BMC Proceedings

Proceedings

Open Access

Novel approach for genome scan meta-analysis of rheumatoid arthritis: a kernel-based estimation procedure

Laurent Briollais*[1,2], Gilles Durrieu[3] and Ranodya Upathilake[1]

Address: [1]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 60 Murray Street, Toronto, Ontario MT5 3L9, Canada, [2]Public Health Sciences Department, University of Toronto, 155 College Street, Toronto, Ontario M5T 3M7, Canada and [3]Université Bordeaux 1 and Laboratoire GEMA, UMR CNRS 5805-EPROC, Place du Dr. Peyneau, Arcachon 33120, France

Email: Laurent Briollais* - laurent@mshri.on.ca; Gilles Durrieu - g.durrieu@epoc.u-bordeaux1.fr; Ranodya Upathilake - ranodya@gmail.com

* Corresponding author

from Genetic Analysis Workshop 15
St. Pete Beach, Florida, USA. 11–15 November 2006

Published: 18 December 2007

BMC Proceedings 2007, **1**(Suppl 1):S96

This article is available from: http://www.biomedcentral.com/1753-6561/1/S1/S96

© 2007 Briollais et al; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Genome scan meta-analysis (GSMA) can prove very useful in detecting genetic effects too small to be detected in an individual linkage study and can also lead to more consistent results. In this paper, we propose a new kernel-based estimation procedure for GSMA. Instead of estimating identity by descent between markers, as performed in interval mapping approaches, we estimated directly the nonparametric linkage score between markers using a kernel procedure. The GSMA is then extended to take into account the kernel estimate of the nonparametric linkage score and its variance at a given chromosomal position. The method is applied to the rheumatoid arthritis genome scan data (Genetic Analysis Workshop 15 Problem 2).

Background

Rheumatoid arthritis (RA) is a chronic inflammatory disease that primarily affects the synovial tissues of multiple joints in the body. The etiology of the disease remains unknown, but it appears to have a complex genetic component. Several genome scans for RA studies have been performed to identify susceptibility loci, but most of the results have not been replicated [1]. These inconsistencies could arise from the small sample size, low statistical power, and clinical or genetic heterogeneity of these studies. Genome scan meta-analysis (GSMA) that combines the results from several linkage studies can have greater statistical power to detect small genetic effects and can lead to more consistent results. A general difficulty in GSMA is the heterogeneity across studies due to different marker maps, marker informativeness, sample sizes, sampling plans, and linkage tests. Loesgen et al. [2] proposed a meta-analytic test that computed a weighted average estimate of score statistics. Recently, Etzel et al. [3] used this method in a genome-wide meta-analysis of RA. Because of differences in marker maps across studies, they decided to align the marker maps after performing some interval mapping and combining nonparametric linkage (NPL) scores obtained from GeneHunter2 for markers in a pre-specified interval. Their method requires the estimation of identity-by-descent (IBD) sharing probabilities through the interval between two markers [4,5], which can be somewhat inaccurate and imprecise. The variability of the IBD estimate is difficult to measure and often not reflected in the GSMA. In this paper, we propose an alter-

native approach that estimates the NPL score between markers directly using a kernel-based estimation procedure. The GSMA is then extended to take into account the kernel estimate of the NPL score and its variance at a given chromosomal position.

Methods

Data

We have included three linkage studies in the meta-analysis of RA: NARAC (North American Rheumatoid Arthritis Consortium), ECRAF (European Consortium on Rheumatoid Arthritis Families) and United Kingdom (UK). Only microsatellite scans and RA affection status (binary outcome) were considered. NARAC has performed microsatellite scans for 511 mutliplex families, decomposed into 757 smaller families. ECRAF had high-density microsatellite data from 88 families, including 105 sib pairs typed with 1089 microsatellite markers. The UK group performed two screens: an initial screen of the entire genome using 369 markers analyzed on 175 families and a second screen performed on 197 families using 89 markers in regions showing evidence for linkage. The two screens were combined in our analyses. A summary of the data analyzed is given in Table 1.

GSMA

We first performed a multipoint linkage analysis of each individual study and estimated the NPL score and marker information content (IC) at each marker location as well as at each systematic 2-cM interval using the program MERLIN [6]. We then performed the GSMA by calculating the weighted average of the NPL scores at a given chromosomal position [2]:

$$Z_{MA_j} = \frac{\sum_{i=1}^{k} w_{ij} Z_{ij}}{\sqrt{\sum_{i=1}^{k} w_{ij}^2}},$$

where Z_{ij} is the NPL score from the i^{th} study at the j^{th} position, k is the number of studies, and w_{ij} is the weight given to each study. To perform the GSMA, we used three different strategies that differ in the way the NPL score and IC are estimated between markers and the definition of the weight.

