目次

1　本マニュアルの概要

本業務は、厚生労働省科学研究費補助金「ヒト幹細胞を用いた再生医療の臨床実用化のための基盤構築に関する研究」の一環として実施し、国内の 7 拠点から日々産出される遺伝子発現データに対して、既存の Public Database から発現データを収集し比較するための解析ツール(プロトタイプシステム)を開発することを目的とする。そのようなツールを解析することで、iPS 細胞、ES 細胞、体性幹細胞のプロトコルによる発現の違い、分化のステージによる発現の違い、施設間差異などを明らかにすることを目標とする。

本業務では、既存データをアーカイブする機能と、新規データを問合せとしたアーカイブに対する検索機能を実装し、各々「ゲノムマッピング・発現量正規化ツール」、「発現量データ検索・比較ツール」を開発した。

本マニュアルでは、これら 2 つの解析ツールの機能概要と、それらを構成するファイル構成、システムの管理方法について説明する。

2　システムの機能概要

本章では、本業務で開発した「ゲノムマッピング・発現量正規化ツール」と「発現量データ検索・比較ツール」の機能の概要を説明する。

A　ゲノムマッピング・発現量正規化ツール

本ツールは、RNA-Seq により得られたリードをゲノム配列に対してマッピングし、そのマッピング結果に基づいて発現量の計算を行うツールである。このツールでは、RNA-Seq データを提供する一次データベースからのデータ取得、マッピング精度向上のための前処理、マッピング、発現量計算、発現量の正規化がパイプライン化されており、幾つかのコマンドを実行するだけで、大規模データに対して上述の処理を実行することができる。また、RNA-Seq のデータに関する各種情報（後述のメタデータ）と、解析の進捗情報を閲覧するための管理者用 WEB ページも用意されている。

上述解析パイプラインの流れを表 1 に示す。個々の処理については、表の後に詳述する。

表 1　ゲノムマッピング・発現量正規化ツールの処理の流れ

No.	処理名	処理の概要
1	SRA メタデータ取得	SRA からのメタデータのダウンロード及びそのフォーマット変換を行う。このフォーマットは、本システム独自形式。Study, experiment, sample, run を対象に、このフォーマット変換を行い、別個のファイルに格納する。
2	リードデータ取得	SRA から sra-lite を取得し、そこから FASTQ 形式でリードデータを抽出する。
3	ゲノム配列データ取得	NCBI からヒト、マウス、ラットのゲノム配列データをダウンロードし、bowtie 及び bowtie2 の index を作成する。
4	遺伝子構造アノテーションデータ取得	NCBI からヒト、マウス、ラットの遺伝子構造アノテーションをダウンロードし、cufflinks が取り扱える GFF に変換する。
5	マッピング前処理	リードデータに対して、クオリティチェック、低クオリティリードの除去、3’端低クオリティ領域のトリム、不明塩基(N)の割合が大きなリードの除去、短いリードの除去、ペアが揃っていないリードの除去を行う。
6	マッピング	リードデータがシングルリードかペアリードかチ

		ェックし、そのチェック結果を受けて、bowtie/bowtie2 を適切な設定で実行する。前処理を受けたリードの最大長を参照し、50base より短い場合は bowtie、それ以上の場合は bowtie2 を使用する。
7	発現量計算・正規化	No.4 で得られる遺伝子構造アノテーションデータと、No.6 で得られるマッピング結果を基に、既知転写産物及びマッピング結果から求められる転写産物予測結果に対しての発現量を計算する。この計算には cufflinks を使用し、FPKM として正規化する。
8	管理データ表示	メタデータ及び解析進捗状況を WEB ページで表示させる。

【1.　SRA メタデータ取得】

NCBI SRA（以下、SRA）には、RNA-Seq を含む各種リードデータと、それらデータに関わる情報として、使用されたサンプルや実験手法に関する情報がメタデータとして格納されている。本処理の対象となるのは、SRA から提供されている次のメタデータである。

- Study（研究単位）
- Experiment（実験単位）
- Run（シークエンシング単位）
- Sample（サンプル単位）

SRA の FTP サイトには、これらのメタデータが、それぞれのために設計されたスキーマに基づいて XML 形式でファイルに保存されている。本処理では、そのメタデータを取得し、パイプライン内における処理の簡便化を目的として、本システムで使用する値だけを抽出してタブ区切り形式で保存することとした。この抽出結果は、後述の「管理データ WEB ページ」で閲覧することができる。

【2.　リードデータ取得】

SRA の FTP サイトでは、リードデータや各種付随情報を独自の圧縮形式である sra-lite で提供している。本処理では、指定した sra-lite データをダウンロードし、ダウンロードされた sra-lite から FASTQ 形式でリードデータを抽出する。この抽出には、NCBI から提供されている専用ツールである fastq-dump を用いる。

【3.　ゲノム配列データ取得】

NCBI の MapViewer（以下、MapViewer）から提供されているゲノム配列データを取得する。

ゲノム配列データには複数のバージョンがあるが、本システムではリファレンスとして提供されているゲノムデータを使用する。

後続処理であるマッピングでは bowtie/bowtie2 を使用するため、それらツールで必要となるゲノム配列データに対するインデックスを作成する。それには、bowtie/bowtie2 に付属する bowtie-build/bowtie2-build を使用する。

