Beyond Endometriosis Genome-Wide Association Study: From Genomics to Phenomics to the Patient

Krina T. Zondervan, DPhil1,2  Nilufer Rahmioglu, PhD2  Andrew P. Morris, PhD2,3  Dale R. Nyholt, PhD4  Grant W. Montgomery, PhD5  Christian M. Becker, MD1  Stacey A. Missmer, ScD6,7,8

1_Nuffield Department of Obstetrics and Gynaecology, Endometriosis Care Centre, University of Oxford, Oxford, United Kingdom  
2_Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom  
3_Dept of Biostatistics, University of Liverpool, Liverpool, United Kingdom  
4_Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Australia  
5_QIMR Berghofer Medical Research Institute, Brisbane, Australia  
6_Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts  
7_Chaming Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts  
8_Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts

Address for correspondence Krina T. Zondervan, DPhil, Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Level 3, Women’s Centre, John Radcliffe Hospital, Oxford OX3 9DU, United Kingdom (e-mail: krina.zondervan@obs-gyn.ox.ac.uk).

Semin Reprod Med 2016;34:242–254

Abstract

Endometriosis is a heritable, complex chronic inflammatory disease, for which much of the causal pathogenic mechanism remains unknown. Genome-wide association studies (GWAS) to date have identified 12 single nucleotide polymorphisms at 10 independent genetic loci associated with endometriosis. Most of these were more strongly associated with revised American Fertility Society stage III/IV, rather than stage I/II. The loci are almost all located in intergenic regions that are known to play a role in the regulation of expression of target genes yet to be identified. To identify the target genes and pathways perturbed by the implicated variants, studies are required involving functional genomic annotation of the surrounding chromosomal regions, in terms of transcription factor binding, epigenetic modification (e.g., DNA methylation and histone modification) sites, as well as their correlation with RNA transcription. These studies need to be conducted in tissue types relevant to endometriosis—in particular, endometrium. In addition, to allow biologically and clinically relevant interpretation of molecular profiling data, they need to be combined and correlated with detailed, systematically collected phenotypic information (surgical and clinical). The WERF Endometriosis Phenome and Biobanking Harmonisation Project is a global standardization initiative that has produced consensus data and sample collection protocols for endometriosis research. These now pave the way for collaborative studies integrating phenomic with genomic data, to identify informative subtypes of endometriosis that will enhance understanding of the pathogenic mechanisms of the disease and discovery of novel, targeted treatments.

Keywords

► endometriosis  
► genes  
► genome-wide association study  
► epigenetic  
► phenotype  
► EPHeCt

Issue Theme Lessons from Genome-Wide Association Studies in Reproduction; Guest Editor, Joop S. E. Laven, MD, PhD

Copyright © 2016 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel: +1(212) 584-4662.


ISSN 1526-8004.
Endometriosis is a chronic inflammatory condition affecting an estimated 176 million women worldwide in their reproductive years, associated with pelvic pain (dysmenorrhea, dyspareunia, and chronic pelvic pain) as well as reduced fertility. It is characterized by the presence of tissue resembling endometrium outside the uterine cavity. This is present mainly on pelvic surfaces such as the ovaries, uterosacral ligaments, peritoneal surface, and in deeper tissue such as the rectovaginal septum or the bladder. Definitive diagnosis can only be established through surgical (laparoscopic) visualization of the lesions ideally with histological verification, resulting in reported average diagnostic delays worldwide of 7 to 10 years from onset of symptoms. The disease is most commonly staged using the revised American Fertility Society (rAFS)/American Society for Reproductive Medicine (ASRM) stages I to IV classification, based on lesion number and size, presence of adhesions, and ovarian versus peritoneal involvement. Higher stages are assigned with the occurrence of ovarian cysts and adhesions, but do not correlate well with pain severity. Current treatment options are limited to surgical removal (often repeated) of lesions and/or hormone medication with considerable side effects.

The exact prevalence of endometriosis in the population is unknown, because of the need for a surgical diagnosis, but it is estimated to affect approximately 5 to 10% of women during the reproductive years; the prevalence increases to 35 to 50% in women having a laparoscopy for endometriosis-associated symptoms of pelvic pain and subfertility. The condition has substantial long-term impact on the quality of life of sufferers and their families, affecting general wellbeing, daily life, and work productivity. It also carries a substantial economic burden, with direct and indirect costs combined estimated at €9,500/woman/year, and treatment costs similar to those of type II diabetes and rheumatoid arthritis. Therefore, endometriosis is an area of major unmet personal, clinical, and societal need, for which novel treatments and methods of noninvasive diagnosis urgently need to be identified.

