nature genetics

Genome-wide association meta-analysis identifies new endometriosis risk loci

Dale R Nyholt[1,16], Siew-Kee Low[2,16], Carl A Anderson[3], Jodie N Painter[1], Satoko Uno[2,4], Andrew P Morris[5], Stuart MacGregor[1], Scott D Gordon[1], Anjali K Henders[1], Nicholas G Martin[1], John Attia[6,7], Elizabeth G Holliday[6,7], Mark McEvoy[6,8,9], Rodney J Scott[7,10,11], Stephen H Kennedy[12], Susan A Treloar[13], Stacey A Missmer[14], Sosuke Adachi[15], Kenichi Tanaka[15], Yusuke Nakamura[2], Krina T Zondervan[5,12,17], Hitoshi Zembutsu[2,17] & Grant W Montgomery[1,17]

We conducted a genome-wide association meta-analysis of 4,604 endometriosis cases and 9,393 controls of Japanese[1] and European[2] ancestry. We show that rs12700667 on chromosome 7p15.2, previously found to associate with disease in Europeans, replicates in Japanese ($P = 3.6 \times 10^{-3}$), and we confirm association of rs7521902 at 1p36.12 near *WNT4*. In addition, we establish an association of rs13394619 in *GREB1* at 2p25.1 with endometriosis and identify a newly associated locus at 12q22 near *VEZT* (rs10859871). Excluding cases of European ancestry of minimal or unknown severity, we identified additional previously unknown loci at 2p14 (rs4141819), 6p22.3 (rs7739264) and 9p21.3 (rs1537377). All seven SNP effects were replicated in an independent cohort and associated at $P < 5 \times 10^{-8}$ in a combined analysis. Finally, we found a significant overlap in polygenic risk for endometriosis between the genome-wide association cohorts of European and Japanese descent ($P = 8.8 \times 10^{-11}$), indicating that many weakly associated SNPs represent true endometriosis risk loci and that risk prediction and future targeted disease therapy may be transferred across these populations.

Endometriosis (MIM 131200) is a common gynecological disease associated with severe pelvic pain that affects 6–10% of women in their reproductive years[3,4] and 20–50% of women with infertility[5]. Endometriosis risk is influenced by genetic factors and has an estimated heritability of around 51% (ref. 3).

Two large endometriosis genome-wide association studies (GWAS)[1,2] have reported associations at genome-wide significance. The first, in a Japanese sample of 1,423 cases and 1,318 controls obtained from BioBank Japan (BBJ), with 484 cases and 3,974 controls for replication, implicated a SNP (rs10965235) in the *CDKN2B-AS1* gene on chromosome 9p21.3 (overall odds ratio (OR) = 1.44, 95% confidence interval (CI) = 1.30–1.59; $P = 5.57 \times 10^{-12}$)[1]. The second GWAS, by the International Endogene Consortium (IEC) in a sample of European ancestry from Australia (2,270 cases and 1,870 controls) and the UK (924 cases and 5,190 controls), with 2,392 cases and 2,271 controls from the United States for replication, identified an intergenic SNP (rs12700667) at 7p15.2 (overall OR = 1.20, 95% CI = 1.13–1.27; $P = 1.4 \times 10^{-9}$)[2]. These two studies did not report replication of each other's top locus, partly because rs10965235 is monomorphic in populations of European ancestry. The study of individuals of European ancestry did find association with rs7521902 (OR = 1.16, 95% CI = 1.08–1.25; $P = 9.0 \times 10^{-5}$) near the *WNT4* gene at 1p36.12, which was reported to be suggestively associated in Japanese (OR = 1.20, 95% CI = 1.11–1.29; $P = 2.2 \times 10^{-6}$).

Encouraged by the *WNT4* association and with accumulating evidence for many complex traits that the number of discovered variants is strongly correlated with experimental sample size[6], we sought to increase the ratio of controls to cases in the Australian GWAS cohort and to perform a formal meta-analysis of the Australian (Queensland Institute of Medical Research, QIMR), UK (OX) and Japanese (BBJ) GWAS data.

To increase the power of the Australian GWAS data set, we matched the existing QIMR cases and controls[2] on the basis of ancestry to

[1]Queensland Institute of Medical Research, Brisbane, Queensland, Australia. [2]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan. [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [4]First Department of Surgery, Sapporo Medical University, School of Medicine, Sapporo, Japan. [5]Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [6]Centre for Clinical Epidemiology and Biostatistics, School of Medicine and Public Health, University of Newcastle, Newcastle, New South Wales, Australia. [7]Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine, Hunter Medical Research Institute, Newcastle, New South Wales, Australia. [8]School of Medicine and Public Health, University of Newcastle, Newcastle, New South Wales, Australia. [9]Public Health Research Program, Hunter Medical Research Institute, Newcastle, New South Wales, Australia. [10]School of Biomedical Sciences and Pharmacy, University of Newcastle, Newcastle, New South Wales, Australia. [11]Division of Genetics, Hunter Area Pathology Service, Newcastle, New South Wales, Australia. [12]Nuffield Department of Obstetrics and Gynaecology, University of Oxford, John Radcliffe Hospital, Oxford, UK. [13]Centre for Military and Veterans' Health, University of Queensland, Mayne Medical School, Brisbane, Queensland, Australia. [14]Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [15]Department of Obstetrics and Gynecology, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan. [16]These authors contributed equally to this work. [17]These authors jointly directed this work. Correspondence should be addressed to D.R.N. (dale.nyholt@qimr.edu.au), K.T.Z. (krina.zondervan@well.ox.ac.uk), H.Z. (zembutsh@ims.u-tokyo.ac.jp) or G.W.M. (grant.montgomery@qimr.edu.au).