Method 1

Following Etzel et al. [3], the first method tries to align the marker maps. After the 2-cM interval mapping was completed with MERLIN, the NPL scores that were within 1 cM of each other were combined and the statistic ZMA was computed in each interval. The weight w_{ij} is the product of the number of sib-pairs equivalents (SPE) from the i^{th} study and the IC estimated from MERLIN for the i^{th} study at the j^{th} interval:

$$w_{ij} = SPE_i * IC_{ij}.$$

Method 2

This approach is not based on marker alignment. Instead of using an interval mapping estimate of the NPL score and IC between markers, we used a kernel regression method. The statistic ZMA is then computed at all marker positions available after merging the three data sets (i.e., if one marker is present in one study but missing in the other two, its associated NPL score and IC are estimated by the kernel regression). The weight is identical to Method 1 except that the IC is now replaced by its kernel estimate (ICK):

$$w_{ij} = SPE_i * ICK_{ij}.$$

Method 3

The third approach is identical to Method 2 but now takes into account in the weight the precision of the kernel estimator, more precisely the inverse of standard deviation of NPL kernel estimator (*SDnpl*):

$$w_{ij} = SPE_i * ICK_{ij} * I/SDnpl_{ij}.$$

Model

In Methods 2 and 3, the relationship between the NPL score (or the information content) (Y) and the marker location (T) is modelled using a nonparametric model given by:

$$Y_i = m(T_i) + \varepsilon_i, \text{ for } i = l,..., n,$$

where $m(\cdot)$ and ε denote, respectively, the regression function to be estimated and the model error; term n is the number of observations. The stochastic distribution $f(\cdot)$

Table 1: Summary of studies included in the meta-analysis

		No. of families			
Study	Population	Total	2 siblings	>2 siblings	No. of microsatellite markers
NARAC	U.S. Caucasian	757	208	535	396
ECRAF	French	88	16	72	1089
UK	U.K. Caucasian	372	158	213	369

of ε is typically unknown and is unlikely to follow any familiar distribution such as the normal distribution. Hence, we decided to use nonparametric statistics. The random variable enables to characterize the variation of Y around $m(t)$, the mean regression curve with:

$$m(t) = E(Y / T = t) = \frac{\int \gamma f(\gamma, t) d\gamma}{f(t)}.$$

So, the regression function $m(\cdot)$ depends on the joint and the marginal densities, which are both unknown. A density estimator allows the analysis of data sets that could exhibit skewness and multimodality due to different factors (for example, mixture of several distributions and clusters). A histogram type estimator is the most frequently used but could be strongly affected by the choice and number of classes chosen. So we decided to use a nonparametric kernel estimator that behaves much better statistically. A kernel estimator of a function $f(\cdot)$ is defined by:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - T_i}{h} \right),$$

where n is the sample size, h is the bandwidth (the smoothing parameter) to be determined, and K is the continuous fixed kernel function with finite variance generally satisfying $K > 0$, $K(-t) = K(t)$, and $\int K(t)\, dt = 1$. Here we considered the Gaussian kernel. This raises the question of determining the bandwidth parameter. The estimator $\hat{f}_h(.)$ is wiggly when h is small and very flat when h is large. Different procedures have been previously proposed to determine h, for example cross-validation. The problem in our application is that we need to estimate the NPL score function at different marker locations using the kernel procedure but also its variance. Both the kernel estimator and the variance depend on the same smoothing parameter and an optimal choice for the kernel estimator might not be optimal for the variance. To our knowledge, there is no optimal procedure for this problem. For that reason, we could not apply the classical cross-validation procedure, so we decided to choose the bandwidth empirically. The bandwidth was chosen inversely proportional to the number of markers of each individual study on each chromosome. Therefore, h was not constant in our study but depended on the genetic background. More exactly, we chose: $h = cst * \sum_{i=1..k} M_i / M_i$, where M_i is the number of markers of each individual linkage study on one particular chromosome and the constant was fixed to 4.0. Intuitively, we understand that a study with less markers yields more variable results. Applying a larger h leads to a smoother function and thus to a decreased variability. This determination of h provided a good estimation of both the NPL score function and its variance.

