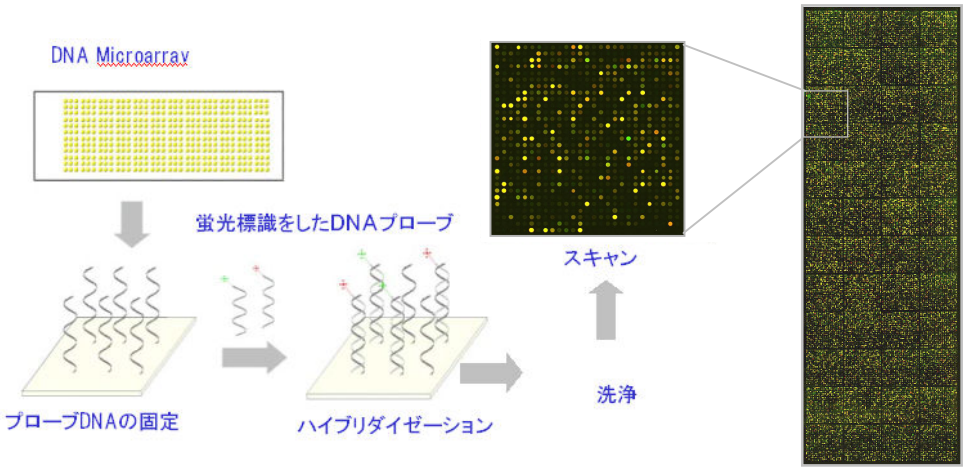
Filgen®
biochip & nanofilm technologies

Microarray Data Analysis Tool データ解析の進め方 Vol.1

－ 解析を始める前に －

まず、はじめに

マイクロアレイ(下図参照)とは数千から数万といった膨大な遺伝子情報をもつDNAプローブが貼り付けられた基板のことをいいます。実験対象となる細胞から抽出されたmRNAを蛍光標識をし、基板のDNAプローブに対して、ハイブリダイゼーションを行います。その後、スキャナーで蛍光強度を読み取り、数値化を行います。これにより、一度の実験で膨大な遺伝子の発現情報を得ることが出来ます。これまでのような個々の遺伝子を対象とした解析から網羅的な解析、ネットワーク的な解析が行えるようになりました。

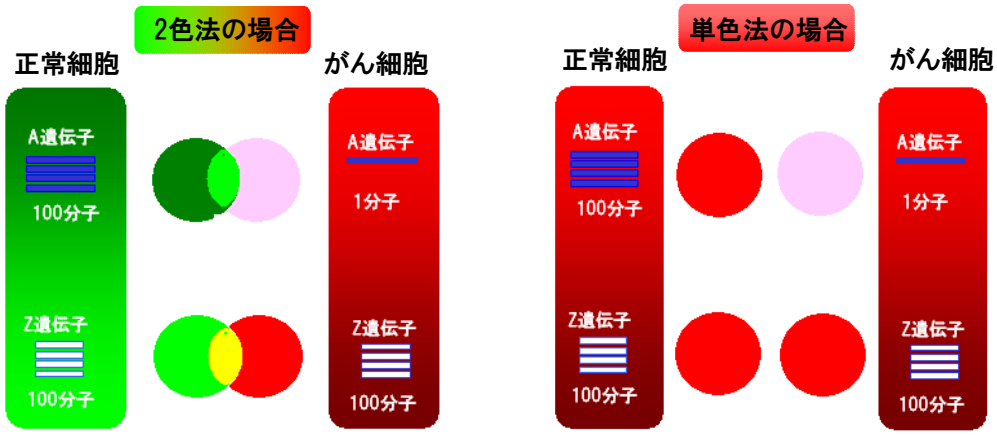


マイクロアレイの解析とは.....

実験後、スキャナーで読み取られた画像データは上図のように丸いスポットが規則正しく配置されたイメージとなります。この、丸いスポットの1つ1つがDNAプローブに対応し、これらの蛍光強度を数値化します。上図は、異なるサンプルを異なる蛍光色素(Cy3,Cy5)でラベリングし、1枚のアレイ上で競合的ハイブリダーゼーションを行った2色法の場合の画像データになります。黄色のスポットが多く、ところどころ緑と赤のスポットがあります。黄色のスポットは赤と緑の比率が1となり、発現量に差がないスポットとなります。つまり、スポットの色が赤または緑になるほど、発現量に差がある(発現変動している)スポットとなります。2色法の他に、異なるサンプルを同じ蛍光色素でラベリングし、1アレイで1サンプルのハイブリダーゼーションを行い、アレイ間のデータを比較する単色法があります。

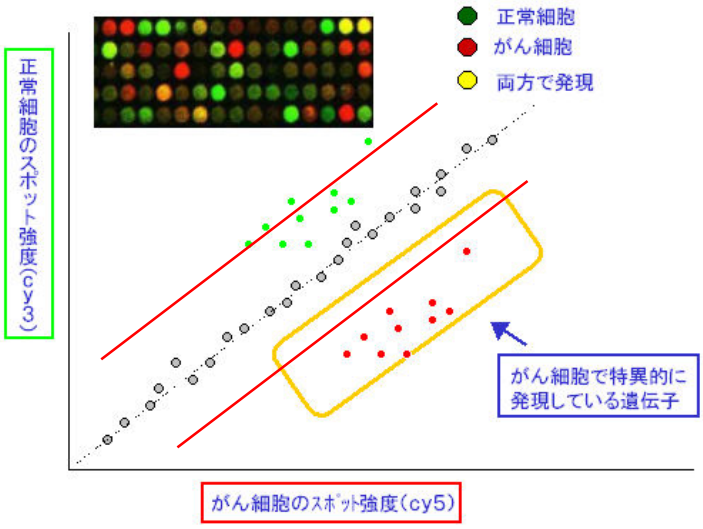
下記が2色法と単色法の概要となります。

2色法の場合は同一スポットに対して異なる蛍光色素でラベリングしたmRNAをハイブリダイズします。そのため、赤と緑という2つのシグナルデータを一つのデータとして表示するため緑のシグナル/赤のシグナルから計算したRatio値(比率)を用います。一方、単色法では別々のアレイに対して、同じ蛍光色素でラベリングしたmRNAを使用するため、Ratio値(比率)だけでなく、シグナル値の比較も可能となります。



