マイクロアレイデータ解析（4 日目、5 日目、6 日目）

アレイデータのファイルや、実習の連絡事項を共有できる wiki を用意しました。
http://bit.ly/microarray2014

〈4 日目〉

ｉ）Excel によるデータ解析

1. 全サンプルのデータを 1 つのワークシートにまとめる
行番号とプローブ番号の対応関係はどのサンプルも同じ。したがってサンプル X のデータの隣にサンプル Y のデータをコピー&ペーストすれば 1 枚のワークシートにまとまる。上記を繰り返して、全サンプルの gProcessedSignal, gIsWellAboveBG の値を 1 枚のワークシートにまとめる。この状態で一度ファイルを保存しておくこと。

最初のサンプルで残す列：
・FeatureNum - スポットの番号
・ControlType - コントロールのスポットかどうか？
・ProbeName - プローブ名
・GeneName - 遺伝子名
・SystematicName - 遺伝子名
・gProcessedSignal - green（Cy-3）のシグナル強度
・gIsWellAboveBG - シグナルがバックより十分高いか？

第 2 サンプル以降で残す列：
・gProcessedSignal
・gIsWellAboveBG

2. 発現量が少なくアレイで検出できなかったものは、gIsWellAboveBG の値が 0 となっている。今回は、全てのサンプルで gIsWellAboveBG の値が 1 となっているもののみを解析の対象とする。また、ControlType が 0 以外のものはコントロールのプローブなので、解析の対象からはずす。各自、関数を利用したり「並べ替え」機能を用いて解析の対象となる行を抽出すること。

3. X-Y プロット
サンプル X の gProcessedSignal の値（以下、X）を横軸、サンプル Y の gProcessedSignal の値（以下、Y）を縦軸にプロット。

4. MA プロット
log10(XY)を横軸、log2(Y/X)を縦軸にプロット。
・縦軸が 0 とはどういう意味か？ → 発現量に変化なし
・縦軸が 1 とはどういう意味か？ → 発現量が 2 倍に上昇
・縦軸が-1 とはどういう意味か？ → 発現量が半分に減少

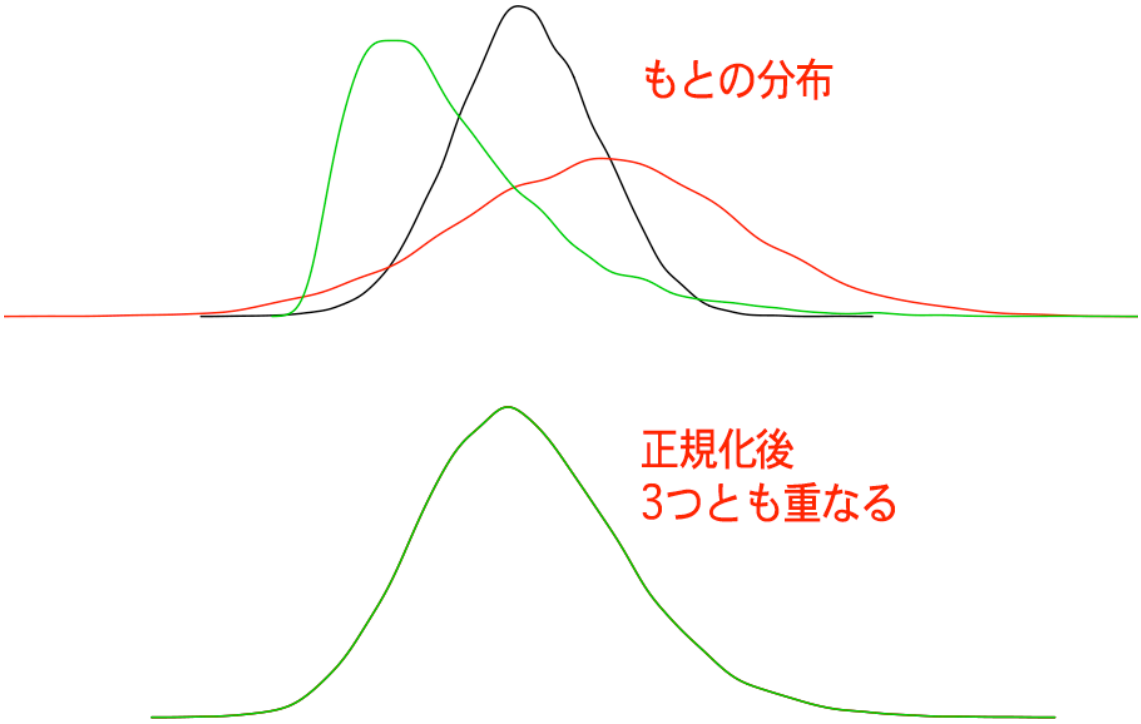
5. 解析
replicate 間で X-Y プロットおよび MA プロットを作成。
◎ mock-1 対 mock-2
・fold change が 1 以上のものは全体の何%か。
・fold change が-1 以下のものは全体の何%か。
・fold change が 0.1 以上のものは全体の何%か。
・fold change が-0.1 以下のものは全体の何%か。

