Appendix 1. Methods Supplement

We built a relational database for preeclampsia (dbPEC) that was implemented on MySQL version 5.1.66 running on a Linux server. All data retrieval, user interface and data output used PERL, PHP and HTML scripts. (Scripts are available upon request to authors). We used SciMiner, a semantic data mining and natural language processing program, to extract published articles from PubMed, focusing on preeclampsia-associated genes and protein information. The queries used are shown in Appendix 2, Supplemental Table 1. The web-based version of SciMiner provides robust methods for identification, filtering and extraction. However, users of the web-based version of SciMiner are restricted to the number of queries and/or retrievals they can pull out at a time. Thankfully, however, SciMiner allows users to download their entire package to a Linux-based operating system for extended and unrestricted use and retrievals. The functionality that SciMiner provides, different from customary manual search strategies of PubMed, is that filters can be created to identify articles with specific content and then SciMiner also has intelligent algorithms that will extract the desired information, in this case genes, proteins or sources, and deposit them into a form that can then be reviewed.

In order to mine the literature, we created a broad set of queries (Appendix 2, Supplemental Table 1), which identified different definitions of preeclampsia as well as additional preeclampsia-related phenotypes (e.g., HELLP, gestational hypertension (GH)). Our search strategy was designed to result in a more sensitive, rather than specific, pool of articles for

Triche EW, Uzun A, DeWan AT, Kurihara I, Liu J, Occhiogrosso R, et al. Bioinformatic approach to the genetics of preeclampsia. Obstet Gynecol 2014;123.

The authors provided this information as a supplement to their article.

the subsequent curation process. Drop-down menus and check-boxes in a web-based aggregation tool facilitated curation in a systematic fashion.

A curation team consisting of medical students formally trained in molecular biology, cell biology and genetics read and evaluated each article extracted by SciMiner. Study investigators (ET, AU, AD, JFP) met weekly with the curation team to go over questions about the curation process, discuss any articles with unclear findings, and share particularly interesting articles. Complicated articles, or any publication the curators felt would be useful to discuss and evaluate, were reviewed in the curation meeting. We assessed inter-rater reliability using kappa scores among the trained curators. [12,13] According to well-defined protocols and documentation (see Documentation Manual available upon request), the curators "accepted" or "rejected" each publication and each gene and/or genetic variant in the article based on the data and statistics presented. If no significant associations were found between genes or genetic variants and preeclampsia-related phenotypes, or the article was not an original scientific article (i.e. a review paper), it was rejected and no further documentation was recorded. 'Accepted' articles presenting original scientific research that contained any statistically significant associations between genes/variants and preeclampsia or related phenotypes were deposited into the database along with its unique PMID. Each article's reference list was further reviewed by the curator to identify additional potentially relevant articles that may have been missed by SciMiner. Identified articles were imported into the database and curated as described.

From accepted articles, each individual gene or variant examined in the article had the possibility of being "accepted" or "rejected" based on the statistical significance of its association with a preeclampsia-related phenotype. Each particular gene or variant from the article was extracted by SciMiner and pre-populated onto an article curation page. If a particular gene/variant was not significantly associated with preeclampsia or a related phenotype, it was rejected. Since primary data were not available to the curation team, we used the author's pre-specified levels of confidence for statistical testing. Genes/variants significantly associated with a preeclampsia-related phenotype were 'accepted.' Each was curated according to its association with the phenotypic characteristics, source (maternal, fetal or both), and presence of co-occurring conditions including intrauterine growth restriction (IUGR), HELLP syndrome, and/or gestational hypertension (GH).Information about single nucleotide polymorphisms (SNP), alternative mutation names, information or *rs* numbers and their associations with particular phenotypes were recorded.

The preeclampsia phenotype was coded to reflect the wide range of definitions we encountered for preeclampsia in the published literature (see the table below). Combinations of phenotypes (e.g., mild + severe) were only chosen when distinct analyses revealed significant associations for each phenotype separately. We also recorded the species for which the associations were significant (Yes/No for each of H. sapiens, M. musculus, R. norvegicus, other). Sometimes there were additional genes/variants in the article that were overlooked by SciMiner during the initial extraction. These variants were manually added to the pre-populated list and individually assessed according to the same protocol described above.

Triche EW, Uzun A, DeWan AT, Kurihara I, Liu J, Occhiogrosso R, et al. Bioinformatic approach to the genetics of preeclampsia. Obstet Gynecol 2014;123.

The authors provided this information as a supplement to their article.

Additional information on the genes and their associations were recorded in a "Notes" section. Curators reported any inconsistencies or interesting observations about the findings. The curators recorded pre-specified information, including a "verification", which was the curator's judgment of whether the findings are certain, plausible, or unlikely based on assessment of the design and analysis. Curators also recorded in the notes section any gene-gene interactions examined in the study and whether "pathway information" was discussed or identified in the article.

In addition, because the timing of the onset of preeclampsia is a potentially important phenotypic characteristic, we also recorded information on early (<34 weeks) or late onset (≥36 weeks) preeclampsia when specifically (separately) examined in the study. Choices were early, late, or early + late (early and late analyzed separately and significant associations found with each). Following a consistent file-naming protocol, text files containing the HGNC numbers of the genes significantly associated with each onset were specified in this section.

To avoid duplication, all genes were listed by their HGNC identification number. If SciMiner extracted a gene in a different format or if the annotations in the article used common names, the curators used HGNC to search for aliases in order to locate and input the unique HGNC ID. If the gene in the paper was listed by accession number or other names that could not be located in HGNC, curators checked in the NCBI Gene Finder (http://www.ncbi.nlm.nih.gov/gene) for the HGNC ID. When multiple choices were available for a gene name (e.g., EPHX1, 2, 3, 4 but THE paper only referred to EPHX), the curators

searched the article for additional information on chromosomal location or primers used to genotype the gene; this information was then searched in BLAT (http://genome.UCSC.edu/cgi-bin/hgBLAT?command=start). Such discrepancies were reconciled at the weekly curation meeting. For articles that studied specific variants (e.g., SNPs), curators listed the *rs* numbers in a separate section on the article curation database, but identified the relevant gene by its HGNC ID.

Triche EW, Uzun A, DeWan AT, Kurihara I, Liu J, Occhiogrosso R, et al. Bioinformatic approach to the genetics of preeclampsia. Obstet Gynecol 2014;123.

The authors provided this information as a supplement to their article.

Phenotypic Descriptions of Preeclampsia Identified in the Literature During the Curation Process From Extracted Publications

PE Phenotype Definition	Genes*	PMIDs*
Not specified	51	115
Mild only	15	7
Severe only	140	106
Greater than or equal to 140 or 90 + proteinuria	447	546
Mild and† severe	51	39
Greater than or equal to 140 or 90 + proteinuria and† severe	40	31
Severe and† not specified	3	2
Greater than or equal to 140 or 90 + proteinuria and† not specified	3	3
Eclampsia	2	2
Severe only and† eclampsia	2	2
Greater than or equal to 140 or 90 + proteinuria and† severe and† eclampsia	7	4
TOTAL	761	857

The columns show the number of genes 'accepted' for the specific phenotype and the number of 'accepted' articles (PMIDs) for the defined phenotype.

*Several genes and articles appear more than once.

† Indicates that the 2+ definitions were analyzed separately and each was found to be associated with the particular gene in a given article (PMID).

Triche EW, Uzun A, DeWan AT, Kurihara I, Liu J, Occhiogrosso R, et al. Bioinformatic approach to the genetics of preeclampsia. Obstet Gynecol 2014;123.

The authors provided this information as a supplement to their article.

