ORIGINAL ARTICLE

Kazuki Shimo-onoda · Toshihiro Tanaka
 Kozo Furushima · Toshiaki Nakajima · Satoshi Toh
 Seiko Harata · Kazunori Yone · Setsuro Komiya
 Hiroki Adachi · Eiji Nakamura · Hitoshi Fujimiya
 Ituro Inoue

Akaike's information criterion for a measure of linkage disequilibrium

Received: July 3, 2002 / Accepted: October 2, 2002

Abstract The extent and distribution of linkage disequilibrium (*LD*) in humans is a current topic especially for gene mapping of complex diseases. Akaike's information criterion (*AIC*) was applied to estimate *LD* and compared with other standard *LD* measures, D' and r^2 . By comparison of an independent model (IM; linkage equilibrium) and a dependent model (DM; linkage disequilibrium), the parsimonious model is the one with the smaller *AIC* score. Therefore, the extent of *LD* by *AIC* is expressed as $AIC(IM) - AIC(DM)(AIC(LD))$. A total of 39 single-nucleotide polymorphisms on a 1.6-Mb region of chromosome 21q22 were identified, and genotyped in 192 Japanese individuals. All possible pairs were analyzed to estimate *LD* and the analyses were compared. *AIC(LD)* became highly positive as the D' value increased and was negative at D' values of around 0.2. Because a negative value of *AIC(LD)* implies linkage equilibrium, D' values below 0.2 should be regarded as linkage equilibrium. The *LD* estimate by *AIC* yielded results similar to those obtained by r^2 , indicating that *AIC(LD)* would be useful for fine gene mapping.

Key words *AIC* · Linkage disequilibrium · Gene-mapping

K. Shimo-onoda · T. Tanaka · K. Furushima · T. Nakajima ·
 I. Inoue (✉)
 Division of Genetic Diagnosis, The Institute of Medical Science,
 The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo
 108-8639, Japan
 Tel. +81-3-5449-5325; Fax +81-3-5449-5764
 e-mail: ituro@ims.u-tokyo.ac.jp

K. Shimo-onoda · K. Yone · S. Komiya
 Department of Orthopedic Surgery, Faculty of Medicine, Kagoshima
 University, Kagoshima, Japan

T. Tanaka · K. Furushima · S. Toh · S. Harata
 Department of Orthopedic Surgery, School of Medicine, Hirosaki
 University, Hirosaki, Japan

H. Adachi · E. Nakamura · H. Fujimiya
 Dynacom Co. Ltd., Mobara, Japan

Introduction

Owing to the recent emphasis on genome-wide association studies to identify susceptibilities to common metabolic diseases, the evaluation of linkage disequilibrium (*LD*) in humans has become a key prerequisite for the design of such studies (Risch and Merikangas 1996; Lander 1996; Collins 1995; Jorde 2000). Biallelic single-nucleotide polymorphisms (SNPs) are the markers of choice because of their high frequency, low mutation rates, and amenability to automation (Landegren et al. 1998; Wang et al. 1998). Two *LD* measures, D' and r^2 , are commonly used. However, both *LD* measures potentially have shortcomings for the evaluation of *LD*: D' is known to be inflated when the sample size is small, even for SNPs with common alleles, and r^2 is very sensitive to allele frequencies (Ardlie et al. 2002). Therefore, an alternative *LD* measure is desirable.

In the present study, we applied Akaike's information criterion (*AIC*) as an alternative estimate for *LD*. *AIC* was developed by Akaike (Akaike 1974) and is useful for determining a better-fitting model in comparison with nonhierarchical models. For several alternative nonhierarchical models, the better-fitting model was considered to be that with a lower *AIC* value. By applying *AIC* as an *LD* measure, linkage equilibrium was clearly separated from linkage disequilibrium, and *LD* by *AIC* showed a pattern similar to that of r^2 , indicating that *AIC* could serve as an alternative method for evaluating *LD*.

Subjects and methods

Study subjects

Because the current study was initially designed to identify susceptibilities to ossification of the posterior ligament of the spine (OPLL), we examined 96 unrelated OPLL patients and 96 unrelated non-OPLL patients recruited in Kagoshima, Japan. All of the patients gave written informed consent, and the study was approved by the Ethical

Committee of the Faculty of Medicine, Kagoshima University, and the Ethical Committee of the Institute of Medical Science, Tokyo University. The structures of *LD* were the same among patients and controls; thus, we present the combined data.

