 為 Q 對數量性狀 y_i 的作用，ε_i 假設服從 N(0,σ^2)。透過檢驗 a 是否為零的假設

$$H_0: \quad a = 0$$

$$H_1: \quad a \neq 0$$

來決定 Q 是否是一個 QTL。

- IM 統計模式的概度函數(likelihood function)

由於 IM 模式中所假設的基因Q(x_i^*)位於區間內,故其基因型無法直接觀測得到。Q的基因型有兩種，可能為QQ或Qq（$x_i^*=½$或$-½$），所以Q的基因型(x_i^*)服從二項分布

$$x_i^* = \begin{cases} ½\text{的機率為 } p_{i1} \\ -½\text{的機率為 } p_{i2} = 1 - p_{i1} \end{cases}$$

p_{i1}（p_{i2}）是Q為QQ（Qq）的條件機率。在$\varepsilon_i \sim \mathrm{N}(0,\sigma^2)$的假設下，IM 模式的概度函數為一常態混合函數(normal mixture model)

$$L(\theta|Y,X) = \prod_{i=1}^{n}\left[\sum_{j=1}^{2} p_{ij}\mathrm{N}\left(\mu_{ij},\sigma^2\right)\right]$$

上式之 $\mu_{i1} = \mu + \frac{1}{2}a$ 和 $\mu_{i2} = \mu - \frac{1}{2}a$

- MLE(maximum likelihood estimates)

已觀測資料(observed data)：數量性狀(y_i)、標識 (X_i)。

缺失資料 (missing data)：QTL (x_i^*)。

EM algorithm(Dempster et al. 1977)：

E-step：更新 QTL 基因型的事後條件機率

$$\pi_{ij} = P\left(x_i^* | y_i, X_i\right) = \frac{p_{ij}\mathrm{N}\left(\mu_{ij}, \sigma^2\right)}{\sum_{j=1}^{2} p_{ij}\mathrm{N}\left(\mu_{ij}, \sigma^2\right)};$$

$$i = 1,2...n; \quad j = 1,2.$$

M-step: 求 μ ，a 和 σ^2 滿足下列式子

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(\pi_{i1} - 1/2\right)a\right]$$

$$a = \frac{\sum_{i=1}^{n}(y_i - \mu)\left(\pi_{i1} - 1/2\right)}{1/4}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \mu)^2 - 2(y_i - \mu)(\pi_{i1} - 1/2)a + \frac{1}{4}a^2]$$

- 漸近變方和共變方矩陣（asymptotic variance-covariance matrix）

  可利用 Louis(1982)的方法求得漸近變方與共變方矩陣。詳細推導過程可參考 Kao & Zeng（1997）。

- 利用概度比測驗（likelihood ratio test）

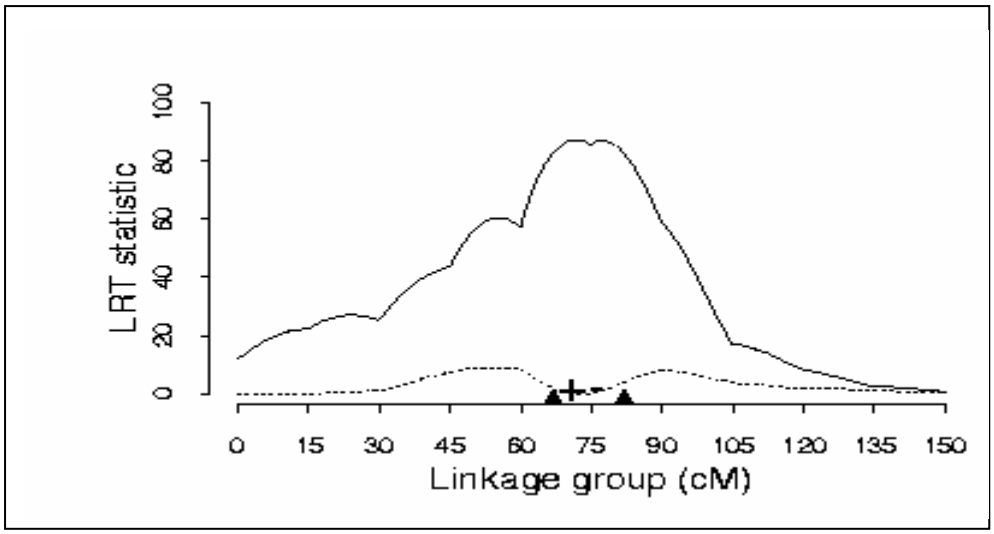
  利用 LRT 檢驗 QTL 的作用 a 是否為零

$$LRT = -2\log\left[\frac{Sup_{\theta_0} L(\theta|Y,X)}{Sup_{\theta} L(\theta|Y,X)}\right]$$