【4.　遺伝子構造アノテーションデータ取得】

MapViewer から提供されている遺伝子構造アノテーションデータを取得する。後続処理である発現量計算・正規化では、遺伝子構造アノテーションデータが GFF 形式でファイルに保存ざれている必要がある。MapViewer から提供されている遺伝子構造アノテーションは GFF 形式となっているが、後続処理で使用されるプログラムの例外終了を引き起こすため、それを避けるためのフィルタリング処理を行う。具体的には、ゲノム配列中で複数個所に同一 ID の tRNA が出現しており、それが例外終了を引き起こすため、本システムにおいては、蛋白質をコードした遺伝子コード領域以外をフィルタリングで落とすこととした。この処理には、cufflinks に同梱されている gffread を用いる。

【5.　マッピング前処理】

マッピングの精度を上げるために、リードに対する前処理を行う。具体的には以下に示す処理が行われる。

1. クオリティチェック
   fastqc により、もともとのリードのクオリティを取得する。
2. 低クオリティリードの除去
   Fastx 同梱プログラムである fastq_quality_filter により、クオリティが 20 以上の塩基がリード全長の 80%に満たない場合は、そのリードを除去する。
3. 3’端低クオリティ領域のトリム
   Prinseq 同梱プログラムである prinseq-lite.pl により 3’端のクオリティが 20 未満の領域を取り除く。
4. 不明塩基(N)の割合が大きなリードの除去
   prinseq-lite.pl により、N がリード全長の 1%を超える場合、そのリードを除去する。
5. 短いリードの除去
   前処理を受ける前のリードの長さに対して 90%未満のリード長さになったリードを除去する。
6. クオリティチェック
   fastqc により、本処理を受けたリードのクオリティを取得する。
7. ペアが揃っていないリードの除去
   ペアが揃っていないリード（ペアエンドであるにも関わらず、前処理でその片側の

リードが削除されてしまったリード）を除去する。

【6.　マッピング】

bowtie/bowtie2 を用いてマッピングする。リードデータがシングルリードの場合は、シングルリードとしてマッピングし、ペアリードの場合は、ペア情報を使ったマッピングを行うように設定する。また、前処理を受けたリードの最大長を参照し、50base より短い場合は bowtie、それ以上の場合は bowtie2 を使用する。

【7.　発現量計算・正規化】

cufflinks を用いて発現量計算とその正規化を行う。No.4 で得られる遺伝子構造アノテーションデータを cufflinks に与えることで、既知 CDS 以外にマッピングされたリードから遺伝子構造を予測する。更に、既知遺伝子構造アノテーション及び新規予測結果に対して発現量を計算する。この発現量は FPKM と呼ばれる正規化された値である。

【8.　管理データ表示】

No.1 で得られたメタデータ、解析の進捗情報及び解析の統計情報を管理者が閲覧するための WEB ページを作成した。Study, experiment, run, sample のページがあり、それぞれのページでメタデータが閲覧できる。Run のページでは、解析の進捗情報・統計情報が表示される。個々の run は、一つの FASTQ ファイルに対応するため、各 run に、解析の進捗情報・統計情報が表示されることとなる。図 1 は、その表示画面である。Original sequence, Preprocessing, Mapping, Expression という列名が与えられている列が、進捗状況・統計情報のために設けられた列で、解析が終了した後には、水色で塗りつぶされ、前処理で残ったリードの本数や、マッピングされたリードの本数が表示される。fastqc の列は特別扱いされており、チェックの結果、リードのクオリティが高かったならば緑、悪かったならば赤、それらの中間であったならばオレンジ色の背景色が与えられるようになっている。それと同時に、fastqc が出力するリードのステータスの統計も表示される。更に、この欄をクリックすると fastqc によって作成される、チェック結果を詳細にレポートしたページにジャンプする。

