Human Reproduction Vol.22, No.3 pp. 717–728, 2007 doi:10.1093/humrep/del446
Advance Access publication December 11, 2006.

Significant evidence of one or more susceptibility loci for endometriosis with near-Mendelian inheritance on chromosome 7p13–15

Krina T.Zondervan[1,9], Susan A.Treloar[3,4], Jianghai Lin[2,5], Daniel E.Weeks[6], Dale R.Nyholt[3,4], Jon Mangion[7], Ian J.MacKay[7], Lon R.Cardon[1], Nicholas G.Martin[3,4], Stephen H.Kennedy[2], Grant W.Montgomery[3,4] and Study Group[8]

[1]Wellcome Trust Centre for Human Genetics, and [2]Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Oxford, UK, and [3]Cooperative Research Centre for Discovery of Genes for Common Human Diseases, Melbourne, and [4]Queensland Institute of Medical Research, Brisbane, Australia, and [5]College of Life Sciences, Sun Yat-sen University, Guangzhou, China, and [6]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, UK, and [7]Oxagen, Abingdon, UK and [8]Study Group: Queensland Institute of Medical Research, Brisbane, Australia: Jacqueline Wicks, Brandon J.Wainwright, and Anjali Henders; Oxagen, Abingdon, UK: Delilah Zabaneh, Gary Dawson, Vicki Smith, Alisoun Carey, and Simon T.Bennett; Queensland Endometriosis Research Institute, Brisbane, Australia: Daniel T.O'Connor; Nuffield Department of Obstetrics and Gynaecology, University of Oxford, UK: David Barlow, and Ann Lambert; Australian Genome Research Facility, Melbourne, Australia: Kelly R.Ewen-White.

[9]To whom correspondence should be addressed at: E-mail: krinaz@well.ox.ac.uk

BACKGROUND: Endometriosis is a common disease with a heritable component. The collaborative International Endogene Study consists of two data sets (Oxford and Australia) comprising 1176 families with multiple affecteds. The aim was to investigate whether the apparent concentration of cases in a proportion of families could be explained by one or more rare variants with (near-)Mendelian autosomal inheritance. METHODS AND RESULTS: Linkage analyses (aimed at finding chromosomal regions harbouring disease-predisposing genes) were conducted in families with three or more affected (Oxford: $n = 52$; Australia: $n = 196$). In the Oxford data set, a non-parametric linkage score (Kong & Cox (K&C) Log of ODds (LOD)) of 3.52 was observed on chromosome 7p (genome-wide significance $P = 0.011$). A parametric MOD score (equal to maximum LOD maximized over 357 possible inheritance models) of 3.89 was found at 65.72 cM (D7S510) for a dominant model with reduced penetrance. After including the Australian data set, the non-parametric K&C LOD of the combined data set was 1.46 at 57.3 cM; the parametric analysis found an MOD score of 3.30 at D7S484 (empirical significance: $P = 0.035$) for a recessive model with high penetrance. Critical recombinant analysis narrowed the probable region of linkage down to overlapping 6.4 Mb and 11 Mb intervals containing 48 and 96 genes, respectively. CONCLUSIONS: This is the first report to suggest that there may be one or more high-penetrance susceptibility loci for endometriosis with (near-)Mendelian inheritance.

Key words: chromosome 7/endometriosis/family study/linkage

Introduction

Endometriosis (MIM 131200) is a complex disease caused by a combination of genetic and environmental factors, and is increasingly recognized as a major women's health issue (Berkley *et al.*, 2005). It is characterized by the presence of endometrial-like tissue in sites outside the uterus, mainly the pelvic peritoneum, ovaries and rectovaginal septum (Giudice and Kao, 2004), causing severe dysmenorrhoea, dyspareunia, chronic pelvic pain and subfertility. The diagnosis can only be made reliably by visualizing the pelvis during surgery; histological confirmation is often—but not always—obtained as well. The general population prevalence is uncertain: estimates range from 1 to 10% because of the lack of both a non-invasive diagnostic tool and a consistent phenotypic definition (Eskenazi and Warner, 1997). Familial aggregation has firmly been established in humans (Zondervan *et al.*, 2002; Stefansson *et al.*, 2001) and non-human primates (Zondervan *et al.*, 2004), with heritability in humans estimated to be ~52% (Treloar *et al.*, 1999).

Endometriosis produces peritoneal inflammation and fibrosis and can lead to the formation of ovarian cysts and pelvic adhesions. Disease severity tends to be measured on a four-point scale using the revised American Fertility Society (rAFS) classification system (The American Fertility Society, 1985),

based on the size and location of peritoneal lesions and the presence of any endometriomas/adhesions. The rAFS system presupposes that endometriosis is a single, progressive disease; other classification systems, which distinguish- for example- between peritoneal and ovarian forms, imply that different phenotypes exist, possibly with different aetiologies (Koninckx *et al.*, 1999). The origin of endometriomas remains controversial, but it is generally accepted that peritoneal deposits arise from retrograde menstruation, the transport of viable endometrial cells through the Fallopian tubes into the pelvic area during menses (Sampson, 1927). As the process occurs to some extent in 90% of women (Halme *et al.*, 1984), research has focused on the apparent increased ability of endometrial cells in women with endometriosis to escape immunological surveillance, adhere to the peritoneal surface, proliferate and—in some cases—invade the surrounding tissues. Consequently, genetic variants involved in cell adherence and matrix degradation (Shan *et al.*, 2005), cell proliferation (Kado *et al.*, 2002), angiogenesis (Bhanoori *et al.*, 2005), and immunological dysfunction (Ishii *et al.*, 2003; Vigano *et al.*, 2003) have been studied as disease-susceptibility candidates.

It is unlikely, however, that the candidate gene approach will ultimately be productive given how little detailed knowledge exists about the underlying, aberrant cellular and molecular mechanisms in endometriosis. An alternative approach, a positional cloning strategy (the identification of disease-susceptibility genes through linkage analysis), has been adopted by the collaborative International Endogene Study (IES) (Treloar *et al.*, 2002). Linkage analyses aim to find chromosomal regions that are shared among multiple cases in families more often than expected on the basis of independent Mendelian inheritance and that may therefore harbour disease-predisposing genes. The IES brought together two independent groups, the UK-based Oxford Endometriosis Gene (OXEGENE) Study and the Australian Genes Behind Endometriosis Study. From 1995 to 2002, these two groups collected data on 1176 families (UK = 245; Australian = 931) containing at least two affected members with surgically confirmed endometriosis for the purpose of positional cloning. The sample size was sufficiently large to detect linkage to genetic loci of modest effect (sibling recurrence risk ratio $\lambda_s \geq 1.3$, compared with the general population) with 80% power (Treloar *et al.*, 2002). The results of the full genome-wide linkage scan, in which one significant linkage signal on chromosome 10q26 and several suggestive signals were found, were published recently (Treloar *et al.*, 2005).

The present report discusses findings of linkage analyses on families with three or more affecteds that were pursued in the Oxford data set and subsequently in the Australian data set. The aim of these analyses was to investigate whether a rare genetic variant with a (near-) Mendelian segregation pattern might be responsible for the apparent concentration of closely related cases in a proportion of families, similar to, for example, the scenario of BRCA1 and BRCA2 in breast cancer (Miki *et al.*, 1994; Wooster *et al.*, 1995).

