12. 協議

本仕様に記載のある事項および記載の無い事項について疑義が生じた場合には、受注者は本学関係者と協議の上、その決定に従うものとする。

13. 著作権

本調達の実施により新たに開発された部分の著作権は本学に帰属するものとする。

14. 機密保持条項

受注者は、本開発により直接または間接に知り得た情報について、その機密を保ち、漏洩、開示、発表をしてはならない。ただし、あらかじめ本学の承認を得た場合、および受注者が以前から保有しているものに関してはこの限りではない。

15. その他の必要事項

(1) 作業の進捗に応じて、適宜、打合せを開催するものとする。
(2) 納入物の検収、システムの運用開始時には、その支援作業を行うこと。

スケジュール

開発項目	1月	2月	3月
類似検索機能の開発 • 正規化ロジック • メール通知機能あるいはジョブ管理機能 • 検索結果一覧表示機能	→	→	
GSEA実行・表示機能の開発 • 類似性検索結果を受けてのGSEA • ユーザ設定データセットを用いての再解析機能		→	
チューニング及びサンプル問い合わせデータの探索		→	→
システム設計のドキュメンテーション • 概念設計 • ファイル・ディレクトリ構成、設定・プラグインに関するドキュメンテーション	→		→

厚生労働科学研究費補助金
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への機能追加

2014/2/12 打ち合わせ資料

株式会社三菱総合研究所

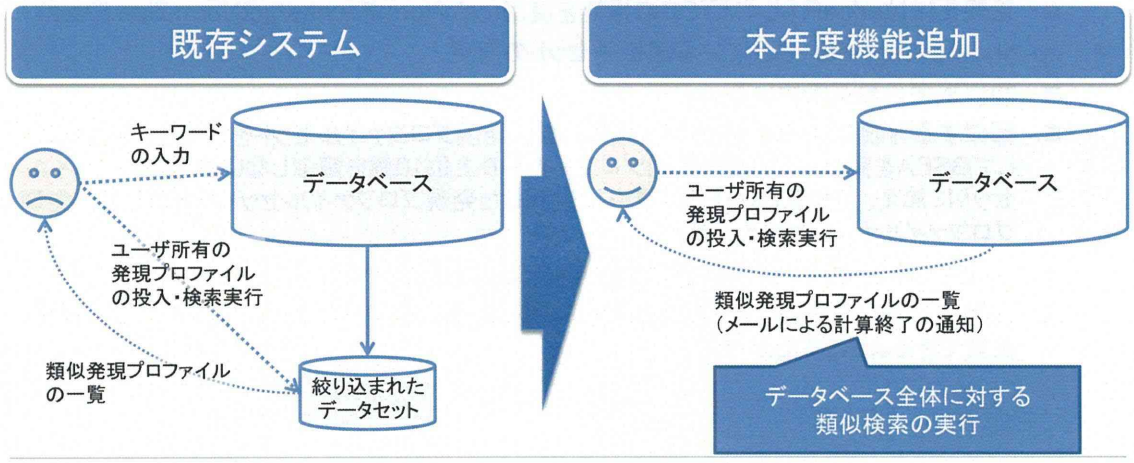
打ち合わせ内容

- 開発物についての確認（昨年12/20打ち合わせ資料）
- 本開発の重要ポイントの確認(本年1/09打ち合わせ資料)
- システム設計・開発
  - WEBインターフェースの開発
  - 類似性検索機能の開発
- スケジュール

本年度開発：操作性の簡便化(1/3)

◎ データベース全体に対する類似プロファイル検索機能

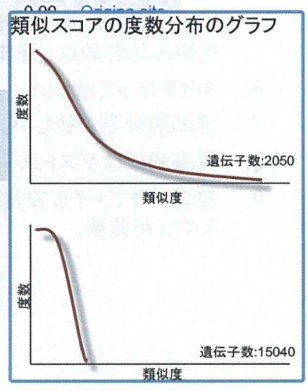
- キーワード検索による検索対象の絞り込みを廃止
- 長時間計算ジョブ・同時検索要求を適切に処理できるようにシステムを改修
- 計算結果が確実にユーザに届けられる仕組みを実装



本年度開発：検索結果表示機能の改良(2/3)