The path to providing new therapeutic options that benefit patients lies in improved understanding of the fundamental pathogenic processes underlying endometriosis. Although advances have been made in the past decade shedding more light on mechanisms that are likely to promote the maintenance and growth of lesions, the actual causes remain largely unclear. The most widely accepted origin of peritoneal lesions is through retrograde menstruation, in which menstrual debris containing viable endometrial cells enter the pelvic cavity via the Fallopian tubes. This hypothesis is supported by prospective epidemiological studies showing that increased “exposure to menstruation” (e.g., early age at menarche, increased menstrual flow, shorter cycle length) increases disease risk. As retrograde menstruation is common, research has focused on providing explanations for the selective adherence of endometrial cells and progressive growth into lesions, including the potential role of endometrial aberrations in women with endometriosis, altered peritoneal environment, impaired immune surveillance, and increased angiogenic capacity. Certainly, inflammatory and immune responses, angiogenesis, and apoptosis are altered in affected women favoring the survival and maintenance of endometriotic tissue. In particular, lesions produce estradiol, in a feedback cycle that favors the overexpression of P450 aromatase and decreases the expression of 17β-HSD2, while failing to respond to progesterone (“progesterone resistance”) required to counteract estradiol effects. This mechanism is considered one key process through which lesion maintenance and growth may be promoted in endometriosis, but, crucially, it is not known whether these processes are a necessary cause, or rather an effect of the disease process.

One method to improve understanding of complex disease pathogenesis is through the identification of genetic factors (DNA sequence variants) underlying disease risk. As germ-line (inherited) genetic variants in DNA are not subject to environmental influences, or a result of the disease process, a significant association between a variant and disease implies a role in its pathophysiology—provided the study is well designed and there are no systematic biases in genetic profiles between cases and controls. The involvement of genetic factors in the development of endometriosis is supported by numerous studies, and its heritability (the proportion of variability in disease risk in the population due to genetic factors) is estimated at 47 to 52%.

The first types of study design to unravel genetic factors underlying disease were so-called “linkage” studies in families containing multiple affected members. Linkage studies considered the statistical evidence of sharing of chromosomal regions between multiple cases within a family and between families. They assume that most of the disease risk in a family is attributable to the same genetic variant(s) shared by cases, and that other affected families have disease-causing variants in the same gene (“monogenic disease”). Linkage studies were extremely successful in the 1980s/1990s to uncover single rare variants responsible for monogenic diseases (e.g., cystic fibrosis, Huntington disease); however, they proved to be problematic in elucidating the genetic basis for complex diseases because of their multifactorial origin (multiple cases in the same and between families may not have the same underlying etiology of disease). Although significant linkage to chromosomes 10 and 7 was observed for endometriosis, the (likely rare) variants responsible for these linkages remain to be identified. Subsequently, candidate gene association studies based on biological hypotheses became commonplace to identify specific DNA variants associated with disease. Indeed, in endometriosis, hundreds such studies have been conducted but due to inherent issues in study design (e.g., validity of the biological hypothesis; few variants tested per gene; varying case definitions; small sample sizes), very few of their results have been replicated.

Given the disappointments with linkage and candidate gene association studies in complex disease, enthusiasm then grew for genome-wide association studies (GWAS). GWAS are hypothesis-free and test genetic variants across the entire genome for association with a disease or trait. From the year 2000 onward, several major parallel developments paved the way for GWAS to become feasible: (1) the sequencing of the human genome (the Human Genome Project);
(2) the discovery of common single base-pair variants that are highly abundant across the genome (single nucleotide polymorphisms [SNPs]) and that can be used for genome-wide association testing; and (3) the development of technology to genotype these SNPs at large scale and ever decreasing cost. Subsequent developments included the International HapMap project and recently, the 1000 Genomes Project that provided data on the correlation between SNPs in close proximity to each other (linkage disequilibrium [LD]) in different ethnic populations, thus allowing the statistical inference (“imputation”) of most of the known common SNPs in the human genome based on genotyping of only a few 100K SNPs. In addition, statistical methods and guidelines were developed for the appropriate analysis of GWAS, which arise from a large multiple-testing issue that could lead to many false-positive results. After quite some debate, a general genome-wide significance threshold per SNP of a p-value of less than $5 \times 10^{-8}$ was generally accepted by journals to indicate a statistically robust association.

Early GWAS findings were from analyses performed in individual and/or relatively small datasets (e.g., ~2,000 cases vs. 2,000 controls), and were therefore required to be supplemented by evidence of replication in independent datasets. However, the high statistical significance threshold coupled with the current practice of meta-analysis of multiple large independent GWAS datasets makes the possibility of a false-positive association signal being produced by random sampling variation extremely low. Furthermore, studies have shown that joint analysis of GWAS is more powerful than two-stage (discovery and replication) GWAS designs.

These developments led to wide-scale adoption of GWAS for the investigation of common genetic variants underlying complex disease. Indeed, since the first study in 2005, GWAS have been successful in identifying thousands of common DNA sequence variants associated with hundreds of human diseases and traits. Given the considerable uncertainty about the causes of endometriosis, and its substantial heritability, GWAS also have provided a promising approach in improving the understanding of the genetic contribution to the disease.

**Genome-Wide Association Study in Endometriosis**

**Overview of Studies to Date**

Five GWAS analyses have been published for endometriosis, comprising four independent datasets—one on women of Japanese ancestry and three on women of European ancestry. The first, smallest, GWAS published in 2010 including 696 cases and 825 controls yielded no genome-wide significant associations. This was quickly followed in the same year by Uno et al. including 1,907 Japanese cases and 5,292 controls from Biobank Japan (BBJ). They reported one genome-wide significant association in CDKN2B-AS1 ($p = 5.57 \times 10^{-12}$; odds ratio [OR] = 1.44; 95% confidence interval [CI]: 1.30–1.59) and two suggestive associations (rs13271465 in an intergenic region on chromosome 8, $p = 9.84 \times 10^{-6}$ and rs16826658 16.6 kb away from wingless-type MMTV integration site family member 4 [WNT4], $p = 1.66 \times 10^{-6}$).