Selection of SNP markers and real-time pyrophosphate DNA sequencing

A genome region covering 1.6Mb on chromosome 21q22 was screened for SNPs. Thirty-nine SNPs with minor allele frequencies >0.05 were identified from the public database, as summarized in Table 1. Gap regions ($>50\text{kb}$) were sequenced, and two SNPs were identified between C21orf62 and PRKCBP, and PRKCB2 and INFAR2.

A polymerase chain reaction (PCR) analysis was performed with a standard protocol, with the exception that biotin-labeled primers were used. The single-stranded PCR samples, which were annealed with their respective sequencing primers, were transferred to a translucent 96-well plate for real-time pyrophosphate DNA sequencing (Ronaldi et al. 1996, 1998). The DNA sequence was read using a PSQ 96 instrument (Pyrosequencing AB, Uppsala, Sweden) and PSQ 96 SNP reagent kit (Pyrosequencing AB).

LD measure by D' and r^2

The pairwise haplotype frequencies were estimated by the expectation-maximization (EM) method, by use of the Arlequin (Schneider et al. 2000) or SNPalyze program (Dynacom, Mobara, Japan). The two methods gave near-identical results. Lewontin's coefficient D' and r^2 (Hill and Robertson 1968) were estimated as described (Nakajima et al. 2002).

LD measure by *AIC*

In the comparison of nonhierarchical models, Akaike's information criterion (*AIC*) (Akaike 1974) was used, $-2\log(\text{maximum likelihood}) + 2k$, where k is the number of free parameters estimated in the model. When the independent model (IM; linkage equilibrium) and dependent model (OM; linkage disequilibrium) are compared, the parsimonious model is the one with the smaller *AIC* score.

AIC(LD) as defined in this manuscript can be written as follows:

$$AIC(LD) = AIC(IM) - AIC(DM)$$

AIC with linkage disequilibrium, *AIC(LD)*

There are four categories, and the probability that a single trial chooses each category is p_{11} , p_{12} , p_{21} , p_{22} , where $p_{22} = 1 - p_{11} - p_{12} - p_{21}$. Trials are repeated n times and the observed frequencies in the four categories are a , b , c , d , where $d = n - a - b - c$.

$L_{\text{dependent}}$ is the likelihood function for three parameters (p_{11} , p_{12} , p_{21}) given the observed data. This is the likelihood of the observed data when linkage disequilibrium is present.

$$L_{\text{dependent}} = p_{11}^a p_{12}^b p_{21}^c p_{22}^d = p_{11}^a p_{12}^b p_{21}^c (1 - p_{11} - p_{12} - p_{21})^d \quad (1)$$

We define $l_{\text{dependent}}$ as the natural log of $L_{\text{dependent}}$:

$$l_{\text{dependent}} = a \log p_{11} + b \log p_{12} + c \log p_{21} + d \log(1 - p_{11} - p_{12} - p_{21}).$$

By partially differentiating $l_{\text{dependent}}$ with regard to p_{11} , p_{12} , p_{21} and setting the three formulas obtained equal to 0, we easily get the following solution:

$$p_{11} = \frac{a}{n}, p_{12} = \frac{b}{n}, p_{21} = \frac{c}{n} \quad (2)$$

These are the maximum likelihood estimates of the three parameters.

The maximum likelihood will be obtained as a function of a , b , c , and d by substituting these values into equation (1).

$$L_{\text{max,dependent}} = \left(\frac{a}{n}\right)^a \left(\frac{b}{n}\right)^b \left(\frac{c}{n}\right)^c \left(\frac{d}{n}\right)^d \quad (3)$$

AIC(DM) is defined as equal to $-2\log L_{\text{max,dependent}} + 6$

AIC without linkage disequilibrium, *AIC(IM)*

Under the independent condition, the number of independent parameters becomes two instead of three. The number of independent parameters can be reduced by solving the following equation:

$$p_{11}p_{22} = p_{11}(1 - p_{11} - p_{12} - p_{21}) = p_{12}p_{21} \quad (4)$$

By solving this equation, we get

$$p_{21} = \frac{p_{11}(1 - p_{11} - p_{12})}{p_{11} + p_{12}} \quad (5)$$

Therefore, the likelihood function under the independent condition should be

$$L_{\text{independent}} = p_{11}^a p_{12}^b \left(\frac{p_{11}(1 - p_{11} - p_{12})}{p_{11} + p_{12}} \right)^c \times \left(\frac{p_{12}(1 - p_{11} - p_{12})}{p_{11} + p_{12}} \right)^d \quad (6)$$