異なるサンプル間で X-Y プロットおよび MA プロットを作成。
◎ mock-1（または mock-2）対 RNA-1（または RNA-2）で X-Y プロットおよび MA プロットを作成。
・fold change が 1 以上のものは全体の何%か。
・fold change が-1 以下のものは全体の何%か。
・fold change が 0.1 以上のものは全体の何%か。
・fold change が-0.1 以下のものは全体の何%か。
◎ mock-1 対 2OMe で X-Y プロットおよび MA プロットを作成。
◎ mock-1 対 LNA で X-Y プロットおよび MA プロットを作成。
（以下同様）

1'. 参考：アレイ間の normalization (75%tile 法)
アレイ間の全体的なシグナル強度の差を補正するために必要な操作で、方法もいくつかある。ここでは 75%tile 法でおこなう。
各サンプルごとに、コントロールを除く全プローブでの発現量の値を順番に並べ替え、順位 75%に位置するもの(発現量が高いほうから数えて 25%目の遺伝子)の値を求める。この 75%tile の値は通常サンプルごとに異なるが、それらの相乗平均値 a を求める。各サンプルごとに、全プローブの値に（a / そのサンプルにおける 75%tile 値）を掛ける。全サンプルで 75%tile の値を a に揃える。

1". 参考：アレイ間の normalization (quantile) → 余裕がある人は Excel で実装
シグナル強度の分布を、サンプル間で強制的に合わせてしまう normalization の手法。
各サンプルごとに、コントロールを除く全プローブでの発現量の値を順番に並べ替える。1 位になっているプローブでシグナル値の相乗平均を求め、その値で置き換える（log2 を求めて相加平均を取るのが簡単)。2 位以下も同様に、同順位のプローブでシグナル値の相乗平均を取って値を置き換える。結果、全サンプルで分布が同一になる。

例）発現量が n 位の log2(シグナル値) が (-0.4, 0.1, 0.2, -0.3)なら、(-0.1, -0.1, -0.1, -0.1) に置き換えてしまう。



〈５日目〉

ii）GeneSpring GX による解析

0. GeneSpring GX のインストール

講師からライセンスコードを受け取り、GeneSpring GX をインストールする。詳細は wiki を参照。なお、帰る前に必ずライセンスを Surrender すること。忘れると別の端末で使えなくなる。Automatic Software Update の画面が出る場合があるが、今回の実習では必要ないので Cancel を選択（実行するとかなり時間がかかり、ディスク容量も消費する）。

1. データの読み込み
GeneSpring GX を起動（初回は時間がかかる）
Project → New Project → 適当な名前をつける
Create new experiment
　　Experiment name → 適当な名前をつける
　　Analysis type → Expression

Experiment type → Agilent Expression Single Color
Workflow type → Data Import Wizard
Experiment notes → メモ。空欄でも可
Load Data → Choose Files から数値データ（TXT）を全部選択。8 班分。
/home08/sirna/Microarray/2014-06-13/*

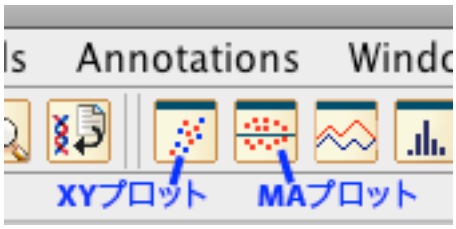
途中で、初回のみ Technology Agilent 何とかが not found. などというエラーがでる場合があるが、「Yes」と答えれば Agilent の 1 色法の解析に必要なファイルが自動的にダウンロードされる。