In 2011, the International Endogenous Consortium (IEC) published the first GWAS in women of European ancestry (3,194 cases and 7,060 controls, with replication in 2,392 cases and 2,271 controls). They observed one genome-wide significant locus, in an intergenic region on chromosome 7 (rs12700667; $p = 1.4 \times 10^{-9}$; OR = 1.20; 95% CI: 1.13–1.32), and meta-analyzing their results for WNT4 with those reported by Uno et al. reported genome-wide significant association also for this locus (rs7521902; $p = 4.2 \times 10^{-8}$; OR = 1.19; 95% CI: 1.12–1.27). A subsequent collaborative genome-wide meta-analysis of the IEC and BBJ datasets in 2012 identified seven genome-wide significant loci, confirming 7p15.2, WNT4, and CDKN2B-AS1 while finding four new loci in GREB1 at 2p25.1 (rs13394619; $p = 6.1 \times 10^{-8}$; OR = 1.15; 95% CI: 1.09–1.20), near VEGF at 12q22 (rs10859871; $p = 5.5 \times 10^{-9}$; OR = 1.20; 95% CI: 1.14–1.26), near ID4 (rs7739264; $p = 3.6 \times 10^{-10}$; OR = 1.17; 95% CI: 1.11–1.23), and rs4141819 in an intergenic region on 2p14 ($p = 8.5 \times 10^{-8}$; OR = 1.15; 95% CI: 1.09–1.21). A recent further analysis of these datasets using 1,000 genomes imputation of genotypes additionally identified IL1A (OR: 1.21; 95% CI: 1.13–1.29; $p = 3.43 \times 10^{-8}$). The fourth GWAS, in women of European ancestry (2,019 cases and 14,471 controls), was published in 2013, confirming WNT4 and reporting a new intergenic locus on 2q23.3 ($p = 4.7 \times 10^{-8}$; OR: 1.20; 95% CI: 1.13–1.29).

Meta-analysis of the reported loci across all four independent GWAS and four replication datasets, totaling >11,000 cases and >30,000 controls, confirmed consistency of effects across datasets and populations for six of the reported loci (rs12700667 on 7p15.2, rs7521902 near WNT4, rs10859871 near VEGF, rs1537377 near CDKN2B-AS1, rs7739264 near ID4, rs13394619 in GREB1), with an additional two (rs1250248 in fibronectin 1 [FN1] and rs4141819 on 2p14) showing evidence close to the genome-wide significance threshold of $p < 5 \times 10^{-8}$. Details of all the loci and evidence of association are provided in Table 1.