By partially differentiating $\log L_{\text{independent}}$ with regard to p_{11} , p_{12} , and solving the two equations obtained after setting the formulas equal to 0, we get

$$p_{11} = \frac{(a+b)(a+c)}{n^2}, p_{12} = \frac{(a+b)(b+d)}{n^2} \quad (7)$$

Table 1. Informations of 39 single-nucleotide polymorphisms (SNPs) on chromosome 21q22

Gene name	Nucleotide position NT_011512	SNP	Frequency	Marker name	JSNP ID(JST) abSNP(NCBI) ID (rs)	F-primer	R-primer
HUNK	18951435	C/T	0.34	2F/54	JST030955	tagcatctgactggaaacaagcg	actagatacttgttctttccceggg
	18951724	A/T	0.19	2F/345		tagcatctgactggaaacaagcg	actagatacttgttctttccceggg
	18952196	A/C	0.35	3F/79		gatgtcaacaacagatcaagcgg	cagatgtctggaaagctcccttaac
	19228432	C/A	0.05	3F309		cttcagctgacagccctgaggt	ggcagcgtcactcaggaaggt
	19271910	C/T	0.38	130		Bio-cccgtatctcactggtggg	tgcactctcctccctaccc
	19283272	T/C	0.05	12F/84		actggcagccctgacaactg	gctgcaacataatctcttggg
	19285532	G/A	0.21	7F/140		ggatttcaggaggagtgca	ggagggttgggagatagctt
	19286958	G/A	0.34	5F/382		cgggatcaggagggaagt	attcccccccccaaatgta
	19522155	C/A	0.31	6F/229		cccccttaggaaggacttt	caccatttatctccgtgc
	19526247	A/G	0.33	1F/352		cattttctttgtaaaatagctct	ttaacaattgatggtttcaagg
	19526263	A/C	0.15	1F/368		cattttctttgtaaaatagctct	ttaacaattgatggtttcaagg
	19550956	C/T	0.13	12F/233		ctcaagctagcaggggacag	ttttacagaggcagtaagtgg
	19652717	C/T	0.48	1F/244		tgtttttaaaggaaaggtgtga	tctcagcagacaaggcatalga
	19756077	T/C	0.08	8F134		atgctctcattgcttagca	ggggcttctcttgtggttta
PRKCBP2	19800149	T/C	0.48	C-P(C21orf62-PRKCBP2)NO1-5	caatattctgtctggaccaa	gacctgctgggggaaaagtaa	
	19800231	C/T	0.35	C-P(C21orf62-PRKCBP2)NO1-5	caatattctgtctggaccaa	gacctgctgggggaaaagtaa	
	19842333	G/A	0.46	C-P(C21orf62-PRKCBP2)NO2-6	tggattgtggtaacagcaga	gggcttagagggtgtgacca	
	19842969	C/A	0.44	C-P(C21orf62-PRKCBP2)NO2-6	tggattgtggtaacagcaga	gggcttagagggtgtgacca	
	19886173	C/T	0.34	C-P(C21orf62-PRKCBP2)NO3-3	gcgggttgaatgcacagtgtg	ctactctccattgggttgaat	
	19929976	A/G	0.40	C-P(C21orf62-PRKCBP2)NO4-2	aaggatgtgctggactcac	cggaggaaaggaggagaaagg	
	19975634	A/G	0.14	PRKCBP2	gagcccgatgacattttct	cgtggcgatcttggaaag	
	20021465	G/A	0.07	P-I(PRKCBP2-IFNAR2)NO1-6	agaccagctgtttggagagc	cggctactaccaacaacaaa	
	20063926	A/G	0.38	P-I(PRKCBP2-IFNAR2)NO2-2	ctgggcttgggatactagt	gaatcatttggcattcagc	
	20109792	A/G	0.12	P-I(PRKCBP2-IFNAR2)NO3-5	ggggcaacaacaccctct	caggcaaatagagactgtga	
IFNAR2	20153134	G/A	0.20	P-I(PRKCBP2-IFNAR2)NO4-3	gggtcgcccttgacctctaa	gatggatttggccctgtgtt	
IFNAR2	20180527	A/C	0.37	h	JST056061	tggctacataattgacaatgg	Bio-ccaacctctctccaattaccctg
IFNAR2	20190902	G/T	0.45	i	rs1051393	caggctctctattctgtgt	Bio-ttggcttccacttaactctgac
IFNAR2	20197349	T/C	0.09	j'	JST009668	tagaccaggtatcactgagct	aatcacaactctgggtgttcaa
IFNAR2	20197448	T/C	0.14	j		Bio-tagaccaggtatcactgagct	aatcacaactctgggtgttcaa
IL10RB	20220729	T/G	0.47	c	JST034428	catattaccttataatggcgggag	Bio-ttggcttccacttaactctgac
IL10RB	20245394	A/G	0.21	d	rs1058867	ctctttccattgtcggatgag	acatcaagaatggcaactagagc
IL10RB	20246028	G/A	0.21	d'		ctctttccattgtcggatgag	acatcaagaatggcaactagagc
IFNAR1	20300309	C/A	0.42	e	JST031526	cagacaacagctatgctacc	Bio-aaacaccttgggtcagctct
IFNAR1	20303978	G/A	0.38	f	JST053186	Bio-ttggcattttaacctatactgga	tcagtggtctacaaaacgattat
IFNAR1	20303978	G/A	0.38	g	rs1041868	gagtggttaggagcaaaaggagtt	Bio-aatgtaagagctgtcccgta
IFNGR2	20363960	T/C	0.48	a	JST005184	Bio-cgaagattcgctgtgacacg	atgaaagagcagagataatgca
IFNGR2	20386341	A/G	0.16	b	rs1059293	Bio-ggctggtatctgaactggtga	ttagaaagcgtgaacactctgcag
GART	20483341	T/G	0.31	1	JST056070	tgtactctccagatactagctgc	taacacatggggagcgtctcg
SON	20517711	G/A	0.22	4	JST010747	cggagcagctctctggaacagc	tggataattctgacagcactcg