- 検索結果に対する付加情報の追加
  - 検索結果として得られる類似発現プロファイル一覧に細胞種や実験条件などのアノテーションを追加する
  - 発現プロファイルの特徴に関わる情報として、up- or down-regulate されている遺伝子セットの情報を追加する(GSGE)
  - 類似検索に問題がないか（次元の呪いに陥っていないか）を確認するために、類似スコアの度数分布をグラフ表示

Similar set	Dissimilar set	Rank	Sample ID	Annotation	Corr. Coef.	P-value	Link
☐	☐	1	GSM189751	Intestinal mucosa biopsy from healthy human individuals other: pair 01 gender: female age: 30 genotype/variation: monozygotic twin	1.000	0.00	Origina site Vizualization
☐	☐	2	GSM189754	Intestinal mucosa biopsy from healthy human individuals other: pair 02 gender: female age: 39 genotype/variation: monozygotic twin	0.998		
☐	☐	3	GSM189753	Intestinal mucosa biopsy from healthy human individuals other: pair 02 gender: female age: 39 genotype/variation: monozygotic twin	0.998		
☐	☐	4	GSM189763	Intestinal mucosa biopsy from healthy human individuals other: pair 07 gender: female age: 39 genotype/variation: monozygotic twin	0.997		



本年度開発:操作性簡便化・検索結果表示機能の改良に伴う計算方法の変更(3/3)

- 発現プロファイルの類似検索
  - 正規化して相関係数を計算、相関係数が降順となるようにソートし、上位をヒットしたものとして表示
  - データベース全体に対して上述計算を実行
  - GEを用いてスケジューリング
  - 計算終了をメール通知、あるいはユーザ別ジョブリスト(次世代シーケンサ解析システムで実現されている機能)で管理
  - 正規化にはedgeRもしくはTCCの機能を使用(calcNormFacotrs/getNormalizedData)
- up- or down-regulate されている遺伝子セットの同定
  - 類似検索の結果を用いる
  - 類似する発現プロファイルセットと類似しない発現プロファイルセットを用意し、それらに対してGSEAを実行する(類似発現プロファイルの上位10個を類似しない発現プロファイルセットに加え、下位50%からランダムに選択した発現プロファイルセットを類似しない発現プロファイルセットに加える等)
  - GSEAの結果を類似検索の結果に付与し、ユーザに提供する
  - 検討中:類似する発現プロファイルセット、類似しない発現プロファイルセットをユーザが設定できるようにするか?

本開発の重要ポイントの確認(1/3)

昨年12月20日の打ち合わせにおいてHGC側から示された重要ポイントは以下の通り。

1. ユーザインターフェース(WEBページ)の見た目を美しくしてほしい
2. 操作系を、簡便且つ直感的に操作できるものにしてほしい
3. システムに柔軟性を持たせてほしい。柔軟性とは、システムの挙動を簡単に変更できることであり、設定ファイルの変更や、プラグインの追加で実現されることが望ましい。
4. 計算ログ(実行されたコマンドのログ)をユーザが見れるようにしてほしい
5. 管理ページが欲しい。管理ページとは、データベースの更新を行うための管理者用のページの事で1次データの所在を入力して「更新ボタン」を押すとデータベース更新が行える機能を有するページのことである。更に、データやプログラムのバージョンが一覧できる必要もある。ただし、まずはデータやプログラムのバージョン一覧が 表示されることとし、更新については、更新方法をドキュメント化し、管理ページに更新ボタンを組み込むのは、優先度的には後とする。
6. APIを作ってほしい。
7. 詳細設計書が欲しい。
8. 将来的にはゲストユーザが利用することを想定してシステムを作成してほしい。
9. 発現プロファイル検索は、根本的に難しいので、本開発でその性能を追求する必要はない。それよりも1.～3.や5.が重要。

本開発の重要ポイントの確認(2/3)

前項にしめしたHGCからの要望に対するMRIの見解は以下の通り。

- 基本的に全ての要望に対応する。
- ただし、3.、5.及び7.に関しては、ここまで想定していなかったことと、どこまでちゃんとやるかがコストに直結することとから、HGC側と相談しながら進めることとする。
- 3.に関しては、プラグインの追加機能のようなものは後回しにして、まず発現プロファイルの類似検索機能を実現させる。その外観や操作性が満足いくものになってから、プラグイン追加機能の実装に取り掛かる。
- 7.に関しては、完全な詳細設計書を用意するのではなく、まず、打ち合わせでの議論に用いれる程度のある程度ざっくりしたものを用意する。どのファイルがどのような機能を有しているかや、インプット・アウトプットは何かといった程度の物は用意する。
- つまり、3.及び7.に関しては、Agile型開発とする。