  在常態混合分布下，LRT 在 H_0: $a=0$ 下的漸近分布並不服從 χ^2（Mclachlan 1987）。然而，混合比例（mixing proportion）p_{ij} 若是已知，漸近分布會是 χ^2（Goffinet & Loisel 1992）。

- 傳統的方法檢驗 QTL 是否與標識連鎖，IM 則是檢驗特定位點上是否有 QTL 的存在。

- IM 視 LRT 顯著且最大為 QTL 所在位置



- 臨界值（critical value）的選擇

（一）連鎖圖上標識（基因）間的相關性為 $1-2r$。若標識未連鎖（$r=0.5$），則無相關。若標識連鎖（$r<0.5$），則有正相關，且標識越近，相關性越強。

（二）利用 IM 逐一檢查連鎖圖上每一區間內的每個位置是否存有 QTL 的步驟，實際上是進行許多互相獨立和彼此相關的多重檢定（multiple test），故假設檢定的臨界值需適當調整，以保證第一型錯誤的正確性。

（三）Lander & Bostein（1989）建議當標識相當密時,LRT 的臨界值可以用 Orenstein-Uhlenbeck diffusion 過程來決定。當標識較稀疏時，可用 Bonferroni argument 來調整。此外，一般認為臨界值必需隨著標識的數目、區間的大小、數量性狀的遺傳率、QTL 作用的大小和方向來調整（Jansen 1994)。Vesscher & Haley (1996)建議當一個作用大的 QTL 被偵測到後，對下一個 QTL 偵測的臨界值應調整(變小)。Churchill & Doerge(1994)則提議利用 permutation 來決定臨界值。

- IM 與傳統方法相比較下的優點（Lander & Botstein 1989）：
  - （一）Likelihood 剖面圖可以很清楚地呈現染色體組上每個位置是否有 QTL 的證據強弱。
  - （二）對 QTL 作用的估計是無偏的。
  - （三）QTL 可能的位置可由 support interval 表示。
  - （四）偵測 QTL 的能力較強。如果 QTL 位於區間內的中央位置，IM 可節省 $\psi \times 100\%$ 的樣本數。例如區間大小為 10，20，30 和 40 cM， IM 可分別節省 9%，16%，23% 和 28% 的樣本數。

- Permutation test(Doerge & Churchill 1996)

（一）洗牌（shuffle）：將原有資料內個體之基因型（標識）與外表型（性狀值）以隨機結合的方式打破，產生“新”的資料組。之後，對此“新”的資料組進行 IM 分析，得每一位點之 LRT 統計值。記錄其中最大的 LRT。

（二）洗 N 次牌得 N 組“新”資料，對每一組“新”資料進行 IM 分析，得 N 個最大的 LRT。

（三）算出這 N 個 LRT 的（1-α）百分位數（percentile）。

（四）原資料位點的 LRT 統計值若大於步驟（三）之$(1-\alpha)$百分位數，則表示該位點存在有一 QTL（推翻H_0）。

- 迴歸區間定位法（Regression interval mapping, Haley & Knott 1992; Martinez & Curnow 1992）

（一）他們認為 IM 利用 EM 演算法經多次循環求 MLE 的方式，在推導上和程式寫作上頗為複雜，而提出迴歸區間定位法來近似 IM。

(二)迴歸區間定位法的回交族群統計模式為

$$y_i = \mu + aE\left(x_i^*\right) + \varepsilon_i$$

上式之

$$E\left(x_i^*\right) = \frac{1}{2}\left(p_{i1}\right) - \frac{1}{2}\left(p_{i2}\right) = p_{i1} - \frac{1}{2}$$