These are the maximum likelihood estimates under the independent condition, and they are expressed as $\hat{p}_{11}$ and $\hat{p}_{12}$.

The maximum likelihood will be obtained by substituting equation (7) into equation (6)

$$L_{max, independent} = \left(\frac{(a+b)(a+c)}{n^2} \right)^a \left(\frac{(a+b)(b+d)}{n^2} \right)^b \times \left(\frac{(a+c)(c+d)}{n^2} \right)^c \left(\frac{(b+d)(c+d)}{n^2} \right)^d \quad (8)$$

$AIC(IM)$ is defined as equal to $-2\log L_{max, independent} + 4$

Likelihood ratio criterion

Therefore, from equations (3) and (8), Pearson's likelihood ratio criterion will be

$$LRC = L_{max, independent} / L_{max, dependent} = \left(\frac{(a+b)(a+c)}{na} \right)^a \left(\frac{(a+b)(b+d)}{nb} \right)^b \times \left(\frac{(a+c)(c+d)}{nc} \right)^c \left(\frac{(b+d)(c+d)}{nd} \right)^d. \quad (9)$$

This can be written as

$$LRC = \prod_{i=1}^4 \left(\frac{e_i}{x_i + e_i} \right)^{x_i + e_i}, \quad (10)$$

where

$$e_1 = \frac{(a+b)(a+c)}{n}, e_2 = \frac{(a+b)(b+d)}{n}, e_3 = \frac{(a+c)(c+d)}{n}, e_4 = \frac{(b+d)(c+d)}{n}, \quad (11)$$

and $x_1 = a - e_1$, $x_2 = b - e_2$, $x_3 = c - e_3$, $x_4 = d - e_4$

The relationship between $AIC(LD)$ and LRC can be written as follows:

$$\begin{aligned} AIC(LD) &= AIC(IM) - AIC(DM) \\ &= -2\log L_{max, independent} + 4 \\ &\quad - (-2\log L_{max, dependent} + 6) \\ &= -2\log \frac{L_{max, independent}}{L_{max, dependent}} - 2 \\ &= -2\log LRC - 2 \end{aligned}$$

