Towards Automatic Discovery of Genes Related to Human Placenta

Laritza M. Rodriguez, MD, PhD[1], Stephanie M. Morrison, MPH[1], Kathleen Greenberg, PhD[1,2], Dina Demner Fushman, MD, PhD[1]

[1]National Library of Medicine, Bethesda, MD; [2]ICF International, Fairfax, VA

One essential step in understanding complex pathways has been the compilation of disease-specific gene candidates extracted from the literature. Data analysis then makes it possible to create candidate gene assays and translate the results to the wet bench for biochemical research.

Our goal is to extract genes, gene pathways, biomarkers, and related events from the human placenta literature to create a specialized human placenta gene repository, and to identify pathways that can uncover target genes and gene therapies for pregnancy-related diseases. Here we present the first phase of the study: extraction of gene mentions from text.

METHODS

The document collection was retrieved from PubMed® using search filters and the following search terms: placenta, gene, biomarker, polymorphism, enzyme, preeclampsia, hypertension, diabetes, growth restriction. The tailored search returned 428 papers. To identify specific gene mentions from titles and abstracts we used the NLM MetaMap 2014 restricting processing to semantic type Gene. We then manually classified the extracted gene names into:

- Specific gene mentions (e.g. CDKN1A)
- General genetic terms (e.g. growth factor gene
- Errors (e.g. preeclampsia susceptibility).

RESULTS

In the 428 retrieved abstracts, MetaMap identified 413 documents with gene mentions. Overall, MetaMap identified 826 distinct gene names in titles and abstracts in the collection. Of these, we classified

- 753 as specific gene mentions
- 71 as general genetic terms
- 2 as errors

Table 1: Examples of the most frequent mentions

Gene Names	**# Mentions**
NLRP5 gene	178
NCR3 wt allele	83
CD8A wt allele	83
CD69 wt allele	71
PGF gene	26
LEP gene	18
General Genetic Terms	
Genes	246
Alleles	51
Genome	37
Human gene	34
Errors	
Genes, vif	50
Preeclampsia, susceptibility to	2

CONCLUSIONS

Our preliminary analysis shows that the placenta literature contains enough specific gene mentions to warrant further text mining on the genes of interest to identify pathways, biomarkers, and relationships between placenta gene expression and maternal and/or fetal diseases, ultimately identifying predictors of diseases that may clinically manifest only later in life.

REFERENCES

1. Roseboom TJ, Watson ED. The next generation of disease risk: are the effects of prenatal nutrition transmitted across generations? Evidence from animal and human studies. Placenta. 2012;33 Suppl 2:e40-e44. doi:10.1016/j.placenta.2012.07.018.
2. Cohen PR. DARPA's Big Mechanism program. Phys Biol. 2015;12(4):045008. doi:10.1088/1478-3975/12/4/045008.
3. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc JAMIA. 2010;17(3):229-236. doi:10.1136/jamia.2009.002733.