**Insight into Disease Origin Heterogeneity**

GWAS have to be large in sample size (typically >2,000 cases) to detect, with genome-wide significance, the modest effect sizes that common genetic variants exert in complex disease. Most case datasets are therefore collected retrospectively and often lack detailed phenotypic information. Of the endometriosis GWAS conducted to date, two studies involving three datasets included surgically confirmed cases (IEC QIMR; IEC Oxford; Utah). In the IEC datasets, retrospective evaluation of records only allowed broad classification into stage B (rAFS stage III/IV defined by large lesion volume and/or large ovarian disease with many adhesions) and stage A (stage I/II defined by small lesion volume and allowing for only small ovarian disease with a few adhesions). The IEC used various statistical methods to show that stage B disease is more “genetically driven” than stage A: the relative contributions of common genetic variation assayed in the GWAS to stage B versus stage A disease were 31 versus 15%.
<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Location (chr bp)</th>
<th>Genomic annotation</th>
<th>Nearest gene (kb distance)</th>
<th>Meta-analysis p-value all endo (stage III/IV)</th>
<th>All endometriosis</th>
<th>Stage III/IV endometriosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OR (95% CI)</td>
<td>Sample size</td>
<td>OR (95% CI) Sample size</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(cases:controls)</td>
<td>(cases:controls)</td>
<td>(cases:controls)</td>
</tr>
<tr>
<td>rs7521902</td>
<td>1:22,490,474</td>
<td>Intergenic</td>
<td>WNT4 (21)</td>
<td>$1.8 \times 10^{-15}$ ($1.8 \times 10^{-10}$)</td>
<td>1.18 (1.13–1.23)</td>
<td>9,114:30,407 1.25 (1.16–1.33)</td>
</tr>
<tr>
<td>rs10859871</td>
<td>12:95,711,626</td>
<td>Intergenic; regulatory region variant</td>
<td>VEZT (17)</td>
<td>$4.8 \times 10^{-15}$ (6.8 × 10^{-7})</td>
<td>1.18 (1.13–1.22)</td>
<td>7,680:26,866 1.19 (1.11–1.27)</td>
</tr>
<tr>
<td>rs7739264</td>
<td>6:19,785,338</td>
<td>Intergenic; noncoding transcript variant</td>
<td>ID4 (52)</td>
<td>$6.2 \times 10^{-10}$ ($1.2 \times 10^{-8}$)</td>
<td>1.11 (1.08–1.15)</td>
<td>10,072:29,137 1.20 (1.13–1.28)</td>
</tr>
<tr>
<td>rs12700667</td>
<td>7:25,901,389</td>
<td>Intergenic</td>
<td>NFE2L3 (290/331)</td>
<td>$1.6 \times 10^{-9}$ (3.6 × 10^{-8})</td>
<td>1.13 (1.08–1.17)</td>
<td>11,506:32,678 1.22 (1.14–1.31)</td>
</tr>
<tr>
<td>rs7798431</td>
<td>7:25,860,562</td>
<td>Intronic splice variant</td>
<td>GREB1 (0)</td>
<td>$5.4 \times 10^{-9}$ (9.7 × 10^{-8})</td>
<td>1.13 (1.09–1.18)</td>
<td>8,734:24,633 1.24 (1.14–1.33)</td>
</tr>
<tr>
<td>rs1537377</td>
<td>9:22,169,450</td>
<td>Intergenic</td>
<td>CDKN2BAS1 (48/4)</td>
<td>$1.0 \times 10^{-8}$ (2.3 × 10^{-7})</td>
<td>1.12 (1.08–1.17)</td>
<td>7,680:26,866 1.18 (1.11–1.26)</td>
</tr>
<tr>
<td>rs1333049</td>
<td>9:22,125,253</td>
<td>Intronic</td>
<td>IL1A (2.3)</td>
<td>0.25 (0.60)</td>
<td>1.04 (0.98–1.10)</td>
<td>6,681:21,531 1.03 (0.94–1.12)</td>
</tr>
<tr>
<td>rs13394619</td>
<td>2:11,727,257</td>
<td>Intronic splice variant</td>
<td>GREB1 (0)</td>
<td>$2.9 \times 10^{-8}$ (2.1 × 10^{-3})</td>
<td>1.13 (1.07–1.20)</td>
<td>8,376:27,691 1.18 (1.11–1.24)</td>
</tr>
<tr>
<td>rs6542095</td>
<td>2:113,529,183</td>
<td>Downstream gene variant</td>
<td>IL1A (2.3)</td>
<td>$3.6 \times 10^{-7}$ (3.4 × 10^{-8})</td>
<td>1.15 (1.09–1.22)</td>
<td>5,300:10,218 1.21 (1.13–1.29)</td>
</tr>
<tr>
<td>rs4141819</td>
<td>2:67,864,425</td>
<td>Intronic; noncoding transcript variant</td>
<td>ETAA1 (227)</td>
<td>$8.8 \times 10^{-6}$ (9.2 × 10^{-8})</td>
<td>1.08 (1.04–1.12)</td>
<td>7,680:26,866 1.16 (1.09–1.24)</td>
</tr>
<tr>
<td>rs1250248</td>
<td>2:216,286,843</td>
<td>Intergenic</td>
<td>FNI (0)</td>
<td>$1.1 \times 10^{-4}$ (8.0 × 10^{-8})</td>
<td>1.11 (1.04–1.18)</td>
<td>8,443:14,190 1.26 (1.16–1.38)</td>
</tr>
<tr>
<td>rs6734792</td>
<td>2:151,624,632</td>
<td>Intergenic</td>
<td>RND3 (281)</td>
<td>$2.2 \times 10^{-6}$ (2.7 × 10^{-5})</td>
<td>1.10 (1.06–1.16)</td>
<td>6,636:22,849 1.10 (1.05–1.15)</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; OR, odds ratio.

*Human Genome Build 38.*

*Annotation according to Ensembl genome browser: [www.ensembl.org](http://www.ensembl.org) and UCSC genome browser: [https://genome.ucsc.edu/](https://genome.ucsc.edu/).*

*Meta-analysis $p$-value is from Rahmioglu et al.33 For rs4141819, rs1250248, and rs6542095, the meta-analysis $p$-values including only stage III/IV cases is given in parenthesis, which reach borderline significance.*
respectively, results that were supported by subsequent analyses. The genetic distinction between origins of stage A versus stage B became further apparent when considering the association at individual genetic loci ( Fig. 1 ).

Unfortunately, no other phenotypic information was collected in the individual case datasets, hampering further dissection of genetic heterogeneity between the surgically defined stages. Whether these differences can be attributed to the pathogenesis of ovarian disease or adhesion development cannot be determined. Furthermore, this rAFS stage-based dichotomization does not provide insight into genetic heterogeneity by different surgically or symptomatically defined case definitions.

Insights into Biology

Of the 12 “index” SNPs (defined as the SNP with the lowest p-value at a GWAS locus) that have been associated with endometriosis, only two are located within genes, in introns, while the remaining are intergenic at varying distances from known genes ( Table 1 ). Because of this, it has been difficult
to highlight specific causal molecular mechanisms through which endometriosis-associated variants impact on disease, and the genes through which their effects are mediated. For a more detailed review of the biology implicated by these loci, refer to the study by Rahmioglu et al.\textsuperscript{33}

Genetic variants in the identified loci are likely to have roles in regulation of expression of near or distantly located genes. Genomic annotation provided by the Encyclopedia of DNA Elements (ENCODE) project\textsuperscript{42} on a wide variety of cell types and tissues is useful to investigate the putative regulatory function of these variants. ENCODE is an international collaboration of research groups funded by the National Human Genome Research Institute (United States) to build a comprehensive list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active (https://www.encodeproject.org/). Regulatory elements are typically investigated through DNA hypersensitivity assays, assays of DNA methylation, and immunoprecipitation of proteins that interact with DNA and RNA, that is, modified histones, transcription factors, chromatin regulators, and RNA-binding proteins, followed by sequencing. Though information on endometrial tissue and cells is very limited, a major limitation for endometriosis research, many of the genomic annotation features are shared across different cell and tissue types.
The following genomic loci have been associated with endometriosis in GWAS:

- **Rs13394619** is an intronic splice variant (affecting the formation of different transcripts) in GREB1 (growth regulation by estrogen in breast cancer 1). GREB1 encodes for an early response gene involved in estrogen-induced growth of breast cancer cells. Its role in estrogen-induced endometri-um proliferation remains to be uncovered.

