Endo-Mining: herramienta web para la búsqueda automatizada de genes potencialmente relacionados con la endometriosis a través de minería de textos en PubMed

Reproducibilidad, revisionabilidad y facilidad de uso potenciadas a través del lenguaje de programación R

Jorge Vallejo Ortega

24 de February, 2021

1 Palabras clave

Endometriosis, Text mining, Shiny.

2 Temática escogida

La endometriosis es una enfermedad del sistema reproductor femenino que afecta alrededor del 10% de mujeres en edad reproductora XXXnecesita citaXXX. Sus principales síntomas son infertilidad y dolor (Rolla 2019) por lo que, sin ser una enfermedad que amenace directamente la vida del paciente, sí supone una importante disminución en su calidad de vida.

El origen y patogénesis de esta enfermedad son desconocidos, y aunque existen varias teorías sobre su causa, ninguna ha sido probada de forma concluyente. El método de diagnóstico de referencia es la laparoscopia. Los métodos de diagnóstico por imagen (ultrasonidos, resonancia magnética) son menos invasivos pero no son capaces de alcanzar un diagnóstico en todos los casos. No se le conocen biomarcadores confiables, ni existe consenso sobre posibles factores de riesgo ambientales o genéticos(Rolla 2019).

El tratamiento medicamentoso de la endometriosis involucra el uso de analgésicos, hormonas y reguladores hormonales. De esta forma se tratan el dolor y la progresión de la enfermedad. La única forma de tratar las lesiones producidas por la enfermedad es mediante cirugía(Rolla 2019).

Una de las causas que complican la investigación de nuevos tratamientos, biomarcadores y factores de riesgo es la dificultad para modelizar esta enfermedad in vivo. La endometriosis se produce de forma natural sólo en animales con ciclo menstrual, lo que deja fuera de esta aproximación los animales más usados en laboratorio (ratones, ratas y conejos). Los primates no humanos, como los macacos, sí desarrollan endometriosis pero en un porcentaje muy escaso. Como consecuencia, los modelos en uso se limitan principalmente a estudios in silico, cultivos celulares, y animales transgénicos en los que está afectado algún gen supuestamente implicado en el desarrollo de la endometriosis.XXX nedds citation XXX

3 Problemática a resolver

La generación de modelos animales transgénicos, el estudio de biomarcadores, y la investigación de nuevas dianas terapéuticas dependen en gran medida de la propuesta de genes implicados en los mecanismos

del desarrollo de la endometriosis, gravedad de la enfermedad y respuesta a tratamientos. Existen varias estrategias complementarias para encontrar genes implicados en enfermedades o procesos biológicos en general: ensayos de expresión diferencial (mediante estudios de exomas o paneles de genes específicos), estudios de asociación del genoma completo (GWAS, *genome wide association study*) y minería de texto a partir de estudios previos relacionados (artículos científicos, tesis, patentes).

La estrategia económicamente más accesible y logísticamente más simple es la minería de texto, ya que aprovecha trabajos científicos ya publicados y permite resaltar los genes más prometedores de entre miles de candidatos. Ya existen artículos publicados que utilizan minería de textos para descubrir genes posiblemente implicados en endometriosis(Bouaziz et al. 2018), pero estos utilizan diferentes herramientas informáticas para los diferentes pasos del análisis, con el investigador tomando importantes decisiones entre paso y paso. Este esquema de trabajo, aunque funcional, genera varios problemas:

1. Reproducibilidad. Los pasos concretos seguidos en cada etapa del análisis no están señalados al detalle en las publicaciónes científica, lo que complica la repetición exacta del protocolo original.
2. Dificultad de revisión. Relacionado con el problema anterior. Al no tener un registro de cada paso y cada variable usada en los diferentes programas durante el análisis, es más difícil detectar errores o sesgos en el propio protocolo.
3. Conocimientos especializados. Utilizar diferentes herramientas informáticas para cada paso del análisis implica que quien quiera reproducir el análisis, o adaptarlo a su propia investigación, debe aprender a su vez a usar todas esas herramientas. El problema aumenta de magnitud cuando tenemos en cuenta que no todos los grupos de investigación cuentan con conocimientos básicos en bioinformática, necesitando aprender determinadas técnicas desde cero.

4 Objetivos

Este trabajo pretende alcanzar tres objetivos:

1. Reproducir estudios anteriores de minería de textos para la búsqueda de genes relacinados con la endometriosis. Se realizará mediante el desarrollo y uso de una herramienta propia.
2. Mejorar la reproducibilidad y facilitar la revisión de futuros estudios. Esto se realizará generando una herramienta en forma de script en lenguaje que permita reproducir fácilmente y de forma automática todo el proceso de búsqueda de datos, análisis, y presentación de resultados. El código estará disponible públicamente para que cualquier persona interesada pueda revisarlo, mejorarlo, y adaptarlo a sus propios proyectos.
3. Facilitar la repetición de este estudio a personas sin conocimientos en el lenguaje R, o en general sin grandes conocimientos de programación. Esto se conseguirá generando una interfaz web mediante la herramienta Shiny para usar el script anterior. Se buscará que el uso de la interfaz sea sencillo e intuitivo.

Referencias

Bouaziz, J., R. Mashiach, S. Cohen, A. Kedem, A. Baron, M. Zajicek, I. Feldman, D. Seidman, and D. Soriano. 2018. "How Artificial Intelligence Can Improve Our Understanding of the Genes Associated with Endometriosis: Natural Language Processing of the PubMed Database." *BioMed Research International* 2018 (March). doi:10.1155/2018/6217812.

Rolla, Edgardo. 2019. "Endometriosis: Advances and Controversies in Classification, Pathogenesis, Diagnosis, and Treatment." *F1000Research* 8. doi:10.12688/f1000research.14817.1.