図 1　管理者ページ

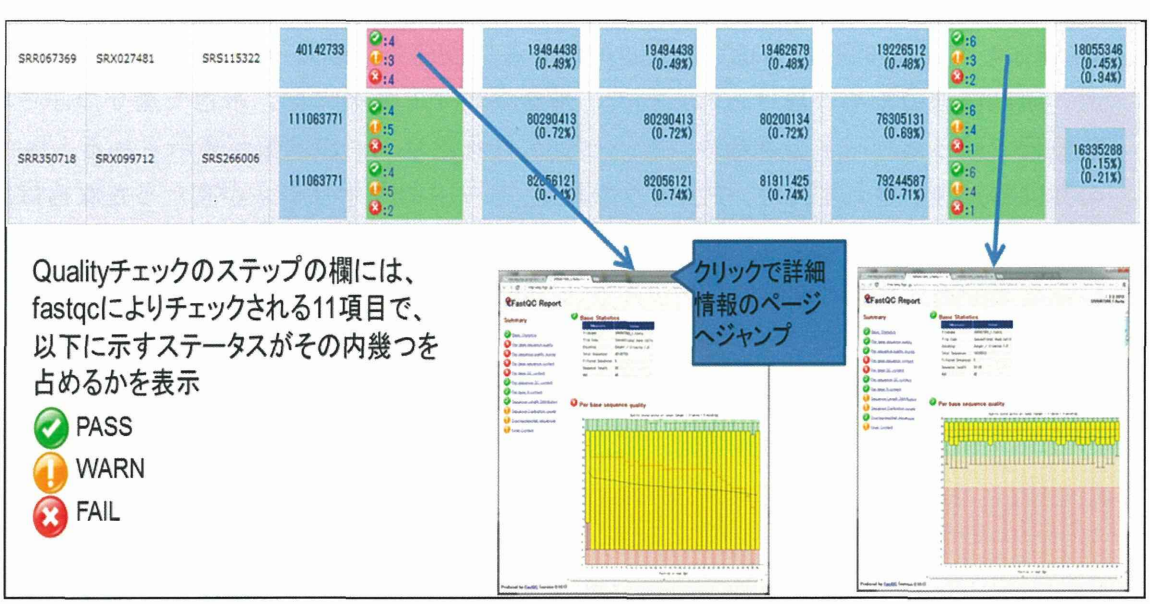


図 2　管理者ページにおけるクオリティ情報の表示

図 3　fastqc によるチェック結果の詳細情報の表示（前処理前）

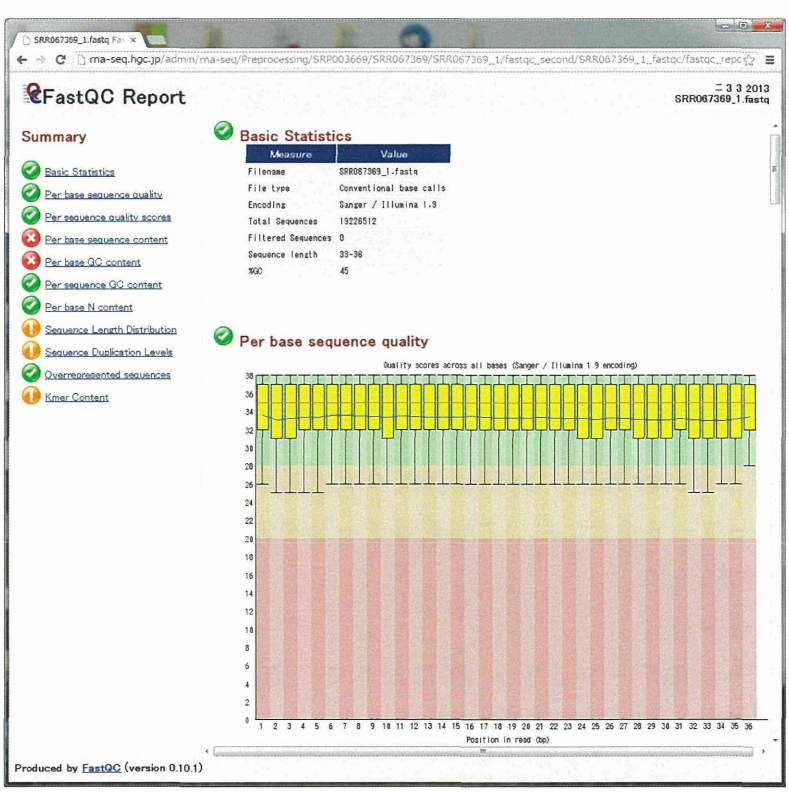


図 4　fastqc によるチェック結果の詳細情報の表示（前処理後）

B　発現量データ検索・比較ツール

本ツールは、ゲノムマッピング・発現量正規化ツールによって得られた発現量データに対して、キーワード検索や、データ間の比較を行うためのツールであり、本事業における研究拠点の研究者が利用することを想定したものである。

本ツールの実現のためには、まずデータの整備として、ゲノムマッピング・発現量正規化ツールによって得られた発現量データに対して、キーワード検索用のインデックスファイルを作成する。また、比較用に適した発現量データの変換も行う。

これらの整備したデータに対して、ユーザが検索や比較を容易に実行するためのユーザインタフェースを構築した。

上記の整備の流れを表 2 に整理する。

表 2 発現量データ検索・比較ツール整備の流れ

No.	処理名	処理の概要
1	発現量データの検索用インデックスファイルの作成	1)データセット、2) データセットに含まれるサンプル、3) データセットの中で測定された遺伝子、等で検索できるように、元情報を加工することで、これらのインデックスファイルを作成し、DB 化する。
2	発現量データの比較用データファイルの作成	発現量データ間の有意差検定や相関係数計算を実施する段階でデータ変換等が不要で高速に処理が可能となるように、比較用に適したデータ形式のファイルを作成する。
3	ユーザインタフェースの構築	ウェブブラウザ上で、発現量データのキーワード検索や、データ間の比較、データの登録ができるユーザインタフェースを構築する。