Materials and methods

The data set

Family collection and recruitment protocols for each of the studies have been described in detail elsewhere (Treloar *et al.*, 2002). Briefly, in the Oxford study, Caucasian women with endometriosis who had affected sisters or other affected relatives were identified by collaborating clinicians in Belgium, Britain, Ireland, Russia and the USA. In addition, affected relative pairs were recruited through the media, European and American Endometriosis Associations and the OXEGENE website (www.medicine.ox.ac.uk/ndog/oxegene/oxegene.htm). In the Australian study, affected relative pairs of Caucasian origin were recruited using similar methods and by accessing the twin databases of the Queensland Institute of Medical Research. In both studies, disease severity was assessed from the surgical records using the rAFS classification system (The American Fertility Society, 1985). Since records did not always allow an accurate distinction between Stages I and II, or III and IV, these were regrouped into Stage 'A' and Stage 'B', respectively. Both studies also collected information on self-reported age at symptom onset and first diagnosis, and details of associated symptoms, such as pelvic pain (as a symptom underlying endometriosis diagnosis) and subfertility (inability to conceive for at least 12 months).

In total, 1176 families containing at least two affected members with surgically confirmed endometriosis were recruited in the two studies (Oxford = 245; Australian = 931) (Treloar *et al.*, 2005)). Three further families had been recruited in the Oxford study since the genome-wide linkage analysis. The present analysis was limited to families containing more than the minimum number of two cases per family required for inclusion in the IES, i.e. three or more affecteds (Oxford: $n = 52$; Australia: $n = 196$). Such increase in the size of the sampling unit can achieve dramatic improvements in statistical power and location estimation (Williams *et al.*, 1997; Williams and Blangero, 1999). More importantly, families containing larger numbers of affected members may be more genetically homogenous, which is of particular relevance in the context of a complex trait when investigating the presence of a relatively rare segregating variant.

Linkage analyses

Linkage analyses can be separated into 'non-parametric' and 'parametric' analyses. Non-parametric linkage (NPL) analysis makes no assumptions about the underlying inheritance model of the putative disease-causing genetic variant. The method works by calculating the inherited genetic allele sharing between cases within families and summing this across all families (Ott, 1999). Markers that show a higher degree of allele sharing among cases than expected on the basis of independent inheritance provide increased evidence for linkage. Allele sharing is expressed as an NPL score (either calculated using all affected members—NPL_{all}—or using affected sibpairs only—NPL_{pairs}). Since NPL scores as evidence for linkage can be conservative when genotype information is incomplete (because not all pedigree members are typed), Kong & Cox (K&C) recommended transformation of NPL scores to Log of ODds (LOD) scores (Kong and Cox, 1997), referred to as 'K&C' LOD scores. A LOD score is a likelihood ratio statistic used to represent evidence for linkage, and it has traditionally been used as the outcome of interest in parametric linkage studies. These require a number of parameters relating to the inheritance of the disease to be specified, e.g. allele frequency of unknown disease allele, disease risk (penetrance) of heterozygotes and homozygotes and penetrance of non-carriers (phenocopy rate). Although more powerful in detecting linkage when the model specified is close to being correct, parametric methods can lose power

when compared with non-parametric methods when the parameters included in the model are incorrect.

Single point linkage analysis only provides evidence of the linkage of each marker with the disease trait; multipoint methods use the expected recombination rates from a specified marker map to infer evidence of linkage in positions between markers (Ott, 1999). For the present article, we used both multipoint and single point analyses; however, since conclusions from the two analyses were similar, only multipoint results are presented. NPL analysis uses affecteds only, whereas parametric linkage analysis uses both affecteds and unaffecteds. All parametric analyses presented in this article were run twice, assigning both 'unaffected' and 'unknown' disease status to known unaffecteds, because negative surgical findings do not completely exclude the possibility that a woman might have small lesions or might develop the disease in the future. However, re-running the analyses in this way did not materially influence the results.

The data analysed consisted of microsatellite (genetic variants that are highly polymorphic) genotyping information obtained as part of the main genome-wide linkage scan (Treloar *et al.*, 2005). Chromosomal distance between the microsatellites was sufficiently large to ensure independent inheritance of the markers in the general population (i.e. they were in 'linkage equilibrium'). Pedigree errors had already been checked extensively in the main linkage scan but were re-checked for genotypes resulting in unlikely recombinations (which could indicate genotyping errors) using Merlin v.1.0-alpha, based on sex-specific information from the most recent Rutgers linkage-physical microsatellite marker map (Kong *et al.*, 2004).

Both groups continued to pursue independent analyses after the collaborative genome-wide scan. Initial analyses for the present article were therefore conducted in the Oxford data set only. A genome-wide non-parametric multipoint linkage analysis of the 52 Oxford families was conducted using Merlin v1.0-alpha and Minx (Merlin in X) (Abecasis *et al.*, 2002). K&C LOD scores were calculated from NPL_{all} following a linear model (Kong and Cox, 1997). The significance of the results was assessed empirically in simulations in which genotypes in the 52 families were assigned randomly under the null hypothesis of no linkage, using gene-dropping algorithms implemented in Merlin (Abecasis *et al.*, 2002).

For subsequent parametric linkage analysis, a likely disease inheritance model was required. Segregation analysis (a method used to find the most likely disease inheritance model using disease information in a pedigree) could not be employed to assess this model reliably because it requires specification of the method of proband ascertainment, which varied because of the wide range of family recruitment methods employed. In addition, studies have shown that using parameters derived from segregation analyses, which look for the overall rather than locus-specific pattern of inheritance, can lead to a major reduction in the power to detect linkage at a particular disease locus (Dizier *et al.*, 1996). However, the maximized maximum LOD (MOD) score was found by iterating across a range of different inheritance models without prior assumptions (Hodge and Elston, 1994). With a fixed phenocopy rate (absolute risk of disease for non-carriers of the rare putative genetic variant) of 0.06 (close to the estimated population prevalence of endometriosis), 17 parameter values for the disease allele frequency were used: 0.0001, 0.0005, 0.001, 0.002, 0.004, 0.006, 0.008, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6 and 0.8. Six parameter values were used for both heterozygous and homozygous penetrance (absolute risk of disease for heterozygote and homozygote carriers, respectively): 0.1, 0.3, 0.5, 0.7, 0.9 and 1.0. Heterozygous penetrance was restricted to a value smaller than or equal to homozygous penetrance, giving 21 combinations for heterozygous and homozygous penetrance values. In total, this resulted in 357 disease models with different values for disease allele frequency and heterozygous and homozygous penetrance.

The main finding on chromosome 7 was subsequently followed-up by incorporating the Australian data set. NPL analyses were conducted on the 196 Australian families alone and combined with the 52 Oxford families by calculating K&C LOD scores. Subsequent parametric linkage models, using the same MOD score method as described earlier, were run using the combined data set. The significance of the parametric MOD score analysis was assessed empirically by data simulation, thus correcting for the multiple tests performed (Weeks *et al.*, 1990). Simulations had to reflect the methods used to obtain the results, which were derived after a genome-wide significant result in the Oxford data set was followed up by adding the second, Australian, data set and conducting an MOD score analysis on chromosome 7 iterating across 357 different disease models. Therefore, 1000 simulations were conducted by randomly generating genotypes for the Australian pedigrees under the null hypothesis of no linkage, using the gene-dropping algorithms implemented in Merlin (Abecasis *et al.*, 2002). Each simulated data set was added, in turn, to the original Oxford data set and, for each combined data set created, an MOD score analysis was run for chromosome 7 using the 357 parametric models. The simulations answered the question how often an MOD score greater or equal to the observed result would be expected by chance in the combined data set, conditional on the result already observed on chromosome 7 for the Oxford data.