Received 16 May; accepted 24 September; published online 28 October 2012; doi:10.1038/ng.2445

Table 1 Summary of the endometriosis case-control cohorts

Cohort	Ancestry	Number of cases (stage B)	Number of controls
QIMR-HCS GWAS	European	2,262 (905)	2,924
OX GWAS	European	919 (452)	5,151
BBJ GWAS	Japanese	1,423	1,318
GWAS meta-analysis		4,604	9,393
Replication	Japanese	1,044	4,017
Total		5,648	13,410

individuals from the Hunter Community Study (HCS)[7]. After stringent quality control, the combined QIMR-HCS GWAS cohort consisted of 2,262 endometriosis cases and 2,924 controls, increasing the number of controls by 1,054 and the Australian effective sample size by 24%. We also performed more stringent quality control, incorporating the OX data set, resulting in a revised OX GWAS cohort of 919 endometriosis cases and 5,151 controls. All cases in the QIMR-HCS and OX studies have surgically confirmed endometriosis and disease stage from surgical records using the revised American Fertility Society (rAFS) classification system[8]; subjects are grouped into stage A (stage 1 or 2 disease or some ovarian disease with a few adhesions; n = 1,680, 52.8%), stage B (stage 3 or 4 disease; n = 1,357, 42.7%) or unknown stage (n = 144, 4.5%). Details of the final GWAS and independent replication case-control cohorts are summarized in **Table 1**, and a schematic of our study design is provided in **Figure 1**.

Meta-analysis of all 4,604 endometriosis cases and 9,393 controls for the 407,632 SNPs that were represented in the QIMR-HCS, OX and BBJ GWAS data showed that the A allele of rs12700667 at the 7p15.2 locus in individuals of European ancestry (OR = 1.22, 95% CI = 1.13–1.31; $P = 7.2 \times 10^{-8}$) also replicates in the Japanese GWAS data (OR = 1.22, 95% CI = 1.07–1.39; $P = 3.6 \times 10^{-3}$), producing an overall OR of 1.22 (95% CI = 1.14–1.30) and $P = 9.3 \times 10^{-10}$ in the GWAS meta-analysis; we also confirmed association with allele A of rs7521902 at the 1p36.12 *WNT4* locus (OR = 1.18, 95% CI = 1.11–1.25; $P = 4.6 \times 10^{-8}$) (**Table 2**).

The GWAS meta-analysis identified a previously unknown associated locus at 12q22 near the *VEZT* gene (allele C of rs10859871: OR = 1.18, 95% CI = 1.12–1.25; $P = 5.5 \times 10^{-9}$). We also established association with allele G of rs13394619 in the *GREB1* gene at 2p25.1 (OR = 1.12, 95% CI = 1.06–1.18; $P = 2.1 \times 10^{-5}$), previously reported (OR = 1.35, 95% CI = 1.17–1.56; $P = 3.8 \times 10^{-5}$) in a small independent Japanese GWAS of 696 cases and 825 controls[9]. The association for the G allele of rs13394619 approached conventional genome-wide significance ($P \leq 5 \times 10^{-8}$) in combined analysis of the QIMR-HCS, OX, BBJ, Adachi 500K and Adachi 6.0

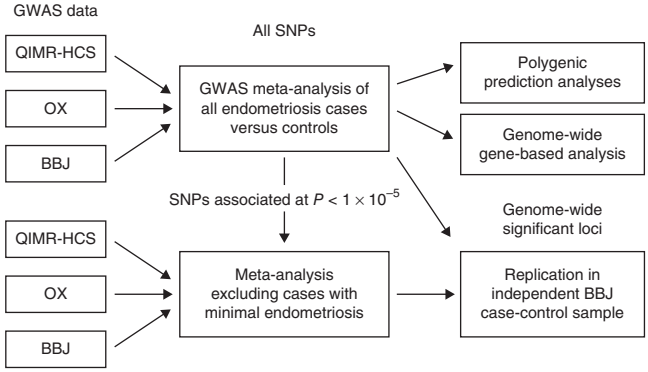

Figure 1 Study design.

Table 2 Summary of the GWAS and replication results for the seven genome-wide significant loci

					GWAS										Replication			Total	
					QIMR-HCS		OX		BBJ		Meta-analysis								
Chr.	SNP	Position (bp)	RA	OA	RAF_{case}	$RAF_{control}$	RAF_{case}	$RAF_{control}$	RAF_{case}	$RAF_{control}$	P_{all}	$P_{stage\ B}$			RAF_{case}	$RAF_{control}$	P	P_{all}	$P_{stage\ B}$
1	rs7521902	22490724	A	C	0.265	0.236	0.259	0.238	0.570	0.514	4.6×10^{-8}	2.3×10^{-9}			0.568	0.521	6.5×10^{-5}	3.2×10^{-11}	7.6×10^{-13}
2	rs13394619[a]	11727507	G	A	0.538	0.514	0.551	0.521	0.485	0.449	6.1×10^{-8}	7.0×10^{-8}			0.489	0.455	3.5×10^{-2}	6.1×10^{-9}	6.7×10^{-9}
2	rs4141819	67864675	C	T	0.331	0.298	0.343	0.309	0.226	0.203	4.0×10^{-7}	6.5×10^{-8}			0.220	0.203	5.1×10^{-2}	8.5×10^{-8}	4.1×10^{-8}
6	rs7739264	19785588	T	C	0.545	0.512	0.556	0.515	0.772	0.742	1.3×10^{-7}	5.8×10^{-8}			0.778	0.744	6.9×10^{-4}	3.6×10^{-10}	2.1×10^{-10}
7	rs12700667	25901639	A	G	0.769	0.730	0.776	0.744	0.221	0.189	9.3×10^{-10}	3.8×10^{-11}			0.197	0.191	2.6×10^{-1}	3.6×10^{-9}	1.1×10^{-9}
9	rs1537377	22169700	C	T	0.424	0.395	0.436	0.401	0.410	0.379	2.5×10^{-6}	1.0×10^{-8}			0.402	0.359	1.3×10^{-4}	2.4×10^{-9}	5.8×10^{-12}
12	rs10859871	95711876	C	T	0.332	0.299	0.332	0.295	0.373	0.328	5.5×10^{-9}	3.7×10^{-7}			0.377	0.328	1.1×10^{-5}	5.1×10^{-13}	2.6×10^{-11}

Chr., chromosome, RA, risk allele; OA, other allele; RAF, risk allele frequency. Genomic position is shown relative to GRCh37 (hg19). P_{all} includes all available endometriosis cases. $P_{stage\ B}$ excludes endometriosis cases of unknown and minimal (rAFS 1–2) stages where detailed stage data was available.
[a]GWAS meta-analysis and total P values for rs13394619 include results published in Adachi *et al.*[9], consisting of $P = 6.1 \times 10^{-4}$ ($RAF_{case} = 0.517$, $RAF_{control} = 0.414$) and $P = 1.0 \times 10^{-2}$ ($RAF_{case} = 0.488$, $RAF_{control} = 0.429$) obtained in their 500K array (290 cases, 262 controls) and 6.0 array (406 cases, 563 controls) cohorts, respectively.