【1.　発現量データの検索用インデックスファイルの作成】

ユーザがキーワード検索において発現量データを特定するための以下の 3 種類のインデックスを作成する。

① データセット

発現量データのデータセット ID に対して、そのタイトルや説明文(Description)、等を含めたインデックスファイル。

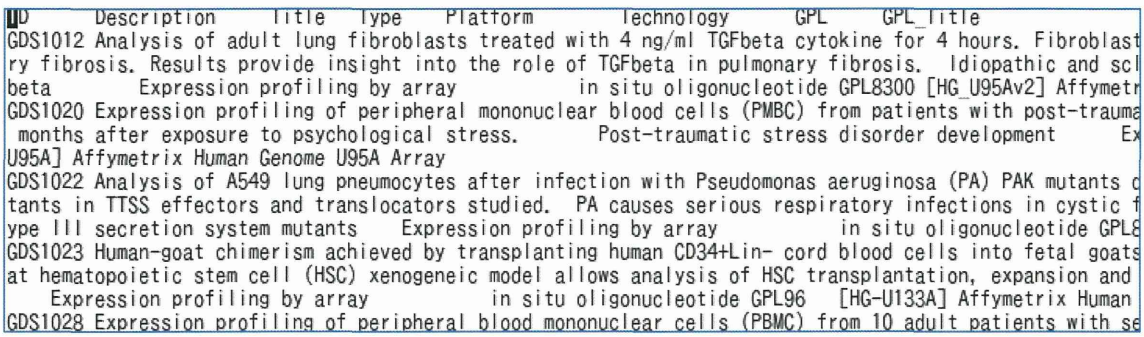
② データセットに含まれるサンプル

発現量データのデータセット ID に対して、データセットに含まれるサンプルデータの ID や説明文(Description)、等含めたインデックスファイル。

③　データセットの中で測定された遺伝子

発現量データのデータセット ID に対して、データセットに含まれる遺伝子の ID(Entrez Gene、Refseq 等)、遺伝子シンボル、名称、等を含めたインデックスファイル。

これらのインデックスファイルを DB 化し検索できるようにする。



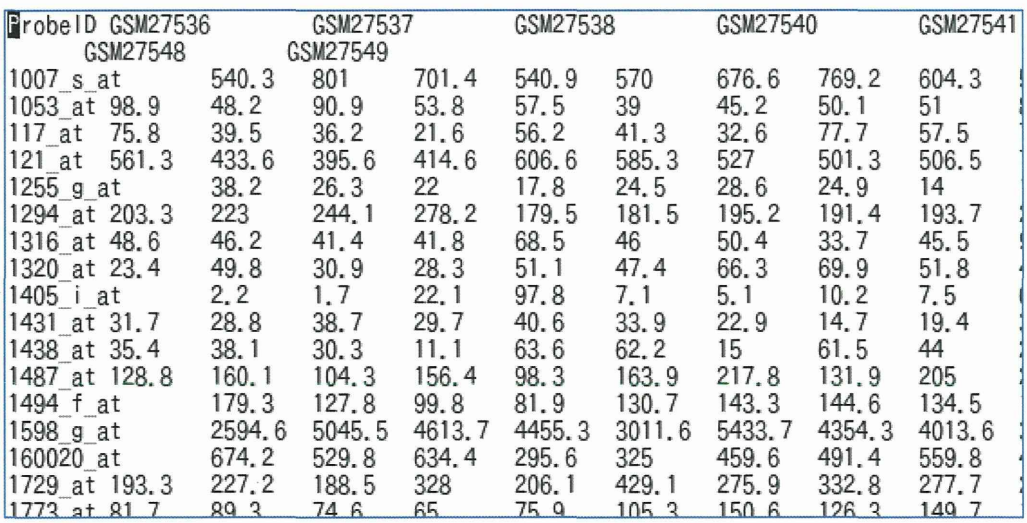
ID	Description	Title	Type	Platform	Technology	GPL	GPL_Title
GDS1012	Analysis of adult lung fibroblasts treated with 4 ng/ml TGFbeta cytokine for 4 hours. Fibroblast						
ry fibrosis. Results provide insight into the role of TGFbeta in pulmonary fibrosis. Idiopathic and scl							
beta	Expression profiling by array			in situ oligonucleotide	GPL8300	[HG_U95Av2] Affymetr	
GDS1020	Expression profiling of peripheral mononuclear blood cells (PMBC) from patients with post-trauma						
months after exposure to psychological stress.		Post-traumatic stress disorder development				Ex	
U95A] Affymetrix Human Genome U95A Array							
GDS1022	Analysis of A549 lung pneumocytes after infection with Pseudomonas aeruginosa (PA) PAK mutants c						
tants in TTSS effectors and translocators studied. PA causes serious respiratory infections in cystic f							
ype III secretion system mutants	Expression profiling by array			in situ oligonucleotide	GPL8		
GDS1023	Human-goat chimerism achieved by transplanting human CD34+Lin- cord blood cells into fetal goats						
at hematopoietic stem cell (HSC) xenogeneic model allows analysis of HSC transplantation, expansion and							
Expression profiling by array			in situ oligonucleotide	GPL96	[HG-U133A] Affymetrix Human		
GDS1028	Expression profiling of peripheral blood mononuclear cells (PBMC) from 10 adult patients with se						

