Hum Genet
DOI 10.1007/s00439-006-0278-y

ORIGINAL INVESTIGATION

Residual linkage: why do linkage peaks not disappear after an association study?

Scott Gordon · Peter M. Visscher

Received: 1 August 2006 / Accepted: 5 October 2006

Abstract Family-based candidate gene and genome-wide association studies are a logical progression from linkage studies for the identification of gene and polymorphisms underlying complex traits. An efficient way to analyse phenotypic and genotypic data is to model linkage and association simultaneously. An important result from such an analysis is whether any evidence for linkage remains after fitting polymorphisms at candidate genes (residual linkage), because this may indicate locus and allelic heterogeneity in the population and will influence subsequent molecular strategies. Here we report that substantial residual linkage is to be expected, even under genetic homogeneity and when the underlying causal polymorphisms are genotyped and fitted in the model. We simulated a powerful design to detect linkage to quantitative trait loci, with 5, 10 or 20 causal SNPs spread throughout the genome. These SNPs were responsible for all genetic variation, and hence for both linkage and association. Residual linkage at the largest linkage peak from a genome-wide scan was substantial, with mean LOD scores of 0.4, 0.7, and 1.4 for the case of 5, 10 and 20 underlying causal SNPs, respectively. For less powerful designs, the proportion of the original LOD scores that remains after association will be even larger. All cases of 'significant' residual linkage are false positives. The reason for the apparent paradox of detecting residual linkage after fitting causal polymorphisms is that the linkage signals at the largest peaks in a genome-scan are severely inflated, even if all peaks correspond to true linkage. Our findings are general and apply to linkage mapping of any phenotype and to any pedigree structure.

S. Gordon · P. M. Visscher (✉)
Queensland Institute of Medical Research,
300 Herston Road, Herston, 4029 Brisbane, Australia
e-mail: peter.visscher@qimr.edu.au

Introduction

Many linkage studies have been performed in the last decade, for rare and common diseases and quantitative traits. The usual paradigm for follow-up is to focus on the chromosome regions with the strongest evidence for linkage and perform a fine mapping and candidate gene study. Recently the prospect of whole genome association studies has become feasible because of the abundance of well characterised SNPs and high-density genotyping platforms, so that all candidate regions can be simultaneously investigated for association.

Researchers who have spent time and effort to collect family data for linkage studies are likely to use the same collections for association analysis. The statistical analysis of family based association studies takes into account that there are two sources of information contained in the pedigrees: the cosegregation of genotypes and phenotypes within families (linkage) and the association between genotypes and phenotypes across families (association). A commonly asked question in such analyses is ''does the association explain my linkage peak?'' (Almasy et al. 2005; Ichikawa et al. 2005; Kammerer et al. 2004; Soria et al. 2000; Zhu et al. 1999), and a number of statistical tests have been proposed to address this (Li et al. 2004, 2005; Sun et al. 2002). In this report we define linkage after association as 'residual linkage'. There are a number of reasons why association between phenotypes and genotypes will not eliminate

the linkage peak. If there are multiple alleles (or haplotypes) at the same susceptibility locus in the population that cause the observed linkage peak then association between a single marker (haplotype) and phenotype will not explain all of the variation. Multiple linked genes in the same broad linkage region and different families segregating for causal variants at different genes will have the same effect. In the absence of such heterogeneity, the linkage peak may not disappear because of imperfect linkage disequilibrium between the genotyped markers and an ungenotyped causal variant. Finally, if the initial linkage peak was wholly spurious then true association will reduce the test statistic by reducing the amount of residual variation but will not eliminate it.

In this study, we show that there is an additional simple reason why linkage peaks are unlikely to disappear, even when there is a single genotyped causal variant that is responsible for the linkage peak. The reason is that linkage peaks for complex traits are generally artificially inflated in terms of their apparent effect size. We quantify the expected amount of residual linkage as a function of the number of polymorphisms that underlie genetic variation in the population, and show that it is substantial.