本開発の重要ポイントの確認(3/3)

本開発の重要ポイントについてまとめる。

- ユーザフレンドリーな外観にする
  - 見た目の美しさ
  - 操作性の高さ(GEO2R等が参考になる)
  - 使ってみて有用性が伝わること(予てからの中井教授からの要望)
- 拡張性を持たせること
  - コードの見通しをよくする(美しいコードにするというよりは、設定の変更や機能拡張がしやすい状態に保つことが主旨)
    - システムの設計をドキュメンテーションする(ただし、詳細設計を懇切丁寧に書き下すようなことはしない)
    - システムの挙動に関わるパラメータを、できるだけ集約する(設定ファイル・ディレクトリに集約する)
    - 解析フロー中の機能追加・変更が見込まれる工程に関しては、プラグイン化する
- 運用しやすくすること
  - システムに含まれている解析ツールやデータのバージョンが一覧できるようにする
  - データ更新がWEBインターフェースからできるようにする(基盤システムネットワーク内からは外部サイトのデータをダウンロードできないことから、現状では実現不可能)

システム設計(1/10)

- ✓ まずは、画面遷移や処理の流れなど、システムの大枠を示す。
- ✓ 今後、開発と同時進行する形でドキュメント化する。個別の画面設計や、入出力データのフォーマット、使用されるライブラリ等については、その過程でドキュメントに含めることとする。

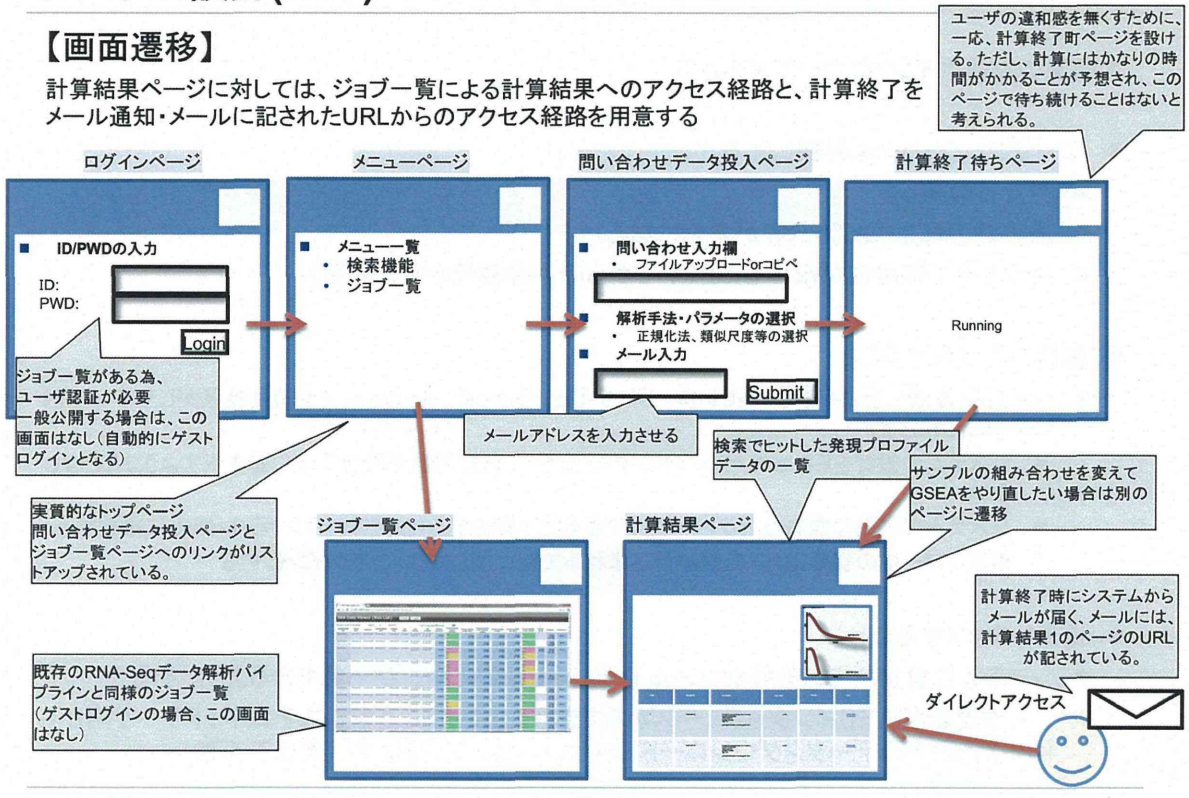
以下は、ドキュメントの目次案。■は確定次項、●は本打ち合わせの相談次項、◆は今後の開発の進捗とともにドキュメント化される予定。

