S-PLUSを用いた次世代シーケンサーの実験デザインの最適化

東京農工大学

緒方法親

まとめ

- 序論：次世代シーケンサーとは
- 背景１：次世代シーケンサーによるデータ解析行程
- 背景２：RNA-Seqによる遺伝子発現定量解析の原理
- 目的：最小限データ量によるデータ解析条件の最適化
- 方法：実験データを用いたシミュレーション
- 結果１：MA-Plotによる評価結果
- 結果２：遺伝子発現差解析による評価結果
- 結論

次世代シーケンサーとは

同時並行にDNAの塩基配列を大量に解読する装置

ヒトゲノム解読プロジェクトで活躍
したサンガーシーケンサー

DNA解読量

<u>１日あたり40万塩基対</u>

イネゲノムの解読に

7年、200億円

次世代シーケンサー（NGS）

DNA解読量

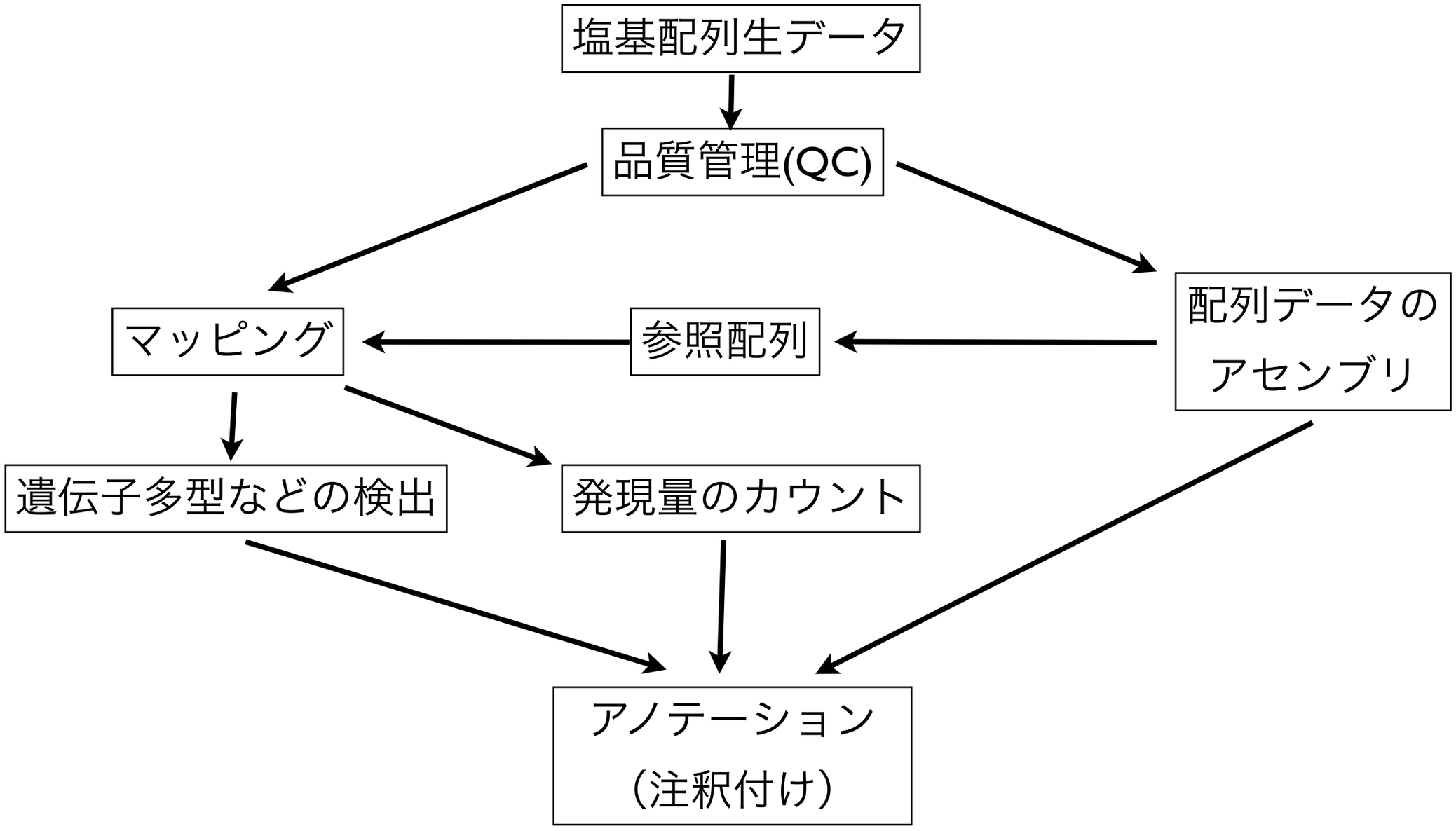
<u>１日あたり50億塩基対</u>

イネゲノムの解読に

3週間、300万円

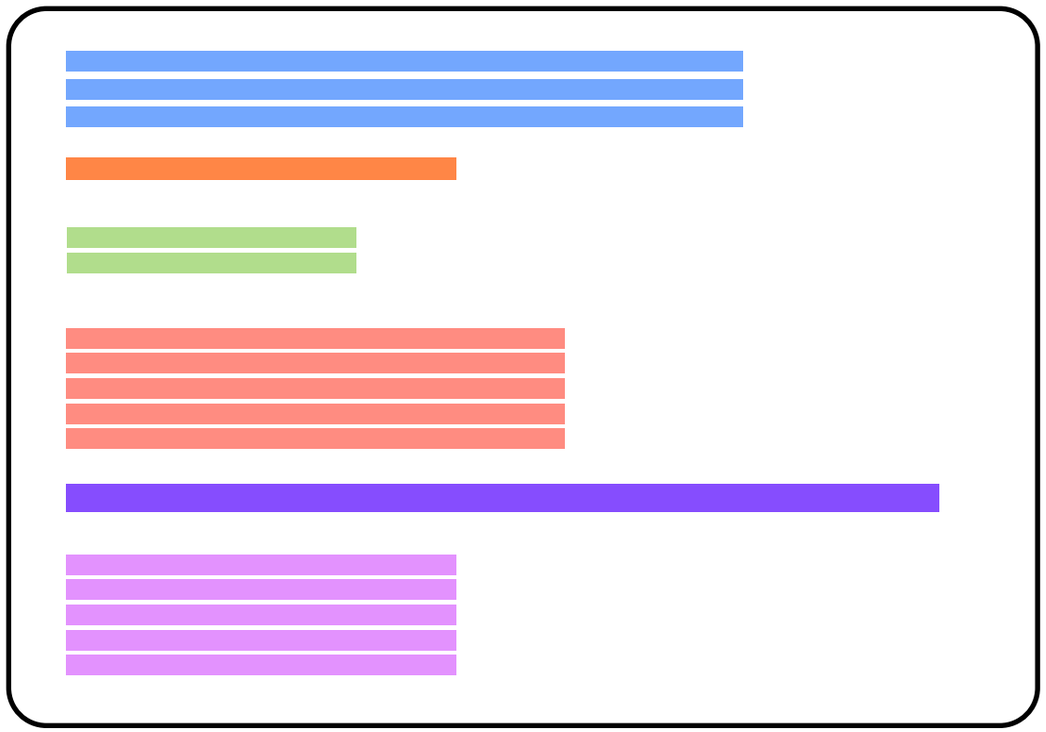
従来より解読スピードが大幅に上昇し、大量のデータ解析が低価格で可能となった。

次世代シーケンサーによるデータ解析行程



- 一回の実験で数十億単位の大量の塩基配列データが産生される
- RNA-Seqの場合は数万個の遺伝子の発現定量データを得ることが可能である
- 塩基配列データから、生物学的に有用な知識発見を行うまでに多くの情報解析が必要