Methods

Simulations

We simulated a powerful linkage study for a quantitative trait, with 5,000 sibling pairs, 738 autosomal microsatellite markers and a heritability of 0.8. For each replicated sample, microsatellite marker genotypes were simulated on the parents of 5,000 sibling pairs, using the locations of markers and their allele frequencies from actual markers that have been used for linkage studies in our laboratory (Cornes et al. 2005; Morley et al. 2006; Visscher et al. 2006). The average marker spacing was 4.7 cM. Markers were simulated to be in linkage equilibrium in the population of parents. Gametes from the parents and genotypes of the sibling pairs were simulated assuming Haldane's mapping function and random mating. Markers were simulated on all parents and all progeny, i.e., there were no missing data.

Genetic variation in each replicated sample was simulated by creating 5, 10 or 20 independent causal SNPs that each explained an equal proportion of the genetic variance, i.e., 16, 8 and 4% of the phenotypic variance, respectively. Hence, the amount of genetic variance in the population, and corresponding linkage and association, was entirely determined by the simulated SNPs. The simulated samples lack heterogeneity, and are examples of best-case scenarios for linkage and association. A polymorphic SNP with alleles B and b was simulated by randomly choosing one of the microsatellites and assigning allele B to the most common microsatellite allele and allele b to all other microsatellite alleles. In this way, the appropriate segregation within families and the correct amount of linkage disequilibrium between the SNP and the microsatellite from which it was simulated is achieved. Hence, we simulated each SNP in complete linkage and in complete linkage disequilibrium with a randomly chosen microsatellite. The simulated multiple SNPs were required to be at least 50 cM distant from each other, to avoid the creation of a 'super locus' that would explain more variance than was intended. Additive gene action at each of the 5, 10 or 20 SNPs was simulated, and the effect size of SNP i was $a_i = \sqrt{[h^2/(2mp_i(1-p_i))]}$, with h^2 the total additive heritability of the trait (=0.8), m the number of SNPs (5, 10 or 20) and p_i the frequency of allele B. The phenotype for individual k in the sample was simulated as

$$Y_k = \sum [x_i a_i] + e_k,$$

with x_i an indicator variable for SNP i, which is –1, 0, 1 for genotypes bb, Bb and BB, respectively. The residual e was drawn from a standard normal distribution, $e \sim N(0,1-h^2)$, and no other sources of family resemblance were simulated. Phenotypes were only simulated on the sibling pairs.

For a fully informative marker, the expected LOD score for linkage at the location of a causal SNP is 0.6, 1.7 and 6.0, for loci explaining 4, 8 and 16% of the phenotypic variance, respectively (Purcell et al. 2003). The expected LOD score for association (which is not the focus of this study) is much larger, 83.5, 172.2 and 359.6, respectively, for a SNP explaining 4, 8 and 16% of the phenotypic variance (Purcell et al. 2003).

Statistical analyses

Variance component linkage analyses were performed using Merlin (Abecasis et al. 2002). For each simulated data set, two genome-scans were performed. In the first genome-scan, a test statistic for linkage was calculated at all 738 marker locations, fitting an average sibling (polygenic) effect and a quantitative trait locus (QTL) effect in the model. The chromosome with the largest test statistic was selected and the causal SNP closest to the peak location identified. In the second joint linkage and association genome scan, this causal SNP was fitted

in the model of analysis as a covariate (with one degree of freedom), in addition to the family and QTL effects. The rationale for this sequence of analyses is that it mimics a linkage analysis followed by a positional candidate gene study. The test statistic was a standard likelihood-ratio-test (LRT) from comparing a full model, which included random QTL and polygenic effects and a reduced model, which excluded the QTL effects. The test statistic for residual linkage was also a LRT, fitting the fixed effect of the SNP in both the full and reduced models. Hence we are testing the effect of the QTL after fitting the SNP, not the effect of the SNP. Under the null hypothesis of no residual linkage, the asymptotic distribution of the LRT statistics is a 50:50 mixture of zero and a χ^2 with one degree of freedom (Self and Liang 1987). LRT statistics were converted to LOD scores. The 5% significance threshold for the mixture distribution is a LRT statistic of 2.71, which is equivalent to a LOD score of 0.59.