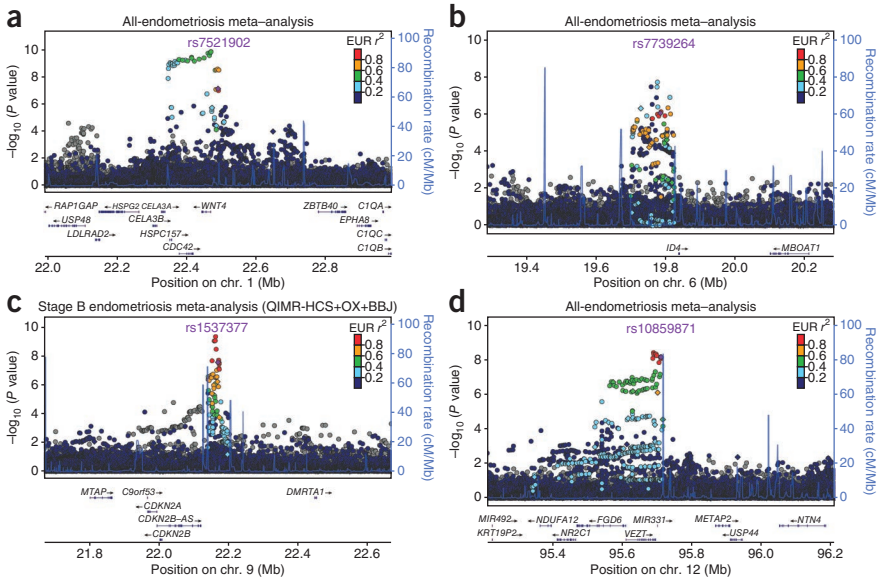

Figure 2 Annotated plots for loci where imputation helped resolve the associated region. (**a**–**d**) Evidence for association with endometriosis from the QIMR-HCS, OX and BBJ genome-wide association meta-analysis across the 1p36.12 (**a**), 6p22.3 (**b**), 9p21.3 (**c**) and 12q22 (**d**) regions after imputation using the 1000 Genomes Project reference panel. Diamond and circle symbols represent genotyped and imputed SNPs, respectively. The most significant genotyped SNP is represented by a purple diamond. All other SNPs are colored according to the strength of LD with the top genotyped SNP (as measured by r^2 in European (EUR) 1000 Genomes Project data).

GWAS data (OR = 1.15, 95% CI = 1.09–1.20; $P = 6.1 \times 10^{-8}$) (**Table 2**). In addition to the 3 SNPs reaching genome-wide significance on chromosomes 1, 7 and 12 (rs7521902, rs12700667 and rs10859871, respectively), the Manhattan plot of all endometriosis genome-wide association meta-analysis results (**Supplementary Fig. 1**) showed that 34 SNPs had suggestive evidence of association ($P \leq 1 \times 10^{-5}$).

Given the substantially greater genetic loading (or liability) of moderate-to-severe (stage B) endometriosis (rAFS stage 3 or 4 disease) compared to minimal (stage A) endometriosis (rAFS stage 1 or 2 disease)[2], a secondary analysis was performed for the SNPs with suggestive genome-wide association, with meta-analysis performed on the association results from QIMR-HCS and OX stage B cases versus controls with the BBJ association results (for which stage information not available).

After excluding endometriosis cases with minimal (rAFS stage 1 or 2) or unknown severity in the QIMR-HCS and OX cohorts, GWAS meta-analysis implicated new loci at 2p14 (allele C of rs4141819: OR = 1.22, 95% CI = 1.14–1.32; $P = 6.5 \times 10^{-8}$), 6p22.3 (allele T of rs7739264: OR = 1.21, 95% CI = 1.13–1.30; $P = 5.8 \times 10^{-8}$) and 9p21.3 (allele C of rs1537377: OR = 1.22, 95% CI = 1.14–1.30; $P = 1.0 \times 10^{-8}$) (**Table 2**, **Supplementary Fig. 2**, **Supplementary Tables 1** and **2** and **Supplementary Note**).

Annotated plots showing evidence for association in the combined QIMR-HCS, OX and BBJ GWAS data of genotyped SNPs across the seven implicated loci from the analysis of all cases and stage B cases only are provided in **Supplementary Figures 3–9**. Imputation using the 1000 Genomes Project reference panel resulted in more significant *P* values and helped resolve the associated region at the 1p36.12 (rs56318008: $P_{all} = 1.3 \times 10^{-10}$), 2p25.1 (rs77294520: $P_{stage\ B} = 8.6 \times 10^{-8}$), 2p14 (rs2861694: $P_{stage\ B} = 7.9 \times 10^{-9}$), 6p22.3 (rs6901079: $P_{all} = 1.9 \times 10^{-8}$), 9p21.3 (rs7041895: $P_{stage\ B} = 5.1 \times 10^{-10}$) and 12q22 (rs11107968: $P_{all} = 3.9 \times 10^{-9}$) loci (**Fig. 2** and **Supplementary Figs. 10–16**). Of particular note, the imputed SNPs at 1p36.12 with the most significant association, rs56318008 and rs3820282 ($P_{all} = 1.6 \times 10^{-10}$), are located 22 bp 5′ to *WNT4* and within the gene, respectively.

