次世代シーケンサーの運営について

次世代シーケンサーの運営について

○鈴木智大[A)], 道羅英夫[B)]
A) 静岡大学技術部静岡分室教育研究支援部門, B) 静岡大学グリーン科学技術研究所

１．はじめに

塩基配列を決定することは、生物学的研究に必要不可欠なことである。1977 年のサンガー法の開発[1)]は、科学者に多くの遺伝的情報をもたらし、広く世界に普及したが、この技術はスループット性や解析スピードなど多くの制限があった。近年開発された次世代シーケンサーは全く異なったシーケンステクノロジーを利用することで、これら問題を克服し、ゲノムやトランスクリプトーム、エピゲノム解析などにおいて画期的な発見を生み出している。しかし、次世代シークエンサーが抱える重大な問題として、得られた膨大なオミックスデータを如何に正確に解析するかといった点が挙げられる。また、次世代シーケンサーの網羅的解析は、これまでの手法では得られなかった、低発現遺伝子の発現解析などから、検体間で発現差の見られた大量の遺伝子を産出する。しかし得られた候補遺伝子から、信頼性の高い遺伝子を選出し・絞り込む方法を確立することが非常に重要な課題となる。

図 1.　MiSeq 装置

平成 25 年度より静岡大学グリーン科学技術研究所研究支援室では、共同利用研究設備として新たに illumina 社のデスクトップ型次世代シーケンサーMiSeq（図 1）を導入し、順調にその運用を行っている。本発表では特にゲノム及びトランスクリプトームの解析に焦点を当てて静岡大学でのその運営方法、大規模データの解析パイプライン作成等について紹介する。

２．ゲノム・RNA の抽出

良好な解析結果を得るためには、高品質のゲノム・RNA を抽出する事が非常に重要である。静岡大学ではゲノム・RNA の抽出に関しては依頼者本人に行ってもらい、品質評価を厳密に行うことで解析の精度を上げることとした。一般的な方法ではあるが、ゲノムの場合は 1）OD 測定、2）電気泳動による RNA 混入の有無の確認、3）蛍光定量を行い、OD 測定と蛍光定量での濃度に極端な乖離がないことを確認した。また、RNA の場合は 1）OD 測定、2）電気泳動によるゲノム混入の有無の確認、3）バイオアナライザを用いて RNA の分解の程度の確認を行っている。特に RNA の品質に関しては、実験作業の時間や試料の保存状態等による影響を受けやすい。また、実験手法による違いで RNA 解析のデータに影響がでるという報告[2)]もあり、如何に品質の高い RNA を精製できるかが重要なステップとなる。

３．ライブラリの作成

ゲノム及び RNA 解析のライブラリ作成の手順の概略を図 2、3 に示す。ライブラリ作成の鍵となる手順の一つとして、断片化が挙げられる。MiSeq の解析では最長でライブラリの両末端 300 bp の塩基配列が解読できるため、両端のデータを最大限活かすためには 500～550 bp 程度の鎖長のライブラリ作成が必要とされる。ゲノム解析に関しては超音波エネルギーアコースティックソルビライザー（Covaris 社）を用いた

断片化で切断が可能である。しかしRNA解析に関してはビーズ上で、2価の陽イオンを用いた断片化を行う。RNAの断片化は150 bp程度の鎖長のライブラリしか得られないが、今回反応時間の検討など様々な条件件等を行い比較的長鎖のライブラリの作成に成功した。今後は逆転写反応を先に行なった後、Covarisを用いたcDNAの断片化を行うなどの検討を行う予定である。基本的にライブラリ作成のステップに関しては試薬のマニュアルに沿って行うこととなる。現行のマニュアルは以前よりかなり簡略化されてきたとはいえ、まだ実験ステップが複雑でないとは言い切れない。作業者の違いによるビーズ混入の恐れや、出来上がるライブラリの量の違いによって混合して解析を行うマルチプレックス解析に影響が出る可能性もあるため、ライブラリ作成に関しては、研究支援室で行なう事とした。

