遺伝子発現データベース

nebion ヒトRNA-seq データリリース

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2016年7月にリリースされたヒトRNA-seqデータについて紹介します。

GEOに登録された遺伝子発現データ

NCBIのGene Expression Omnibus (GEO)やEBIのArrayExpressには世界中の研究者がマイクロアレイや次世代シーケンサー(NGS)で解析を行った生データを登録しています。GEOには2016年8月1日時点で71,851実験1,886,673サンプルのデータが登録されています。このうち遺伝子発現データは56,395実験1,300,683サンプルです。GEOに登録されている生データはさまざまなプラットフォームを使って解析されています。GEOに登録されているプラットフォームの総数は16,205になります。表1はサンプル数が多い上位10個のプラットフォームです。2000年代半ばに最初のNGSがリリースされてから10年以上が経ちますが、登録サンプル数は現時点でもアフィメトリクス社やイルミナ社のマイクロアレイが上位を占めます。5位と8位はイルミナ社のNGSであるHiSeq 2000を使ったヒトとマウスの解析結果になります。それぞれ約3,000実験37,000 ～ 49,000サンプルのデータが登録されています。

表1 GEOの上位10個のプラットフォーム(太字がNGS)

プラットフォーム名	サンプル数
Affymetrix Human Genome U133 Plus 2.0 Array	122,684
Broad Institute Human L1000 epsilon	115,209
Illumina HumanHT-12 V4.0 expression beadchip	53,541
Affymetrix Mouse Genome 430 2.0 Array	49,035
Illumina HiSeq 2000 (Mouse)	**48,921**
Illumina HumanMethylation450 BeadChip	47,682
Affymetrix Human Genome U133A Array	38,369
Illumina HiSeq 2000 (Human)	**37,145**
Affymetrix Human Gene 1.0 ST Array	28,467
Illumina HumanHT-12 V3.0 expression beadchip	22,287

RNA-seqデータのキュレーション

GENEVESTIGATORで組織やガン、発生段階ごとの遺伝子発現を比較したり、さまざまな実験条件での発現比較において有意に発現変動している遺伝子を見つけたりするためには元になるデータが公共データベースに十分登録されている必要があります。HiSeq 2000などイルミナ社製NGSで測定されたヒトのNGSデータは60,000サンプル以上が登録されており十分なデータ量です。そこで、Nebion社ではイルミナ社製NGSで測定されたヒトRNA-seqデータのキュレーションを行いGENEVESTIGATORへの収載を開始しました。第1弾として2016年7月に34実験2,443サンプルのデータがリリースされました(注1)。ヒトRNA-seqデータは今後も定期的にリリースされます。

図1はNebion社のRNA-seqデータ解析パイプラインの概要です。まず、FastQCプログラムを使ってリード配列の品質を確認します。品質が基準を満たさないサンプルは後の処理を行いません。NGSはマイクロアレイと比べて品質のばらつきが大きく、基準を満たさないサンプルの割合が多いです。また、この時にアノテーション情報なども確認し、生データが登録されていない実験やbiological replicatesがない実験も処理を行いません。次に、Bowtieプログラムを使ってトランスクリプトームデータベース(Ensembl Release 75)にマッピングを行います。個々の転写産物にマッピングした後に遺伝子ごとに集約します。マッピング済みのリード配列からRSEMプログラムを使って転写産物ごとにリード数を算出します。転写産物ごとのリード数はTPM (Transcripts Per Million)で正規化します。TPMはRPKM (Reads Per Kilobase Million)やFPKM (Fragments Per Kilobase Million)と比べてサンプル間での発現量の比較に適しています。正規化されたリード数からBioconductorのLimmaパッケージのVoom関数あるいはedgeRパッケージを使って発現量の比較を行います。

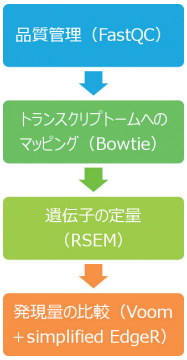

図1 RNA-seqのデータ解析パイプライン

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をお試しください。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ください。

(注1) 具体的な実験のアクセッション番号については弊社ウェブサイトよりお問い合わせください。