Notably, the most associated genotyped SNP at 9p21.3 (rs1537377) is 55 kb centromeric to the SNP associated with genome-wide significance that was reported in the original BBJ GWAS[1] (rs10965235) located in the *CDKN2B-AS1* gene and 49 kb 3′ to the transcriptional end site of *CDKN2B-AS1*. The rs10965235 SNP is monomorphic in populations of European ancestry, and we investigated the independence of the associations at rs10965235 and rs1537377 in the BBJ GWAS data. First, in the BBJ GWAS data, alleles of rs10965235 and rs1537377 are very weakly correlated, with linkage disequilibrium (LD) metrics of $r^2 = 0.028$ and $D' = 0.461$. Second, the allelic association *P* values for rs10965235 and rs1537377 are 1.6×10^{-4} and 1.8×10^{-2}, respectively. After conditioning on rs10965235, weak residual association remained at rs1537377 ($P = 9.0 \times 10^{-2}$). Consequently, the data suggest that there may be two independent genetic risk factors near the *CDKN2B-AS1* locus at 9p21.3. *CDKN2B-AS1* encodes a long non-coding RNA adjacent to and transcribed from the opposite strand of *CDKN2B* (p15), *CDKN2A* (p16) and *ARF* (p14). Loss of heterozygosity for *CDKN2A* and hypermethylation of the *CDKN2A* promoter have been reported in endometriosis[10,11].

To further validate the seven SNPs implicated by the meta-analysis, we carried out a replication study using a cohort of 1,044 cases and 4,017 controls obtained from BioBank Japan independent of the BBJ GWAS cohort. As shown in the forest plots of risk allele effects estimated using all cases versus controls (**Fig. 3**), the effects (ORs) were in the same direction for all seven implicated SNPs across the GWAS and replication cohorts. With the exception of rs12700667, which was previously replicated ($P = 1.2 \times 10^{-3}$) in 2,392 cases and 2,271 controls from the United States[2], and rs4141819 (with marginal $P = 5.1 \times 10^{-2}$), all SNPs were replicated with nominal significance at $P < 0.05$ (**Table 2**). All seven SNPs surpassed the conventional genome-wide significance threshold of $P \leq 5 \times 10^{-8}$ after combined analysis of the GWAS and replication cases and controls (**Table 2**). A conservative adjustment of the total *P* values for rs4141819 ($P_{all} = 8.5 \times 10^{-8}$; $P_{stage\ B} = 4.1 \times 10^{-8}$) for performing two independent GWAS (all and stage B endometriosis cases versus controls) would give $P > 5 \times 10^{-8}$ ($P_{all\ adjusted} = 1.7 \times 10^{-7}$; $P_{stage\ B\ adjusted} = 8.2 \times 10^{-8}$). However, the accurately imputed ($R^2 > 0.95$) SNP rs2861694 ($P_{stage\ B} = 7.9 \times 10^{-9}$), in strong LD with rs4141819 ($r^2 = 0.981$, $D' = 1.0$, and $r^2 = 0.867$, $D' = 1.0$, in the 379 European and 286 Asian 1000 Genomes Project reference samples, respectively), would retain genome-wide significance ($P_{stage\ B\ adjusted} = 1.6 \times 10^{-8}$).

The quantile-quantile plots for the QIMR-HCS, OX and BBJ GWAS data (**Supplementary Fig. 17a–c**) reflect our stringent quality control, whereas the GWAS meta-analysis quantile-quantile plot (**Supplementary Fig. 17d**) shows a significant preponderance of SNPs with small *P* values of $<1 \times 10^{-3}$, suggesting that many of these nominally significant SNPs are likely to represent true signals[12].

To further examine the shared genetic risk across our populations of European and Japanese ancestry, we performed polygenic

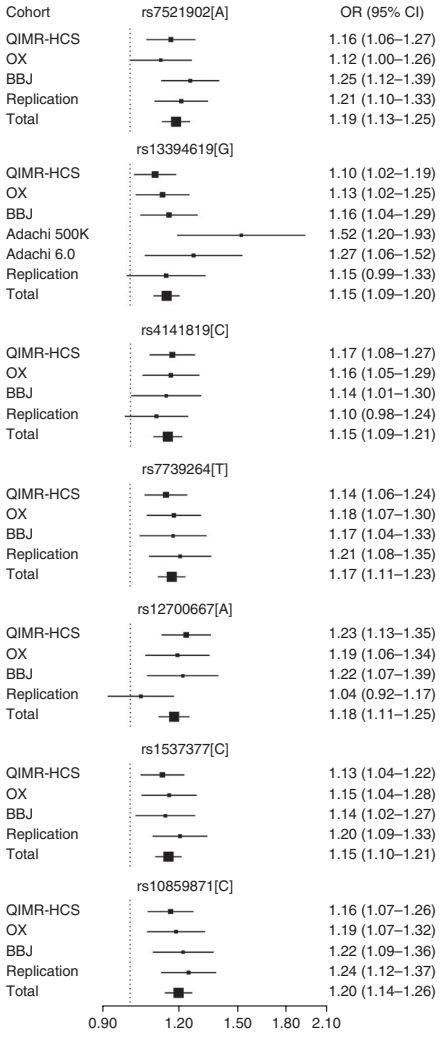

Figure 3 Forest plots of risk allele effects for the seven SNPs reaching genome-wide significance in the individual and total endometriosis case-control cohorts.

prediction analysis[13] to evaluate whether the aggregate effects of many variants of small effect in the BBJ GWAS cohort could predict affected status in the GWAS cohorts of European descent. The BBJ-derived risk scores significantly predicted affected status in the QIMR-HCS (R^2 = 0.0064; $P = 6.9 \times 10^{-7}$), OX ($R^2 = 0.0057$; $P = 9.6 \times 10^{-6}$) and combined QIMR-HCS and OX all-endometriosis case-control sets ($R^2 = 0.0054$; $P = 8.8 \times 10^{-11}$). For the individual and combined QIMR-HCS and OX case-control sets, the variance explained peaked in the SNP sets with BBJ GWAS P of <0.1, using all genome-wide association meta-analysis SNPs (**Fig. 4a**) and after excluding all SNPs within 2,500 kb of the seven implicated SNPs listed in **Table 1** (**Fig. 4b**). Analogously, performing the prediction in reverse, the risk scores from the combined QIMR-HCS and OX sample significantly predicted affected status in the BBJ case-control set ($R^2 = 0.0106$; $P = 3.3 \times 10^{-6}$) (**Supplementary Fig. 18** and **Supplementary Note**).