- **Rs1250248** is located in an intron of FN1, which is a transcription factor-binding site (ENCODE). FN1 mediates various cellular interactions with the extracellular matrix and has important roles in cell adhesion, migration, growth, and differentiation. It has been shown that SOX2, a transcription factor that targets FN1, is involved in mediating cell migration in ovarian cancer.

- **Rs7521902** is an intergenic variant that is located 21 kb downstream of WNT4. WNT4 encodes for a protein imperative for development of the female reproductive tract. Moreover, WNT4 is shown to be expressed in normal peritoneum, suggesting that endometriosis can arise through metaplasia using developmental pathways involved in development of the female reproductive tract.

- **Rs7739264** is an intergenic variant located 52 kb downstream of ID4 (inhibitor of DNA binding 4). ID4 is an oncogene that is overexpressed in most primary ovarian cancers but not in normal ovary and fallopian tubes. Moreover, it has been shown to be overexpressed in most ovarian, endometrial, and breast cancer cell lines. ID4 can potentially regulate HOXA9 and CDKN1A genes, which are mediators of cell proliferation and differentiation. More specifically, HOXA genes are involved in differentiation of fallopian tubes, uterus, cervix, and vagina.

- **Rs1537377 and rs1333049** are intergenic variants located 48 kb and 4 kb upstream of cyclin-dependent kinase inhibitor 2B antisense RNA (CDKN2B-AS1). CDKN2B-AS1 has been shown to mediate tumor suppressor genes, namely, CDKN2B, CDKN2A, and ARF expression. CDKN2A has previously been reported in endometriosis, adenomyosis, endometrial cancer through loss of heterozygosity or hyper-methylation of its promoter.

- **Rs10859871** is an intergenic variant, located in a regulatory region (DNase hypersensitivity cluster present in 71/125 cell types, ENCODE), 17 kb upstream of vezatin (VEZT). VEZT encodes an adherens junction transmembrane protein. It is also a putative tumor suppressor gene that targets cell migration, cell adhesion genes, and TCF19, a cell cycle progression gene which has been associated with lymphocyte count, mean cell hemoglobin, white blood cell count, hematocrit count, and eosinophil count. The collective evidence suggests VEZT can have a role in maintaining immunological balance through regulation of TCF19.

- **Rs6542095** is a downstream region variant located 2.3 kb of IL1A. IL1A (interleukin 1A) encodes for a member of the interleukin 1 cytokine family that is involved in proinflam-matory immune processes and hematoipoiesis. This protein may have a role in the inflammatory component of endometri-osis that has been suggested by co-occurrence of autoim-mune disease and endometriosis as well as success of anti-inflammatory drugs on relieving endometriosis-related pelvic pain.

The remaining four variants are in three independent intergenic regions that have no genes within 200 kb.

- **Rs12700667** is an intergenic variant in chromosome 7p15.2, and located in histone modification (acetylation and methylation) marks H3K27ac and H3K4me1 (ENCODE). The region contains multiple potentially relevant genes including (1) a transcription factor called NEF2L3 (nuclear factor erythroid derived 2-like 3) that regulates cell differentiation, inflammation, and carcinogenesis; (2) a microRNA, miRNA-148a which is involved in regulation of Wnt/β-catenin signaling pathway that has a documented role in communication of epithelial and stromal cells of the endometrium; and (3) two members of a transcription factors family called homeobox A, HOXA10 and HOXA11, that have a role in uterine development.

The intergenic signal on 2p14 is 227 kb away from ETAA1 (Ewing tumor-associated antigen 1) gene, which encodes for a Ewing family tumor-specific cell surface antigen. Likely more relevant is that this variant sits in a 188 kb long noncoding RNA called, AC007422.1, the function of which remains unknown.

The intergenic signal on 2q23.3 is located 280 kb upstream of RND3 (rho family GTPase 3) gene, encoding a member of Rho family of small guanosine triphosphate–binding proteins that is involved in regulating actin cytoskeleton and stress fiber formation. It is also a regulator of cell cycle progression, transformation, and cell migration.

We consider the magnitude of the p-value testing association for each of these loci with stage A versus stage B endometriosis in the IEC GWAS imputed to the latest 1000 Genomes panel (version 3). Of the 10 loci, 6 showed a strong signal for stage B disease, but no statistically significant association at all with stage A (Fig. 1): ID4, GREB1, 7q15.2, CDKN2B-AS1, FN1, and 2p14. WNT4 showed weaker association for stage A, much reduced compared with stage B (not shown). VEZT showed a similar strength of association for stage A and B, as did IL1A albeit with different SNP sets (not shown). Finally, the SNPs reported on 2q23.3 showed limited evidence of association in the IEC dataset (not shown), nor did it show consistent evidence in the meta-analysis by Rahmio et al. 33 Thus, the majority of GWAS loci identified to date are implicated most strongly in the origins of stage B (rASRM stage III/IV) disease. The discrepancy in association signals with stage I/II may be driven by inclusion of stage I cases, as our previous work showed stages II and III to be similar in terms of genetic burden.