為 QTL 基因型的條件期望值。

(三)他們以特定的模擬比較了 IM 和迴歸區間定位法，發現兩方法間之差異相當微小。因此，後續有許多人(Whittaker 1996; Xu 1996,1998; Lebreton 等 1998; Goffinet & Mangin 1998; Dupuis & Siegmund 1999; Rebai & Goffinet 200 等）用 REG IM 去研究更複雜的 QTL 定位問題。

- ML 和 REG 區間定位法的差異

（一）Jansen（1993）指出迴歸法的模式從遺傳的觀點很難解釋。

（二）Kao（2000）以分析和模擬的方式指出 ML 和 REG 兩種方法間的差異是存在的，且可能是顯著的。從估計上比較，兩種方法的相似程度取決於條件事後機率 π_{ij} 和條件機率 p_{ij} 的相似程度

$$\pi_{ij} = \frac{p_{ij}\mathrm{N}(\mu_{ij},\sigma^2)}{\sum p_{ij}\mathrm{N}(\mu_{ij},\sigma^2)}$$

π_{ij} 和 p_{ij} 越相似，迴歸法近似 IM 的程度越好。反之，則越差。REG 近似 IM 的程度隨(1)遺傳率增高，(2) QTL 的位置偏向區間的中央，(3)QTL 位於較寬的區間，(4)QTL 間的交感作用變強，(5)QTL 的作用差異變大，(6)QTL 的連鎖程度變強而變差。

- REG 區間定位法雖然在精確和精準度上無法與 ML 法相比，但仍不失做為一先前預備分析的好方法。

- 貝式（Bayesian）區間定位法

  Satagopan et al. 1996; Hoeschele & Vanrandan 1993; Sillanpaa & Arjas 1999; Yi & Xu 2000 ……

- 推廣至其他試驗族群和對多元性狀(multiple trait)、分類性狀（categorical trait）的分析(Jiang & Zeng 1995；Hackett & Weller 1995；Xu & Atchley 1996；Song et al. 1999；Henshall & Goddard 1999 ; Zeng et al. 2000……)。

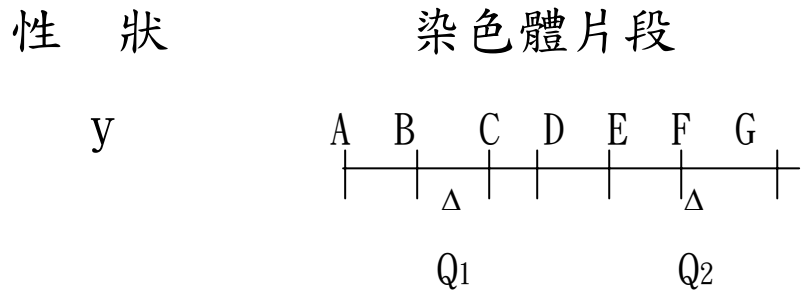
組合區間定位法（Composite interval mapping）

- IM 的方法可以深入標識和標識的區間內對 QTL 位置進行估計，解決了傳統方法無法估計 QTL 位置的問題。

- QTL 若（緊密）連鎖，IM 的估計是偏差的。

- IM 效率不高。

-



Jansen（1993）和 Zeng（1993,1994）分別提出結合 IM 和迴歸分析，以提昇 QTL 定位的準確性和效率的方法。Zeng 稱之為組合區間定位法(CIM)。Jansen 稱之為 MQM(multiple QTL model)。

- CIM 的回交族群模式

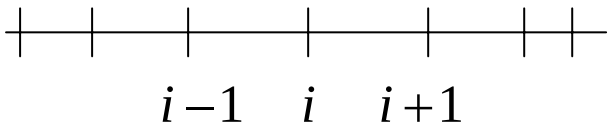
$$y_i = \mu + ax_i^* + \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \text{，}$$

x_i^*：待測的 QTL。

x_{ij}：其他適當選取的標識。

選取適當的標識納入 x_{ij} 做為 covariates，可阻止其他連鎖 QTL 對待測之 QTL(x_i^*)的干擾，並可減少遺傳殘差而增加 QTL 估計的準確性和偵測能力（Zeng 1993）。