- □ ユーザインターフェース
  - ■ 画面遷移
    - ■ 案①メール送信できない場合
    - ■ 案②メール送信できる場合
  - ● 画面設計
  - ◆ ユーザインターフェース作成に使用するライブラリ、使用方法
  - ● 入出力データのフォーマット
  - ● CGIパラメータ
- □ 処理フロー
  - ● 全体像
  - ● 解析に関わる処理フロー
  - ● その他の処理フロー(メール送信等)
- □ ディレクトリ・ファイル構成・拡張性に関わる設計
  - ■ ディレクトリ・ファイル構成
  - ◆ 拡張性に関わる設計(プラグインの仕組み等)

システム設計(2/10)

【画面遷移】

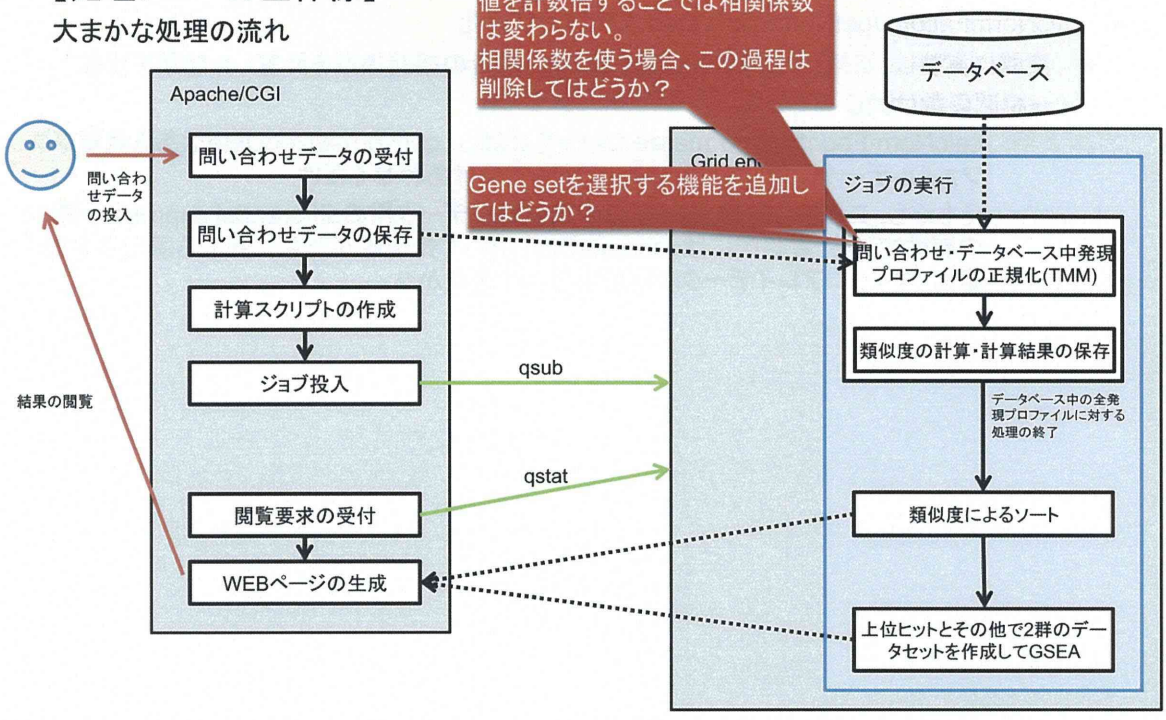
計算結果ページに対しては、ジョブ一覧による計算結果へのアクセス経路と、計算終了をメール通知・メールに記されたURLからのアクセス経路を用意する



システム設計(3/10)

【処理フローの全体像】

大まかな処理の流れ



システム設計(4/10)