Two of the signals, WNT4 and 7p15.2 point at WNT signaling as a key pathway involved in endometriosis pathogenesis. Interestingly, the 7p15.2 locus was also found to be genome-wide significantly associated with fat distribution (waist-to-hip ratio adjusted for body mass index) in an entirely independent GWAS. This led to a study leveraging the two GWAS datasets showing, for the first time, evidence of shared genetic origins and pathways between endometriosis and obesity.
endometriosis and fat distribution, mostly through the WNT/β catenin signaling pathway. In a similar type of analysis, shared genetic origins between endometriosis and ovarian cancer were explored, showing strong genetic correlations between endometriosis and clear-cell, endometrioid, and low-grade serous ovarian cancer.

The GWAS SNPs associated with endometriosis to date explain less than 2% of disease risk variability. In contrast, half of the estimated 50% heritability (26%) is estimated to be due to common variants that can be captured through GWAS. Successful examples from other complex diseases have shown that by increasing sample sizes, many more common genetic loci (with population frequencies > 1%) can be discovered. For instance, the most recent GWAS analyses in breast cancer involving 62,000 cases identified more than 90 GWAS loci, explaining 16% of disease risk. For Crohn disease, 140 common genetic loci have been found, explaining 14% of disease risk, in meta-analyses involving 20,000 cases. Extended endometriosis consortium efforts are currently ongoing to conduct further meta-GWAS studies including datasets with endometriosis diagnostic information worldwide, increasing sample size fourfold to around 20,000, as well as the first “exome-wide” analyses focusing on variants that directly impact gene protein products and allow more direct biological interpretation.

Utility of Endometriosis GWAS Results: What Are the Big Questions?

Broadly speaking, there are three areas of potential utility for GWAS results for a disease/trait: (1) improve understanding of pathogenesis; (2) dissection of phenotypic heterogeneity (discovery of “subtypes”); and (3) risk prediction. We will discuss the endometriosis GWAS results for each of these areas later.

Understanding the Pathogenesis

Making headway with inferring causality and functional mechanisms for pathogenesis from any GWAS results is hampered by two issues. First, the most significantly associated SNP with a disease may not be the causal variant but may be correlated (in LD) with the causal variant. Due to common ancestry of individuals in a population, SNPs located near each other are typically correlated (“in LD”). LD can stretch over many thousands of DNA base pairs and include 10 to 1,000s of SNPs. Although there are various ways to narrow down the region of association, including sequencing and fine mapping as well as statistical methods (e.g., through transethnic mapping or calculation of so-called “credible sets” of SNPs that have the highest probability to include the causal variant), these methods need to be supplemented by other functional evidence that associate a specific variant to actual transcription variation. Second, as mentioned earlier, most variants at GWAS loci are now known to be located in distal intergenic or intronic (regulatory) regions, with unknown effects on transcription of protein-coding genes and pathways. This feature, characteristic of GWAS variants, represents a major challenge we face in translating our GWAS signals into clinically meaningful results. Understanding the effects of genetic variants on perturbation of protein-coding “effector” transcripts and pathways requires integration with genomic annotation data, including gene expression (RNA) and DNA modification profiles, in tissues/cells relevant to the disease.

Recently, there have been major advances in understanding the regulatory architecture of the human genome, and, in particular, in the development of techniques for assessing the relationships between regulatory elements and the genes they control. Studies integrating gene expression data from a diverse range of tissues with genetic data have provided important insights into DNA variants that regulate gene expression (“expression quantitative trait loci SNPs [eQTL SNPs]”), showing that many eQTLs are tissue specific and that this tissue specificity is an important feature of disease susceptibility variants. Studies have similarly shown that genetic variants can also affect transcription through altering DNA methylation profiles (metQTLs), and—of particular significance—can affect how DNA is “packaged” by altering chromatin signatures that can create or destroy transcription factor binding sites and thereby alter the chromatin signature with long range (100 kb) effects on gene expression transcription factor binding. Many GWAS associations map to “DNase I hypersensitive” sites (DHS) that mark chromatin accessibility for transcription factor binding, which in turn have been associated with changes in RNA expression.

The importance of identifying which variant controls which gene was recently highlighted by analyses of regulatory control at the FTO locus (influencing BMI and obesity). The causal variants map intronic to FTO and it had been widely assumed that they regulate expression of this gene. However, it was subsequently shown that these FTO intronic SNPs lie within a region that regulates expression of a gene 500 kb away (IRX3), demonstrating that enhancers (DNA sequences affecting levels of transcription) of one gene can lie in the introns of another.