Using the kernel estimator of the marginal and joint densities function, the regression estimator of $m(\cdot)$ is:

$$\hat{m}_h(t) = \frac{\sum_{i=1}^{n} K_h(t - T_i) Y_i}{\sum_{i=1}^{n} K_h(t - T_i)},$$

where $K_h(\cdot)$ was chosen to be the Gaussian kernel function here. This form of the estimate of the regression curve was proposed by Nadaraya [7] and Watson [8]. The estimator $\hat{m}_h(.)$ is a consistent estimator $m(t)$ and normally distributed when h tends toward 0. It is also shown [9,10] that under regularity conditions when h tends toward 0, the variance of the estimator $\hat{m}_h(.)$ can be approximated by:

$$\hat{V}ar\left[\hat{m}_h(t) \right] \approx \frac{1}{nh} \frac{\sigma^2(t)}{f(t)} \|K\|_2^2, \text{ where } \|K\|_2^2 = \int K^2(t) dt$$

and $\hat{\sigma}^2(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_h(t - T_1)}{\hat{f}_h(t)} \left(Y_i - \hat{m}_h(t) \right)^2$ is the estimator of $\hat{\sigma}^2(t)$.

In our method 3 above, we took $SDnpl = \sqrt{Var\left[\hat{m}_h(t) \right]}$.

All our computations were preformed with the computer program *R* for Linux.

Results

The results of the Z_{MA} test statistic on the whole genome are presented in Figure 1. Our results confirm the role of the HLA region in the susceptibility to RA with a Z_{MA} test statistic close to 8.0 on chromosome 6 (Fig. 1). The three methods gave consistent results for this chromosome. We also found some suggestion of linkage on chromosomes 1 (240 to 260 cM), 2 (200 to 250 cM), 8 (10 cM), 16 (40 cM), 18 (80 to 90 cM), and 21 (45 cM), with a Z_{MA} test statistic close to 2.0. For most chromosomes, the three methods gave very similar results. Method 2 performs identically to Method 1, and this was expected because the kernel estimation is only used to smooth the information content and NPL score functions. Some differences, however, were observed between Method 3 and the two others, especially on chromosomes 2, 3, 13, 21, and 22. To