【ディレクトリ・ファイル構成・拡張性】

- 現在、RNA-Seq解析パイプラインをデファクタリング中。ディレクトリ構成、設定ファイルが整理された後、それらをそのままコピーして、本開発に流用する計画。
  - 昨年度の開発においては、RNA-Seq解析システムと発現プロファイル検索システムを本番環境で統合する経緯があり、それがそのままディレクトリ構成・URLに反映されていた。URLとしても美しくなかったため、機能の系統関係がそのまま反映された形に変更。
  - 設定ファイルが分散していたため機能別に集約。以下の3ファイルになる予定。
    - WEBインターフェース用
    - RNA-Seq解析システム用
    - 発現プロファイル検索機能用
- 具体的なディレクトリ構成・設定ファイルについては別資料を参照。

システム設計(5/10)

【類似性検索機能の検討(1/4)】

- calcNormFacotrs/getNormalizedDataを用いた正規化
  - 実際に実装し、正規化した場合と正規化しない場合の相関係数を計算した結果を比較
    →相関係数は同じ
    - calcNormFacotrsはnormalize factorを計算し、getNormalizedDateはある発現プロファイルデータに対してnormalize factorを掛けあわせるだけ
    - そもそも、これらはRNA-Seq等のタグカウントデータ用のdifferential expression用の関数である(最終的にはestimateDEを用いて発現値が有意に異なる遺伝子を同定)。マイクロアレイデータに使用するには適切か？

システム設計(6/10)

【類似性検索機能の検討(2/4)】

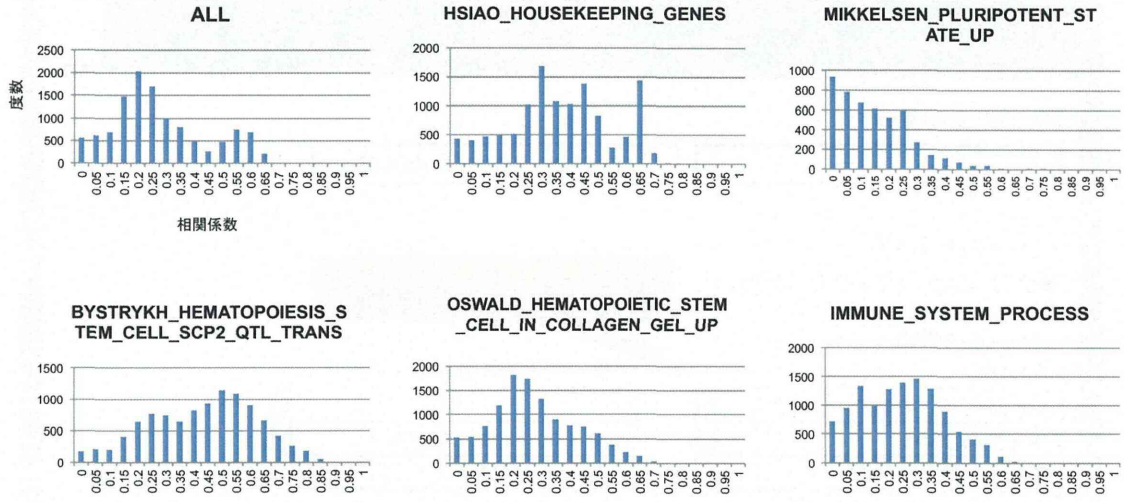
- 正規化せずにピアソンの相関係数を用いてはだめか？
  - mRNAのカウント同士(マイクロアレイ、RPKM)、タグカウント同士(RNA-Seqの生データ)であれば、そこそこうまく機能するはず→実際に試してみる
    - 以下のデータを問い合わせとして、GEOのdata setに登録されているデータに対して相関係数を計算

      GSM603052 Analysis of iPSCs generated from fibroblasts from patients with Hutchinson-Gilford progeria syndrome (HGPS), a rare and fatal premature aging disease. Premature aging was recapitulated by differentiation of the HGPS-iPSCs. Results provide insight into molecular mechanisms underlying premature aging. Induced pluripotent stem cell-based accelerated aging model Expression profiling by array Homo sapiens in situ oligonucleotideGPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array Homo sapiens
    - 今回の計算では、遺伝子の区別にEntrez Gene IDを使用。現状、GEOのdata setには、32,199の発現プロファイルデータが登録されており、Entrez Gene IDが付与されているものは、23,518つ。
    - 2つの発現プロファイルで共有する遺伝子のみを相関係数の計算に使用。全遺伝子の相関係数の計算の他、MSigDBに登録されいてるGene setでの相関係数も計算。
    - MSigDBに登録されいてる10,295のGene setのうち、77つを対象に計算。Shirokaneでの計算時間は5時間30分程度。