Genetic heterogeneity (the likelihood that several disease-predisposing genetic variants at different chromosomal locations were present in the families) was investigated through the heterogeneity LOD (HLOD) and comparing the likelihoods of linkage models fitted under homogeneity and heterogeneity inheritance models using HOMOG (Ott, 1986).

Association analyses

Associations between family-based LOD scores and phenotypic characteristics were analysed using analysis of variance for unadjusted, and general linear modelling for adjusted, associations. The association between alleles of marker D7S484 and endometriosis was assessed using family-based association tests (Lake *et al.*, 2000).

Critical recombinant analysis

To narrow down the most likely region of linkage, haplotypes (allele combinations of adjacent markers on the same chromosome) were constructed for individuals in the pedigrees using the 'best estimate' maximum likelihood-based method in Merlin (Abecasis *et al.*, 2002). Haplotype configurations shared by affecteds within the families were compared to identify sites where recombinations were likely to have occurred (thus narrowing the chromosomal segment of interest).

Results

Analysis of the Oxford data set

Of the 52 Oxford families, 41 contained three affecteds; ten four affecteds, and one five affecteds. In the study, 28 of the families were recruited from the UK, 17 from the USA, 3 from Ireland, and 1 family each from Canada, Germany, Norway and Russia. Table I shows the phenotypic characteristics of the Oxford families.

Six K&C LOD score peaks >1 were found (Figure 1) in the genome-wide linkage analysis of the Oxford data set, with one of the peaks—on chromosome 7p—reaching a multipoint

Table I. Phenotypic characteristics of the Oxford versus Australian families

Mean (SD) phenotypic characteristic per family	Oxford families ($n = 52$)	Australian families ($n = 196$)	P-value[a]
Total number of individuals	9.67 (7.53)	7.54 (2.81)	**0.001**
Number of genotyped individuals	5.46 (1.59)	5.90 (2.81)	0.11
Number of generations	2.75 (0.68)	2.65 (0.59)	0.29
Number of affecteds	3.23 (0.47)	3.32 (0.59)	0.25
Kinship coefficient between affecteds	0.215 (0.094)	0.203 (0.092)	0.44
Number of affected relative pairs (%)			
Sib	96 (52.2)	392 (53.0)	
Half-sib	2 (1.1)	4 (0.5)	
Cousin	17 (9.2)	120 (16.2)	
Parent–child	43 (23.4)	116 (15.7)	
Grand-parent	2 (1.1)	1 (0.1)	
Avuncular	24 (13.0)	106 (14.3)	0.008[b]
Number of affected sisters per sibship	2.27 (0.84)	2.41 (0.68)	0.20
Number of members with Stage A disease[c]	1.21 (0.94)	2.19 (1.01)	<0.001
Number of members with Stage B disease[c]	1.87 (0.97)	1.11 (0.92)	<0.001
Age at onset	22.6 (6.5)	20.7 (4.9)	0.05
Age at diagnosis	30.5 (5.2)	28.4 (5.1)	0.009
Number with subfertility	1.13 (0.95)	1.30 (0.93)	0.25
Number with pelvic pain	2.54 (0.85)	2.53 (0.96)	0.93

[a]*T*-test.
[b]Fisher's exact test.
[c]Stage A, rAFS stage I or II; Stage B; rAFS stage III or IV.

K&C LOD of 3.52 (Figure 2A) at 63.6 cM Haldane (between D7S484 and D7S510). This result was genome-wide significant at $P = 0.011$. The parametric MOD score analysis (see Materials and Methods) produced an MOD of 3.89 at 65.72 cM (D7S510); the optimal model found was dominant with reduced penetrance (allele frequency = 0.04, phenocopy rate 0.06, heterozygous penetrance = 0.5, homozygous penetrance = 0.5).

Analysis of the combined Oxford–Australia data set

To follow-up the finding on chromosome 7, the analysis was extended to incorporate the 196 Australian families with three or more affecteds (three affecteds = 145, four affecteds = 40, five affecteds = 10 and six affecteds = 1). Table I shows the phenotypic characteristics of the Australian families compared with those in the Oxford study. Although the Oxford families were on average larger in size, there was no difference in the number of genotyped individuals, the number of affecteds or the average kinship coefficient between affecteds. However, the Oxford data set contained more affected mother–child pairs (23.4%) compared with the Australian data set (15.7%). In addition, the Oxford sample had a significantly greater number of more severely affecteds with rAFS Stages III and IV (Stage B) disease when compared with the Australian sample. The mean age of diagnosis in the Oxford sample was also significantly greater than in the Australian sample.

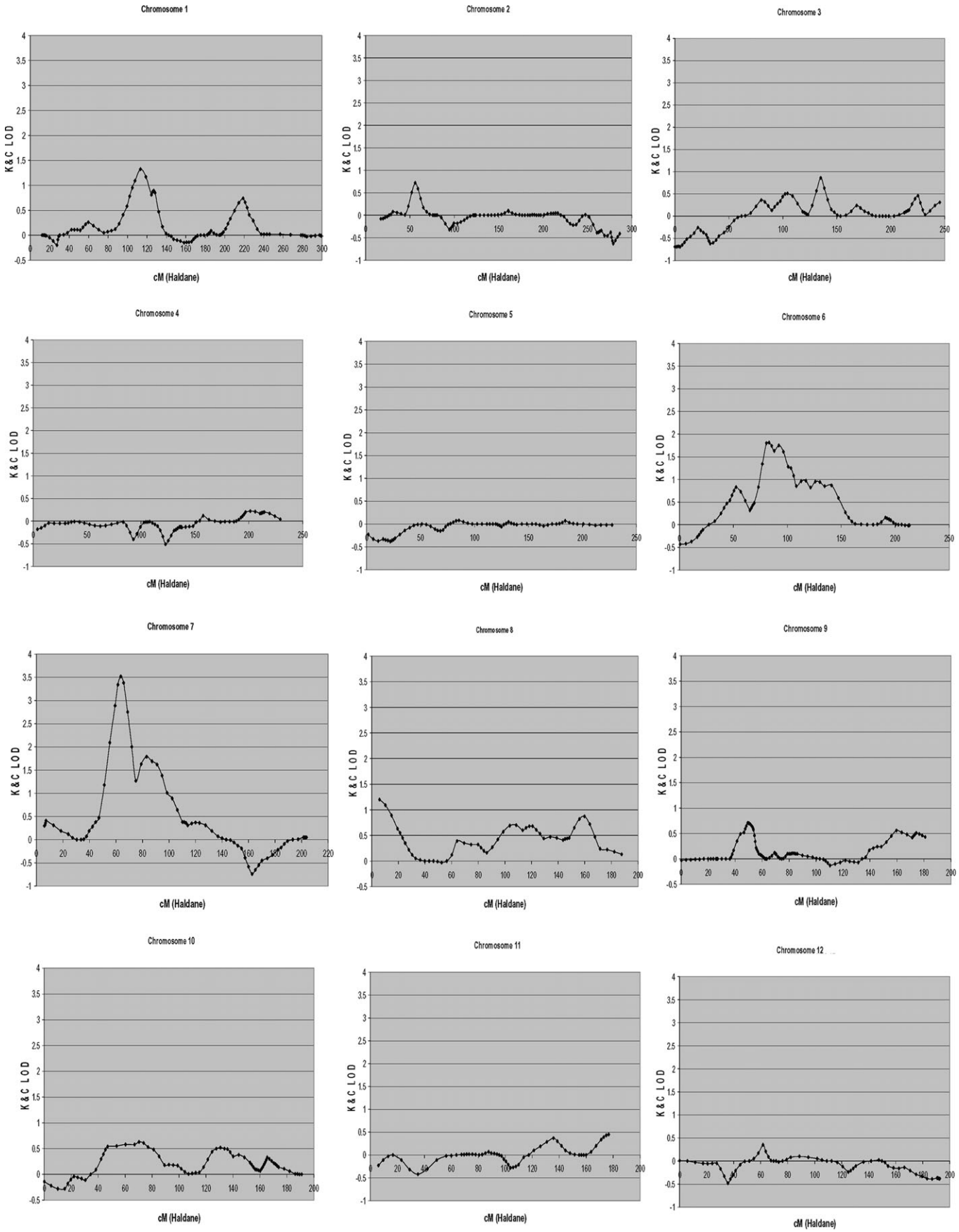
Figure 2A shows the multipoint non-parametric results on chromosome 7 for the Australian and for the combined data set when compared with the result found for the Oxford families. The maximum multipoint K&C LODs for the 196 Australian families were 0.78 at 151.9 cM Haldane and 0.44 at 47.1 cM Haldane (the location of D7S516). The combined data set showed a maximum multipoint K&C LOD of 1.46 at 57.3 cM Haldane (between D7S516 and D7S484).