A gene-based genome-wide association analysis using the VEGAS (versatile gene-based association study) program[14], which accounts for gene size and LD between SNPs, identified 1,184 genes with combined P of ≤0.05 and determined that the top 3 ranked genes associated with endometriosis were *WNT4* at 1p36.12 ($P = 5.0 \times 10^{-9}$), *VEZT* at 12q22 ($P = 5.7 \times 10^{-7}$) and *GREB1* at 2p25.1 ($P = 2.5 \times 10^{-5}$) (**Supplementary Table 3).** In addition to identifying SNPs that reached genome-wide significance near the top three genes, we found that the *WNT4* and *VEZT* genes easily surpassed our conservative gene-based threshold for significant association of $P \leq 2.85 \times 10^{-6}$ (calculated as P = 0.05/17,538 independent genes). *WNT4* encodes wingless-type MMTV integration site family, member 4, and is important for the development of the female reproductive tract[15] and steroidogenesis[16]. *VEZT* encodes vezatin, an adherens junctions transmembrane protein that is downregulated in gastric cancer[17]. *GREB1* encodes growth regulation by estrogen in breast cancer 1, an early response gene in the estrogen regulation pathway that is involved in hormone-dependent breast cancer cell growth[18]. For the four remaining implicated regions at 2p14, 6p22.3, 7p15.2 and 9p21.3, no genes showed significant association ($P \leq 1.3 \times 10^{-3}$) after adjusting VEGAS results for testing 37 genes across all 7 regions (**Table 2**, **Supplementary Figs. 3–9** and **Supplementary Table 4**).

In conclusion, given their high gene-based ranking, proximity to genome-wide significant SNPs, known pathophysiology and reported gene expression (**Supplementary Fig. 19** and **Supplementary Note**), the *WNT4*, *VEZT* and *GREB1* genes are strong candidates for further studies aimed at understanding the molecular pathogenesis of endometriosis. Our results also suggest that a considerable number of SNPs that were nominally implicated (for example, at $P < 0.1$) in the GWAS cohorts of individuals of European and Japanese descent represent true endometriosis risk loci. Moreover, the significant overlap in common polygenic risk for endometriosis indicates that genetic risk prediction and future targeted disease therapy may be transferred across these populations.

URLs. Catalog of Published Genome-Wide Association Studies, http://www.genome.gov/gwastudies/; Gene Expression Omnibus (GEO) database, http://www.ncbi.nlm.nih.gov/gds/; Genevar database, http://www.sanger.ac.uk/resources/software/genevar/; GWAMA, http://www.well.ox.ac.uk/gwama/; MaCH, http://www.sph.umich.edu/csg/abecasis/MaCH/; Mammalian Gene Expression Uterus database (MGEx-Udb), http://resource.ibab.ac.in/cgi-bin/MGEXdb/microarray/scoring/interface/Homepage.pl;

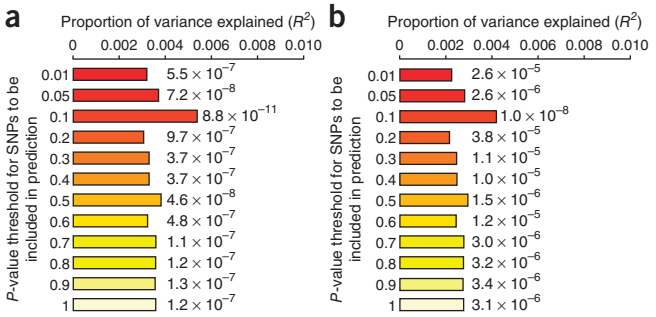

Figure 4 Allele-specific score prediction for endometriosis, using the BBJ population as the discovery data set and the combined QIMR-HCS and OX population as the target data set. (**a**,**b**) The variance explained in the target data set on the basis of allele-specific scores derived in the discovery data set for 12 significance thresholds. The *y* axis indicates Nagelkerke's pseudo R^2, representing the proportion of variance explained. The number above each bar is the P value for the target data set prediction analysis (R^2 significance). Prediction was performed using all GWAS meta-analysis SNPs (**a**) and after excluding all SNPs within 2,500 kb of the seven implicated SNPs listed in **Table 1** (**b**). The results were not driven by a few highly associated regions, indicating that a substantial number of common variants underlie endometriosis risk.

METAL, http://genome.sph.umich.edu/wiki/METAL_Program; METASOFT, http://genetics.cs.ucla.edu/meta/index.html; minimac, http://genome.sph.umich.edu/wiki/Minimac; 1000 Genomes Imputation Cookbook, http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook; 1000 Genomes Project, http://www.1000genomes.org/; PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/; SNPSpD, http://genepi.qimr.edu.au/general/daleN/SNPSpD/; Wellcome Trust Case Control Consortium, http://www.wtccc.org.uk/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge with appreciation all the women who participated in the QIMR, OX and BBJ studies. We thank Endometriosis Associations for supporting study recruitment. We also thank the many hospital directors and staff, gynecologists, general practitioners and pathology services in Australia, the UK and the United States who provided assistance with confirmation of diagnoses. We thank Sullivan and Nicolaides Pathology and the Queensland Medical Laboratory Pathology for pro bono collection and delivery of blood samples and other pathology services for assistance with blood collection. The HCS team thanks the men and women of the Hunter region who participated in the study.

We thank B. Haddon, D. Smyth, H. Beeby, O. Zheng, B. Chapman and S. Medland for project and database management, sample processing, genotyping and imputation. We thank Brisbane gynecologist D.T. O'Connor for his important role in initiating the early stages of the project and for confirmation of diagnosis and disease stage from clinical records of many cases, including 251 in these analyses. We are grateful to the many research assistants and interviewers for assistance with the studies contributing to the QIMR collection. The QIMR study was supported by grants from the National Health and Medical Research Council (NHMRC) of Australia (241944, 339462, 389927, 389875, 389891, 389892, 389938, 443036, 442915, 442981, 496610, 496739, 552485 and 552498), the Cooperative Research Centre for Discovery of Genes for Common Human Diseases (CRC), Cerylid Biosciences (Melbourne) and donations from N. Hawkins and S. Hawkins. D.R.N. was supported by the NHMRC Fellowship (339462 and 613674) and Australian Research Council (ARC) Future Fellowship (FT0991022) schemes. S.M. was supported by NHMRC Career Development Awards (496674 and 613705). E.G.H. (631096) and G.W.M. (339446 and 619667) were supported by the NHMRC Fellowship scheme. The HCS was funded by the University of Newcastle, the Gladys M Brawn Fellowship scheme and the Vincent Fairfax Family Foundation in Australia.