For each of the primary linkage analyses, the test statistics pertaining to the peak location and the proportion of variance explained at the peak location were recorded, as were the estimates at the location of the nearest causal SNP. For the secondary joint linkage and association analysis, the test statistic and proportion of variance explained by linkage were recorded at the initial peak location and at the location of the nearest causal SNP. If there was no SNP within 50 cM of the peak location, then the peak was attributed to a false positive signal.

Results

Tables 1 and 2 summarise the results for analyses at the linkage peak location (Table 1) and at the nearest causal SNP (Table 2). Clearly, the residual linkage effects are substantial, both at the peak location and at the location of a causal SNP. At the peak location, the residual linkage test statistic (LOD) was 0.4, 0.7 and 1.4, for 5, 10 and 20 simulated causal SNPs, respectively. These test statistics are 4, 16 and 47% of the mean test statistic for linkage when the nearest causal SNP was not fitted. Hence, in the case of 20 causal SNPs of equal effect, the test statistic for linkage reduces, on average, by only one half when the true causal SNP is identified and fitted in the model of analysis. In addition to the large average residual linkage test statistics, the standard deviations are also large. For example, the SD of the LOD score at the location of a causal SNP nearest to the largest linkage peak is 0.5 when 10 causal SNPs were simulated (Table 2), so that many residual test statistics exceed a LOD of, say, 1.5. This would most likely lead to the

Table 1 LOD score and proportion of variance explained by linkage at the location of the genome-wide linkage peak, without and with fitting the nearest causal SNP

# Causal SNPs	Linkage analysis		Residual linkage after association		
	LOD	h^2(%)	LOD	h^2(%)	False positive rate (%)[a]
5	10.0 (2.3)	22.4 (2.7)	0.4 (0.4)	3.7 (2.4)	22
10	4.5 (1.2)	15.0 (2.1)	0.7 (0.6)	5.8 (2.2)	48
20	2.9 (0.6)	12.0 (1.5)	1.4 (0.5)	8.2 (1.4)	98

Mean (SD) of 50 simulated replicate genome scans

[a] Proportion of residual LOD scores larger than 0.59

Table 2 LOD score and proportion of variance explained by linkage at the location of the causal SNP nearest to the genome-wide linkage peak, without and with fitting the nearest causal SNP

# Causal SNPs	Linkage analysis		Residual linkage after association		
	LOD	h^2(%)	LOD	h^2(%)	False positive rate (%)[a]
5	9.9 (2.3)	22.4 (2.7)	0.3 (0.4)	3.5 (2.5)	20
10	4.4 (1.3)	14.8 (2.3)	0.6 (0.5)	5.3 (2.0)	40
20	2.3 (1.1)	10.5 (3.2)	0.8 (0.5)	6.0 (2.8)	68