システム設計(7/10)

【類似性検索機能の検討(3/4)】

- GSM503052を問い合わせとした相関係数の計算



システム設計(8/10)

【類似性検索機能の検討(4/4)】

- GSM503052を問い合わせとした相関係数の計算

問い合わせ	DB	遺伝子数	相関係数(ALLの平均値を使用)	p値	相関係数	p値	遺伝子セット
GSM603052	GSM603054	146	0.122098	7.10E-02	0.995225	0.00E+00	PECE_MAMMARY_STEM_CELL_DN
GSM603052	GSM603054	388	0.122074	8.07E-03	0.994772	0.00E+00	HSIAO_HOUSEKEEPING_GENES
GSM603052	GSM603050	146	0.122120	7.10E-02	0.994443	0.00E+00	PECE_MAMMARY_STEM_CELL_DN
GSM603052	GSM603054	32	0.1221	2.53E-01	0.994076	0.00E+00	VANOVA_HEMATOPOIESIS_STEM_CELL_SHORT_TERM
GSM603052	GSM603054	200	0.122081	4.02E-02	0.993372	0.00E+00	RAMALHO_STEMNESS_UP
GSM603052	GSM603050	200	0.122118	4.02E-02	0.991625	0.00E+00	RAMALHO_STEMNESS_UP
GSM603052	GSM603054	888	0.122094	1.35E-04	0.991615	0.00E+00	REACTOME_IMMUNE_SYSTEM
GSM603052	GSM603050	886	0.122122	1.34E-04	0.991075	0.00E+00	REACTOME_IMMUNE_SYSTEM
GSM603052	GSM603054	274	0.122070	2.17E-02	0.990976	0.00E+00	OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL_DN
GSM603052	GSM603050	124	0.122108	8.83E-02	0.990965	0.00E+00	KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS
GSM603052	GSM603054	70	0.122084	1.47E-01	0.99096	0.00E+00	KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM
GSM603052	GSM603054	24	0.122104	2.85E-01	0.990937	0.00E+00	BYSTRYKH_HEMATOPOIESIS_STEM_CELL_SCP2_QTL_TRANS
GSM603052	GSM603050	220	0.122133	3.53E-02	0.988964	0.00E+00	ENDOMEMBRANE_SYSTEM
GSM603052	GSM603054	38	0.122081	2.33E-01	0.98893	0.00E+00	JIANG_HEMATOPOIESIS_STEM_CELL_NUMBER_SMALL_VS_HUGE_UP
GSM603052	GSM603050	874	0.122131	1.48E-04	0.988308	0.00E+00	BYSTRYKH_HEMATOPOIESIS_STEM_CELL_QTL_TRANS
GSM603052	GSM603050	66	0.12213	1.64E-01	0.988037	0.00E+00	FEIER_GLIOMA_STEM_CELL_DN
GSM603052	GSM603054	316	0.12208	1.50E-02	0.98787	0.00E+00	JAATINEN_HEMATOPOIETIC_STEM_CELL_UP
GSM603052	GSM603054	21366	0.987286	0.00E+00	0.987286	0.00E+00	ALL
GSM603052	GSM603050	21366	0.987105	0.00E+00	0.987105	0.00E+00	ALL
GSM603052	GSM603050	672	0.122126	7.57E-04	0.987045	0.00E+00	VANOVA_HEMATOPOIESIS_STEM_CELL_AND_PROGENITOR
GSM603052	GSM603054	94	0.122083	1.21E-01	0.986987	0.00E+00	HOEBEKE_LYMPHOID_STEM_CELL_UP
GSM603052	GSM603050	332	0.122133	1.30E-02	0.986705	0.00E+00	IMMUNE_SYSTEM_PROCESS
GSM603052	GSM603054	232	0.122104	3.17E-02	0.986482	0.00E+00	OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL_UP
GSM603052	GSM603054	718	0.122094	5.31E-04	0.98594	0.00E+00	WONG_ADULT_TISSUE_STEM_MODULE
GSM603052	GSM603054	296	0.122078	1.79E-02	0.985134	0.00E+00	VANOVA_HEMATOPOIESIS_STEM_CELL_LONG_TERM
GSM603052	GSM603050	274	0.122101	2.17E-02	0.984803	0.00E+00	OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL_DN