We thank L. Cotton, L. Pope, G. Chalk and G. Farmer. We also thank P. Koninckx, M. Sillem, C. O'Herlihy, M. Wingfield, M. Moen, L. Adamyan, E. McVeigh, C. Sutton, D. Adamson and R. Batt for providing diagnostic confirmation. The work presented here was supported by a grant from the Wellcome Trust (WT084766/Z/08/Z) and makes use of Wellcome Trust Case Control Consortium 2 (WTCCC2) control data generated by the WTCCC. A full list of the investigators who contributed to the generation of these data is available at the Wellcome Trust website (see URLs). Funding for the WTCCC project was provided by the Wellcome Trust under awards 076113 and 085475. C.A.A. was supported by a grant from the Wellcome Trust (098051). A.P.M. was supported by a Wellcome Trust Senior Research Fellowship. S.H.K. is supported by the Oxford Partnership Comprehensive Biomedical Research Centre, with funding from the Department of Health National Institute for Health Research (NIHR) Biomedical Research Centres funding scheme. K.T.Z. is supported by a Wellcome Trust Research Career Development Fellowship (WT085235/Z/08/Z).

We thank the members of the Rotary Club of Osaka-Midosuji District 2660 Rotary International in Japan for supporting our study. This work was conducted as part of the BioBank Japan Project that was supported by the Ministry of Education, Culture, Sports, Science and Technology of the Japanese government.

AUTHOR CONTRIBUTIONS

Manuscript preparation and final approval: D.R.N., S.-K.L., C.A.A., J.N.P., S.U., A.P.M., S.M., S.D.G., A.K.H., N.G.M., J.A., E.G.H., M.M., R.J.S., S.H.K., S.A.T., S.A.M., S.A., K.T., Y.N., K.T.Z., H.Z. and G.W.M. Study conception and design: D.R.N., S.M., Y.N., K.T.Z., H.Z. and G.W.M. **GWAS data collection, sample preparation and clinical phenotyping:** J.N.P., S.U., A.K.H., N.G.M., J.A., E.G.H., M.M., R.J.S., S.H.K., S.A.T., K.T.Z., H.Z. and G.W.M. **Replication data collection, sample preparation and clinical phenotyping:** S.A., K.T. and H.Z. **Replication genotyping:** H.Z. **Data analysis:** genome-wide association analysis: D.R.N., C.A.A. and S.-K.L.; imputation and replication analysis: D.R.N. and S.-K.L.; polygenic prediction, gene-based analysis and meta-analysis: D.R.N. **Obtaining study funding:** D.R.N., S.M., N.G.M., S.H.K., S.A.T., S.A.M., Y.N., K.T.Z. and G.W.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2445.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Uno, S. *et al.* A genome-wide association study identifies genetic variants in the *CDKN2BAS* locus associated with endometriosis in Japanese. *Nat. Genet.* **42**, 707–710 (2010).
2. Painter, J.N. *et al.* Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. *Nat. Genet.* **43**, 51–54 (2011).
3. Treloar, S.A., O'Connor, D.T., O'Connor, V.M. & Martin, N.G. Genetic influences on endometriosis in an Australian twin sample. *Fertil. Steril.* **71**, 701–710 (1999).
4. Montgomery, G.W. *et al.* The search for genes contributing to endometriosis risk. *Hum. Reprod. Update* **14**, 447–457 (2008).
5. Gao, X. *et al.* Economic burden of endometriosis. *Fertil. Steril.* **86**, 1561–1572 (2006).
6. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
7. McEvoy, M. *et al.* Cohort profile: The Hunter Community Study. *Int. J. Epidemiol.* **39**, 1452–1463 (2010).
8. American Society for Reproductive Medicine. Revised American Society for Reproductive Medicine classification of endometriosis: 1996. *Fertil. Steril.* **67**, 817–821 (1997).
9. Adachi, S. *et al.* Meta-analysis of genome-wide association scans for genetic susceptibility to endometriosis in Japanese population. *J. Hum. Genet.* **55**, 816–821 (2010).
10. Goumenou, A.G., Arvanitis, D.A., Matalliotakis, I.M., Koumantakis, E.E. & Spandidos, D.A. Loss of heterozygosity in adenomyosis on hMSH2, hMLH1, p16Ink4 and GALT loci. *Int. J. Mol. Med.* **6**, 667–671 (2000).
11. Martini, M. *et al.* Possible involvement of hMLH1, $p16^{INK4a}$ and PTEN in the malignant transformation of endometriosis. *Int. J. Cancer* **102**, 398–406 (2002).
12. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
13. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
14. Liu, J.Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).
15. Vainio, S., Heikkila, M., Kispert, A., Chin, N. & McMahon, A.P. Female development in mammals is regulated by Wnt-4 signalling. *Nature* **397**, 405–409 (1999).
16. Guo, X. *et al.* Down-regulation of *VEZT* gene expression in human gastric cancer involves promoter methylation and miR-43c. *Biochem. Biophys. Res. Commun.* **404**, 622–627 (2011).
17. Boyer, A. *et al.* WNT4 is required for normal ovarian follicle development and female fertility. *FASEB J.* **24**, 3010–3025 (2010).
18. Rae, J.M. *et al.* GREB 1 is a critical regulator of hormone dependent breast cancer growth. *Breast Cancer Res. Treat.* **92**, 141–149 (2005).

ONLINE METHODS

GWAS samples and phenotyping. Initially, 2,351 surgically confirmed endometriosis cases were drawn from women recruited by the QIMR study[19], and a further 1,030 cases were obtained from women recruited by the Oxford Endometriosis Gene (OXEGENE) study. Australian controls consisted of 1,870 individuals recruited by QIMR[2] and 1,244 individuals recruited by the HCS[7]. UK controls encompassed 6,000 individuals provided by the WTCCC2. Approval for the studies was obtained from the QIMR Human Ethics Research Committee, the University of Newcastle and Hunter New England Population Health Human Research Ethics Committees and the Oxford regional multicenter and local research ethics committees. Informed consent was obtained from all participants before testing[2].