Mean (SD) of 50 simulated replicate genome scans

[a] Proportion of residual LOD scores larger than 0.59

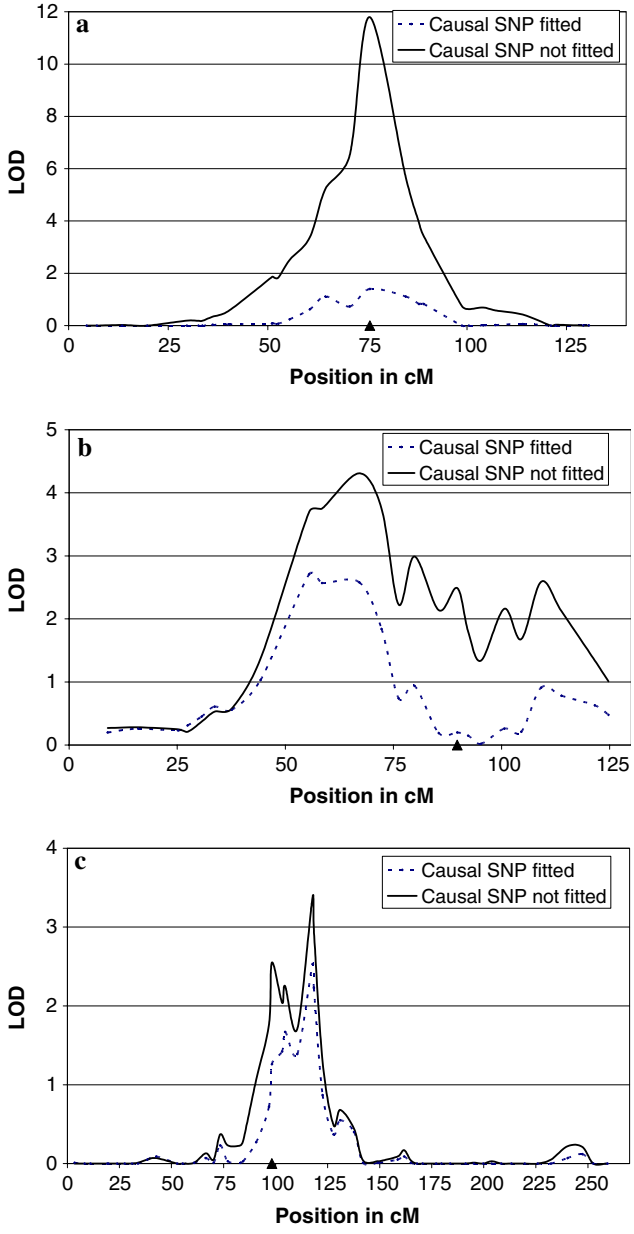

Fig. 1 LOD score for linkage before and after fitting the causal SNP. Selected examples of one simulated replicate of 5 SNPs (**a**), 10 SNPs (**b**) and 20 SNPs (**c**). *Triangles* denote the location of the causal SNP. The three examples correspond to simulated chromosomes 17, 13 and 2, respectively

erroneous conclusion that there are additional polymorphisms in the linkage region that require discovery.

At the location of the linkage peak, the LOD scores for residual linkage was significant at the 5% level, i.e., exceeding the critical value of 0.59, for 5, 10 and 20 simulated causal variants in 22, 48 and 98% of replicates, respectively (Table 1). At the position of the causal variants, these proportions were 20, 40 and 68%, respectively (Table 2). The proportion of replicates significant at the causal SNPs closest to the linkage peak for 20 simulated SNPs was 'only' 68% because in a number of replicates the test statistic for linkage (without fitting the SNP) was not significant. These results demonstrate that the false positive rate for residual linkage is extremely large.

To obtain 50 genome scans that resulted in the linkage peak being within 50 cM of a simulated SNP, 63 samples needed to be simulated for the case of 20 simulated SNPs. Hence, the false positive rate of the linkage peaks from the genome scan was 13/63 = 21%. The 13 false positive replicates were discarded, but in reality such false peaks may be pursued, with obviously no success in identifying causal variants.

Figure 1 give examples of chromosome-wide linkage and residual linkage for 5 (Fig. 1a), 10 (Fig. 1b) and 20 (Fig. 1c) causal SNPs. For each Figure, we selected the most extreme simulated data set (1 out of 50, or 2%) in terms of the magnitude of the residual linkage test statistic. These results, in particular for 10 and 20 simulated SNPs (Fig. 1b, c), would most likely lead to the incorrect conclusion that there are other variants to be identified in the region of interest.

Discussion

We have shown that a substantial amount of residual linkage is to be expected in a combined linkage-association scan and that testing for residual linkage will lead to a high false positive rate. We have quantified the amount of residual linkage for a fairly powerful linkage analysis of quantitative traits under genetic homogeneity, i.e., under a best-case scenario. Most linkage studies for quantitative traits are severely underpowered, and this causes the apparent effect sizes at linkage peak to be over-estimated more than in our study and will result in even larger amounts of residual linkage. Smaller studies with reduced power will also result in more false positive linkage findings, but that is outside the focus of this study. Hence, researchers should expect that association at candidate genes will not eliminate their observed linkage, even if they have genotyped causal variants that were responsible for the linkage signal in the first place. The explanation of the results is straightforward, and is a direct consequence of performing multiple tests in a genome scan and focussing on the largest test statistic (Beavis 1994; Goring et al. 2001; Lynch and Walsh 1998). If the true proportion of variance explained at a trait locus is 10% but the estimate of this proportion at linkage peak is 25%, then after fitting a causal variant there will be an apparent linkage peak that explains 15% of the

variance. This residual linkage is a false positive. The bias in the estimate of residual linkage occurs because the focus of genome-wide scans is on detection of trait loci and not on (unbiased) estimation of their effect size.