In the subsequent parametric MOD score analysis of the combined data set (see Materials and Methods), an MOD score of 3.30 was found at 59.3 cM Haldane (the location of D7S484) for a recessive model, with allele frequency of 0.02, heterozygous penetrance of 0.1 and homozygote penetrance of 0.9. The Oxford and Australian samples both supported the peak when analysed separately using the specified model: a peak LOD of 1.88 was found in the Oxford sample at 59.3 cM, whereas the Australian families showed a peak LOD of 1.42 at 59.3 cM. The multipoint results for the separate and combined data sets are shown in Figure 2B. Significance of the MOD score of 3.30 for the combined data set was assessed empirically by simulation (see Materials and Methods). It showed that a MOD of 3.30 or greater—in the linkage peak region originally observed in the Oxford data (47–87 cM)—was observed in 35 out of the 1000 simulations, providing a P-value of 0.035 (95% confidence interval: 0.024–0.048). On the entire chromosome 7, the result was observed in 47 of the 1000 simulations, a P-value of 0.047 (95% confidence interval: 0.035–0.062).

Genetic heterogeneity analysis

The HLOD at 59.3 cM Haldane under the recessive model was identical to the MOD, suggesting that there was no evidence for genetic heterogeneity. It also represented the maximized HLOD across all inheritance models, indicating that there were no other models for which a higher LOD score could be obtained even when allowing for genetic heterogeneity.

However, the model that produced the MOD score for the combined data set was recessive, in contrast to the analyses in the Oxford data set in which the optimal model was a dominant one. When the combined data set was analysed using this dominant model, there was no evidence for linkage under the assumption of genetic homogeneity (assuming all families


Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
K&C LOD
cM (Haldane)

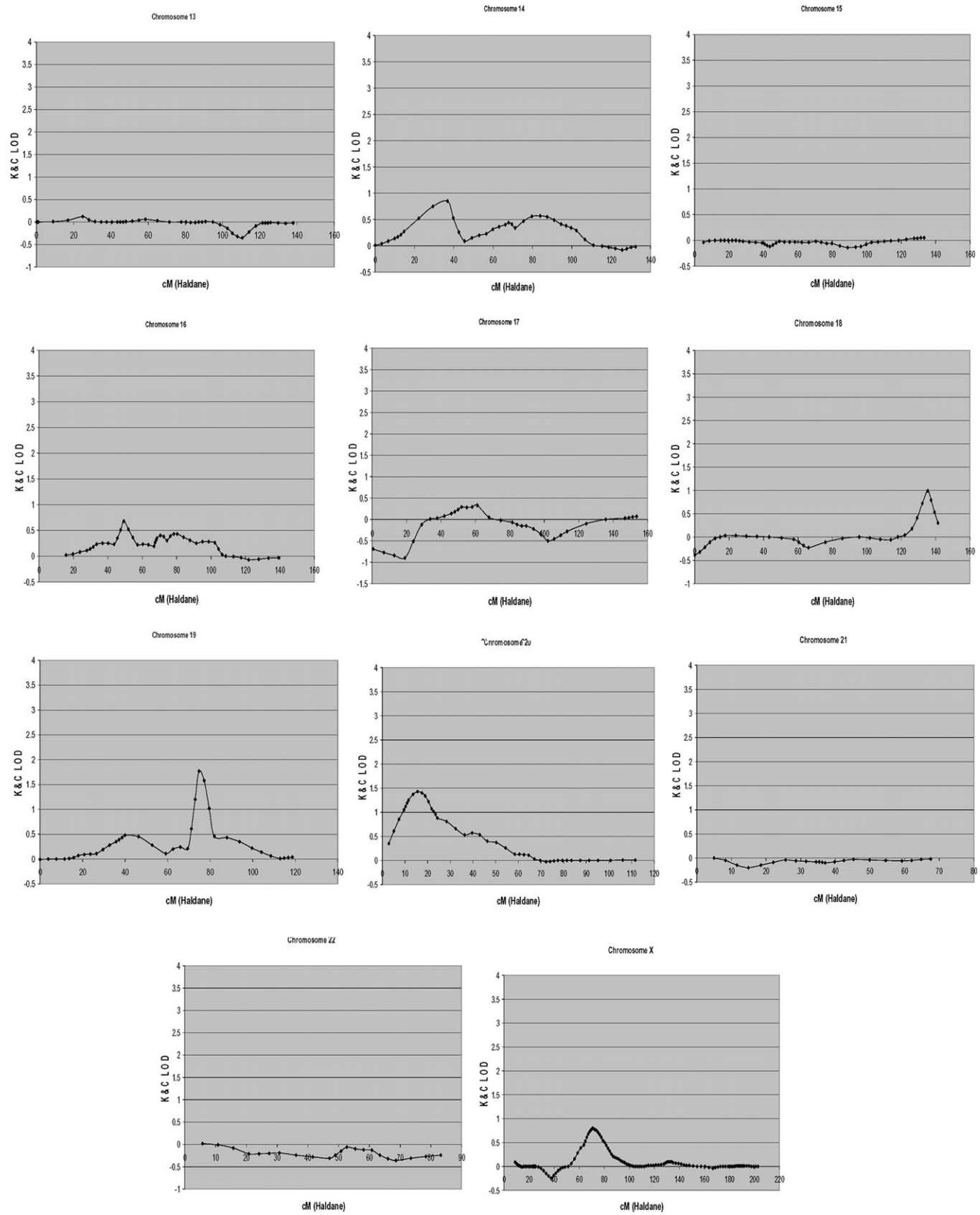


Figure 1. Results of the non-parametric whole genome linkage scan in 52 families with three or more cases of endometriosis, expressed as Kong & Cox (K&C) Log of ODds (LOD) scores (Kong and Cox, 1997). K&C LOD on *y*-axis is plotted against cM (Haldane) on *x*-axis.