All Japanese GWAS case and control samples were obtained from BioBank Japan at the Institute of Medical Science at the University of Tokyo. A total of 1,423 cases were diagnosed with endometriosis by the presence of multiple clinical symptoms, physical examinations and/or laparoscopy or imaging tests. We used 1,318 female control samples from healthy volunteers from the Osaka-Midosuji Rotary Club (Osaka, Japan) and women in BioBank Japan who were registered to have no history of endometriosis. All participants provided written informed consent to this study. The study was approved by the ethical committees at the Institute of Medical Science at the University of Tokyo and the Center for Genomic Medicine at the RIKEN Yokohama Institute.

GWAS genotyping and quality control. QIMR and OX cases and QIMR controls were genotyped at deCODE genetics on Illumina 670-Quad (cases) and 610-Quad (controls) BeadChips. HCS controls were genotyped at the University of Newcastle on 610-Quad BeadChips (Illumina). The WTCCC2 controls were genotyped at the Wellcome Trust Sanger Institute using Illumina HumanHap1M BeadChips. Genotypes for QIMR cases and controls were called with Illumina BeadStudio software. Standard quality control procedures were applied as outlined previously[20]. Briefly, individuals with call rate of <0.95 and SNPs with mean BeadStudio GenCall score of <0.7, call rate of <0.95, Hardy-Weinberg equilibrium P value of $<1 \times 10^{-6}$ or minor allele frequency (MAF) of <0.01 were excluded. Cryptic relatedness between individuals was identified through a full identity-by-state (IBS) matrix. Ancestry outliers were identified using data from 11 populations from HapMap 3 and 5 Northern European populations genotyped by the GenomeEUtwin Consortium using EIGENSOFT[21,22]. To increase the power of the Australian GWAS data set, we matched the existing QIMR cases and controls[2] by ancestry to individuals from the HCS[7] genotyped on Illumina 610-Quad chips. After stringent quality control, the resulting QIMR-HCS cohort consisted of 2,262 endometriosis cases and 2,924 controls, increasing the Australian effective sample size by 24% (ref. 2).

Quality control procedures for the OX genotype data resulted in the removal of SNPs with genotype call rate of <0.99 and/or heterozygosity of <0.31 or >0.33. Genome-wide IBS was estimated for each pair of individuals, and one individual from each duplicate or related pair (IBS > 0.82) was removed. Genotype data were combined with data from the Utah residents of Northern and Western European ancestry (CEU), Han Chinese in Beijing, China (CHB) and Japanese in Tokyo, Japan (JPT), and Yoruba from Ibadan, Nigeria (YRI) HapMap 3 reference populations, and individuals who did not have Northern European ancestry were identified using EIGENSOFT and subsequently removed. SNPs with genotype call rate of <0.95 were removed, and this threshold was increased to 0.99 for SNPs with MAF of <0.05. In addition, SNPs showing (i) deviation from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$); (ii) difference in call rate between the 1958 British Birth Cohort (58BC) and National Blood Service (NBS) control groups ($P < 1 \times 10^{-4}$); (iii) difference in allele and/or genotype frequency between control groups ($P < 1 \times 10^{-4}$); (iv) difference in call rate between cases and controls ($P < 1 \times 10^{-4}$) and (v) MAF of <0.01 were removed[2].

The BBJ cases and controls were genotyped using the Illumina HumanHap550v3 Genotyping BeadChip. Quality control filtering required sample call rate of ≥0.98, IBS analysis was used to exclude samples with close relatedness and principal-component analysis was used to exclude non-Asian samples. We also performed SNP quality control (call rate of ≥0.99 in both cases and controls and Hardy-Weinberg equilibrium P of $\geq 1 \times 10^{-6}$ in controls). In total, 460,945 SNPs on all chromosomes passed the quality control filters and were further analyzed[1].

Genome-wide association meta-analysis. For SNPs passing quality control, tests of allelic association (–assoc) were performed using PLINK[23] in the separate QIMR-HCS, OX and BBJ GWAS data sets. The primary meta-analysis of all endometriosis cases versus controls in the QIMR-HCS, OX and BBJ GWAS data was performed using a fixed-effect (inverse variance–weighted) model, where the effect size estimates, β coefficients, are weighted by their estimated standard errors using GWAMA software[24].

The P-value threshold of 7.2×10^{-8} for GWAS of dense SNPs and resequencing data[25,26] was used to define association at genome-wide significance, and SNPs with association at $P \leq 1 \times 10^{-5}$ were considered to show a suggestive association (this threshold is also used in the online Catalog of Published Genome-Wide Association Studies).

Given the substantially greater genetic loading of moderate-to-severe (stage B) endometriosis (rAFS stage 3 or 4 disease) compared to minimal (stage A) endometriosis (rAFS stage 1 or 2 disease)[2], a secondary analysis was performed for suggestive SNPs (associated at $P \leq 1 \times 10^{-5}$), where we performed meta-analysis of the association results from QIMR-HCS and OX stage B cases versus controls with the BBJ association results. As previously shown[2], the exclusion of minimal endometriosis cases has the potential to enrich true genetic risk effects, even taking into account the reduced sample size.

Consistency of allelic effects across studies was examined using the Cochran's Q test[27]. Between-study (effect) heterogeneity was indicated by Q statistic P values of <0.1 (ref. 28). Meta-analysis of SNPs associated at fixed-effect $P \leq 1 \times 10^{-5}$ that showed evidence of effect heterogeneity was also carried out using the recently developed Han and Eskin random-effects model (RE2) implemented in METASOFT software[29]. In contrast to the conventional DerSimonian-Laird random-effects model[30], the RE2 model increases power under heterogeneity[29].

Genotype imputation analysis. To assess the impact of variants not present on the Illumina BeadChips and better define the associated regions, we imputed genotypes in the region 2,500 kb upstream and downstream of the most significant genotyped SNP using the full reference panel from the 1000 Genomes Project Interim Phase 1 Haplotypes (2010–2011 data freeze, 2011–2006 haplotypes). Imputation was performed separately for the QIMR-HCS, OX and BBJ GWAS data sets with only the overlapping genotyped SNPs within 2,500 kb of the most significant genotyped SNP, using the MaCH and minimac programs[31,32] and following the two-step approach outlined in the online Minimac: 1000 Genomes Imputation Cookbook (see URLs). Analysis of imputed genotype dosage scores was performed using mach2dat[31,32] and PLINK. The quality of imputation was assessed by means of the R^2 statistic. Results for poorly imputed SNPs, defined as having R^2 of <0.3, were subsequently removed. The results from association analysis of imputed data in the QIMR-HCS, OX and BBJ data sets were then combined via meta-analysis of the β coefficients weighted by their estimated standard errors using GWAMA.