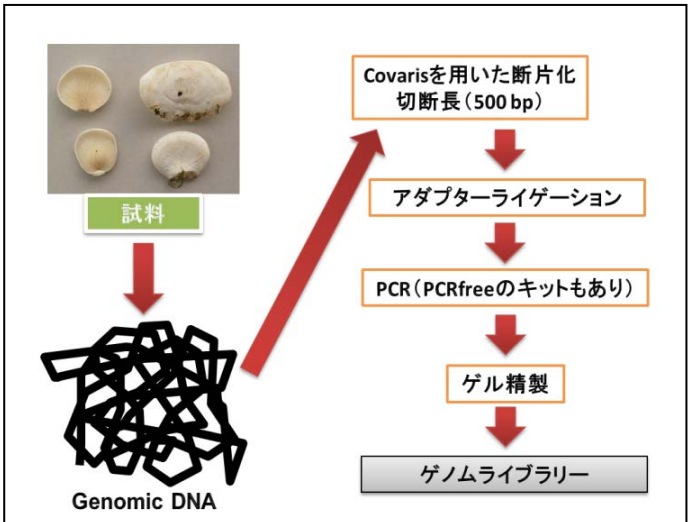

図2. ゲノムライブラリの作成方法

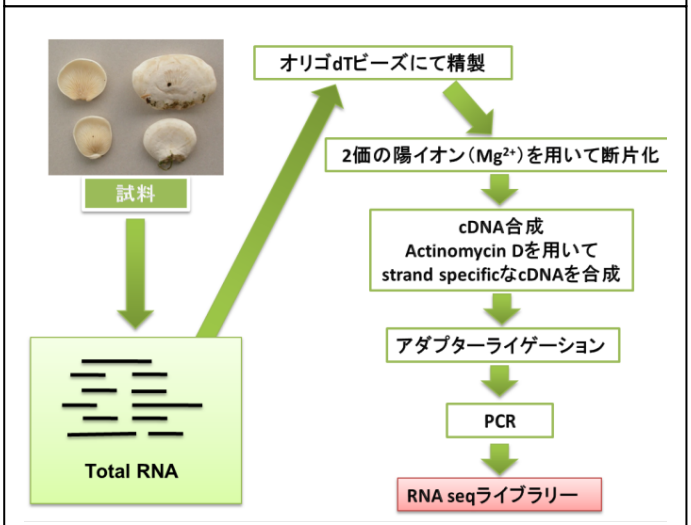

図3. RNAライブラリの作成方法

4．ライブラリの定量

静岡大学では特にゲノム解析の場合など、マルチプレックス解析を行う計画が非常に多い。マルチプレックス解析とは、ライブラリ作成時に異なるインデックス配列をつけてサンプル調製を行うことで、シーケンスの後に各サンプルを個別のデータとして出力する方法である。本方法は一度のランで多数のサンプルのシーケンスを行うことができるため、サンプル当たりに必要なランニングコストを下げることなど、多くの利点を有する。しかしマルチプレックス解析を行うにあたり必要なのは、どのくらいデータが必要なのか（推定のゲノムサイズなど）を綿密に計画しサンプルの混合比を決定しておくことや、作成したライブラリの濃度を正確に定量することが必須である。また定量を正確に行えば、シーケンスの際に作成されるクラスター数も正確にコントロールすることが可能である。クラスター数は、多すぎるとシーケンスのクオリティーを低くし、少なすぎれば出力されるデータも少なくなることから、定量の重要性が伺える。

我々はライブラリの定量にはリアルタイムPCR法を使用し、試薬はNGSライブラリ定量キット（KAPA biosystems社）を用いて定量した。リアルタイムPCRの定量値からライブラリのシーケンスに供す量を決定し、シーケンスで得られたクラスター濃度を確認したところ、正確な定量値が算出されていることを確認した。

5．データ解析

膨大なデータを一度に解析する必要があるため、以下の解析サーバーを運用した。CPUはIntel(R) Xeon (2.13 GHz/6 core)、メモリは72 GB（4 GB×18本）、ストレージは12 TB、OSはRed Hat Enterprise Linux Serverを用いた。一般的なデータ解析の概要図4に示す。全てのサンプルにおいてまず得られたReadからクオリ