図 5 検索用インデックスファイルの例

【2.　発現量データの比較用データファイルの作成】

発現量データ間の有意差検定や相関係数計算を実施する段階でデータ変換等が不要で高速に処理が可能となるように、比較用に適したデータ形式のファイルを作成する。

図 6 のように、縦に遺伝子、横にサンプルを並べ、発現量をタブ区切りで並べたデータ形式とする。



ProbeID	GSM27536	GSM27537		GSM27538		GSM27540		GSM27541
GSM27548		GSM27549						
1007_s_at	540.3	801	701.4	540.9	570	676.6	769.2	604.3
1053_at	98.9	48.2	90.9	53.8	57.5	39	45.2	50.1	51
117_at	75.8	39.5	36.2	21.6	56.2	41.3	32.6	77.7	57.5
121_at	561.3	433.6	395.6	414.6	606.6	585.3	527	501.3	506.5
1255_g_at		38.2	26.3	22	17.8	24.5	28.6	24.9	14
1294_at	203.3	223	244.1	278.2	179.5	181.5	195.2	191.4	193.7
1316_at	48.6	46.2	41.4	41.8	68.5	46	50.4	33.7	45.5
1320_at	23.4	49.8	30.9	28.3	51.1	47.4	66.3	69.9	51.8
1405_i_at		2.2	1.7	22.1	97.8	7.1	5.1	10.2	7.5
1431_at	31.7	28.8	38.7	29.7	40.6	33.9	22.9	14.7	19.4
1438_at	35.4	38.1	30.3	11.1	63.6	62.2	15	61.5	44
1487_at	128.8	160.1	104.3	156.4	98.3	163.9	217.8	131.9	205
1494_f_at		179.3	127.8	99.8	81.9	130.7	143.3	144.6	134.5
1598_g_at		2594.6	5045.5	4613.7	4455.3	3011.6	5433.7	4354.3	4013.6
160020_at		674.2	529.8	634.4	295.6	325	459.6	491.4	559.8
1729_at	193.3	227.2	188.5	328	206.1	429.1	275.9	332.8	277.7
1773_at	81.7	89.3	74.6	65	75.9	105.3	150.6	126.3	149.7

図 6 比較用発現量データファイルの例

【3.　ユーザインタフェースの構築】

ウェブブラウザ上で、発現量データのキーワード検索や、データ間の比較、デー

タの登録ができるユーザインタフェースを構築する。図 7 はトップ画面の例であり、ユーザは、発現プロファイルの検索、比較、登録の 3 種類の機能を利用することができる。