為什麼？



$$b_{yi\cdot(i-1)(i+1)} = \frac{\sigma_{y_i.(i-1)(i+1)}}{\sigma^2_{i\cdot(i-1)(i+1)}}$$

$$= \sum_{i-1<u<i} \frac{r_{(i-1)u}(1-r_{(i-1)u})(1-2r_{ui})}{r_{(i-1)i}(1-r_{(i-1)i})} + \sum_{i<u<i+1} \frac{r_{u(i+1)}(1-r_{u(i+1)})(1-2r_{iu})}{r_{i(i+1)}(1-r_{i(i+1)})} a_u$$

- CIM = IM + regression

  = one QTL + markers (covariates)

- CIM 模式的概度函數類似 IM，仍為兩種常態分布的混合。CIM 的參數估計並不困難，可利用 EM 演算法求 MLE。
- 使用 CIM 應注意的一些問題
  1. 窗口尺寸(window size)不可太小。
  2. 背景標識(background marker)的選取可依模式選擇(model selection)方式決定。

參考文獻

劉清、陳英勇（1992）利用遺傳標識定位數量性狀基因座。中華農藝，2，1-21

戴政（1995）人類遺傳疾病研究統計方法評介。數學傳播第 29 卷第二期，8-13

歐尚靈（1999）定位二元資料與順序資料的數量性狀基因座。國立台灣大學博士論文

高振宏 （2000）對試驗族群進行數量性狀基因座定位的統計方法評介。中國統計學報第 38 卷第一期，1-36

Churchill, G. A. and R. W. Doerge (1994), Empirical threshold values for quantitative trait mapping. Genetics 138: 967-971.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39: 1-38.

Doerge, R.W. and G. A. Churchill (1996), Permutation test for multiple loci affecting a quantitative character. Genetics 142: 284-294.

Dupuis, J. and D. Siegmund (1999), Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151: 373-386.

Falconer, D. S., and T. F. C. Mackay (1996), Introduction to Quantitative Genetics, Ed 4. Longman, New York.

Goffinet, B., and B. Loisel (1992), Testing in normal mixture models when the proportions are known. Biometrika 79: 842-846

Haley, C. S. and S. A. Knott (1992), A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69: 315-324.

Hoeschele, I. and P. Vanranden (1993a), Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. Theor. Appl. Genet. 85: 953-960.

Hoeschele, I. and P. Vanranden (1993b), Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor. and Appl. Genet. 85: 946-952.

Jansen, R. C. (1993), Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211.

Jiang, C. and Z.-B. Zeng (1997), Mapping quantitative trait loci with dominant and missing markers. Genetica 101: 47-58.

Kao, C.-H. (1995), Statistical methods for locating the positions and analyzing epistasis of multiple quantitative trait genes using molecular marker information. Ph.D. Thesis, North Carolina State University, Raleigh.

Kao, C.-H. and Z.-B Zeng (1997), General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when

using the EM algorithm. Biometrics 53: 359-371.

Kao, C.-H. (2000), On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. Genetics (Oct. issue).

Louis, T. A. (1982), Finding the observed information matrix when using the EM algorithm. Journal of Royal Statistical Society, Series B 44: 226-233.

Lynch, M. and B. Walsh (1998), Genetics and analysis of quantitaive trait. Sinauer Associates, Inc., Sounderland, MA, USA.

McLachlan, G. J. (1987), On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Applied Statistics 36, 318-324.

Satagopan, J. M., S. Brian, M. A.Yandell, A. Newton and T. C. Osborn (1996), A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144: 805-816.

Sax, K. (1923), The association of size difference with seed-coat pattern and pigmentation in Phaseolus ulgaris. Genetics 8: 552-560.

Sillanpaa, M. J. and E. Arjars (1998), Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148: 1373-1388.

Thoday, J. M. (1960), Location of polygenes. Nature 191: 368-370.

Visscher, P. M., B. Thompson, and C. S. Haley (1996), Comfidence interval in QTL Mapping by Bootstrapping. Genetics 143: 1013-1020.

Xu, S. and W. R. Atchley (1996), Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics 143: 1417-1424.

Yi, N and S. Xu (1999), A random model approach to map quantitative trait Loci for complex binary traits in outbred populations. Gnenetics 153: 1029-1040.

Zeng, Z.-B. (1993), Theoretical basis for separation of multiple linked gene effects In mapping of quantitative trait loci. Proc. Natl. Acad. Sci. 90: 10972-10976.

Zeng, Z.-B. (1994), Precision mapping of quantitative trait loci. Genetics 136: 1457-1468.