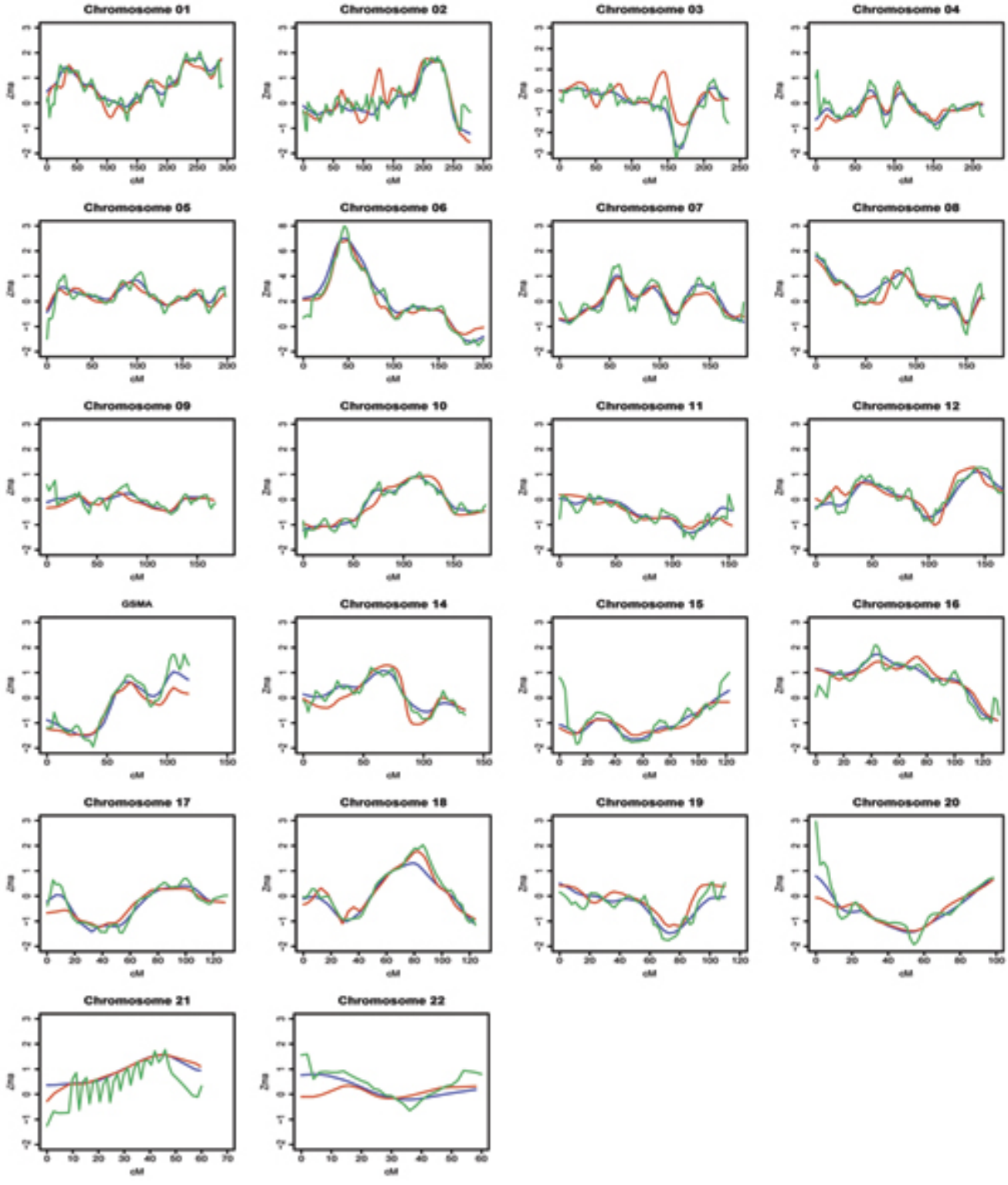

Figure 1
GSMA results on the 22 autosomes. Green line, method 1; blue line, method 2; red line, method 3.

better understand these discrepancies, we describe the GSMA results on chromosome 13, where Methods 1 and 2 gave a larger Z_{MA} test statistic than Method 3 at the location 120 cM (Fig. 2A). The linkage signal was stronger in ECRAF (NPL score > 2.0) than in NARAC and UK studies (NPL score close to 0) (Fig. 2B) and the information content of ECRAF was also larger (Fig. 2C).

However the variance of the NPL score at this location, as estimated by the kernel regression procedure, was higher for ECRAF than for the two other studies (Fig. 2D). Method 3, unlike Methods 1 and 2, weights each study inversely proportionally to this variance and therefore led to a lower Z_{MA} test statistic. Moreover, the peak of linkage in ECRAF is relatively thin, which could be associated with a larger variance of the kernel estimator at this location (Fig. 2B). This is because the variance of the kernel estimator is inversely proportional to the density estimate of the NPL score at one particular location (see variance formula above). In general, denser marker regions and wider peak regions both could contribute to a low variance of the kernel estimator and hence, to a larger GSMA statistic.

Discussion

The use of kernel-based regression methods allow us to estimate the NPL score function at various locations along