Therefore,

$$AIC(LD) = -2\log LRC - 2, \quad (12)$$

and

$$LRC = e^{-\frac{1}{2}AIC(LD)-1} \quad (13)$$

Relation between Pearson's χ^2 statistic and $-2\log LRC$

From equation (10),

$$-2\log LRC = -2 \sum_{i=1}^4 \log \left(\frac{e_i}{x_i + e_i} \right)^{x_i + e_i} = 2 \sum_{i=1}^4 \log \left(1 + \frac{x_i}{e_i} \right)^{x_i + e_i} \quad (14)$$

By applying Taylor's expansion,

$$\begin{aligned} -2\log LRC &= 2 \sum (x_i + e_i) \\ &\quad \times \left(\frac{x_i}{e_i} - \frac{1}{2} \left(\frac{x_i}{e_i} \right)^2 + \frac{1}{3} \left(\frac{x_i}{e_i} \right)^3 - \dots \right) \\ &= 2 \left(\sum \frac{x_i^2}{e_i} - \frac{1}{2} \sum \frac{x_i^3}{e_i^2} + \frac{1}{3} \sum \frac{x_i^4}{e_i^3} - \dots \right. \\ &\quad \left. + \sum x_i - \frac{1}{2} \sum \frac{x_i^2}{e_i} + \frac{1}{3} \sum \frac{x_i^3}{e_i^2} - \dots \right) \end{aligned} \quad (15)$$

Since $\sum e_i = (a+b+c+d)$, $\sum x_i = 0$

Therefore,

$$-2\log LRC = \sum \frac{x_i^2}{e_i} - \frac{1}{3} \sum \frac{x_i^3}{e_i^2} + \dots \quad (16)$$

Since x_i/e_i is sufficiently small compared to 1 when n is large and e_i , $i = 1, 2, 3, 4$ is not small (i.e., when expected frequencies in the cells are not small),

$$-2\log LRC \approx \sum \frac{x_i^2}{e_i} \quad (17)$$

This is simply Pearson's χ^2 statistic for goodness of fit, because e_i is the frequency in a cell expected from the marginal frequencies and x_i is the difference between the expected frequency and the observed frequency (i.e., $a - e_1$, etc).

Therefore, equation (17) can be written as

$$-2\log LRC \approx \frac{(a+b+c+d)(bc-ad)^2}{(a+b)(a+c)(b+d)(c+d)}, \quad (18)$$

which is well known as Pearson's χ^2 statistic for testing the goodness of fit on a 2×2 table.

Relation between r^2 and $-2\log LRC$

r^2 is defined as follows,

$$r^2 = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{(p_{11} + p_{12})(p_{21} + p_{22})(p_{11} + p_{21})(p_{12} + p_{22})} \quad (19)$$

This value is estimated by the maximum likelihood estimation where p_{11} , p_{12} , p_{21} are three independent parameters.

The maximum likelihood estimates for the parameters p_{11} , p_{12} , p_{21} under the dependent condition are given in equation (2). Since r^2 is a function of p_{11} , p_{12} , p_{21} as indicated by equation (19), the maximum likelihood estimate of r^2 can be obtained simply by substituting equation (2) into (19) as follows:

$$\hat{r}^2 = \frac{(bc - ad)^2}{(a + b)(a + c)(b + d)(c + d)} \quad (20)$$

Therefore,

$$\hat{r}^2 \approx \frac{-2\log LRC}{n} \quad (21)$$

Fig. 1. D' and average $AIC(IM) - AIC(DM)$ values are expressed in a bar graph and compared. D' values were divided into five groups (0 – 0.2; 0.2 – 0.4; 0.4 – 0.6; 0.6 – 0.8; and 0.8 – 1.0), and then the average $AIC(IM) - AIC(DM)$ value at each D' group was calculated. AIC, Akaike's information criterion; IM, independent model; OM, dependent model

Fig. 2A,B. Comparison of $AIC(IM) - AIC(DM)$ with r^2 and $nr^2 - 2$. LD estimates by r^2 (A) and $nr^2 - 2$ (B) were compared with $AIC(IM) - AIC(DM)$ and the correlation coefficient (R^2) is shown. AIC, Akaike's information criterion; IM, independent model; DM, dependent model; LD, linkage disequilibrium

Fig. 3A–D. The relationship between $AIC(IM) - AIC(DM)$ and r^2 was evaluated by comparing them with D' and SNP distance, respectively. Comparison between $AIC(IM) - AIC(DM)$ and D' (A), $AIC(IM) - AIC(DM)$ and SNP distance (B), r^2 and D' (C), and r^2 and SNP distance (D) are illustrated. AIC, Akaike's information criterion; IM, independent model; DM, dependent model

Thus,

$$\hat{r}^2 \approx \frac{AIC(LD) + 2}{n}, \quad (22)$$

and

$$AIC(LD) \approx nr^2 - 2, \quad (23)$$

when n is large.