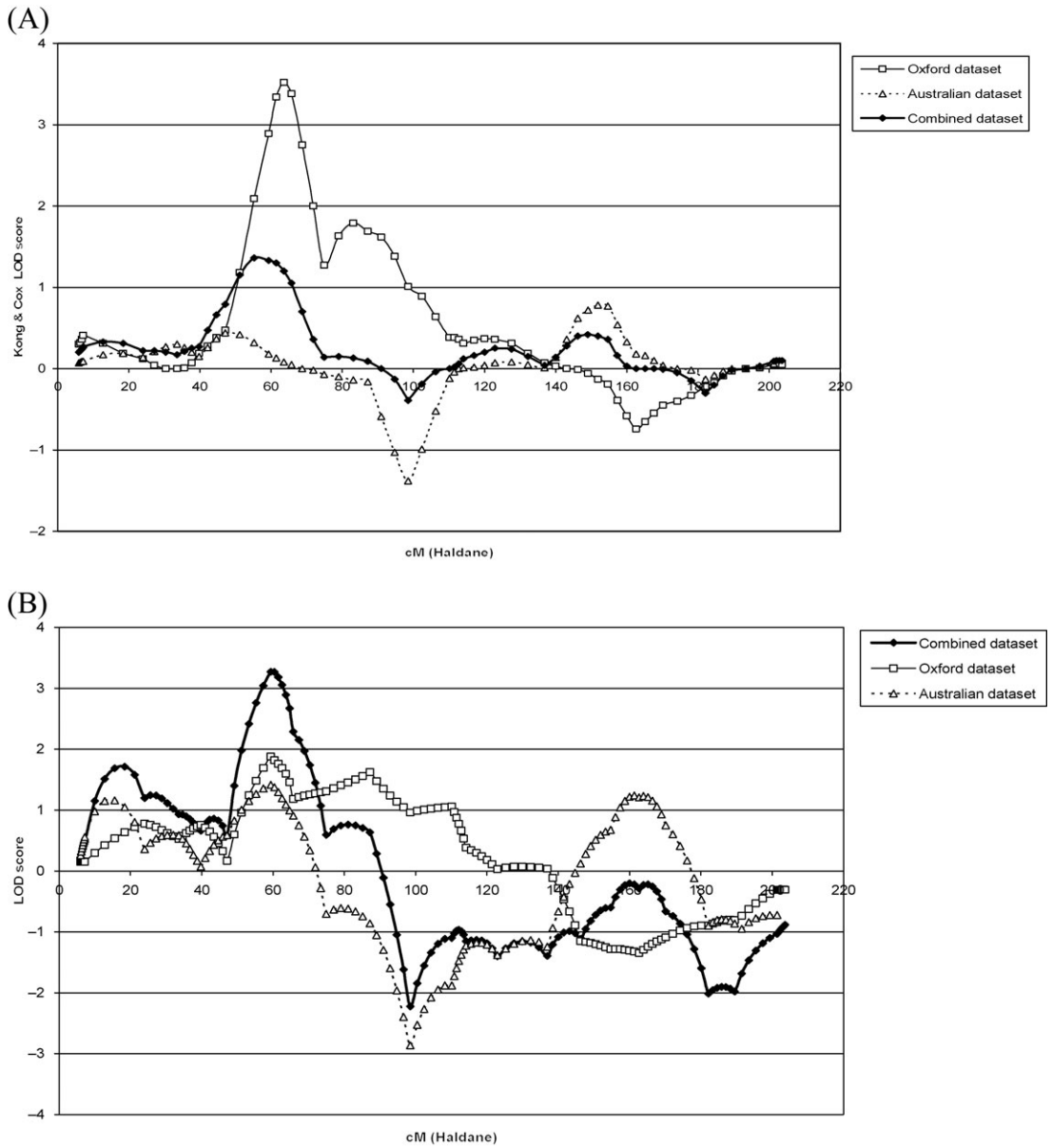


Figure 2. Linkage results on chromosome 7 for the Oxford, Australian and combined data sets: (**A**) non-parametric K&C LOD scores; (**B**) parametric LOD scores derived from the inheritance model that produced the MOD score for the combined data set (allele frequency = 0.02, phenocopy rate = 0.06, penetrance heterozygotes = 0.1, homozygotes= 0.9).

are linked), but linkage under genetic heterogeneity (allowing a proportion of families not to be linked) was suggested (P = 0.005), with an estimated proportion of linked families of 0.30.

Phenotypic influences on the LOD score

Table II shows the association between phenotypic characteristics of the families in the combined data set and contribution to the MOD score under both the dominant and recessive models. For the dominant model that produced the MOD of 3.89 at 65.72 cM for the Oxford data alone, there was a highly significant difference in contribution to the LOD score between the two data sets at that location (Table II). There was an association with the number of relatives who had Stage B disease (equivalent to rAFS Stages III and IV), but this was entirely limited to the Oxford data set and mainly driven by the two families with four Stage B affecteds (data not shown). These two families contributed partial LOD scores of 0.55 and 0.46 to the overall MOD score under the dominant model. Limiting the linkage analysis to families with at least one Stage B affected, however, did not materially improve the combined LOD score. Moreover, even after adjusting for the number of affecteds with Stage B, there remained a significant difference in LOD score at this locus between the Oxford and Australian data sets (suggesting genetic heterogeneity between the two data sets, which was not explained by any of the collected phenotypes).

Under the recessive model, the phenotypic characteristic most associated with an increased family-based LOD score appeared to be the presence of multiple relatives with Stage

Table 2. Associations between families' phenotypic characteristics and mean LOD score at 59.34 cM (recessive model producing the MOD of 3.30 in the combined dataset: AF = 0.02, phenocopy = 0.06, het = 0.1, hom = 0.9) and at 65.72 cM (dominant model producing the MOD of 3.89 in the Oxford dataset alone: AF = 0.04, phenocopy = 0.06, het = 0.5, hom = 0.5)

Phenotypic characteristic per family	Number of families	Mean family LOD (SD) P-value ANOVA (multivariate GLM*)	
		65.72 cM (dominant model)	59.34 cM (recessive model)
Dataset			
Oxford	52	0.0175 (0.243)	0.036 (0.238)
Australia	196	−0.069 (0.258)	0.007 (0.098)
		$P < 0.001$ ($P = 0.002$*)	$P = 0.19$ ($P = 0.70$*)
Number of affecteds			
3	186	−0.022 (0.223)	0.009 (0.064)
4	50	−0.112 (0.304)	−0.004 (0.152)
5 or 6	12	−0.015 (0.502)	0.147 (0.493)
		$P = 0.08$ ($P = 0.08$*)	$P = 0.002$ ($P = 0.001$*)
Number of members with Stage A			
0 or 1	81	0.000 (0.252)	0.037 (0.198)
2 or 3	152	−0.059 (0.250)	0.002 (0.097)
4 or 5	15	−0.052 (0.392)	0.003 (0.199)
		$P = 0.26$ ($P = 0.05$*)	$P = 0.18$ ($P = 0.04$*)
Number of members with Stage B			
0 or 1	153	−0.037 (0.250)	0.007 (0.095)
2 or 3	92	−0.055 (0.272)	0.004 (0.083)
4 or 5	3	0.339 (0.300)	0.652 (0.890)
		P 0.04 ($P = 0.002$*)	$P < 0.001$ ($P < 0.001$*)
Number with known subfertility (≥12 months)			
0	57	0.001 (0.244)	0.037 (0.231)
1	93	−0.046 (0.241)	0.019 (0.113)
2	74	−0.079 (0.279)	−0.015 (0.054)
3 or 4	24	0.018 (0.308)	0.024 (0.117)
		$P = 0.07$	$P = 0.17$
Number with known pelvic pain			
0 or 1	35	−0.039 (0.253)	0.000 (0.065)
2 or 3	188	−0.041 (0.244)	0.017 (0.151)
4 or 5	25	−0.026 (0.383)	0.003 (0.12)
		$P = 0.97$	$P = 0.74$

AF, allele frequency; het, heterozygous penetrance; hom, homozygous prenetrance.
*Significance in multivariate GLM incorporating polymorphic information content, number of affecteds, number of affecteds with Stage B, and Oxford/Australian dataset.