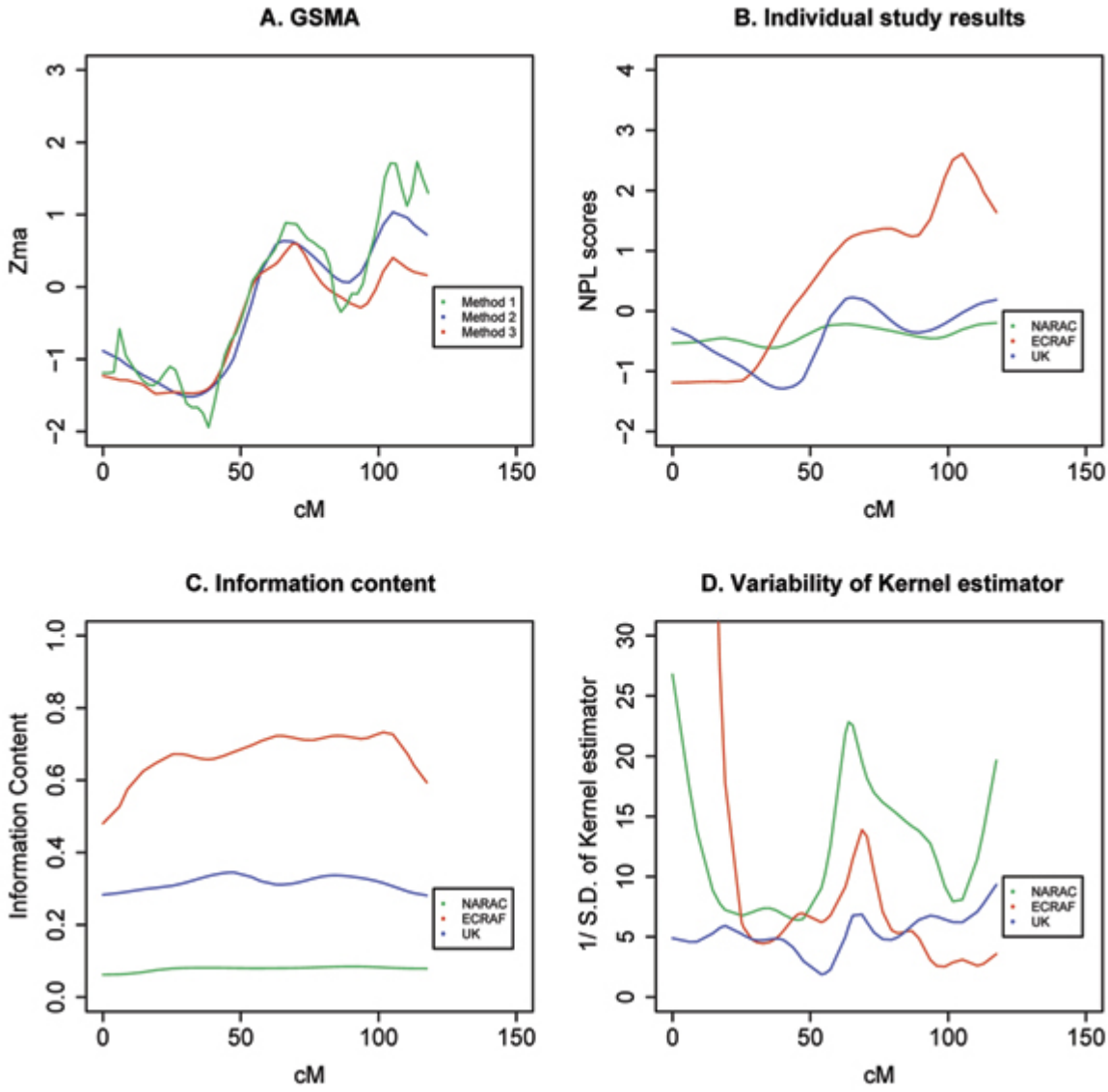

Figure 2
GSMA results on chromosome 13.

the genome and thus make possible the meta-analysis of several linkage studies with different genetic maps. To our knowledge, this is the first kernel-based approach for GSMA studies. Previous GSMAs have tried to perform some map alignment that requires an estimation of the IBD sharing probabilities between markers using interval mapping. However, the variability of this estimate is not reflected in the GSMA statistic. An important advantage of our approach is that it is completely nonparametric and we can obtain a measure of the variability of the NPL score estimate along the genome. Incorporating this variability into the GSMA statistic (Method 3) might improve the consistency of linkage results by over-weighting studies with more precise estimate of the NPL score function. This could reflect, for example, a higher marker density. A larger weight will be given to a study that finds a linkage peak with many markers than to a study that finds the same peak with fewer markers. Therefore, the information about NPL score variability is very useful to weight each individual study. Our procedure can also down-weight thin peaks. Further simulation studies are needed to better understand its properties, in particular in terms of detection of true linkage peaks.

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

This article has been published as part of *BMC Proceedings* Volume 1 Supplement 1, 2007: Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/1?issue=S1.

References

1. Choi SJ, Rho YH, Ji JD, Song GG, Lee YH: **Genome scan meta-analysis of rheumatoid arthritis.** *Rheumatology* 2006, **45:**166-170.
2. Loesgen S, Dempfle A, Golla A, Bickboller H: **Weighting schemes in pooled linkage analysis.** *Genet Epidemiol* 2001, **21(Suppl 1):**S142-S147.
3. Etzel CJ, Chen WV, Shepard N, Jawaheer D, Cornelis F, Seldin MF, Gregersen PK, Amos CI for the North American Rheumatoid Arthritis Consortium: **Genome-wide meta-analysis for rheumatoidarthritis.** *Hum Genet* 2006, **119:**634-641.
4. Fulker DW, Cardon LR: **A sib-pair approach to interval mapping of quantitative trait loci.** *Am J Hum Genet* 1994, **54:**1092-1103.
5. Olson JM: **Multipoint linkage analysis using sib pairs: an interval mapping approach for dichotomous outcomes.** *Am J Hum Genet* 1995, **56:**788-798.
6. Abecasis GR, Cherny SS, Cookson WOC, Cardon LR: **MERLIN-rapid analysis of dense genetic maps using sparse gene flow trees.** *Nat Genet* 2002, **30:**97-101.
7. Nadaraya EA: **On estimating regression.** *Theory Probability Appl* 1964, **10:**186-190.
8. Watson GS: **Smooth regression analysis.** *Sankhy Ser A* 1964, **26:**359-372.
9. Härdle W: *Applied Nonparametric Regression Cambridge: Cambridge University Press*; 1990.
10. Wand MP, Jones MC: *Kernel Smoothing London: Chapman and Hall*; 1995.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral