precisionlife
driving precision medicine

Disease Study

Identification of Genetic Risk Factors for Endometriosis

precisionlife.com

Identification of Genetic Risk Factors for Endometriosis

Executive Summary

Endometriosis is a painful debilitating condition where tissue similar to the lining of the womb starts to grow in other places, such as the ovaries and fallopian tubes. It can affect women of any age, and has a significant impact on quality of life and, in some cases, fertility.

Endometriosis affects 10% of all women, and up to 35% of infertile women. It is, however, poorly detected and often misdiagnosed. Women wait an average of 7.5 years before receiving a clear endometriosis diagnosis, and up to 75% of affected patients may be missed.

There is a great need for novel approaches to identify disease subtypes and understand the mechanisms involved to improve the diagnosis and treatment of this debilitating disease, and reduce costly and potentially harmful surgical interventions and downstream health consequences.

Our analysis identified disease-associated single nucleotide polymorphisms (SNPs) from genes that are strongly associated with the risk of developing endometriosis, and which have a clear mechanism of action connection to the disease symptoms. Four of these have already been directly linked to endometriosis in scientific literature, providing validation for our results. One of the SNPs was present in every endometriosis patient, indicating a potentially strong role in disease.

This analysis has generated the initial patient stratification that may enable the development of better, more personalized diagnostics and novel therapeutic options for a disease that affects 200 million patients around the world.

Methods

The PrecisionLife® platform identifies risk-associated SNPs and genes that are found to be significantly over-represented in a disease population. This platform uses a unique mathematical approach to identify high-order, disease-associated combinations of multimodal (e.g. SNPs, transcriptomic, epidemiological, or clinical) features at whole-genome resolution in large patient cohorts. It has been validated across multiple different disease populations.[1, 2, 3] This type of analysis is intractable to existing methods due to the combinatorial explosion posed by the analysis of large numbers of patients combined with the high numbers of features per patient.

When applied to genomic data, PrecisionLife finds high-order epistatic interactions (multi-SNP genotype signatures, typically of combinatorial order between 3 and 8) that are significantly more predictive of patients' phenotype than those identified using existing single SNP-based methods such as Genome-Wide Association Studies (GWAS). When evaluated in combination with each other, these SNPs can be highly significant in particular disease subpopulations. The phenotype with which the signatures are associated might be disease status, progression rate, therapy response, or other, depending on the data available and study design.

We conducted a hypothesis-free case:control study using genotype data from the UK Biobank, aiming to identify new risk variants associated directly with endometriosis, and gain further insights into the underlying pathology and disease mechanisms in relation to this patient group.

After quality control and removal of samples with missing data, we selected 4,848 women diagnosed with endometriosis (ICD-10 code N80.x) as our case population.

Endometriosis has been estimated to occur in up to 10% of all women, and in as many as 35% of infertile women.[4] As of May 2020, there were 264,810 women registered in the UK Biobank.[5] Based on this incidence rate, we would expect to see approximately 25,000 diagnosed endometriosis cases in the UK Biobank. However, there are only 6,272 women with an endometriosis-related ICD-10 code, indicating that there may be significant mis- or under-diagnosis within the entire UK Biobank female cohort. For this reason, we excluded any controls with the most common co-associated conditions and diseases that endometriosis is often misdiagnosed as, including IBS, uterine fibroids, and high discomfort during menstrual bleed. We selected 9,699 female controls who had not been diagnosed with endometriosis and met this criteria.

Having generated our case:control dataset, we used the PrecisonLife platform to find and statistically validate combinations of SNPs that together are strongly associated with endometriosis. These SNPs were mapped to the human reference genome to identify disease-associated and clinically relevant target genes.

Results

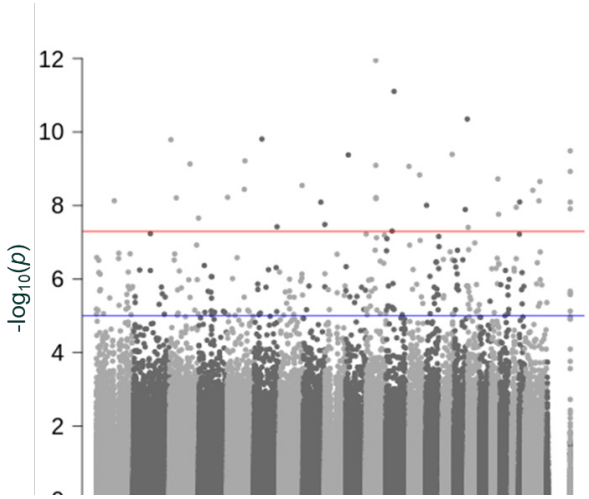
We identified 2,739 combinations of SNP genotypes (n-states) representing different groups of SNP genotypes that were highly associated with the endometriosis patient cohort. The majority (n=2,551) of SNPs were found in combinations of two or more SNP genotypes, and therefore would not have been found using standard GWAS analysis methods (see Table 1, Figure 1). At least one of the most critical (and predictive) SNP genotypes found at the center of the SNP networks shown in Figure 1 is found in every endometriosis case.

All of the SNP genotypes and their combinations were scored using a Random Forest (RF) algorithm based on a five-fold cross-validation method to evaluate the accuracy with which the SNP genotype combinations predict the observed case:control split.

Table 1 Summary of PrecisionLife endometriosis disease study run and results. Number of validated n-states, simple networks, and critical SNPs identified in the endometriosis study using a 5% False Discovery Rate and 250 cycles of fully random permutation.