Although we demonstrated the magnitude of residual linkage for a quantitative trait using a variance component analysis and a sibling pair design, the same principle applies to linkage studies for disease using any family design. For example, for a design in which sibling pairs affected with a disease have been ascertained and applying recently suggested joint linkage and association analysis methods (Li et al. 2005; Sun et al. 2002), excess residual IBD sharing is likely to be found after fitting a causal SNP if only the largest peaks from linkage scans are followed up. In a recent review of the genetics of schizophrenia it was noted that ''In the most convincing cases, the risk haplotypes appear to be associated with small effect sizes and do not fully explain the linkage findings that prompted each study'' (Norton et al. 2006). Our study suggest that at least part of this observation is due to the selection bias that we investigated in this study.

The reason why the residual LOD score is so large can be quantified approximately by considering the distribution of the test statistic at the location of a causal variant. The asymptotic distribution of the LRT statistic for linkage is a non-central χ^2 with non-centrality parameter NCP and one degree of freedom. It has a mean and variance of 1 + NCP and 2(1 + 2NCP), respectively. The LOD score is LRT/4.605. Hence, the expected value of the LOD score (ELOD) is (1 + NCP)/4.605, and the variance of the LOD score is [18.42ELOD –2]/4.605^2. So, for expected LOD scores of 1.0, 2.0 and 3.0 at a particular location, the standard deviation across causal variants (and across replicate studies) is about 0.9, 1.3 and 1.6, respectively. This large standard deviation (relative to the mean) explains why the test statistic at the largest peak is usually much larger than its expected value. It also illustrates the difficulties in drawing inference about replication and non-replication across studies. Because of the large variance of the non-central χ^2 distribution, residual linkage is to be expected even if we had a very powerful design and linkage was only performed at the true locations of multiple causal variants.

If there is no selection among true causal loci from the linkage scan then the estimates of the proportion of variance explained from residual linkage should be unbiased. We verified this by averaging the LOD score for residual linkage for all five SNP locations when five causal SNPs were simulated. The mean LOD score for residual linkage across replicates was 0.10 (SE = 0.02). The expected unbiased value of the residual LOD score is 0.11, because under the null hypothesis of no residual linkage, the LRT has an expected value of 0.5 (and hence a LOD of 0.5/4.605 = 0.11). Hence, there is no excess residual linkage when there is no selection among true causal loci (in our case all causal loci were selected).

How can researchers obtain unbiased estimates of residual linkage? The obvious design would be to have an independent replication sample, in which joint linkage and association parameters can be estimated without multiple testing. However, this implies that both linkage and association will be replicated in other samples. Replicated linkage is difficult to assess, because of the low power of linkage studies to detect trait loci of moderate size and the resulting large confidence intervals of the estimated location of susceptibility loci (Visscher and Goddard 2004). Association studies suffer less from these considerations because they are generally more powerful. If association studies are performed in pedigrees around candidate regions that have been reported independently, then residual linkage is expected to be small if causal variants (or variants in high linkage disequilibrium with them) are genotyped. This is indeed what has been observed in practice (Kammerer et al. 2004; Soria et al. 2000). Therefore, one suggestion is to focus on replicated associations, in particular for polymorphisms which are likely to be functional, as, for example, the polymorphism associated with age-related macular degeneration (Edwards et al. 2005; Haines et al. 2005; Klein et al. 2005), and not to be concerned about the size of the residual linkage peak.

Sun et al. (2002) proposed a statistical method to identify (causal) polymorphisms that explain a linkage signal in an affected sibling pair or trio design, and suggested a simulation approach to adjust the *P*-values for residual linkage when evidence for linkage is only 'suggestive' to avoid the over-estimation of significant residual linkage as described in this study. The proposed simulation is conditional on the observed genotypes and under the null hypothesis that the SNP is the sole cause of excess sharing in the region. Hence, the effect and gene action of the SNP are assumed known when in practice they are only estimated. Simulation results are stored for only those replicates for which the linkage result exceeds suggestive evidence for linkage. It is not obvious how to apply this method to other pedigree designs, trait distributions and analysis methods, and to our knowledge the method proposed by Sun et al. (2002) is not implemented in widely used statistical genetics software tools. Nevertheless, these authors clearly recognised the potential effect of underpowered linkage studies on their proposed test statistic.