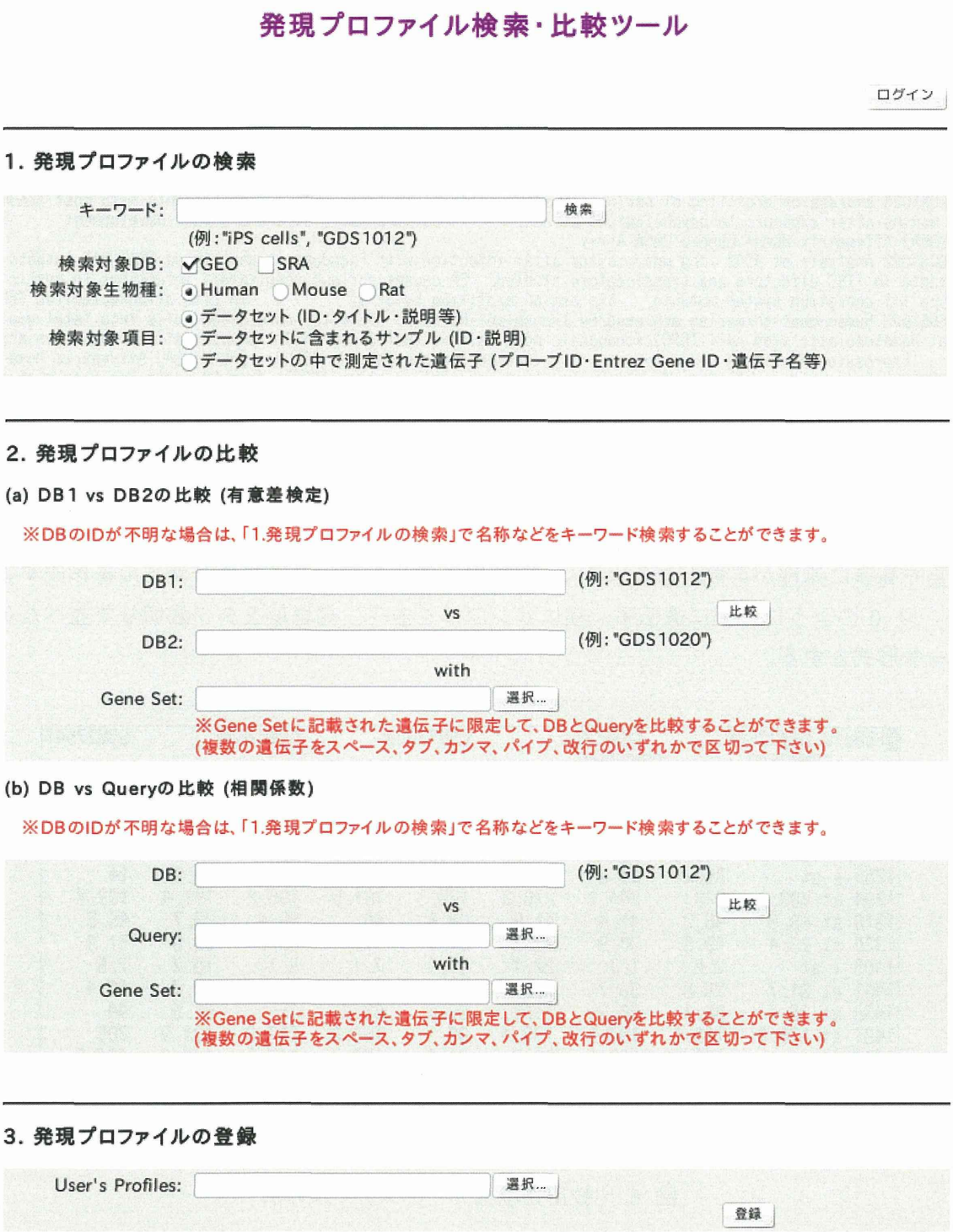


図 7 「発現量データ検索・比較ツール」のユーザインタフェース例

3　システムのファイル構成

本章では、本業務で開発した「ゲノムマッピング・発現量正規化ツール」と「発現量データ検索・比較ツール」のファイル・ディレクトリ構成を説明する。

A　ゲノムマッピング・発現量正規化ツール

本システムのルートディレクトリには、以下のディレクトリが設置される。

表 3　ゲノムマッピング・発現量正規化ツールのルートディレクトリ

ディレクトリ名	内容
SRA	sra-lite、SRA メタデータが格納されるディレクトリ
FASTQ	sra-lite から抽出されたリードデータが格納されるディレクトリ
Genome	ゲノム配列データ・インデックスが格納されるディレクトリ
Preprocessing	前処理結果が格納されるディレクトリ
Mapping	マッピング結果・発現量が格納されるディレクトリ
bin	解析ツールが格納されるディレクトリ

これらのディレクトリ配下に、そこに保存されているファイルの処理に関わるプログラム・設定ファイルが設置される。以下に、その詳細を示す。

【SRA ディレクトリ】

このディレクトリには、sra-lite、SRA メタデータ及びそれら処理に関わるプログラムとその設定ファイルが配置される（表 4）。

表 4　SRA ディレクトリに配置されるプログラム・設定ファイル

プログラム名	処理内容
getSraLiteByWget.pl	sra-lite のダウンロード
genSraStudyTable.pl	Study のメタデータに対するフォーマット変換
genSraExperimentTable.pl	Experiment のメタデータに対するフォーマット変換
genSraRunTable.pl	Run のメタデータに対するフォーマット変換
genSraSampleTable.pl	Sample のメタデータに対するフォーマット変換
genSraAcctable.pl	アクセッション番号一覧のフォーマット変換
conf.txt	設定ファイル

【FASTQ ディレクトリ】

このディレクトリには、リードデータ及びそれら処理に関わるプログラムとその設定ファイルが配置される（表 5）。

表 5　FASTQ ディレクトリに配置されるプログラム・設定ファイル

プログラム名	処理内容
genFastq.pl	sra-lite のダウンロード
conf.txt	設定ファイル

【Genome ディレクトリ】

このディレクトリには、ゲノム配列データ・遺伝子構造アノテーションデータ及びそれら処理に関わるプログラムとその設定ファイルが配置される（表 6）。

表 6　Genome ディレクトリに配置されるプログラム・設定ファイル

プログラム名	処理内容
getGenomeAnnotationByWget.pl	ゲノム配列・遺伝子構造アノテーションのダウンロード
conf.txt	設定ファイル

【Preprocessing ディレクトリ】

このディレクトリには、前処理結果データ及びそれら処理に関わるプログラムとその設定ファイルが配置される（表 7）。

表 7　Preprocessing ディレクトリに配置されるプログラム・設定ファイル

プログラム名	処理内容
genPreprocessing.pl	前処理実行スクリプトの生成
preprocessing.pl	前処理
removeNoCounterpart.pl	ペアが揃っていないリードの除去
count.pl	リード数のカウント
conf.txt	設定ファイル

【Mapping ディレクトリ】

このディレクトリには、マッピング結果、発現量計算結果及びそれら処理に関わるプログラムとその設定ファイルが配置される（表 8）。

表 8　Mapping ディレクトリに配置されるプログラム・設定ファイル

プログラム名	処理内容
genBowtieQue.pl	マッピングスクリプトの作成
genCufflinksQue.pl	発現量計算スクリプトの作成
count.pl	リード数のカウント
conf.txt	設定ファイル

B　発現量データ検索・比較ツール

本ツールのディレクトリには、大きく分けると、発現量データの検索用・比較用データファイルの作成のためのプログラム類を保存したディレクトリと、ユーザインタエース用のディレクトリ、の 2 種類がある。

表 9 発現量データ検索・比較ツールのディレクトリ大区分

ディレクトリ名	ディレクトリ	内容
データ作成	[ホームディレクトリ]/rna-seq/search/	発現量データの検索用・比較用データファイルの作成のためのプログラム類
ユーザインタフェース	/var/www/html/rna-seq/	ユーザインタフェース用の htdocs や cgi-bin 等

【データ作成ディレクトリ】

発現量データの検索用・比較用データファイルの作成のためのプログラム類を保存している。

プログラム名	処理内容
list_GeneID_in_GEO.pl	GEO の GDS に含まれる遺伝子 ID をリストアップする
list_GSM_Description_in_GDS.pl	GEO の GDS に含まれる GSM の Description をリストアップする
list_GDS_Description.pl	GEO の GDS のタイトル、説明文等をリストアップする
conv_GeneID_in_GeneSet.pl	遺伝子セット中の Entrez Gene ID を Probe ID に変換する
make_expressionGDS_for_R.pl	発現データをデータセット単位に分割し、遺伝子 x サンプルの行列形式に変換する

【ユーザインタフェースディレクトリ】

ユーザインタフェース用の htdocs や cgi-bin 等を保存している。

◆　htdocs 類（htdocs/search/）

プログラム名	処理内容
Index.html	トップ画面。但し、実際の表示は CGI を用いているため、自動的に “ProfileSearchAndCompare.cgi” に飛ぶようになっている。