Use spot information in data files to flag the data のみにチェック。以下推奨設定。
Feature is not positive and significant → Not Detected
Feature is not Uniform → Compromised
Feature is not above background → Not Detected
Feature is Saturated → Compromised
Feature is population outlier → Compromised
Normailzation algorithm → Quantile
Baseline Options → Do not perform baseline transformation

2. サンプル名の入力・replicate の指定
画面右側の Experiment Setup タブ → Experiment Grouping → Add Parameter
Parameter name → (例) sample name などと入れておく。
Samples に表示されているデータファイルに対応するサンプル名を、Parameter Values に入力していく（mock, RNA, …）。同じ Parameter Value を入力したものが自動的に replicate として扱われる（後で平均される）。
続いて、画面右側の Experiment Setup タブ → Create Interpretation
先ほど入力したパラメータをチェックして Next → Categorical にチェック → Average over replicates in conditions は Avaraged をチェック。Use Measurements Flagged は Detected のみチェック（Not Detected または Compromised のフラグが立っているものは使用しない）。この操作で、replicate 間の平均の値が算出され、以降ではその値を利用して解析を進めることができる。

3. バックグラウンドに近い値や、異常な値を除去
画面右側の Quality Control タブ → Filter Probesets by Flags → Next
Acceptable Flags: Detected のみチェック
at least [8] out of 8 samples have acceptable values
＝1 サンプルでも Not detected または Compromised と判定された値があるものは除外。

4. X-Y プロット、MA プロットを、Excel と同様に作成する。



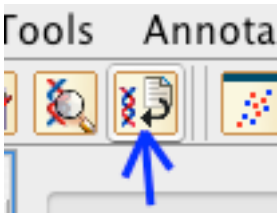
※GeneSpring GX における MA プロットでは、たとえば X-Axis に Wild Type、Y-Axis に mock を選択すると、Wild Type のサンプルで増加している遺伝子が上方（> 0）に、減少している遺伝子が下方（< 0）にプロットされる。

5. 注目している遺伝子の抽出
遺伝子（アクセッション番号）リストを読み込み、その遺伝子に色をつける。
遺伝子リストの例：

Accession NM_003380 NM_014616 NM_000368 NM_177402 ……

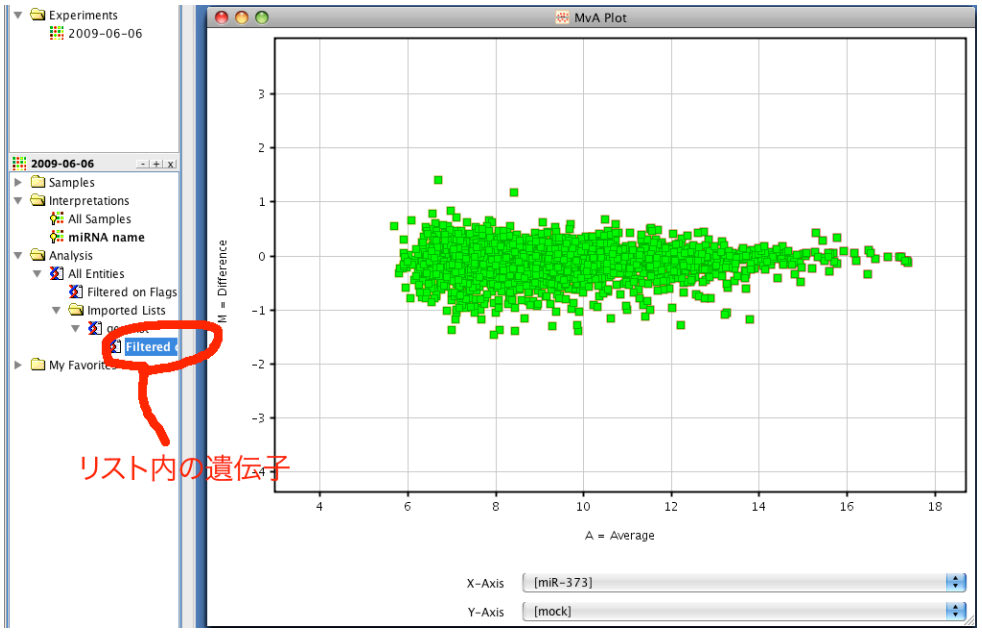
1 行目にタイトル、2 行目以降にアクセッション番号を 1 行につき 1 個記載。
上記をテキストファイルに保存する。