The crucial need for genomic annotation data to advance our knowledge of the mode of action of GWAS variants is now fully recognized. Large-scale initiatives such as ENCODE, the Genotype-Tissue Expression (GTEx) project and the NIH Roadmap Epigenomics Project are underway and aimed at building public databases providing tissue- and cell-specific genomic profiling landscapes (e.g., RNA transcription, DNA methylation, DHS) to enable investigations of the impact of genetic variation on these. However, none of these sources includes endometrium and its constituent cell types—the target tissue for endometriosis—preventing systematic progress on the translation of endometriosis GWAS signals. Also, none of the genome-wide expression studies of endometrium or endometrial cell types published to date has included genomic data that would allow the investigation of eQTLs.

To make headway with understanding the causality and functionality of endometriosis loci, genomic profiling (including eQTLs) of endometrium and endometrial cell types from 100s of women with and without the disease needs to be conducted. Tissue-based profiling has the advantage of assessment of profiles reflecting intercellular communications in their “natural
biological state”; however, cellular heterogeneity means that for this type of study in particular, sufficiently large sample sizes are required to detect signals. Such sample sizes can more easily be achieved in collaborative efforts between research groups; however, differences in case definitions and phenotypic data collected, as well as protocols for tissue sampling, have made combination and comparison of the data difficult. The WERF Endometriosis Phenome and Biobanking Harmonisation Project (EPHect) (see later), which has harmonized and standardized data and sample collection protocols for endometriosis research, now enables these large-scale collaborative studies to be conducted.

**Dissection of Phenotypic Heterogeneity**

The endometriosis GWAS to date have provided the first biological evidence that rAFS stage III/IV disease may arise from distinct pathophysiology compared with stage I/II. However, the limited information available from retrospectively assessed surgical records did not allow analysis of the disease element driving the staging for these cases nor other surgical/clinical phenotypes (e.g., deep infiltrating vs. peritoneal disease), as these data were not systematically collected. Moreover, the limited detail of surgical/clinical data also begged the question as to what the genetically dissected subtypes rAFS “I/II” versus “III/IV” signified. For example, since ovarian endometriosis (endometrioma) typically falls into the rAFS III/IV category, did the GWAS results associate predominantly to this disease entity, or to other disease aspects captured by this categorization? To allow this question to be answered, systematic collection of surgical and clinical data needs to accompany genetic profiling of samples.

Identification of functionally distinct disease subtypes based on molecular profiles will allow much-needed patient stratification for focused endometriosis drug target and biomarker discovery programs. In cancer, there are many successful examples of this approach, for example, the gene expression-based identification of basal, ER⁺/HER2−, HER2⁺/ER−, and “normal” breast cancer subtypes, which have different risk factor/survival rate profiles and targeted drug regimes (e.g., tamoxifen for ER⁺; trastuzumab for HER2 tumors). The Cancer Genome Atlas (TCGA) has embarked on comprehensive, large-scale molecular profiling (DNA genomic, DNA methylation, exome, mRNA/miRNA sequencing) to uncover further cancer subtypes and targeted treatments. To date, only one such study has been conducted in endometriosis—limited to using expression profiling of eutopic endometrium to distinguish cases from controls and rAFS stage III/IV from I/II. There is therefore clear scope for similar detailed molecular dissection studies in endometriosis, and to assess to what extent detailed surgical and clinical data are correlated with—or can capture—such molecularly defined subtypes. Again, to allow such studies to be conducted, we need detailed and standardized data collection on women with biological samples collected through standardized protocols.

**Endometriosis Risk Prediction**

With the successes in GWAS, disease risk prediction based on genetic profiles (“genetic risk scores”) in complex disease has been a rapidly emerging field of interest as the proportion of disease risk explained has been increasing due to larger studies performed. At present, the genetic loci discovered for endometriosis do not have utility in terms of disease risk prediction or screening. This is because, together, they explain less than 2% of all disease risk variance. As the sample size for future planned endometriosis GWAS grows, the proportion of variance explained by new loci will increase. Whether these will have utility in screening tools will depend on the proportion of disease risk they explain at a population level. More likely, screening tools will require combinations of genetic variants with other molecular profile changes between cases and controls (e.g., distinct epigenomic or transcriptomic patterns in eutopic endometrium), though whether such markers will be detectable in readily accessible samples such as blood is also unknown. In addition, inclusion of standardized collection of phenotypic profiles (e.g., clinical, symptomatic, metabolic) in screening tools that correlate with distinct molecular profiles between endometriosis cases and controls may be of benefit.

**Toward the Integration of Genomics with Phenomics: WERF EPHect**

As we highlighted earlier, the translation of endometriosis GWAS results in all three major areas of potential utility—elucidation of causal mechanisms, identification of subtypes of disease, and even risk prediction—will require integration with other molecular profiling as well as detailed phenotypic data. To allow the investigation of the full spectrum of potential disease-related effects, such phenotypic data can (and should) include surgical, clinical (symptomatic), and image-based manifestations of disease, as well as disease correlates such as comorbidities, covariate characteristics (e.g., BMI and fat distribution, cigarette smoking), and demographic factors. Houle et al referred to such deep phenotyping (the acquisition of high-dimensional phenotypic data on an organism-wide scale) as phenomics. Key to the valid acquisition and cross-study comparability of phenomic data are standardization of tools and measurements of data collection. Similarly, robust molecular phenotyping that allows comparisons among studies and centers requires standardization of the protocols with which samples are obtained. The WERF EPHect (endometriosisfoundation.org/ephect) was designed for exactly the following objectives: to standardize and harmonize data collection instruments and sample collection protocols for endometriosis research, reduce data variability, facilitate comparability, and encourage large-scale clinical and basic research collaborations.