◆　cgi 類（cgi-bin/search/）

プログラム名	処理内容
ProfileSearchAndCompare.cgi	トップ画面の、発現プロファイルの検索、比較、登録の全ての表示を行う。引数等により発現する機能が異なるように制御している。
conf	上記 CGI の設定用ファイル。 CGI 等のディレクトリパスや外部サイトの URL、データベースの設定、引数のデフォルト値を設定する。
common_lib.pl	上記 CGI の主要なプログラムをライブラリ化したもの。
lib/cgi-lib.pl	上記 CGI の引数処理のためのライブラリ。
tmp/	発現プロファイルの比較の際にアップロードしたファイルを一時的に保存しておくためのディレクトリ。 ※Apache からの書込みが必要なため、書込み権限を与えておく必要がある。

◆ データ類（admin/search/）

プログラム名	処理内容
GEO/index	検索用インデックスファイル
GEO/expressionGDS/	比較用発現データファイル

4　システムの管理方法

本章では、本業務で開発した「ゲノムマッピング・発現量正規化ツール」と「発現量データ検索・比較ツール」の管理方法を説明する。

A　ゲノムマッピング・発現量正規化ツール

本ツールのパイプラインに含まれている各処理は、3A 章で示したディレクトリに設置されている設定ファイルを適切に記述した後、ターミナルに対するコマンド入力あるいは Grid engine によるジョブ投入で実行されることとなる。以下に、各処理における設定方法と、実行方法について述べる。

【SRA ディレクトリ】

まず、設定ファイル(conf.txt)に適切な設定を記入する。このファイルには、タグと値のペアをスペースあるいはタブ区切りで記述する。以下は、その具体例である。

```
ASCP	/home/tany/.aspera/connect/bin/ascp
ASCP_DSA	/home/tany/.aspera/connect/etc/asperaweb_id_dsa.openssh
WGET	/usr/local/bin/wget
SRA_FTP_WGET	ftp://ftp.ncbi.nlm.nih.gov/sra
SRA_FTP anonftp@ftp-private.ncbi.nlm.nih.gov:/sra
SRA_LITE	/sra-instant/reads/ByStudy/litesra
SRA_ARCHIVE	/home/tany/SRA/sra-lite
SRA_ARCHIVE_WGET	/home/tany/SRA/sra-lite-wget
METADATA_DIR	/home/tany/SRA/NCBI_SRA_Metadata_Full_20130201
ACC_INFO_FILE	SRA_Accessions
STUDY_TYPE	"Transcriptome Analysis"
STUDY_DESCRIPTION	"iPS","induced pluripotent stem","ES","stem"
SAMPLE_DESCRIPTION	"iPS","induced pluripotent stem","ES","stem"
STUDY_FILE	/home/tany/SRA/studyTable.tsv
SAMPLE_FILE	/home/tany/SRA/sampleTable.tsv
EXPERIMENT_FILE /home/tany/SRA/experimentTable.tsv
RUN_FILE	/home/tany/SRA/runTable.tsv
ACCESSION_FILE	/home/tany/SRA/accessionTable.tsv
SRA_GET_LOG	/home/tany/SRA/sra-lite.log
SRA_GET_LOG_WGET	/home/tany/SRA/sra-lite-wget.log
```

この中で、特に重要は設定項目について表 10 に示す。

表 10　SRA ディレクトリの設定ファイル

プログラム名	処理内容
WGET	wget のパス
SRA_FTP_WGET	wget 取得データの保存先パス
SRA_LITE	sra-lite の URI
SRA_ARCHIVE_WGET	sra-lite 保存先パス
STUDY_TYPE	取得対象となる sra-lite の study type
STUDY_DESCRIPTION	取得対象となる sra-lite の study description
SAMPLE_DESCRIPTION	取得対象となる sra-lite の sample description
STUDY_FILE	Stuey メタデータ保存ファイルパス
SAMPLE_FILE	Sample メタデータ保存ファイルパス
EXPERIMENT_FILE	Experiment メタデータ保存ファイルパス
RUN_FILE	Run メタデータ保存ファイルパス
ACCESSION_FILE	Accession リスト保存ファイルパス

以下に、コマンドの実行手順を示す。

```
$ wget ftp://ftp.ncbi.nlm.nih.gov/sra/reports/Metadata/NCBI_SRA_Metadata_Full_20130201.tar.gz
$ tar xfz NCBI_SRA_Metadata_Full_20130201.tar.gz
$ qsub -l ljob -q lljobs.q -S /usr/bin/perl ./getSraLiteByWget.pl
$ ./genSraStudyTable.pl > studyTable.tsv
$ ./genSraExperimentTable.pl > experimentTable.tsv
$ ./genSraRunTable.pl > runTable.tsv
$ ./genSraSampleTable.pl > sampleTable.tsv
$ ./genSraAccTable.pl > accessionTable.tsv
```

【FASTQ ディレクトリ】

SRA ディレクトリの場合同様、設定ファイルに適切な設定を記入し、コマンドを実行する。

```
SRA_ARCHIVE      /home/tany/SRA/sra-lite-wget
FASTQ_ARCHIVE    /home/tany/FASTQ
FASTQ_DUMP       /home/tany/bin/fastq-dump
```

