エンリッチメント解析によるNGSデータの疾患や臓器からの生物学的解釈

情報・システム研究機構 (ROIS) データサイエンス共同利用基盤施設 ライフサイエンス統合データベースセンター (DBCLS)

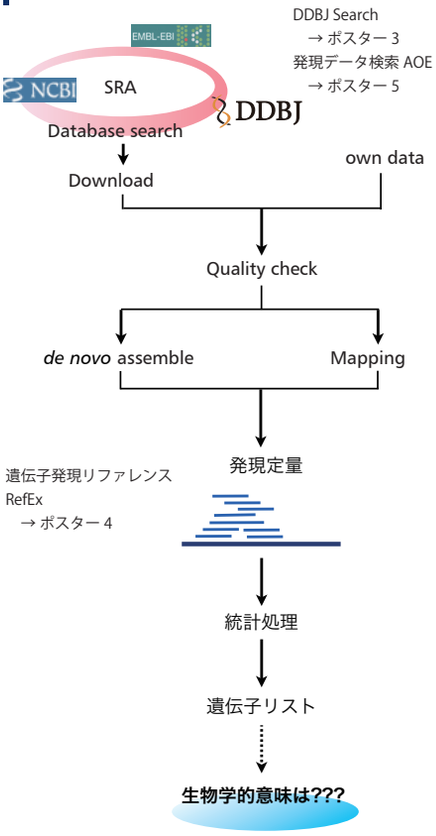

仲里 猛留 ○
Takeru Nakazato


坊農 秀雅
Hidemasa Bono

nakazato@dbcls.rois.ac.jp
@chalkless

ライフサイエンス統合データベースセンター（DBCLS）では、NGSデータ解析の「入り口」として検索エンジンDDBJ Search（DBCLS SRA）や、AOE、RefExを提供している。一方、「出口」としては結果の生物学的な解釈が必要不可欠である。一般的には得られた興味ある遺伝子セットに対してGene Ontology（GO）の用語やパスウェイの対応付けが行われる。我々はこれまで、GOのような細胞レベルの生命現象だけでなく、疾患や臓器のようなより個体レベルでの生命現象でも解釈が行えるようGendooシステムを開発してきた。当該遺伝子の関連文献を網羅的に収集し、付与されたMeSHのキーワードを抽出してスコアリングすることで遺伝子の特徴づけを行っている。現在、本システムを拡張し、個々の遺伝子に対する機能アノテーションから遺伝子セットに対してエンリッチメント解析を行えるよう改良を行っている。このことによりNGSデータの結果について細胞レベルから個体レベルまで幅広い生物学的な解釈が行えるようになる予定である。

RNA-Seq データ解析のフロー



NGS 解析というと、リードをどううまくつなぎ、遺伝子として組み上げ、どのくらいの発現量であったか、という点について、バイオインフォマティクスの課題として議論され、多くの手法が提案されてきた。もちろん、その点は非常に重要なのだが、実際のデータを解析するにあたっては、得られた（たとえば発現がある条件で上昇した、というような）遺伝子リストに対し、生物学的な意味づけを行うことが必要不可欠である。
これまで、BLAST をかけて遺伝子名を対応づける他、Gene Ontology や Pathway に対応づけての生物学的な意味づけが行われてきた。

参考文献

Gendoo: Functional profiling of gene and disease features using MeSH vocabulary
Nakazato T., Bono H., Matsuda H., Takagi T.,
Nucleic Acids Research, 37 (Suppl. 2) (Web Server issue), 2009
doi:10.1093/nar/gkp483

新たな切り口での「生物学的解釈」

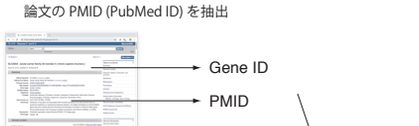
従来の Gene Ontology (GO) や pathway による生物学的解釈は分子や細胞レベルでの解釈であった。他に臓器や疾患という観点からの生物学的解釈も試みたい。そこで MeSH の用語を用いることとした。

MeSH (Medical Subject Headings)
http://www.nlm.nih.gov/mesh/
MEDLINE 収載の文献をインデキシングするためのキーワード集 (controlled vocabulary)
~23,000 語
15 分野 (Disease, Chemicals and Drugs, Anatomy, ...)
階層構造により語を整理
NLM (National Library of Medicine) により管理
MeSH terms

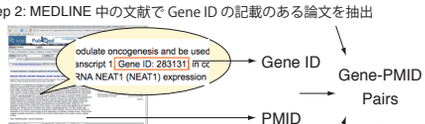
特徴抽出パイプライン

MeSH は遺伝子でなく文献に付与されたキーワード集なので、各遺伝子について関連文献を収集し、そこから MeSH の語を抽出することにより Gene-MeSH ペアを作成している。

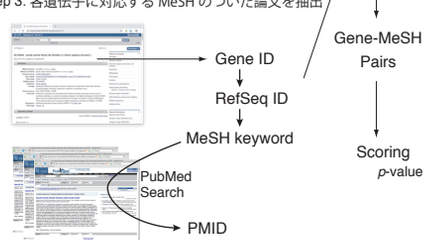
Step 1: Entrez Gene の Bibliography セクションより、論文の PMID (PubMed ID) を抽出



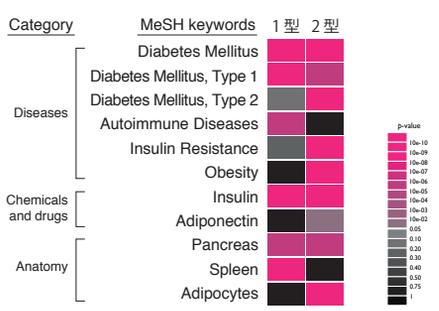
Step 2: MEDLINE 中の文献で Gene ID の記載のある論文を抽出



Step 3: 各遺伝子に対応する MeSH のついた論文を抽出

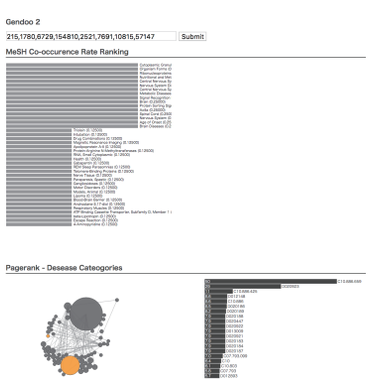


アノテーションからエンリッチメント解析へ



MeSH によるアノテーションの結果例。本図は同じ手法を OMIM の各疾患に対して行い、1 型 /2 型糖尿病について図示したもの。

エンリッチメント解析への応用



既存のウェブサービス Gendoo では、入力となる遺伝子と個々の用語を関連度により並べるにすぎなかった。
これを実際の解析シーンで利用しやすくするべく、エンリッチメント解析を行うためのウェブサービスを構築中である。
ここでは、得られた MeSH によるアノテーション情報や、そこから作成した、あるいは他で用いられる各種遺伝子リストにより実現をしている。
各遺伝子セット（もしくはその遺伝子セットの意味する生物機能情報）を関連に従って並べる他、生物機能情報どうしの関連も可視化できないか試行錯誤を行なっている。

トーゴーの日シンポジウム 2019
日本科学未来館
令和元年 10 月 5 日