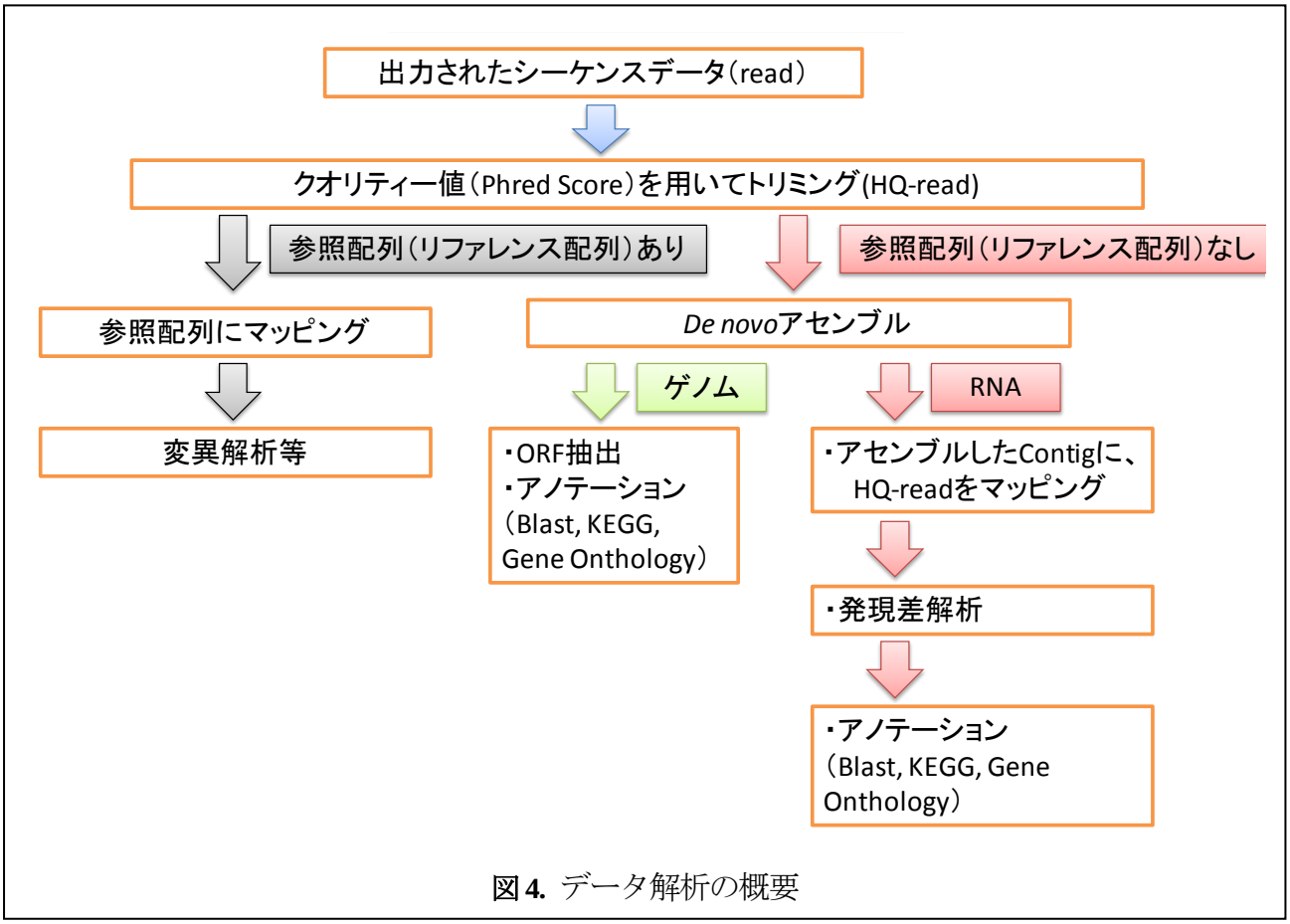

図 4. データ解析の概要

ティー値（Phred Score）の高い配列（HQ-read）を抽出する工程が必要である。我々はソフトウェア FastX tool kit 等を使用して、データのトリミングを行なった。次に必要な工程としては参照配列（リファレンス配列）がデータベース上に存在するか否かであろう。ごく近縁の参照配列がデータベース上に存在するのであれば、参照配列を基に HQ-Read をマッピングして（使用ソフトウェアは BWA[3], Bowtie[4]等）、変異解析等を行うことが可能である。

参照配列の存在しないサンプルの場合、HQ-read からまずシーケンスデータのアセンブル（配列データを貼り合わせて長い Contig 配列データを作成する事）の工程が必要である。アセンブルには k-mer （de novo アセンブリでは全リードを特定の長さのサブシーケンスである k-mer に分解する）やインサート長を指定するなど様々な条件検討を行い、最適化することが必要となる。ゲノム解析の場合アセンブルには Velvet[5], Newbler (Roche), Abyss[6]等が一般的に使用され、RNA 解析の場合 Oases[7], Trinity[8], trans Abyss[9]などのソフトがアセンブルに使用される。またアセンブル後は、ゲノム解析の場合 orf（タンパク質に翻訳される可能性がある）抽出を行った後、BLAST、KEGG 等でのタンパク質機能解析を行うこととなる。RNA 解析の場合はアセンブルした Contig に各サンプルの HQ-read をマッピングしてマップされた Read の数をカウントすることで各遺伝子の発現差解析を行うことも可能である。

６．終わりに

近年、次世代シーケンサーが普及してきたといってもデータベース上に存在しない生物種は未だ数多く残されている。MiSeq から得られるロングリードの情報を有効に利用し、貴重な遺伝的情報を取得できるような場を提供していくことは、今後も必要不可欠である。静岡大学研究支援室では次世代シーケンサーによる大規模データ解析の経験が無い依頼者でも理解しやすいように、重要な情報のみを抽出して視覚的・直感的にもわかりやすいデータを提供するなど、研究サポートを徹底して行きたいと考えている。

参考文献

[1] Sanger, F., et al.：J. Mol. Biol., 94, 441-446. (1975)

[2] McIntyre, L.M., et al.：BMC Genomics, 12, 293 (2011)

[3] Li H., et al.：Bioinformatics, 25:1754-60. (2009)

[4] Langmead, B., et al.：Nat. Methods 9, 357–359 (2012)

[5] Zerbino, DR., et al.：Genome Res.,18: 821-829 (2008)

[6] Simptson JT., et al.：Genome Res., 19,1117-1123 (2009)

[7] Schulz MH., et al.：Bioinformatics, 28, 1086–1092 (2012)

[8] Grabherr, MG., et al.：Nat. Biotechnol., 29, 644-652 (2011)

[9] Birol, I., et al.：Bioinformatics, 21, 2872–2877 (2009).