B disease ($P < 0.001$, adjusted for Oxford/Australian origin, number of affecteds, number of 'Stage A' affecteds, and polymorphic information content). This result was largely driven by one family in the Oxford data set (containing four women with Stage B and one woman with unknown disease stage), which contributed 1.66 to the overall MOD score. An increased contribution of families with multiple Stage B affecteds was further suggested when the recessive parametric analysis was limited to the 189 families who had at least one member with Stage B disease. This resulted in a virtually unchanged LOD score of 3.29 at 59.35 cM.

Critical recombinant analysis of haplotypes

Critical recombinant analysis was adopted to narrow down the linkage region observed for the Oxford data under the dominant model with reduced penetrance (see Materials and Methods). In this study, 32 out of the 52 families had a LOD score >0.1, of which 14 had a LOD >0.2. The haplotypes that were shared by most affecteds within the 14 families all included D7S484 and D7S510, a 17.9 Mb region encompassed by D7S516 and D7S519 (data not shown). Fine mapping of the families with recombinations at either side of these markers, by including markers separated by a maximum of 1 cM, reduced the suggested region of interest to the area between D7S484 and D7S2548, a 6.4 Mb interval containing 48 genes according to the latest assembly of the human genome (NCBI 35) (Wheeler *et al.*, 2002).

For the 16 families in the combined data set contributing most to the MOD score under the recessive model (with family-based LODs >0.1), the location of critical recombinants was also investigated in the haplotypes of affecteds. As expected—because the best fitting model was recessive with high penetrance—virtually all affecteds within each family shared two haplotypes spanning the linkage region (Figure 3). There was a single marker, D7S484 (the location of the linkage peak), which showed allele sharing among affecteds within (but not between) all of the 16 families. There was no association between disease status and D7S484 ($\chi^2 = 5.37$, df = 7, $P = 0.6$), confirming that there was no increased allele sharing among affecteds between families. The linkage region surrounding D7S484 at 35.1 Mb was flanked by D7S516 at 28.0 Mb and D7S510 at 39.0 Mb, an 11 Mb region containing 96 genes.

Discussion

The results presented in this article provide significant evidence that a locus with near-Mendelian autosomal inheritance

Family	LOD at 59.3 cM (D7S484)	No. affecteds in family	No. affecteds with haplotype segment	No. stage A	No. stage B	Markers (D7S…)																				
						531	517	513	507	493	516	484	510	519	502	669	630	657	515	486	530	640	684	661	636	798
1	0.32	4	4	0	4			5	7	8	6	4														
		4	3					14	8	15	5	7														
2	0.26	3	3	1	2	3	8	3	7	1	3	5	8	7												
		3	3			7	9	12	11	15	5	5	6	9	6											
3	0.25	3	3	1	2							7	7	7	4											
		3	3						11	17	8	5	5	5	8	6	10	9	14	9	3	12	11	11	10	6
4	0.39	4	3	4	1							4	4													
		4	4			3	9	5	9	3	6	4	8													
5	0.15	3	3	1	2					1	6	4	8	5	9	11	11	6	14	8	15	12	7	11	9	11
		3	2			9	6	12	9	1	6	4	8	5	9	11										
6	0.28	3	3	1	2					15	5	4														
		3	3			9	8	5	10	11	5	7														
7	0.28	4	4	2	2				5	10	2	5														
		4	3			3	9	13	11	11	5	5	6													
8	0.27	3	3	3	0							5	6													
		3	3								1	2	4	8	6	9	6	10	7	12	11	16				
9	0.15	3	3	1	2					3	5	5	6													
		3	3			6	6	7	10	1	3	5	7	6	4	8	1	9	15	11	17	13	4	11	18	
10	0.28	3	3	3	0							7	6	2	9											
		3	3									8	5	7	7	6	3	9								
11	0.82	4	4	3	1							7	5	6	6											
		4	3						15	14	8	4	9	8	9	12	5									
12	0.27	3	3	2	1					11	5	9	7	7												
		3	3								3	5														
13	0.28	3	3	2	1							8	9	10	8											
		3	3								3	4	6	5												
14	1.66	5	5	0	4			12	15		5	5														
		5	5				7	5	5		5	8	9	8	11	6	3	8	23							
15	0.26	3	3	1	2				6	11	5	4	6	5	10	15	4	9								
		3	3				6	8	5	1	5	5	6	2	8	14	5	9	14	11						
16	0.18	3	3	2	1					13	5	5														
		3	3				9	12	6	9	5	1	8	7												

Figure 3. Chromosome 7 haplotype segments (defined by flanking critical recombinants) that were shared among affecteds within the 16 families contributing most to the MOD score (family-based LOD scores >0.1 at 59.3 cM under the recessive model that generated the MOD score). The single marker D7S484 at 59.3 cM—both copies of which are consistently shared among affecteds within families—is marked in grey.

on chromosome 7p is linked to endometriosis. Traditionally, a LOD ≥ 3 has been considered as evidence of significant linkage; however, more recent recommendations suggest assessing significance by data simulation (Wiltshire *et al.*, 2002). Our simulations found that the original result in the Oxford data set for chromosome 7 was genome-wide significant and that adding the Australian data set upheld this significance, despite a substantially smaller albeit positive linkage signal in the latter data set.

Endometriosis is a condition that affects millions of women world-wide in terms of pain and infertility, and yet little is known about its aetiology. Despite ample evidence of familial aggregation (Bischoff and Simpson, 2000; Zondervan *et al.*, 2001; Stefansson *et al.*, 2002), candidate gene studies have, to date, been largely unsuccessful in uncovering any aetiological targets. To a large extent, this is probably due to inappropriate and underpowered study design (Zondervan *et al.*, 2002; Zondervan and Cardon, 2004), but the complexity of the underlying biological and pathological mechanisms provides an additional challenge in targeting promising candidate genes. Results from linkage analyses—although in themselves problematic because of the large numbers of families with multiple affecteds required for complex traits—could provide vital new clues in the search for predisposing genetic variants and thus underlying biology. For this reason, the Oxford and Australian research groups independently started recruiting families in the mid—1990s, culminating in the collaborative IES which comprised a uniquely large collection of nearly 1200 families with multiple endometriosis cases.

