```
ORGANISM_NAME    HUMAN              "Homo sapiens"
ORGANISM_NAME    MOUSE              "Mus musculus"
ORGANISM_NAME    RAT                "Rattus norvegicus"
BOWTIE_MASK_HUMAN_INDEX        /home/tany/Genome/HUMAN/genome_mask_bowtie
BOWTIE_NOMASK_HUMAN_INDEX      /home/tany/Genome/HUMAN/genome_noMask_bowtie
BOWTIE2_MASK_HUMAN_INDEX       /home/tany/Genome/HUMAN/genome_mask_bowtie2
BOWTIE2_NOMASK_HUMAN_INDEX     /home/tany/Genome/HUMAN/genome_noMask_bowtie2
BOWTIE_MASK_MOUSE_INDEX        /home/tany/Genome/MOUSE/genome_mask_bowtie
BOWTIE_NOMASK_MOUSE_INDEX      /home/tany/Genome/MOUSE/genome_noMask_bowtie
BOWTIE2_MASK_MOUSE_INDEX       /home/tany/Genome/MOUSE/genome_mask_bowtie2
BOWTIE2_NOMASK_MOUSE_INDEX     /home/tany/Genome/MOUSE/genome_noMask_bowtie2
BOWTIE_MASK_RAT_INDEX          /home/tany/Genome/RAT/genome_mask_bowtie
BOWTIE_NOMASK_RAT_INDEX        /home/tany/Genome/RAT/genome_noMask_bowtie
BOWTIE2_MASK_RAT_INDEX         /home/tany/Genome/RAT/genome_mask_bowtie2
BOWTIE2_NOMASK_RAT_INDEX       /home/tany/Genome/RAT/genome_noMask_bowtie2
GFF_HUMAN                  /home/tany/Genome/HUMAN/GFF/scaffolds_transcript.gff
GFF_MOUSE                  /home/tany/Genome/MOUSE/GFF/scaffolds_transcript.gff
GFF_RAT            /home/tany/Genome/HUMAN/RAT/scaffolds_transcript.gff
```

この中で特に重要な設定項目について表 13 に示す。

表 13 Mapping ディレクトリの設定ファイル

プログラム名	処理内容
PERL	perl コマンドのパス
GENOME_DIR	ゲノム配列データ・インデックス保存パス
PREPROCESSING_DIR	前処理結果保存先パス
MAPPING_DIR	マッピング結果保存先パス
BOWTIE	bowtie コマンドのパス
BOWTIE2	bowtie2 コマンドのパス
BOWTIE_OPT	bowtie コマンドのオプション
BOWTIE2_OPT	bowtie2 コマンドのオプション
SAMTOOLS	samtools コマンドのパス
CUFFLINKS	cufflinks コマンドのパス
STUDY_FILE	Study 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
ORGANISM_INDEX	生物種のコード
ORGANISM_NAME	生物種コードと生物種名の対応
BOWTIE_MASK_HUMAN_INDEX	マスクされたゲノム配列に対する bowtie のインデックスファイル(マウス、ラットにも同様のタグあり)
BOWTIE_NOMASK_HUMAN_INDEX	マスクされていないゲノム配列に対する bowtie のインデックスファイル(マウス、ラットにも同様のタグあり)
BOWTIE2_MASK_HUMAN_INDEX	マスクされたゲノム配列に対するbowtie2 のインデックスファイル(マウス、ラットにも同様のタグあり)
BOWTIE2_NOMASK_HUMAN_INDEX	マスクされていないゲノム配列に対する bowtie2 のインデックスファイル(マウス、ラットにも同様のタグあり)
GFF_HUMAN	遺伝子構造アノテーション保存先パス

以下に、コマンドの実行手順を示す。

```
$ ./genBowtieQue.pl > que
$ split -a 3 -d -l 1 que que_
$ for i in que_[0-9][0-9] ; do qsub -l s_vmem=16G,mem_req=16 $i; done
$ ./genCufflinksQue.pl > que2
$ split -a 3 -d -l 1 que2 que2_
$ for i in que2_[0-9][0-9] ; do qsub -l s_vmem=16G,mem_req=16 $i; done
```