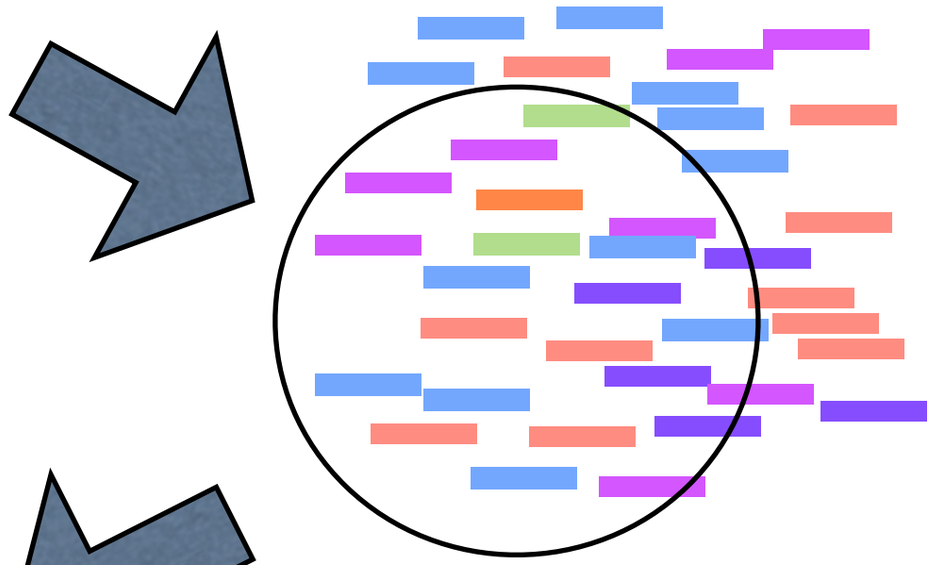
RNA-Seqによる遺伝子発現定量解析の原理



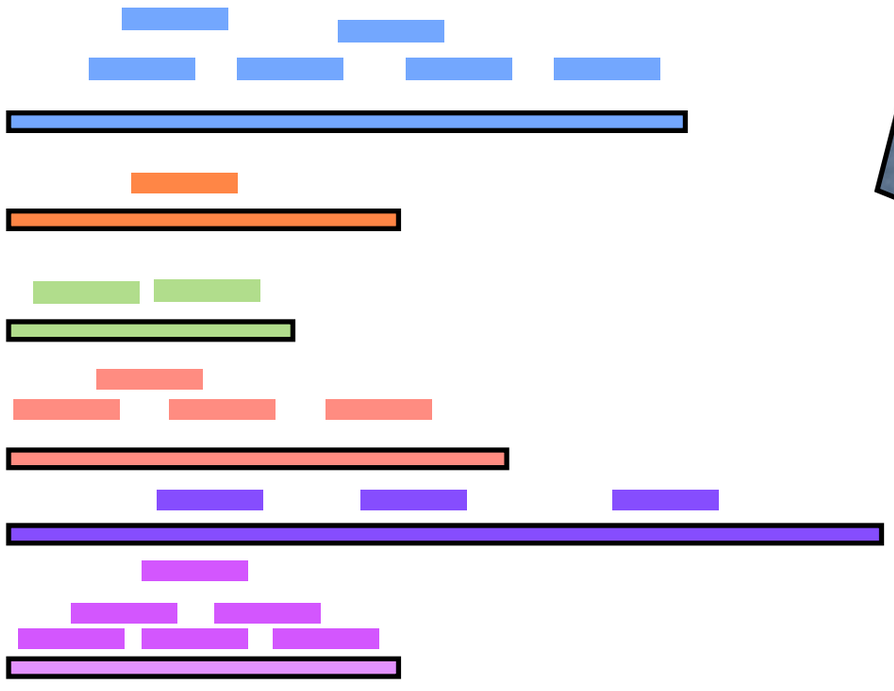
多様な長さ、量のメッセンジャーRNAが細胞内に存在

メッセンジャーRNAの種類はわかっている

各メッセンジャーRNAの発現量が未知

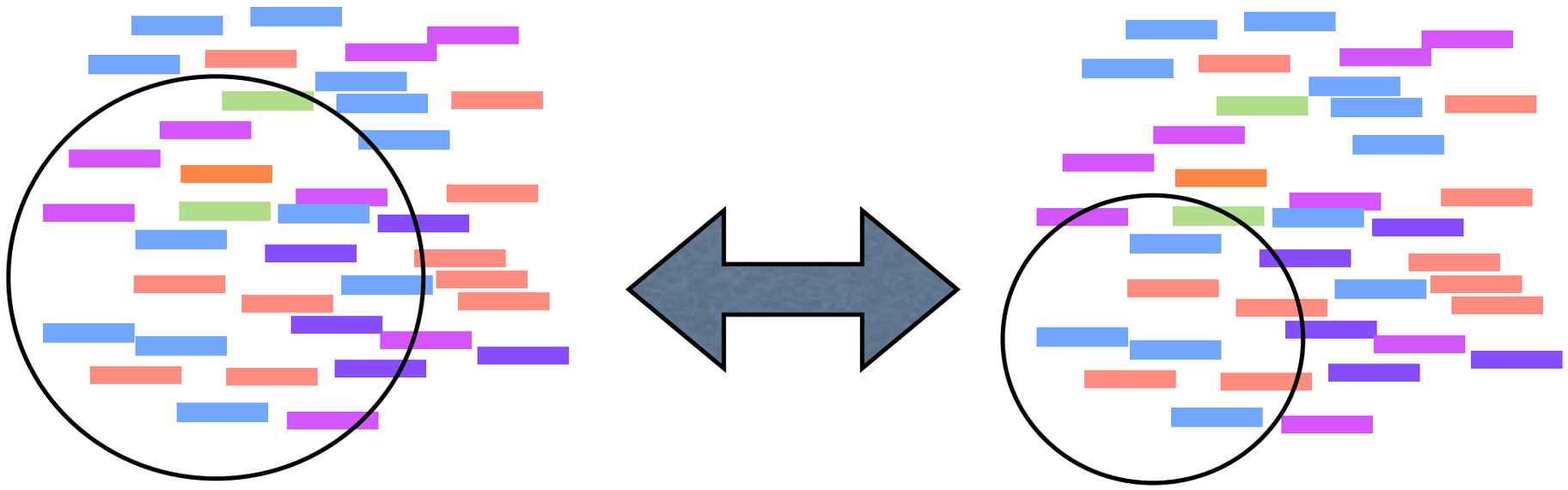


次世代シーケンサーをもちいたRNA-Seq解析は、試料中のRNAを断片化し、その中のメッセンジャーRNAの塩基配列を解読する



得られた塩基配列を用いて相同性解析を行い、各塩基配列がどのメッセンジャーRNAに由来するかを推定する

目的：最小限データ量によるデータ解析条件の最適化



配列解析コスト：　２千万リードあたり15万円

- ガラスプレート上のマイクロ流路内で配列解析を行う
- １本のマイクロ流路に複数のサンプルで配列解析が可能