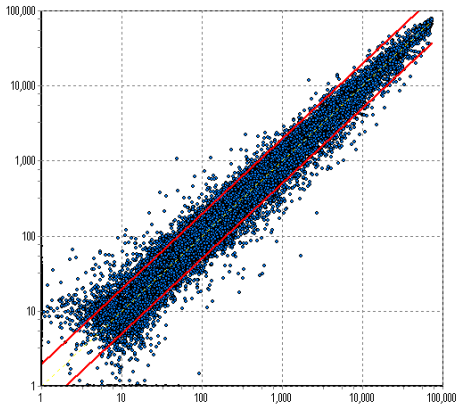
ここまでの話で、マイクロアレイの解析は異なるサンプル間におけるスポットの蛍光強度を比較するということはお分かりになったと思います。しかしながら、実際のアレイのデータは数千から数万という膨大なデータ量であるため、個々のデータを1つ1つみることは非常に大変な作業となります。

そこで、下図のようにスキャッタープロットと呼ばれる散布図がマイクロアレイの解析においてよく使われています。このグラフでは、正常細胞側のスポットの蛍光強度をY軸に、がん細胞側のスポットの蛍光強度をx軸に分布します。各サンプルの発現レベルが同じあれば、プロットは中心45度のラインに乗ってきます。一方、発現レベルの差があるほど、中心から外れてきます。例えば赤いプロットはがん細胞で特異的に高い発現を示したDNAプローブとなります。このように、スキャッタープロットはサンプル間の発現状態の全体像をを視覚的に捉えることができます。



それでは、どのくらいの発現差があれば
有意な差のあるデータといえるのでしょうか？

下図はおよそ3万のデータがプロットされたスキャッタープロットです。先にも述べたように通常、比較サンプル間によるシグナルの比率をみます。スキャッタープロットにおける中心45度のラインがRatio1のラインになります。そして、そのラインを真ん中とし、両方に平行移動した赤の2本のラインが2倍以上発現(Ratio2、Ratio0.5)のラインになります。



アレイの品質にもよりますが、同じサンプルを用いてデータを比較した場合、理論上は45度のラインのRatio1のラインにプロットが収束します。しかしながら、マイクロアレイのデータに中にはアレイの品質、サンプル調製、蛍光標識、ハイブリダイゼーション、洗浄など様々な要因のノイズが含まれ、実際には1.5倍前後のライン付近にプロットが収束されます。そこで、2倍以上の差があれば有意なデータであるとして、データ抽出における閾値の基準に利用されています。この方法には統計的な裏付けはありませんが、マイクロアレイのようにスクリーニング的な要素が強い手法においては、非常によく使用されています。また、この方法で抽出されデータは、後でリアルタイムPCRなどの定量解析法で確認試験を行うことを推奨します。逆に、マイクロアレイだけでデータをまとめようとする場合は、統計的な解析を求められるケースがおおいので、再現性実験が必要となります。

先の説明では発現差という部分でのノイズの影響を説明したのですが、シグナルの大きさでのノイズの影響はどうでしょうか？

上図のスキャッタープロットにおいてシグナルが低いほどバラツキの幅が大きくなっていることがわかります。例えば、シグナルが1と10でも10倍の発現差があります。同じように1000と10000でも10倍の発現差があります。前者は、ノイズレベルのシグナルであり、シグナルが大きい後者の方がデータの信頼度が高くなります。大量のデータから発現差だけを基準にデータを抽出すると、上記のようなノイズ的なデータを多く含め抽出してしまいますので、基準を設け、ノイズデータを除去する必要があります。

それでは、何を基準にしてノイズの除去をすればいいのでしょうか？

使用するアレイやプラットフォームの違いによっても異なりますが、Negative Controlのシグナルを基準にする方法やバックグランドを基準にする方法が多く利用されています。

しかしながら、基準を決めたとしても、すべてのノイズが除去できるとは限りません。逆に除去した中に正しいデータが含まれている可能性もあります。先にも述べましたがノイズであるかどうかも、再現性実験を行うことが一番よい方法であることは確かです。逆に、1比較のデータしかない場合では、基準を高めにし、より厳しい条件でデータを抽出することを推奨します。

次はサンプル間のデータの比較(ノーマライゼーション)について説明します。

従来はbeta-actinやGAPDH等のHousekeeping遺伝子をコントロールとし、それらがサンプル間で一定レベルであるという前提でサンプル間を補正する方法がよく使用されてきました。しかしながら、最近のように全遺伝子を網羅したアレイの補正においては、比較サンプル間での遺伝子の総発現量はほぼ同じであるというグローバルノーマライゼーションが採用されています。刺激によって一部の遺伝子の変動があっても、全体としての発現レベルは変わらないという解釈になります。2色法の場合では異なる蛍光色素の影響を補正するためにloess法が利用されています。

以上、マイクロアレイの概要、解析を進める上でのポイントを紹介してきました。

次はいよいよソフトウェアの使用について説明します。本ソフトウェアは弊社受託サービス専用となっているため、データを読み込むだけですぐに解析に進めることが出来ます。操作も非常に簡単です。ぜひ、トライしてください。