B　発現量データ検索・比較ツール

検索用インデックスの作成方法、比較用発現データの作成方法、ユーザインタフェースの設定方法についてまとめる。

【検索用インデックスの作成方法】

以下の Perl のコマンドを実行する。なお、Perl 内で入力データや出力先の設定を行っているため、適宜修正を行う。

```
$ .list_GeneID_in_GEO.pl
$ list_GSM_Description_in_GDS.pl
$ list_GDS_Description.pl
$ conv_GeneID_in_GeneSet.pl
```

これらで作成されたインデックスファイルを MySQL に登録する。

```
% create database GEO_Human;
%
% create table GDS (GDS_ID varchar(10), DESCRIPTION varchar(1000), TITLE varchar(1000),,,,,);
% create table GSM (GDS_ID varchar(10), GSM_IDs varchar(1000), GSM_DESCRIPTIONs varchar(1000));
% .....
% load data local infile "[GDS_Title_Description_etc.csv]" into table GDS;
% local data local infile "[GDS_to_GeneSymbols.csv]" into table GeneSymbol;
% ....
```

【比較用発現データの作成方法】

元の発現データから比較用のデータに分割・変換するために以下の Perl プログラムを実行する。なお、Perl 内で入力データや出力先の設定を行っているため、適宜修正を行う。

```
$ .make_expressionGDS_for_R.pl
```

【ユーザインタフェースの設定方法】

ユーザインタフェースの設定方法(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)についてまとめる。

```
#--- パスの設定 ---#
$ProfileSearchAndCompare_CGI = “[CGI パスを指定]”;
$Correlation_R                = “”;
$DB_Dir
$DB_List
$UserDB_Dir
$Tmp_Dir

#--- 外部 URL の設定 ---#
$GEO_URL =

#--- mysql の設定 ---#
$mysql_tabe =
$mysql_host =
$mysql_port =
$mysql_user =
$mysql_password =

#--- デフォルト値 ---#
%default  = (“DB”,”GEO”,    “ORAGANISM”,”Human”,    ”COLUMN”,”DataSet”,
 “USERID”,”001
```

平成 24 年度厚生労働省科学研究費補助金
難病・がん等の疾患分野の医療の実用化研究事業
iPS 細胞、ES 細胞、体性幹細胞の解析ツールの開発

(2)ユーザ・マニュアル

株式会社三菱総合研究所

2013 年 3 月

目次

1　本マニュアルの概要

本マニュアルは、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 から収集・解析した発現データと比較するためのツール「発現量データ検索・比較ツール」の使い方をまとめたものです。

本ツールは、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」の一環として開発されました。本ツールはまだプロトタイプシステムですが、今後機能の充実や操作性の向上を図ることで、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研究を支援できることを目標としています。

なお、本ツールと同時に開発された、既存データをアーカイブしゲノムマッピングと発現量の正規化を行う「ゲノムマッピング・発現量正規化ツール」については現在のところ公開はせず、その結果のみを本ツールで検索・閲覧することができます。