- １本のマイクロ流路で解析するサンプル数をどうするか

最小限のコストで、比較解析を行うために必要な最小限のデータ量を調べることによりデータ解析条件の最適化を行いたい。

方法：実験データに基づくシミュレーション

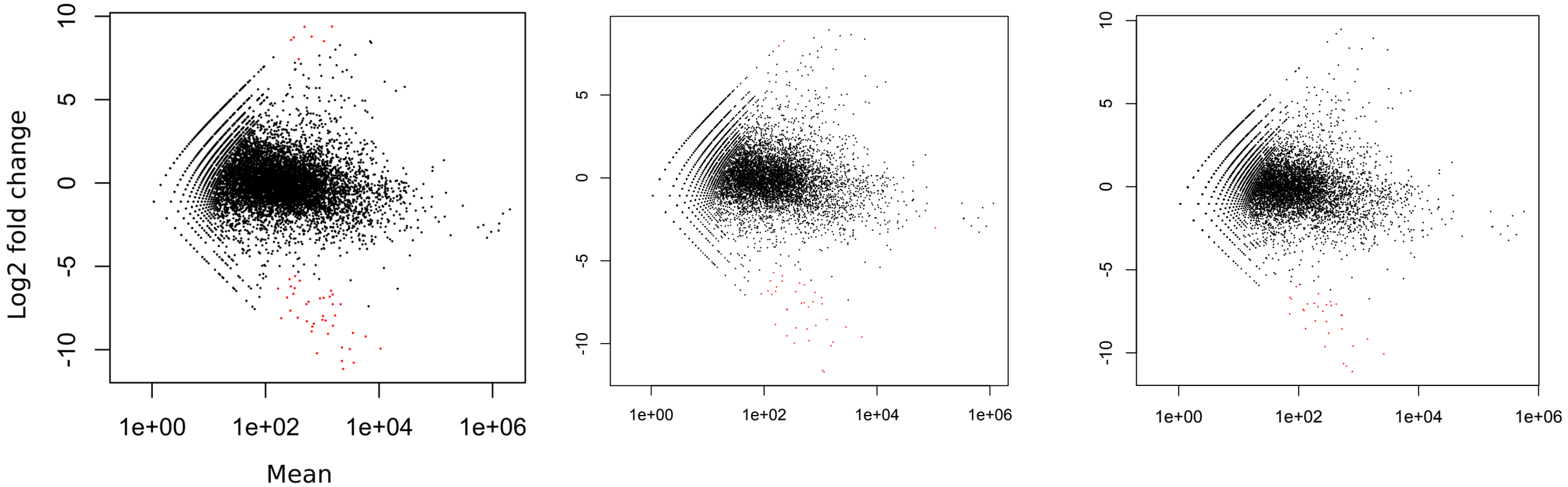
- 1本のマイクロ流路（レーン）全部を用いて配列解析を行ったデータを用いて最適化を検討する。
- 配列解析データから一部のデータを無作為に選択し、少ないデータ量で解析のシミュレーションを行う。つまり、1検体を1/2レーン分、および1/4レーン分のデータ量で解析を行う。
- シミュレーションを行って得られた配列をマッピングし、発現定量解析を行うことにより、比較検討を行う。
- 1レーン全部を用いて比較解析を行ったものを、基準に少ないデータ量で解析のシミュレーション（1/2レーン分、および1/4レーン分の解析データ）を評価する。
- 1レーン全部を用いて比較解析を行ったものに近いデータが得られる最低限度のデータ量を求める。

