

農学系領域におけるゲノム科学教育

石井 一夫



超並列型高速シーケンサー、いわゆる、次世代シーケンサーの実用化と普及に伴い、研究室レベルでのゲノム解析や、パーソナルゲノム解析が身近に語られるようになってきている。体内の細菌叢や環境中のバイオフィームなどの微生物群のゲノム解析を一度に網羅的に行うメタゲノムや、多数のがん患者検体のゲノムリシークエンスにより遺伝子多型プロフィールを網羅的に作成したりすることは、一昔前までは夢のような研究であったが、そのような研究もごく普通の研究室や病院でも可能ではないかと思われるまでになってきた。また、農学系領域では、イネやカイコなどで多種の品種の蓄積があり、これらの多型解析を育種に応用することも期待される。

米国でのゲノム解析

私自身は、次世代シーケンサーを用いてがん細胞における発現定量解析RNA-Seqおよび遺伝子発現を制御する制御タンパク質の結合様式パターンを調べるChIP-Seqなどを行ってきたが、始めた当時はそのような解析が行える環境は国内にはなく、2009年に米国のノースウエスタン大学で職を得て解析を開始した。学内の同部署内を見回すと、中国人とインド人ばかりが目についたのが印象的だ。ChIP-Seqのピーク検出ソフトの開発者も中国人が多く、ミシガン大学やハーバード大学などの中国人バイオインフォマティクソンに連絡をとり、いろいろ情報を得ながら解析を進めたものである。学内の情報システム管理者はDevidと名乗っていたが、会ってみると彼も中国人であった。最終的に職場で日本人職員に会うことはなく、職場の日本人は私一人であった。このような現状だったので、日本国内におけるゲノム科学教育が世界的にもかなり立ち遅れた状況は容易に想像できたが、そのことを心配したり、憂慮したりする余裕はなかったというのが正直な感想だった。

農学系領域でのゲノム解析

縁あって、2011年7月に東京農工大学に職を得てゲノム科学の情報解析を中心に教育を担当する機会を得た。はじめて見ると、農学系領域のゲノム解析は、ヒトやマウスなどのゲノム解析とはかなり様相が異なっていた。それは、イネやシロイヌナズナ、ショウジョウバエなど一部のモデル動物を除き、参照配列がない状態で

データ解析を行う必要があることが多いということである。de novoアセンブリーとRNAの発現定量解析を組み合わせた、いわゆるde novo RNA-Seqや、de novoの配列解析データのアセンブリーによる多型解析などである。RNA-SeqやChIP-Seqで用いられる多くの解析ソフトは、参照配列がありそれにマッピングしていくということが前提になっているので、農学系領域のデータ解析に用いる場合には特別な工夫が必要になる。農学系領域のゲノムデータ解析の教育という点では、さらに別の問題がある。それは、特に農学部という学部のカリキュラムから、情報科学やプログラミング、統計学などの教育をほとんど受けていないという点である。近年のゲノム科学や臨床薬学などの発展から、このような学問の重要性は高まっているが、実際にはこのような教育はほとんどなされておらず、コンピュータリテラシーが不足している状況の中で学生とともに解析を進めていく。

データ解析は主にLinuxをベースにしたフリーソフトウェアが中心である。特に、Velvet, SOAP, TopHatなどの次世代シーケンサーに特化したフリーソフトウェア、awkやPerl, Python, Rubyを中心としたテキスト処理、そしてRによる統計解析などが中心であり、農学系の学生にはしきいの高いものも多い。学生にデータ解析に慣れてもらうためにMacBook Airを配布して端末からのコマンド環境を用いたデータ解析演習を行うなど、導入ができるだけスムーズにいくような工夫を行っている。私自身は、気持ちだけはハッカーなので、データ解析はチャレンジングであればあるほど面白いと感じている。



東京農工大学遺伝子実験施設で使用している次世代シーケンサー（イルミナ社GAIIx）