Import entity list from file ボタンを押す。

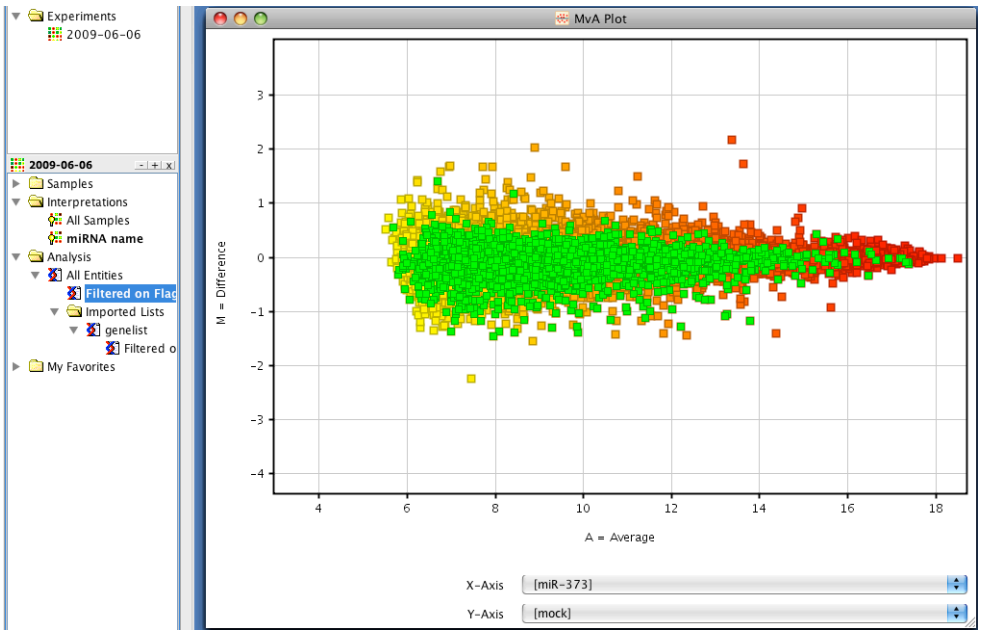


Choose file → 遺伝子リストのテキストファイルを選択
Choose file column to match → 読み込ませた遺伝子リストの 1 行目
Choose technology column to match → Genbank Accession
（アレイのプローブ ID や遺伝子名など、他の項目のリストを読むことも可能）

画面左側の Analysis フォルダ内に Imported Lists フォルダができ、その中に読み込んだリストが保存されている。それを選択すると、リスト内の遺伝子のみがプロットされる。この状態で、「3. バックに近い値や、異常な値を除去」と同じ操作をおこなう。
すべての点を選択した状態で、All Entities 内の Filtered on Flags [Detected]を選択すると、全体のなかでリスト内の遺伝子がどこにプロットされているかわかる。



↓



6. siRNA の seed 部分(*)と相補的な配列が 3' UTR に存在する遺伝子は、siRNA の標的となりうることが報告されている。今回の実験で、RNA-1 (RNA-2)の seed マッチする遺伝子がどのように変動しているか検証せよ。

iii）参考

1. 入力した配列を 3'UTR にもつ遺伝子（アクセッション番号）のリストを表示するページ : http://atlas.rnai.jp/seedmatch/

〈6 日目〉

iii) 統計解析ソフト R を用いたデータ解析

R はオープンソースの統計解析ソフトで、さまざまな OS で利用可能。
Excel や GeneSpring GX と異なりコマンドラインからの操作が基本となるが、大量のデータを高速に扱え、グラフの描画機能も強力である。

実習では Wiki に掲載されているサンプルコードをもとに各種プロットを描画する。