At the start of the collaboration, the phenotypes of all affected members in the families of both data sets were reviewed by the two gynaecologists collaborating in the two studies (S. K. and D. T. O'C.) to ensure that phenotypic definitions were standardized. This highlighted initial differences in recruitment strategies between the two data sets. Oxford focused on families with Stage B (rAFS Stages III/IV) disease at the outset, before including families with Stage A disease (rAFS Stages I/II) later on, whereas the Australian study collected both types from the outset. In addition, even though phenotypic staging criteria were standardized between the two studies, unidentified differences related to practicalities in recruitment protocols and the willingness of women to volunteer for studies or their inclination to be a member of an endometriosis self-help group all could have resulted in phenotypic differences between the data sets. The relative effect of these phenotypic differences on the linkage results in the two studies is unknown.

The results of the present analysis of families with three or more affecteds strongly suggest the presence of one or more genetic variants of low frequency predisposing to endometriosis in the Oxford data set. This is supported by non-parametric as well as parametric linkage results (either under dominant or recessive models). In addition, there appears some evidence for a disease allele at a similar location in the Australian data set, but this result is only supported when a recessive disease model is applied. On the one hand, these results could be interpreted as referring to the same signal. Since the Oxford data contained more affected mother–child pairs than the Australian data set, this would have meant that the former data set was more likely to support a dominant segregation model. This difference in family composition could reflect differences in ascertainment as well as in diagnostic patterns at a population level. Moreover, discrimination between different sets of parameters as estimated by the MOD score method may be difficult and different sets of disease model parameters could provide similar MOD scores (Clerget-Darpoux *et al.*, 1986). On the other hand, the results could represent genuine genetic heterogeneity between the two data sets. This would imply that there are multiple disease variants with both recessive and dominant inheritance patterns in the region of interest on chromosome 7 in the Oxford data set, whereas only one or more variants with a recessive segregation pattern are present in the Australian data set. Alternatively, the predisposing variant(s) could only be present in the Oxford but not in the Australian data set.

For the dominant model that produced the MOD of 3.89 at 65.72 cM for the Oxford data alone, there was a highly significant difference in contribution to the LOD score between the two data sets at that location, which was not explained by information collected on any of the underlying phenotypic characteristics shared by the two data sets. Under the recessive model, there was no evidence of genetic heterogeneity, meaning that a genetic variant at this locus would act in all families with endometriosis. The phenotypic characteristic most associated with an increased LOD score appeared to be the presence of multiple relatives with the Stage B disease. Neither pelvic pain nor infertility (the two main symptom-based phenotypes related to endometriosis) was associated with the strength of the linkage signal. However, the amount of detail collected from the cases on these two subphenotypes was limited to whether or not these were symptoms that led to the diagnosis of endometriosis, and no attempt was made to assess other issues such as the degree or the exact nature of the pain.

Whether or not genetic heterogeneity between the two data sets on the basis of population origin is plausible is a question of debate. Given that both data sets consisted of people who were of Caucasian origin and given that the history of Australian migrants from Western Europe dates only a few centuries back, one would not expect a large diversity in allele frequencies of disease-predisposing genes between the two populations—a notion supported by recent studies (Sullivan *et al.*, 2006; Stankovich *et al.*, 2006). Nevertheless, the family that contributed most to the dominant model in the Oxford data set was of Russian origin, and thus the Oxford data set could possibly contain families of a more diverse population background than the Australian data set. In combination with the various recruitment strategies necessarily employed in family studies of endometriosis of this size, it could therefore be possible that a predisposing variant in the Oxford data set was not present, or only at a very low frequency, in the Australian data set.

This is the second report describing a significant linkage result for endometriosis from the IES (Treloar *et al.*, 2002) and the first to suggest a high penetrance susceptibility locus with (near-)Mendelian autosomal inheritance on chromosome 7p13–15. Interestingly, the first genome-wide linkage report (based on an NPL analysis), including all 1176 Oxford and Australian families combined, did not note that chromosome 7 was linked to endometriosis in this region, although peaks in two other regions on chromosome 7 (at markers D7S517 and D7S2423) had multipoint maximum LOD (MLS) scores $\geq$1.0 (Treloar *et al.*, 2005). Indeed, had a non-parametric analysis been conducted on the combined data set of 248 families with three or more affected women in the first instance, rather than on the Oxford data set alone, the K&C LOD of 1.46 would not have resulted in much enthusiasm to pursue further parametric linkage analyses. In line with this observation, simulation studies previously showed that MMLS methods (which are akin to the MOD score method but only maximize the MLS across two models, recessive and dominant with reduced penetrance) can be more powerful than NPL methods for detecting linkage, with differences in power determined by the underlying true model and linkage information (Abreu *et al.*, 1999; Greenberg and Abreu, 2001).

The locus for which a significant LOD score was found on chromosome 10 in the original linkage scan incorporating about 1176 families did not come up as an area of focus in the present analysis. This is not surprising, however, since the objective of the genome-wide scan of all families was very different from the objective of the current analysis. In the genome-wide scan of 1176 families, the objective was to find evidence of linkage for a complex trait, with underlying assumptions of potentially highly frequent, but low-penetrant genetic variants involved in the aetiology. Such a disease model requires the recruiting of a very large number of families with more than one affected. However, for the present analysis, we were aiming to find linkage due to relatively rare genetic variants with high penetrance, requiring fewer families but with a high number of affecteds. This difference in design meant that a strong result found in one would not necessarily correspond to a similarly strong result in the other. Nevertheless, our genome-wide results using the Oxford families with three or more affecteds (i.e. 52 families out of the 245 in the original genome scan) showed a maximum LOD of $\sim$0.7 on chromosome 10. Given the smaller number of families involved in the analysis, this is not incompatible with the MLS of $\sim$1.3 found in all 245 Oxford families in the original genome scan (Treloar *et al.*, 2005).

Chromosomal linkage areas from genome-wide association scans, such as reported here, are necessarily of relatively low resolution, typically spanning more than 10 million DNA base pairs. When an area of significant linkage is found, it is

subsequently narrowed down by adding more markers (fine-mapping) and by association studies of markers specifically selected to assess genetic variability across the region (possibly focusing on candidate genes). The location of critical recombinants in the haplotypes of affecteds, following the recessive model in the combined data set, narrowed the region of interest with the most evidence of linkage down to a single marker, D7S484 at 35.1 Mb, flanked by D7S516 and D7S510, an 11 Mb region containing 96 genes according to the latest assembly of the human genome (NCBI 35). Critical recombinant analysis in families contributing to the dominant model in the Oxford data set narrowed its most likely area of linkage down to a 6.4 Mb interval between D7S484 and D7S2548, containing 48 genes. Follow-up of these (mostly overlapping) regions through fine-mapping and candidate gene approaches is currently underway. Replication of our results in other populations (e.g. in more genetically homogenous, isolated, populations) will be important, as will be analyses aimed at finding genetic variants in the linkage region with functional relevance to endometriosis pathogenesis.

Acknowledgements

The principal authors' roles were as follows:
K.T.Z.: design (OXF part), analysis and interpretation, drafting article and final approval.
S.A.T.: conception and design (AUS part), interpretation, revising for critical content and final approval.
J.L.: analysis and interpretation and revising for critical content.
D.E.W.: conception and design (OXF and AUS part), interpretation, revising for critical content and final approval.
D.R.N.: conception and design (AUS part), interpretation, revising for critical content and final approval.
J.M.: design (OXF part), interpretation, revising for critical content and final approval.
I.J.M.: conception and design (OXF part), revising for critical content and final approval.
L.R.C.: conception and design (OXF and AUS part), interpretation, revising for critical content and final approval.
N.G.M.: conception and design (OXF and AUS part), revising for critical content and final approval.
S.H.K.: conception and design (OXF and AUS part), interpretation, revising for critical content and final approval.
G.W.M.: conception and design (AUS part), revising for critical content and final approval.

