

スギ・トランスクリプトームデータレポジトリ (SugiExDB)の構築

1. はじめに

トランスクリプトームとは、特定の条件下における細胞や組織内に存在する全ての mRNA（遺伝子の転写産物）の総体を指し、その網羅的な解析をトランスクリプトーム解析と呼びます。近年、次世代シーケンサー（NGS）を用いた RNA sequencing（RNA-seq）がトランスクリプトーム解析の主流となっています。本記事では、スギの RNA-seq データを再解析し、その結果を格納したデータレポジトリ（SugiExDB）の構築について紹介します。

2. RNA-seqの公共データベースへの登録

RNA-seq のコストは年々低下しており、2025 年現在では 1 サンプルあたり数万円以下という現実的な価格で委託解析も可能です。この価格低下に伴い、RNA-seq データの公開数も増加の一途を辿っています。スギにおいては、2024 年 1 月時点で 100 サンプルを超えるデータセットが公開されており、2025 年中には 200 を超える見込みです。RNA-seq データは、論文投稿前に International Nucleotide Sequence Database Collaboration (INSDC) が運営する公共データベースへの登録が義務付けられています。登録されるデータは NGS によって得られた塩基配列情報であり、各サンプルの条件を記したメタデータも同時に登録されます。

3. スギの RNA-seq データの解析

公共データベースに登録される RNA-seq データは、塩基配列情報そのものです。したがって、各遺伝子の発現量を調べるためには、リファレンス配列へのマッピングやリードカウントといった情報解析が必要となります。そこで、我々は 2024 年 1 月までに INSDC に登録されていた 118 サンプルのスギ RNA-seq データと、NGS で新たに解析した 35 サンプルの RNA-seq データについて、スギのリファレンスゲノムへのマッピングとリードカ

ウントを実施しました。

4. データレポジトリ (SugiExDB) の構築

その後、上記で解析した 153 個の RNA-seq データを格納したデータレポジトリ（SugiExDB）を構築しました。レポジトリには、ライフサイエンス統合データベースセンター（DBCLS）が提供する TogoDB を採用しました。TogoDB は、表形式（CSV や TSV）のデータをインポートするだけで、容易にデータベースの公開・管理が行えるため、レポジトリ公開後もデータセットの追加や更新を容易に行うことができます。2025 年 1 月から SugiExDB を https://togodb.org/db/index_rnaseq_cj にて公開しており（図左）、全てのデータを閲覧・ダウンロード可能です。

5. おわりに

今回公開した SugiExDB は、RNA-seq データ（リードカウントおよび正規化されたリードカウント）の格納に特化したリポジトリです。各遺伝子の組織別・条件別の発現パターンを簡易的な図として掲載している点も特徴です（図右）。全てのデータがダウンロード可能であるため、共発現解析など二次的なデータ利用を推奨しています。

最後に、2025 年中に 100 サンプルを超える RNA-seq データの追加を予定しています。スギのトランスクリプトーム解析を実施された研究者の方は、ぜひ高田（naokitakata@affrc.go.jp）までご連絡ください。SugiExDB の更新にご協力いただけると幸いです。

（森林バイオ研究センター 高田 直樹）

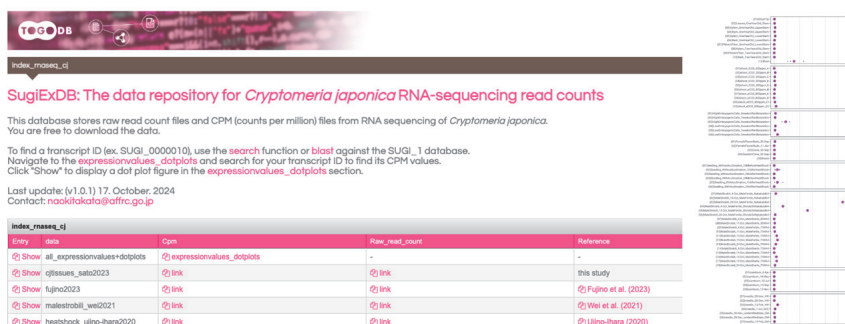


図 SugiExDBのトップ画面(左)と遺伝子発現パターンの一例(右)