The conclusion from this study is that, for complex traits, residual linkage is the norm rather than the exception when the same pedigrees are used for genome-wide linkage and subsequent association for complex traits, even for powerful study designs.

Acknowledgments We thank Naomi Wray, Nick Martin, Dale Nyholt and Grant Montgomery for helpful discussions and comments on the manuscript, and two referees for helpful suggestions. This study was supported by NHMRC grants 389892 and 389891, and NIH grants AA13326-01, AA13446-03, MH66206-01A1 and AA007728.

References

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

Almasy L, Rainwater DL, Cole S, Mahaney MC, Vandeberg JL, Hixson JE, Stern MP, MacCluer JW, Blangero J (2005) Joint linkage and association analysis of the hepatic lipase promoter polymorphism and lipoprotein size phenotypes. Hum Biol 77:17–25

Beavis WD (1994) The power and deceit of QTL experiments: lessons from comparative QTL studies. In: 49th annual corn and sorghum industry research conference 49th annual corn and sorghum industry research conference. American Seed Trade Association, Washington, pp 250–266

Cornes BK, Medland SE, Ferreira MA, Morley KI, Duffy DL, Heijmans BT, Montgomery GW, Martin NG (2005) Sex-limited genome-wide linkage scan for body mass index in an unselected sample of 933 Australian twin families. Twin Res Hum Genet 8:616–632

Edwards AO, Ritter R III, Abel KJ, Manning A, Panhuysen C, Farrer LA (2005) Complement factor H polymorphism and age-related macular degeneration. Science 308:421–424

Goring HH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69:1357–1369

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308:419–421

Ichikawa S, Koller DL, Peacock M, Johnson ML, Lai D, Hui SL, Johnston CC, Foroud TM, Econs MJ (2005) Polymorphisms in the estrogen receptor beta (ESR2) gene are associated with bone mineral density in Caucasian men and women. J Clin Endocrinol Metab 90:5921–5927

Kammerer CM, Gouin N, Samollow PB, VandeBerg JF, Hixson JE, Cole SA, MacCluer JW, Atwood LD (2004) Two quantitative trait loci affect ACE activities in Mexican-Americans. Hypertension 43:466–470

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389

Li C, Scott LJ, Boehnke M (2004) Assessing whether an allele can account in part for a linkage signal: the Genotype-IBD Sharing Test (GIST). Am J Hum Genet 74:418–431

Li M, Boehnke M, Abecasis GR (2005) Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. Am J Hum Genet 76:934–949

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland

Morley KI, Medland SE, Ferreira MA, Lynskey MT, Montgomery GW, Heath AC, Madden PA, Martin NG (2006) A possible smoking susceptibility locus on chromosome 11p12: evidence from sex-limitation linkage analyses in a sample of Australian twin families. Behav Genet 36:87–99

Norton N, Williams HJ, Owen MJ (2006) An update on the genetics of schizophrenia. Curr Opin Psychiatry 19:158–164

Purcell S, Cherny SS, Sham PC (2003) Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 19:149–150

Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610

Soria JM, Almasy L, Souto JC, Tirado I, Borell M, Mateo J, Slifer S, Stone W, Blangero J, Fontcuberta J (2000) Linkage analysis demonstrates that the prothrombin G20210A mutation jointly influences plasma prothrombin levels and risk of thrombosis. Blood 95:2780–2785

Sun L, Cox NJ, McPeek MS (2002) A statistical method for identification of polymorphisms that explain a linkage result. Am J Hum Genet 70:399–411

Visscher PM, Goddard ME (2004) Prediction of the confidence interval of quantitative trait loci location. Behav Genet 34:477–482

Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet 2:e41

Zhu G, Duffy DL, Eldridge A, Grace M, Mayne C, O'Gorman L, Aitken JF, Neale MC, Hayward NK, Green AC, Martin NG (1999) A major quantitative-trait locus for mole density is linked to the familial melanoma gene CDKN2A: a maximum-likelihood combined linkage and association analysis in twins and their sibs. Am J Hum Genet 65:483–492