結果１：MA-Plotによる評価結果

1700万リード　対
1400万リード

1000万リード　対
800万リード

500万リード　対
400万リード



MAplotでは、解析データ数による差はみられれなかった。

全14,623遺伝子中発現が確認されたのは２試料それぞれで

生データ：10,846と12,178

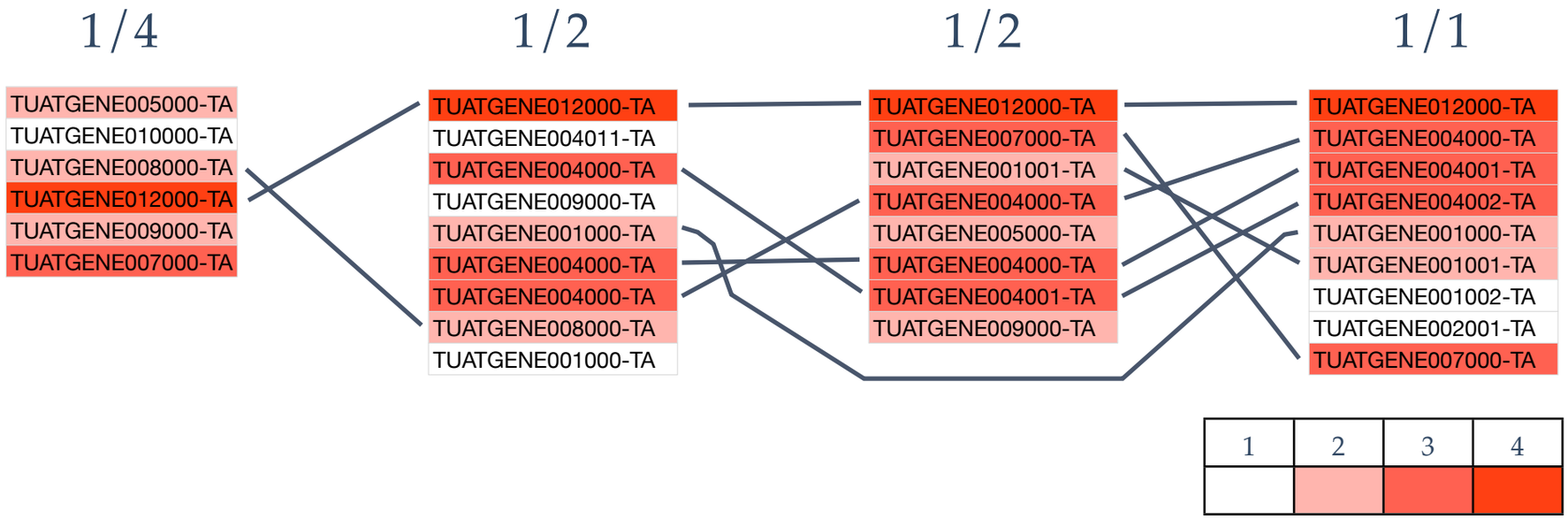
1/2データ：10,285と11,616

1/4データ：9,664と11,020

結果２：遺伝子発現差解析による評価結果

- FDR<0.05 の遺伝子を計数
  - 1/4のデータ量による解析: 6 genes ↑, 34 genes ↓
  - 1/2のデータ量による解析(1): 9 genes ↑, 42 genes ↓
  - 1/2のデータ量による解析(2): 8 genes ↑, 38 genes ↓
  - 1/1のデータ量による解析: 9 genes ↑, 44 genes ↓
- 1/2のデータ量でも発現差の見られた遺伝子数では十分に解析できていると考えられた。

発現が増えた遺伝子は、各データ量で一致していた。



結論

- 次世代シーケンサーの解析コストを下げるために、得られたデータの１/２量、１/４量のデータを用いて配列解析のシミュレーションを行い、少ないデータ量で、データ解析が可能かどうかを評価した。
- １検体について、1/2レーン分（従来の２分の１量）の配列解析データからでも、1レーン分のデータから得られた情報とほぼ同等のデータが得られた。
- このことから、半分のデータ量でも十分に信頼できる発現定量解析を行えると考え、この条件で以後のデータ解析を行うこととした。
- このように、S-PLUSを用いてデータ解析のシミュレーションを行うことによりデータ解析の最適化を行えることが示された。