この設定項目について表 11 に示す。

表 11 FASTQ ディレクトリの設定ファイル

プログラム名	処理内容
SRA_ARCHIVE	sra-lite 保存先パス
FASTQ_ARCHIVE	リードデータ保存パス
FASTQ_DUMP	dastq-dump コマンドのパス

以下に、コマンドの実行手順を示す。

```
$ qsub -S /usr/bin/perl ./genFastq.pl
```

【Genome ディレクトリ】

SRA ディレクトリの場合同様、設定ファイルに適切な設定を記入し、コマンドを実行する。

```
GENOME_FTP ftp://ftp.ncbi.nlm.nih.gov/genomes
GENOME_SEQ_FTP_HUMAN_PATH /H_sapiens/Assembled_chromosomes/seq/
GENOME_ANNOT_FTP_HUMAN_PATH /H_sapiens/GFF/
GENOME_ARCHIVE_HUMAN /home/tany/Genome/HUMAN
GENOME_SEQ_FTP_MOUSE_PATH /M_musculus/Assembled_chromosomes/seq/
GENOME_ANNOT_FTP_MOUSE_PATH /M_musculus/GFF/
GENOME_ARCHIVE_MOUSE /home/tany/Genome/MOUSE
GENOME_SEQ_FTP_RAT_PATH /R_norvegicus/Assembled_chromosomes/seq/
GENOME_ANNOT_FTP_RAT_PATH /R_norvegicus/GFF/
GENOME_ARCHIVE_RAT /home/tany/Genome/RAT
GENOME_GET_LOG_HUMAN /home/tany/Genome/genome_HUMAN.log
GENOME_GET_LOG_MOUSE /home/tany/Genome/genome_MOUSE.log
GENOME_GET_LOG_RAT /home/tany/Genome/genome_RAT.log
```

この中で特に重要な設定項目について表 12 に示す。

表 12 Genome ディレクトリの設定ファイル

プログラム名	処理内容
GENOME_FTP	MapViewer FTP サイトの URI
GENOME_SEQ_FTP_HUMAN_PATH	ヒトのゲノム配列データのパス
GENOME_ANNOT_FTP_HUMAN_PATH	ヒトの遺伝子構造アノテーションデータのパス(マウス、ラットにも同様のタグあり)
GENOME_ARCHIVE_HUMAN	ヒトのゲノム配列・遺伝子構造アノテーションデータの保存先パス(マウス、ラットにも同様のタグあり)

以下に、コマンドの実行手順を示す。

```
$ qsub -l ljob -q lljobs.q -S /usr/bin/perl ./getGenomeAnnotationByWget.pl HUMAN
$ qsub -l ljob -q lljobs.q -S /usr/bin/perl ./getGenomeAnnotationByWget.pl MOUSE
$ qsub -l ljob -q lljobs.q -S /usr/bin/perl ./getGenomeAnnotationByWget.pl RAT
$ qsub -l s_vmem=16G,mem_req=16 bowtieBuildHUMAN.sh
$ qsub -l s_vmem=16G,mem_req=16 bowtieBuildMOUSE.sh
$ qsub -l s_vmem=16G,mem_req=16 bowtieBuildRAT.sh
$ ~/bin/gffread -C -E HUMAN/GFF/ref_GRCh37.p10_scaffolds.gff3 -o- >
HUMAN/GFF/scaffolds_transcript.gff
$ ~/bin/gffread -C -E HUMAN/GFF/ref_GRCh37.p10_top_level.gff3 -o- >
HUMAN/GFF/top_level_transcript.gff
$ ~/bin/gffread -C -E MOUSE/GFF/ref_GRCm38.p1_scaffolds.gff3 -o- >
MOUSE/GFF/scaffolds_transcript.gff
$ ~/bin/gffread -C -E HMOUSE/GFF/ref_GRCm38.p1_top_level.gff3 -o- >
MOUSE/GFF/top_level_transcript.gff
$ ~/bin/gffread -C -E RAT/GFF/ref_Rnor_5.0_scaffolds.gff3 -o- > RAT/GFF/scaffolds_transcript.gff
$ ~/bin/gffread -C -E RAT/GFF/ref_Rnor_5.0_top_level.gff3 -o- > RAT/GFF/top_level_transcript.gff
```

【Mapping ディレクトリ】

SRA ディレクトリの場合同様、設定ファイルに適切な設定を記入し、コマンドを実行する。

```
PERL                /usr/bin/perl
GENOME_DIR          /home/tany/Genome
PREPROCESSING_DIR   /home/tany/Preprocessing
MAPPING_DIR         /home/tany/Mapping
BOWTIE              /home/tany/bin/bowtie
BOWTIE2             /home/tany/bin/bowtie2
BOWTIE_OPT          "--sam --best --strata -p 8 -t"
BOWTIE2_OPT         "-p 1 -t"
SAMTOOLS            /home/tany/bin/samtools
CUFFLINKS           /home/tany/src/cufflinks-2.0.2.Linux_x86_64/cufflinks
STUDY_FILE          /home/tany/SRA/studyTable.tsv
SAMPLE_FILE         /home/tany/SRA/sampleTable.tsv
EXPERIMENT_FILE     /home/tany/SRA/experimentTable.tsv
RUN_FILE            /home/tany/SRA/runTable.tsv
ORGANISM_INDEX      HUMAN,MOUSE,RAT
```