GSM603052	GSM603050	86	0.122115	1.31E-01	0.983898	0.00E+00	HOEBEKE_LYMPHOID_STEM_CELL_DN
GSM603052	GSM603054	420	0.122099	5.83E-03	0.983665	0.00E+00	LIM_MAMMARY_STEM_CELL_DN
GSM603052	GSM603050	420	0.122125	5.82E-03	0.982642	0.00E+00	LIM_MAMMARY_STEM_CELL_DN
GSM603052	GSM603050	108	0.12213	1.04E-01	0.981473	0.00E+00	ZADPANAH_STEM_CELL_ADIPOSE_VS_BONE_DN
GSM603052	GSM603050	36	0.122112	2.33E-01	0.98146	0.00E+00	JIANG_HEMATOPOIESIS_STEM_CELL_NUMBER_SMALL_VS_HUGE_UP
GSM603052	GSM603050	296	0.12211	1.79E-02	0.98101	0.00E+00	VANOVA_HEMATOPOIESIS_STEM_CELL_LONG_TERM
GSM603052	GSM603054	854	0.122088	1.75E-04	0.980313	0.00E+00	SYSTEM_DEVELOPMENT
GSM603052	GSM603050	716	0.122125	5.29E-04	0.979958	0.00E+00	WONG_ADULT_TISSUE_STEM_MODULE
GSM603052	GSM603064	216	0.122098	3.67E-02	0.979396	0.00E+00	BOQUEST_STEM_CELL_DN
GSM603052	GSM603050	232	0.122134	3.16E-02	0.97924	0.00E+00	OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL_UP
GSM603052	GSM603054	118	0.122088	9.39E-02	0.978965	0.00E+00	LEE_NEURAL_CREST_STEM_CELL_DN
GSM603052	GSM603054	382	0.12209	8.49E-03	0.975909	0.00E+00	NERVOUS_SYSTEM_DEVELOPMENT
GSM603052	GSM603050	260	0.122111	2.48E-02	0.975483	0.00E+00	BOQUEST_STEM_CELL_UP
GSM603052	GSM603050	560	0.122115	1.90E-03	0.9752	0.00E+00	SYSTEM_PROCESS
GSM603052	GSM603050	382	0.122122	8.47E-03	0.975152	0.00E+00	NERVOUS_SYSTEM_DEVELOPMENT
GSM603052	GSM603050	120	0.12212	8.66E-02	0.97503	0.00E+00	ZADPANAH_STEM_CELL_ADIPOSE_VS_BONE_UP
GSM603052	GSM603054	560	0.122084	1.91E-03	0.97476	0.00E+00	SYSTEM_PROCESS
GSM603052	GSM603054	124	0.122077	8.84E-02	0.973344	0.00E+00	KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS
GSM603052	GSM603050	376	0.122104	8.93E-03	0.972533	0.00E+00	NEUROLOGICAL_SYSTEM_PROCESS
GSM603052	GSM603050	118	0.12212	9.39E-02	0.972098	0.00E+00	LEE_NEURAL_CREST_STEM_CELL_DN
GSM603052	GSM603054	376	0.122074	8.94E-03	0.97168	0.00E+00	NEUROLOGICAL_SYSTEM_PROCESS
GSM603052	GSM603050	488	0.122122	3.46E-03	0.97113	0.00E+00	LIM_MAMMARY_STEM_CELL_UP
GSM603052	GSM603015	388	0.122139	8.04E-03	0.953604	0.00E+00	HSIAO_HOUSEKEEPING_GENES
GSM603052	GSM603015	146	0.122158	7.09E-02	0.9478	0.00E+00	PECE_MAMMARY_STEM_CELL_DN
GSM603052	GSM603015	200	0.122148	4.01E-02	0.925555	0.00E+00	RAMALHO_STEMNESS_UP
GSM603052	GSM603054	146	0.122088	7.11E-02	0.924646	0.00E+00	LEE_NEURAL_CREST_STEM_CELL_UP
GSM603052	GSM603015	220	0.122163	3.53E-02	0.914435	0.00E+00	ENDOMEMBRANE_SYSTEM
GSM603052	GSM603044	220	0.122072	3.54E-02	0.902592	0.00E+00	ENDOMEMBRANE_SYSTEM
GSM603052	GSM603015	886	0.122152	1.34E-04	0.8886	0.00E+00	REACTOME_IMMUNE_SYSTEM
GSM603052	GSM603044	886	0.12206	1.35E-04	0.887137	0.00E+00	REACTOME_IMMUNE_SYSTEM