WERF EPHect was established in 2013, through a global collaboration between 34 academic endometriosis research centers and 3 industry partners. Through two workshops and several consultation rounds, the Working Group reached a consensus on freely available instruments for surgical and clinical data collection, as well as standard operating protocols (SOPs) for the collection of 10 biological sample types (tissue and fluids in endometriosis research. All instruments have standard, as well as minimum versions.
The WERF EPHect Surgical Standard Form (SSF)\(^98\) was developed with the aim to collect all currently deemed relevant and important information describing the visual endometriosis phenotype and surgical treatment that would allow clinically and scientifically meaningful studies.\(^{102}\) It includes detailed information about clinical covariates: current menstrual cycle, current hormone treatment, and history of previous endometriosis surgery, as well as any imaging findings before the procedure. The second part concentrates on intraoperative findings including the type and duration of the procedure, and the extent, exact location, and color of endometriotic lesions, with a particular focus on size of endometrioma and endometriotic nodules. It allows for an exact description of tissue biopsies, including location and appearance, and surgical treatment of lesions. The WERF EPHect Endometriosis Phenotype Questionnaire Standard (EPQ-S)\(^99\) was developed to capture standardized clinical patient information, and includes validated instruments to assess pelvic pain, subfertility and reproductive history, menstrual history and hormone use, medical and surgical history, medication use, and personal information. For both the SSF and EPQ-S, minimum versions (MSF and EPQ-M) were developed for centers without research support; however, the standard tools are recommended as it allows a much broader phenotypic exploration.

The WERF EPHect standardized biological sample collection protocols were developed by systematic comparison and review of evidence using (1) information provided by the Working Group centers; (2) publicly available SOPs from general large-scale biobanking efforts (e.g., UK Biobank); large biorepositories (International Society for Biological and Environmental Biorepositories; the NCI Biorepositories and Biospecimen Research Branch; and the Australian Biospecimen Network); and (3) systematic literature searches in PubMed, of reference lists of retrieved articles, and of online material from biobanks and biorepositories. Based on review of the evidence, consensus SOPs were developed for the collection, processing, and storage of six biofluid types: blood and its derivatives (serum, plasma, and red/white blood cells), urine, saliva, peritoneal fluid, endometrial fluid, and menstrual effluent.\(^{100}\) Similarly, consensus on SOPs for collection, processing, and storage of five tissue types was reached (ectopic and eutopic endometrium, myometrium, and peritoneum).\(^{101}\) For most steps in the SOPs, two tiers were agreed upon: standard and minimum required. Although following standard protocol steps was recommended whenever possible, the minimum required SOP steps were offered to provide the fundamentals for standardization that need to be adhered to as an absolute minimum requirement given unavoidable logistical and budgetary circumstances.

Feedback on the WERF EPHect instruments is strongly encouraged through its Web site; the tools will be reviewed and updated, first after 1 year and then every 3 years. In addition, to facilitate collaboration, WERF EPHect has implemented a voluntary registry for centers using the tools after having obtained local Institutional Review Board approval (endometriosisfoundation.org/ephect).

**Conclusion**

The path to providing new therapeutic options for endometriosis, targeted to subtypes of disease, lies in improved understanding of the fundamental pathogenic processes underlying the disease. Although GWAS SNP risk loci typically have small effects, they can reveal novel insights into disease pathogenesis and do not imply that their implicated targets will have low therapeutic value.\(^{103}\) GWAS results for endometriosis have delivered novel identification of potential pathophysiologic pathways involved (e.g., WNT signaling). However, as the study sample sizes have been relatively modest, the number of GWAS loci robustly identified to date has been limited to seven; further loci will no doubt be discovered in ongoing meta-GWAS analyses that will increase sample size fourfold.

As is typical for GWAS signals—most of the endometriosis loci reside in intergenic or intronic regions that are likely to regulate gene expression elsewhere. These regulatory pathways, and the causal driver variants at each locus, remain to be uncovered; for example, through integrated gene-expression (eQTL) studies in endometriosis. The loci robustly associated with endometriosis in GWAS to date have also suggested that the pathogenesis of RFS stage III/IV disease is likely to be distinct from stage I/II. To understand this observation, further phenotypic dissection requires the collection of much more detailed, standardized surgical and clinical data, integrated with genomic and other molecular profiling of endometrium and other relevant samples from the same women. The WERF EPHect standardized data collection instruments and sample collection protocols now allow such data to be collected and compared across different endometriosis research centers, paving the way for studies focused on the translation of GWAS results into results that are meaningful for patients and practitioners: novel treatments that target subtypes of disease.

**Funding**

N.R. was supported by an MRC grant (MR/K011480/1). A.P. M. is a Wellcome Trust Senior Fellow in Basic Biomedical Science (WT098017). G.W.M. is supported by the NHMRC Fellowships Scheme (339446, 619667).

**References**

16 Zondervan KT, Cardon LR, Kennedy SH. What makes a good case-control study? Design issues for complex traits such as endometriosis. Hum Reprod 2002;17(6):1415–1423.
30 Wellcome Trust Case Control Consortium. Genome-wide association studies in 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661–678.