2　ツール機能の概要

本章では、「発現量データ検索・比較ツール」の機能の概要を説明します。

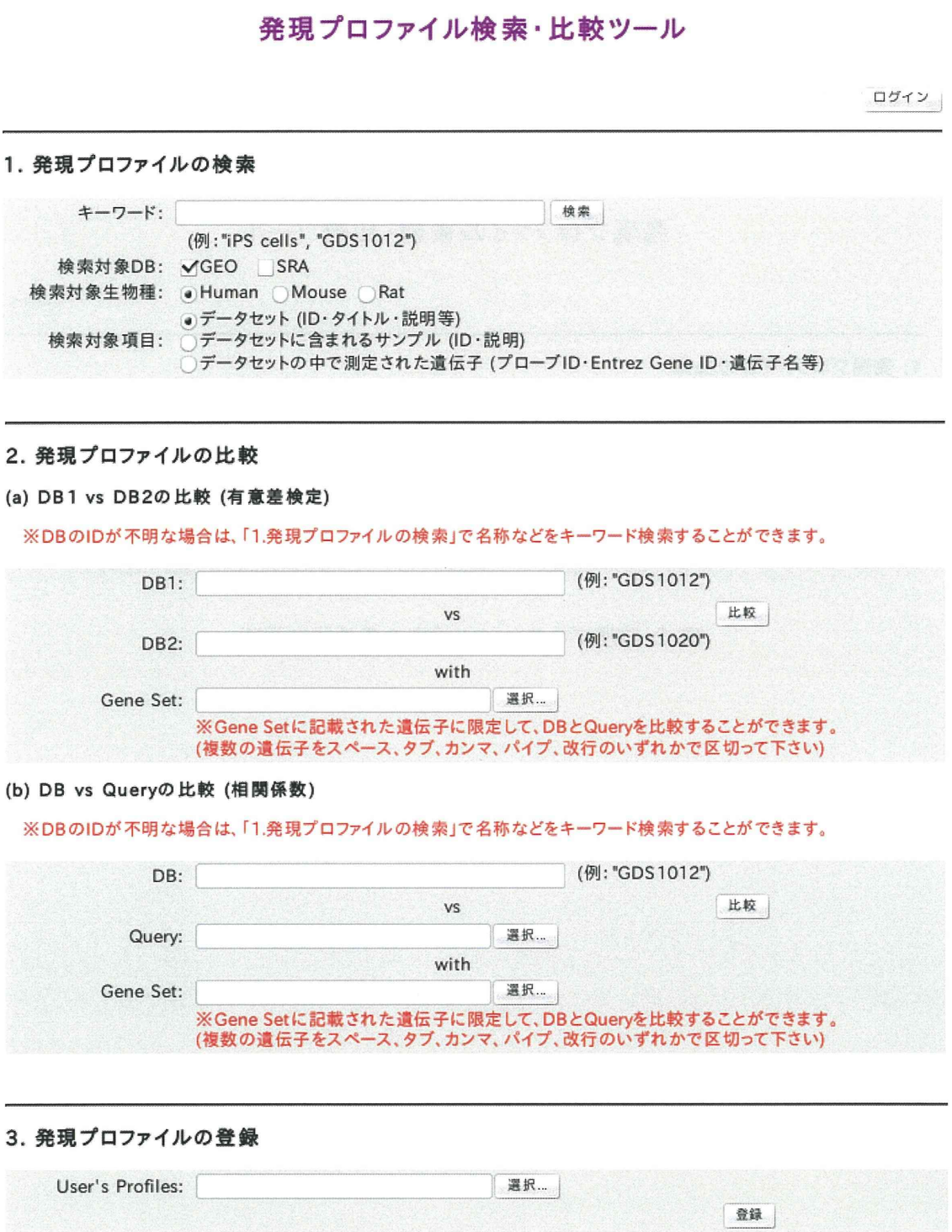

図 1 「発現量データ検索・比較ツール」のトップ画面メニュ

3　発現プロファイル検索の使い方

本章では、「発現量データ検索・比較ツール」の機能のうち、既存の Public Database から収集・解析しデータベース化した発現データをキーワード検索する機能の使い方について解説します。