Results and discussion

Thirty-nine SNPs on chromosome 21q22 were genotyped among 192 Japanese subjects and all the possible pairwise LDs were estimated and compared between AIC and the standard LD measures. LD measured by AIC was expressed as $AIC(IM) - AIC(DM)$, denoted as $AIC(LD)$, as computed in “subjects and methods”. A positive value of $AIC(LD)$ reflects linkage disequilibrium, and a negative value represents linkage equilibrium. D' and average $AIC(LD)$ values were expressed in a bar graph and compared (Fig. 1), which showed that the average $AIC(LD)$

value increased gradually as the value of D' increased. Because $AIC(LD)$ was negative for pairs showing D' values < 0.2 , the threshold of LD by D' could be set at 0.2. The relationship between $AIC(LD)$ and r^2 was evaluated (Fig. 2A), demonstrating that $AIC(LD)$ and r^2 were tightly correlated [correlation coefficient (R^2) = 0.9654]. Because the data points close to the origin may inflate the correlation coefficient, the data points apart from the origin ($r^2 > 0.2$) were extracted. Then, the relationship with $AIC(LD)$ was calculated revealing a sustained high relationship ($R^2 = 0.8501$) (data not shown). As indicated by the likelihood evaluation in “subjects and methods”, the relationship between $AIC(LD)$ and $nr^2 - 2$ was examined and resulted in a high correlation coefficient ($R^2 = 0.9803$) (Fig. 2B). The relationship between $AIC(LD)$ and r^2 was evaluated by comparing D' and SNP distance, respectively. Correlations between LD values, measured by AIC and r^2 , and D' values (Fig. 3A, C) and SNP distances (Fig. 3B, D) were investigated. $AIC(LD)$ was well correlated with D' values ($R^2 = 0.3777$), while $AIC(LD)$ was very sensitive to the SNP distances. Because the similar patterns were observed for the r^2 measure (Fig. 3C, D), and it is reported that r^2 is an appropriate LD measure for fine mapping (Nakajima et al. 2002), $AIC(LD)$ could be considered as an

alternative *LD* measure in gene-mapping of complex diseases. Because the *LD* estimate by *AIC* is not constrained to the interval 0 to 1, a direct comparison of *AIC(LD)* values between data from different sample sizes is difficult. Presumably, this shortcoming of *AIC(LD)* will be overcome by the accumulation of sufficient empirical data in the near future.

Acknowledgments This work was supported in part by a Research for the Future Program Grant of The Japan Society for the Promotion of Science (II). Devoted efforts by Dr. Naoyuki Kamatani to prepare the formula of *AIC* are profoundly appreciated by all the authors. We thank Ms. Yuka Terada for formatting the manuscript, and we are also grateful to Dr. Hiroyuki Takasaki (Okayama University of Science) for inspiring the initial idea of *AIC*.

References

- Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions Automatic Control*, AC-19, 716–723
- Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 3:299–309
- Collins FS (1995) Positional cloning moves from perditional to traditional. *Nat Genet* 9:347–350
- Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theor Appl Genet* 38:226–231
- Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes *Genome Res* 10:1435–1444
- Landegren U, Nilsson M, Kwok PY (1998) Reading bits of genetic information: methods for single-nucleotide polymorphisms. *analysis. Genome Res* 8:769–776
- Lander ES (1996) The new genomics: global views of biology. *Science* 274:536–539
- Nakajima T, Jorde LB, Ishigami T, Umemura S, Emi M, Lalouel JM, Inoue I (2002) Nucleotide diversity and haplotype structure of the human angiotensinogen gene in two populations. *Am J Hum Genet* 70:108–123
- Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–1517
- Ronalgi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* 242:84–89
- Ronalgi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. *Science* 281:363–365
- Schneider S, Kueffer J-M, Roesslerie D, Excoffier L (2000) Arlequin: a software for population genetic data analysis. University of Geneva, Geneva
- Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280: 1077–1082