Replication samples and genotyping. Independent of the BBJ GWAS case-control cohort, a total of 1,044 cases and 4,017 controls were obtained from BioBank Japan and used for replication. We note that 653 of these 1,044 cases were also used in a small GWAS of 696 cases and 825 controls[9]. To maximally use all available association data for rs13394619, given that there is incomplete overlap between the cases in the previous GWAS and our replication cases and no overlap between the controls, we worked with the published results for rs13394619 in the previous GWAS and the results from comparing our non-overlapping 391 replication cases to our 4,017 replication controls.

The seven SNPs (rs7521902, rs13394619, rs4141819, rs7739264, rs12700667, rs1537377 and rs10859871) reaching genome-wide significance in the GWAS meta-analysis were genotyped in the independent Japanese replication cohort using the multiplex PCR-based Invader assay (Third Wave Technologies), as previously described[1].

Replication and total association analyses. Tests of allelic association were performed using PLINK in the independent Japanese replication cohort. Because only associations in the same direction were considered as evidence of replication, one-sided P values were obtained by halving the standard (two-sided) PLINK P values. To determine the total evidence for association, meta-analysis was performed on the one-sided replication P values with the QIMR-HCS, OX

doi:10.1038/ng.2445

and BBJ (and Adachi 500K (290 cases and 262 controls) and 6.0 (406 cases and 563 controls) for rs13394619)[9] GWAS *P* values using METAL[33]. The *P* values observed in each case-control cohort were converted into a signed *Z* score. *Z* scores for each allele were combined across samples in a weighted sum, with weights proportional to the square root of the sample size for each cohort[34]. Given that our cohorts had unequal numbers of cases and controls, we used the effective sample size, where $N_{effective} = 4/(1/N_{cases} + 1/N_{controls})$[33]. We also performed meta-analysis of the β coefficients weighted by their estimated standard errors using GWAMA to estimate the overall ORs and 95% CIs for the SNPs that reached genome-wide significance.

Polygenic prediction. The aim of the prediction analysis was to evaluate the aggregate effects of many variants of small effect. We summarized variation across nominally associated loci into quantitative scores and related the scores to disease status in independent samples. Although variants of small effect (for example, with genotype relative risk of 1.05) are unlikely to achieve even nominal significance, increasing proportions of true effects will be detected at increasingly liberal *P*-value thresholds, for example, $P < 0.1$ (~10% of all SNPs). Using such thresholds, we defined large sets of allele-specific scores in the discovery sample of the Japanese BBJ endometriosis case-control set (1,423 cases and 1,318 controls) to generate risk scores for individuals in the target sample of the QIMR-HCS (2,262 cases and 2,924 controls), OX (919 cases and 5,151 controls) and combined European-ancestry (QIMR-HCS and OX) endometriosis case-control sets (3,181 cases and 8,075 controls). The term risk score is used instead of risk, as it is impossible to differentiate the minority of true risk alleles from the non-associated variants. In the discovery sample, we selected sets of allele-specific scores for SNPs with the following levels of significance: $P < 0.01$, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0. For each individual in the target sample, we calculated the number of score alleles that they possessed, each weighted by the log OR from the discovery sample. To assess whether the aggregate scores reflect endometriosis risk, we tested for a higher mean score in cases compared to controls. Logistic regression was used to assess the relationship between target sample disease status and aggregate risk score. Nagelkerke's pseudo R^2 was used to assess the variance explained. Prediction was performed using all 407,632 SNPs overlapping in the QIMR-HCS, OX and BBJ GWAS data sets, and we then excluded the 6,163 SNPs within 2,500 kb of the 7 implicated SNPs listed in **Table 1**. We also performed the predictions in reverse, using risk scores from the combined QIMR-HCS and OX sample to predict affected status in the BBJ case-control set.

Gene-based association analysis. Gene-based approaches can be more powerful than traditional approaches that are based on data from individual SNPs in the presence of allelic heterogeneity. Therefore, using the QIMR-HCS, OX and BBJ GWAS data, we performed a genome-wide gene-based association study using VEGAS[14]. Briefly, for the 407,632 SNPs present in all three sets, the *P* values from the GWAS of individuals with European ancestry (fixed-effect meta-analysis of QIMR-HCS and OX GWAS data) and the *P* values from the Japanese (BBJ) GWAS were analyzed separately using VEGAS. The VEGAS test incorporates evidence for association from all SNPs across a gene and accounts for gene size (number of SNPs) and LD between SNPs by using simulations from the multivariate normal distribution. We performed meta-analysis on the resulting gene-based *P* values from individuals of European and Japanese descent using Stouffer's *Z*-score combined *P*-value method[34]. A total of 17,538 genes (including 50 kb 5′ and 3′ to their transcriptional start and end sites)[14] contained association results for at least 1 SNP, and a Bonferroni-adjusted significance threshold of $P \leq 2.85 \times 10^{-6}$ (0.05/17,538) was therefore used to indicate significant genome-wide gene-based association.

19. Treloar, S.A. *et al.* 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 (2005).
20. Medland, S.E. *et al.* Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am. J. Hum. Genet.* **85**, 750–755 (2009).
21. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
22. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
23. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
24. Mägi, R. & Morris, A.P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
25. Bajpai, A.K. *et al.* MGEx-Udb: a mammalian uterus database for expression-based cataloguing of genes across conditions, including endometriosis and cervical cancer. *PLoS ONE* **7**, e36776 (2012).
26. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
27. Cochran, W.G. The combination of estimates from different experiments. *Biometrics* **10**, 101–129 (1954).
28. Ioannidis, J.P., Patsopoulos, N.A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE* **2**, e841 (2007).
29. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
30. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).
31. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
32. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
33. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
34. Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. & Williams, R.M. *Adjustment During Army Life.* (Princeton University Press, Princeton, NJ, 1949).