References

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

Abreu PC, Greenberg DA and Hodge SE (1999) Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases. Am J Hum Genet 65,847–857.

Berkley KJ, Rapkin AJ and Papka RE (2005) The pains of endometriosis. Science 308,1587–1589.

Bhanoori M, Arvind BK, Pavankumar Reddy NG, Lakshmi RK, Zondervan K, Deenadayal M, Kennedy S and Shivaji S (2005) The vascular endothelial growth factor (VEGF) +405G>C 5′-untranslated region polymorphism and increased risk of endometriosis in South Indian women: a case control study. Hum Reprod 20,1844–1849.

Bischoff FZ and Simpson JL (2000) Heritability and molecular genetic studies of endometriosis. Hum Reprod Upd 6,37–44.

Clerget-Darpoux F, Bonaiti-Pellie C and Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42,393–399.

Dizier MH, Babron MC and Clerget-Darpoux F (1996) Conclusion of LOD-score analysis for family data generated under two-locus models. Am J Hum Genet 58,p1338–1346.

Eskenazi B and Warner ML (1997) Epidemiology of endometriosis. Obstet Gynecol Clin North Am 24,235–258.

Giudice LC and Kao LC (2004) Endometriosis. Lancet 364,1789–1799.

Greenberg DA and Abreu PC (2001) Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. Genet Epidemiol 21,299–314.

Halme J, Hammond MG, Hulka JF, Raj SG and Talbert LM (1984) Retrograde menstruation in healthy women and in patients with endometriosis. Obstet Gynecol 64,151–154.

Hodge SE and Elston RC (1994) Lods, wrods, and mods: the interpretation of lod scores calculated under different models. Genet Epidemiol 11,329–342.

Ishii K, Takakuwa K, Kashima K, Tamura M and Tanaka K (2003) Associations between patients with endometriosis and HLA class II; the analysis of HLA-DQB1 and HLA-DPB1 genotypes. Hum Reprod 18,985–989.

Kado N, Kitawaki J, Obayashi H, Ishihara H, Koshiba H, Kusuki I, Tsukamoto K, Hasegawa G, Nakamura N and Yoshikawa T *et al.* (2002) Association of the CYP17 gene and CYP19 gene polymorphisms with risk of endometriosis in Japanese women. Hum Reprod 17,897–902.

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

Kong X, Murphy K, Raj T, He C, White PS and Matise TC (2004) A combined linkage-physical map of the human genome. Am J Hum Genet 75,1143–1148.

Koninckx PR, Barlow D and Kennedy S(1999) Implantation versus infiltration: the Sampson versus the endometriotic disease theory. Gynecol Obstet Invest 47(Suppl 1),3–10.

Lake SL, Blacker D and Laird NM (2000) Family-based tests of association in the presence of linkage. Am J Hum Genet 67,1515–1525.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM and Ding W *et al.* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266,66–71.

Ott J (1999) Analysis of Human Genetic Linkage, 3rd edn. John Hopkins University Press, Baltimore, Maryland.

Ott J (1986) Linkage probability and its approximate confidence interval under possible heterogeneity. Genet Epidemiol 1,251–257.

Sampson JA (1927) Peritoneal endometriosis due to the menstrual dissemination of endometrial tissue into the peritoneal cavity. Am J Obstet Gynecol 14,422–469.

Shan K, Ying W, Jian-Hui Z, Wei G, Na W and Yan L (2005) The function of the SNP in the MMP1 and MMP3 promoter in susceptibility to endometriosis in China. Mol Hum Reprod 11,423–427.

Stefansson H, Geirrson RT, Steinthorsdottir V, Jonsson H, Manolescu A, Kong A, Ingadottir G, Gulcher J and Stefansson K (2002) Genetic factors contribute to the risk of developing endometriosis. Hum Reprod 17,555–559.

Stankovich J, Cox CJ, Tan RB, Montgomery DS, Huxtable SJ, Rubio JP, Ehm MG, Johnson L, Butzkueven H and Kilpatrick TJ *et al.* (2006) On the utility of data from the International HapMap Project for Australian association studies. Hum Genet 119,220–222.

Sullivan PF, Montgomery GW, Hottenga JJ, Wray NR, Boomsma DI and Martin NG (2006) Empirical evaluation of the genetic similarity of samples from twin registries in Australia and the Netherlands using 359 STRP markers. Twin Res Hum Genet 9,600–602.

The American Fertility Society (1985) Revised American fertility society classification of endometriosis: 1985. Fertil Steril 43,351–352.

Treloar S, Hadfield R, Montgomery G, Lambert A, Wicks J, Barlow D, O'Connor D and Kennedy S (2002) The international endogene study: a collection of families for genetic research in endometriosis. Fertil Steril v78,679.

Treloar SA, O'Connor DT, O'Connor VM and Martin NG (1999) Genetic influences of endometriosis in an Australian twin sample. Fertil Steril 71,701–710.

Treloar SA, Wicks J, Nyholt DR, Montgomery GW, Bahlo M, Smith V, Dawson G, Mackay IJ, Weeks DE and Bennett ST *et al.* (2005) Genomewide linkage study in 1,176 affected sister pair families identifies

a significant susceptibility locus for endometriosis on chromosome 10q26. Am J Hum Genet 77,365–376.

Vigano P, Infantino M, Lattuada D, Lauletta R, Ponti E, Somigliana E, Vignali M and DiBlasio AM (2003) Intercellular adhesion molecule-1 (ICAM-1) gene polymorphisms in endometriosis. Mol Hum Reprod 9,47–52.

Weeks DE, Lehner T, Squires-Wheeler E, Kaufmann C and Ott J (1990) Measuring the inflation of the lod score due to its maximization over model parameter values in human linkage analysis. Genet Epidemiol 7,237–243.

Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA and Wagner L *et al.* (2002) Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res 30,13–16.

Williams JT and Blangero J (1999) Comparison of variance components and sibpair-based approaches to quantitative trait linkage analysis in unselected samples. Genet Epidemiol 16,113–134.

Williams JT, Duggirala R and Blangero J (1997) Statistical properties of a variance components method for quantitative trait linkage analysis in nuclear families and extended pedigrees. Genet Epidemiol 14,1065–1070.

Wiltshire S, Cardon LR and McCarthy MI (2002) Evaluating the results of genomewide linkage scans of complex traits by locus counting. Am J Hum Genet 71,1175–1182.

Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C and Micklem G (1995) Identification of the breast cancer susceptibility gene BRCA2. Nature 378,789–792.

Zondervan KT and Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5,89–100.

Zondervan KT, Cardon LR and Kennedy SH (2001) The genetic basis of endometriosis. Curr Opin Obstet Gynecol 13,309–314.

Zondervan KT, Cardon LR and Kennedy SH (2002) What makes a good case–control study? Design issues for complex traits such as endometriosis. Hum Reprod 17,1415–1423.

Zondervan KT, Weeks DE, Colman R, Cardon LR, Hadfield R, Schleffler J, Trainor AG, Coe CL, Kemnitz JW and Kennedy SH (2004) Familial aggregation of endometriosis in a large pedigree of rhesus macaques. Hum Reprod 19,448–455.

Submitted on July 21, 2006; resubmitted on September 6, 2006; accepted on September 19, 2006.