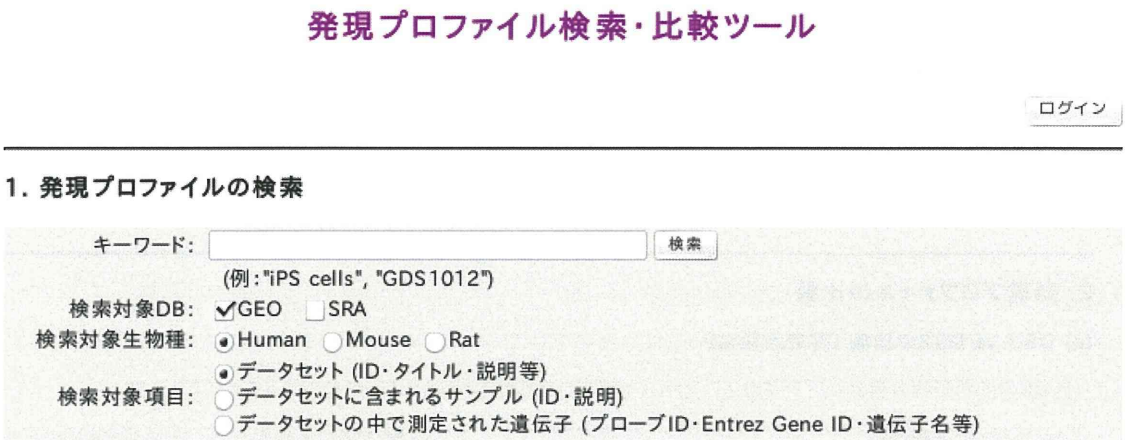

図 2 発現プロファイル検索の条件設定画面

4　発現プロファイル比較の使い方

本章では、「発現量データ検索・比較ツール」の機能のうち、既存の Public Database から収集・解析しデータベース化した発現データを比較するための機能の使い方について解説します。発現プロファイルの比較は、データベース中の 2 つのデータ間の比較と、データベース中のデータとユーザが実験等で得られた発現データとの比較、の 2 種類が可能です。

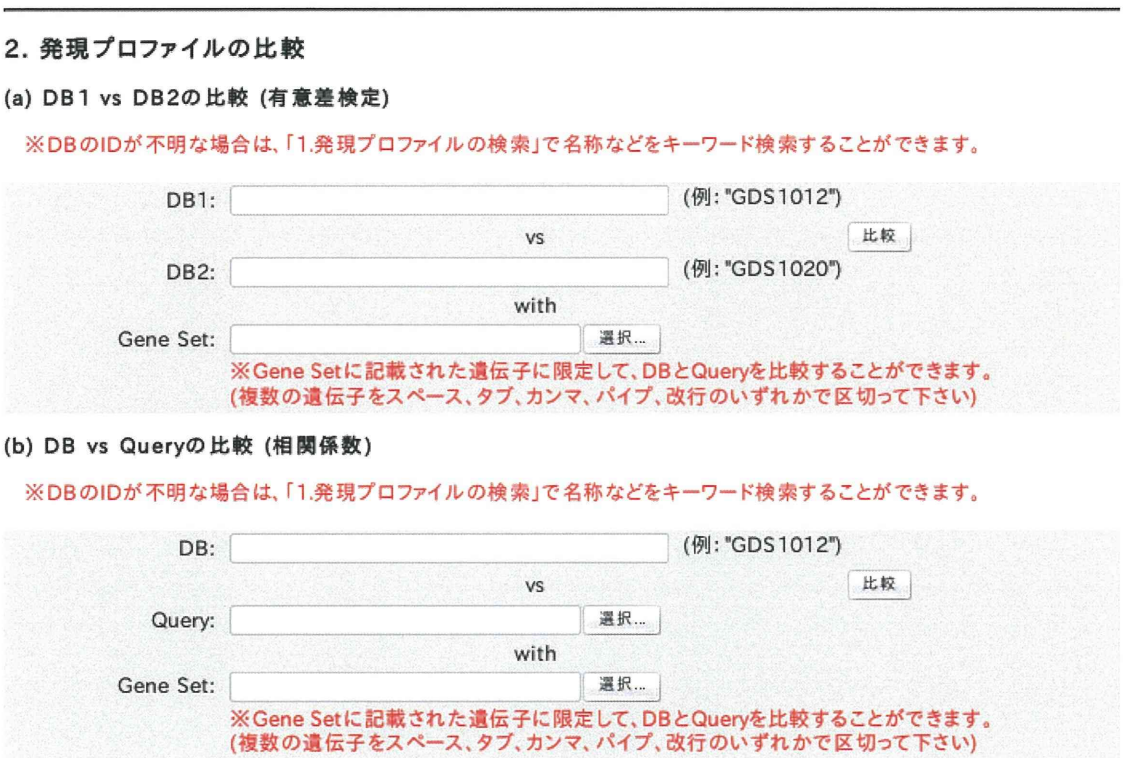

図 3 発現プロファイル比較の条件設定画面

5　発現プロファイルの登録方法

本章では、「発現量データ検索・比較ツール」の機能のうち、ユーザが実験等で解析した発現プロファイルデータをデータベースに登録し、他のデータとの比較や検索に利用するための方法について解説します。