	UK Biobank Endometriosis Study
FDR	5%
Disease signatures or n-states	2,739
Simple networks	1,016
Penetrance (cases represented by all n-states)	100%
RF-scored SNPs	586
RF-scored genes	345

Figure 1 Manhattan plot generated using a standard PLINK[6] GWAS analysis of genome-wide *p*-values of association for the endometriosis UK Biobank cohort. The horizontal red and blue lines represent the genome-wide significance threshold at $p<5e-08$ and $p<1e-05$, respectively.

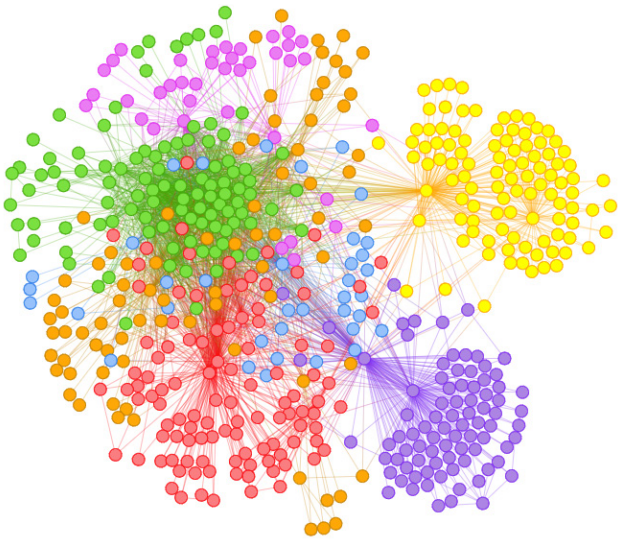


Clustering the SNPs by the patients in whom they co-occur allows us to generate a disease architecture of endometriosis (see Figure 2), providing useful insights into patient stratification. We can use this to find genes and biological pathways that are associated with particular patient subpopulations and comorbidities, enabling the development of disease biomarkers and precision medicine strategies.

When the SNPs were mapped to genes, we identified 345 protein-coding genes that are strongly associated with the risk of developing endometriosis. We identified a large number of genes involved in cell migration, with many linked to cancer in the context of promoting metastases. Several of these genes were also estrogen-responsive, and with differential expression in endometrial and ovarian cancers. Other genes we found regulated key processes such as cell adhesion, angiogenesis, and pro-inflammatory cytokine cascades.

Among the 15 highest RF-scored genes, four of them have already been directly linked to endometriosis in scientific literature, providing validation for our initial hypothesis-free analysis using UK Biobank endometriosis cases. Within those 15 genes, we also identified a glutamate receptor subunit that is involved in the amplification of neuropathic pain, which may indicate a genetic basis for pain associated with endometriosis in a subgroup of patients.

Figure 2 Disease architecture of endometriosis generated by the PrecisionLife platform. Each color identifies a distinct patient subpopulation group.



Future Directions

We are continuing our analysis of population-scale data for endometriosis patients in UK Biobank and as part of the FEMaLe consortium.

FEMaLe is a new, £5.3 million EU Horizon 2020 project that aims to develop precision medicine approaches to improve the treatment and quality of life of patients with endometriosis. The project is led by three researchers from Aarhus University in Denmark, who are heading a consortium of major international research and innovation partners including PrecisionLife, the University of Oxford, the University of Edinburgh, the University of Aberdeen, and Riga Technical University.

We are able to trace back all the significant disease signatures to the cases in which they are found, as well as any additional phenotypic and clinical data these cases have. We can select from a range of factors associated with endometriosis, including infertility, chronic pain, and endometrial and ovarian cancer. This will generate higher resolution stratification of endometriosis patient subgroups, and may provide greater insights into the underlying genetic factors relating to specific phenotypes of the disease.

We will also run adenomyosis (ICD-10 code N80.0) and endometriosis (ICD-10 codes N80.1–8) as two separate cohorts against the same controls used in this study, in order to gain further understanding of the genetic differences underlying both diseases.

Notes and References

1. Taylor, K., Das, S., Pearson, M., Kozubek, J., Strivens, M, & Gardner, S. (2019). Systematic drug repurposing to enable precision medicine: A case study in breast cancer. *Digital Medicine*, *5*(4), 180–186.
2. Gardner, S., Das, S. & Taylor, K. (2020). AI Enabled Precision Medicine: Patient Stratification, Drug Repurposing and Combination Therapies. In *Artificial Intelligence in Oncology Drug Discovery and Development*, Open Access book. https://doi.org/10.5772/intechopen.92594
3. Mellerup, E., Andreassen, O. A., Bennike, B., Dam, H., Djurovic, S., Jorgensen, M. B., Kessing, L., Koefoed, P., Melle, I., Mors, O., & Møller, G. L. (2017). Combinations of genetic variants associated with bipolar disorder. *PLoS One*, *12*(12), e0189739. https://doi.org/10.1371/journal.pone.0189739
4. Klemmt, P. A. B., & Starzinski-Powitz, A. (2018). Molecular and Cellular Pathogenesis of Endometriosis. *Current Women's Health Reviews*, *14*(2), 106–116. https://doi.org/10.2174/1573404813666170306163448
5. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z
6. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

UK

Unit 8b Bankside
Long Hanborough
Oxfordshire
OX29 8LJ

USA

1 Broadway
Cambridge
MA 02142

DENMARK

Agern Allé 3
DK-2970, Hørsholm

POLAND

CIC, Ul. Chmielna 73
00-801, Warszawa

info@precisionlife.com

