遺伝子発現データベース

TCGAデータの収載

GENEVESTIGATORは遺伝子発現データベースのオンライン解析ツールです。公共データベースに登録されたマイクロアレイや次世代シーケンサーの膨大な遺伝子発現データをキュレートすることで、さまざまな研究者により登録された大量の実験結果を統合して解析可能にします。また、GENEVESTIGATORは使いやすいインターフェースと高速な検索エンジンを搭載しているため、研究者が標的遺伝子の探索などの遺伝子発現解析を行う際に、注目する遺伝子の同定や発現変動遺伝子の優先順位付けなどを簡単かつ正確に行うことができます。今回は新しく追加されたTCGAの遺伝子発現データを紹介します。

TCGAについて

TCGA (The Cancer Genome Atlas) は、米国がん研究所 (National Cancer Institute: NCI) と米国ヒトゲノム研究所 (National Human Genome Research Institute: NHGRI) の共同プロジェクト[1]であり、10種類の希少がんを含む33種類のがん種についてゲノム変異／遺伝子発現変動／メチル化異常など7種類のデータタイプを網羅的に解析しています。11,000人以上のがん患者に由来するがんサンプルと正常組織サンプルを比較した解析結果である2.5ペタバイト以上のTCGAのデータセットが公開され、がん研究に幅広く活用されています。

GENEVESTIGATORに収載されたTCGAデータ

Nebionでは2017年からTCGAのデータセットのキュレーションを開始しました。TCGAの遺伝子発現変動のデータは次世代シーケンサー(NGS)とアレイのデータの両方が存在しますが、アレイのデータは一部のがん種のデータしか存在しないので、NGSの遺伝子発現(RNA-seq)データをGENEVESTIGATORに収載する方針です。2018年1月現在、結腸腺がん(Colon adenocarcinoma: COAD)と直腸腺がん(Rectum adenocarcinoma: READ)の2種類のがん種のデータ(667サンプル)が収載されています(図1)。2018年第1四半期に乳がん、肺がん、甲状腺がん、胃がん、胆管がん、膵臓がん、前立腺がんのデータを収載予定です。

TCGAのデータセットをキュレーションする過程で、数多くのアノテーションの間違いが見つかっています。例えば、ICD-O-3 (国際疾病分類腫瘍学第3版) のコードと病理組織所見に不整合がある、生データとメタデータのサンプルIDが一致しない、といった場合です。Nebionではこれらの間違いをキュレーターが修正してGENEVESTIGATORに収載します。TCGAのデータセットは非常に有用ですが、生データをダウンロードして利用する場合はアノテーションに注意が必要です。

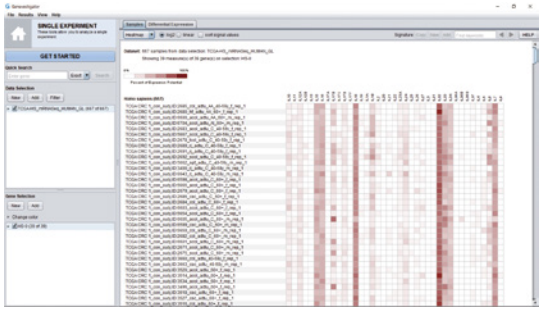

図1. 結腸腺がん(COAD)と直腸腺がん(READ)の667サンプル

RNA-seqのデータ解析パイプライン

Nebionでは2016年からGENEVESTIGATORにRNA-seqのデータを収載していますが、2017年にRNA-seqのデータ解析パイプラインを改良しています。

図2左は2016年のRNA-seqデータ解析パイプラインです。データの品質管理にはFastQCを使います。基準を満たしたサンプルはBowtieを使ってトランスクリプトームデータベース(Ensembl Release 75)にマッピングします。その後、RSEMを使って転写産物の発現量をTPMで算出します。発現量が変動した転写産物の検出にはBioconductorのLimmaパッケージのVoom機能を用います。

図2右は2017年のパイプラインです。2016年のパイプラインと比較して、トランスクリプトームデータベースへのマッピングと転写産物の発現量の算出のステップが変更されています。BowtieとRSEMに代えて新たにSalmonを採用しています。Salmonはquasi-mappingとtwo-phase inference procedureを組み合わせた新しいアルゴリズムにより転写産物の発現量を高速かつ高精度に算出できます。

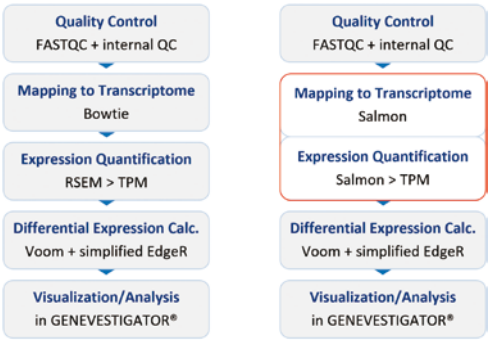

図2. RNA-seqデータ解析パイプライン(2016：左、2017：右)

ご評価

GENEVESTIGATORは、無償でトライアル利用できます。トライアル期間は30日間です。遺伝子発現解析をされる方はぜひGENEVESTIGATORをお試し下さい。トライアルを希望される方は弊社ウェブサイトよりお問い合わせ下さい。

1) https://cancergenome.nih.gov/