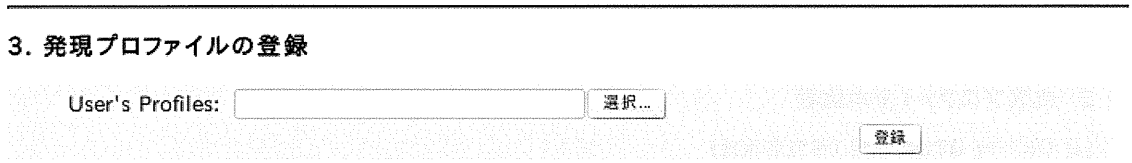

図 4　発現プロファイルの登録画面(データアップロード)

平成 24 年度厚生労働省科学研究費補助金
難病・がん等の疾患分野の医療の実用化研究事業
iPS 細胞、ES 細胞、体性幹細胞の解析ツールの開発

(3)システム・テスト仕様書
(4)システム・テスト結果報告書

株式会社三菱総合研究所

2013 年 3 月

目次

1　本報告書の概要

本報告書は、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に開発された解析ツール(プロトタイプシステム)のシステム・テスト仕様とその結果をまとめたものである。

2　システム・テスト仕様

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」のシステム・テスト仕様をまとめる。

(A) ゲノムマッピング・発現量正規化ツール

No.	テスト名	テスト仕様
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に変換できること。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、ペアがそ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。

(B) 発現量データ検索・比較ツール

No.	テスト名	テスト仕様
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。 プログラムを実行し、インデックスファイルが作成されるか、DB化し検索が可能かを確認する。
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。 比較用データファイルが作成されるかを確認する。
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。 発現プロファイルの検索、比較、登録が可能かを確認する。

3　システム・テスト結果

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」のシステム・テスト結果をまとめる。

(A) ゲノムマッピング・発現量正規化ツール

No.	テスト名	テスト結果	備考
1	SRA メタデータ取得	合格	Grid engine で計算が実行されること、最終的な出力ファイルが適切に生成されることを確認。
2	リードデータ取得	合格	Grid engine で計算が実行されること、最終的な出力ファイルが適切に生成されることを確認。
3	ゲノム配列データ取得	合格	Grid engine で計算が実行されること、最終的な出力ファイルが適切に生成されることを確認。
4	遺伝子構造アノテーションデータ取得	合格	Grid engine で計算が実行されること、最終的な出力ファイルが適切に生成されることを確認。
5	マッピング前処理	合格	Grid engine で計算が実行されること、最終的な出力ファイルが適切に生成されることを確認。
6	マッピング	合格	Grid engine で計算が実行されること、最終的な出力ファイルが適切に生成されることを確認。
7	発現量計算・正規化	合格	Grid engine で計算が実行されること、最終的な出力ファイルが適切に生成されることを確認。
8	管理データ表示	合格	WEB ページ上に、Study, experiment, sample, run のメタデータと、run 毎の解析進捗状況・統計値が表示されることを確認。

(B) 発現量データ検索・比較ツール

No.	テスト名	テスト結果	備考
1	発現量データの検索用インデックスファイルの作成	合格	ログラムを実行し、インデックスファイルが作成されるか、DB 化し検索が可能かを確認。
2	発現量データの比較用データファイルの作成	合格	比較用データファイルが作成されるかを確認。
3	ユーザインタフェースの構築	合格	発現プロファイルの検索、比較、登録が可能かを確認。

iPS細胞、ES細胞、体性幹細胞の解析ツールの開発
―高度化のための技術調査―

2013年3月25日

株式会社三菱総合研究所