論文の内容の要旨

論文題目　　Simulation for long read sequencers
（ロングリードシーケンサーのシミュレーション）

氏　　名　　小野　幸輝

近年、ロングリードシーケンサー（PacBio、Nanopore）を使用した、ゲノム、エピゲノム、トランスクリプトームなどに関する研究が盛んに行われている。ロングリードには、ショートリード（Illumina HiSeq など）よりも多くのエラーが含まれることが知られており、ロングリードを対象とする多くのツールとアルゴリズムが開発されている。実データを使用してツール/アルゴリズムを評価することは一般に困難であるため、様々な条件のリードが生成可能なシミュレーターは非常に有用である。シミュレーターはまた、シーケンス実験のデザインを行う際にも有用である。例えば、ゲノムアセンブリや変異検出に適したシーケンスカバレッジの推定に利用できる。

ロングリードの特徴を理解するために、そして、ロングリードを解析するツール/アルゴリズムの評価や、シーケンス実験のデザインに有用なシミュレーションデータを生成するために、私は 2 つのロングリードシミュレーターを開発した。1 つ目のシミュレーターPBSIM では、13 個の PacBio データセットを統計解析し、得られた特徴（リード長は対数正規分布に従う、エラーはランダム性が高い、など）を正確に模倣したロングリードの生成を、モデルベースの方法と、サンプリングベースの方法で実現した。シーケンス実験のデザインにおける有用性を示すために、PBSIM によって網羅的な条件のシミュレーションデータを生成し、PacBio リード（CLR および CCS）のハイブリッドゲノムアセンブリのテストを行った。その結果、精度の高いハイブリッドゲノムアセンブリを行うためには、CLR は少なくとも 15、CCS は少なくとも 30 のシーケンスカバレッジが必要であることを推定することができた（図 1 ）。

ロングリードはショートリードよりもエラーが多いという欠点を持つが、この欠点はエラーコレクション法の発達によりほぼ解決され、ロングリードの利点であるリードの長さを生かすことが可能になった。加えて、ロングリードはシステマティック（あるいはコンテキスト特異的）なエラーがショートリードよりも少ないという利点もあり、ゲノムアセンブリなどでのエラーバイアスの悪影響を小さく抑えることができる。しかしながら、Nanopore リードにおいては、ホモポリマー配列でエラーが発生しやすいことが報告されている。また、PacBio リードにおいては、リード内のエラー分布に領域的な偏りがあることが報告されており、非常に低品質の領域が観察されている。低品質領域は、キメラと検出漏れのアダプター配列、およびエラーの不均一性が主な原因である。ロングリードのエラーの特徴をより正確に、特にエラーの不均一性をシミュレートするために、最新のモデル選択基準を組み合わせた隠れマルコフモデル

（FIC-HMM）を使用して、quality score の生成モデルを構築し、この生成モデルを実装したロングリードシミュレーターPBSIM2 を開発した。PacBio リードのシミュレーションの性能評価において、PBSIM2 が既存のシミュレーターよりも正確にエラーの不均一性をシミュレートすることが示された（図２）。また、Nanopore リードのシミュレーションにおいても、高い性能を持つことが示された。加えて、PBSIM では十分にシミュレートできなかった、リード長とリード精度の相関、およびエラー率と quality score の関係についても、PBSIM2 では十分に改善することができた。

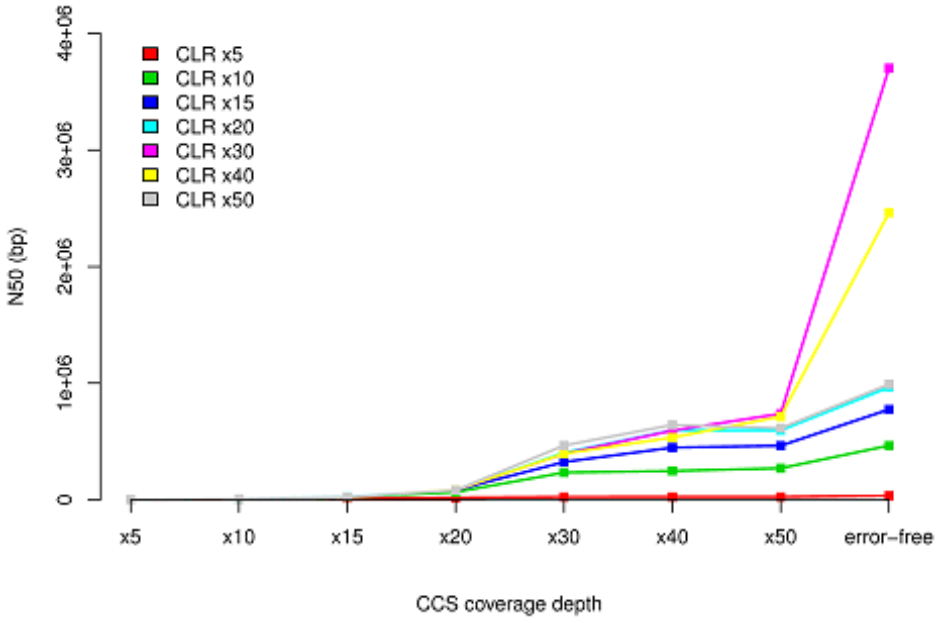

図 1　ハイブリッドアセンブリに必要なシーケンスカバレッジの推定

PBSIM のシミュレーションデータを使って、ショウジョウバエの染色体 Chr2L のハイブリッドアセンブリの評価を行った。縦軸の N50 はコンティグを長い順に並べて上から順に足した時に、全体の長さの半分に達した時のコンティグの長さ。長いほどよいと評価される指標である。横軸は CCS のカバレッジを、グラフの色は CLR のカバレッジを示している。横軸の"error-free"は、CCS による CLR のエラー訂正において、強制的に CLR のエラーをゼロにした場合を表している。このグラフから、精度の高いハイブリッドゲノムアセンブリを行うためには、CLR は少なくとも 15、CCS は少なくとも 30 のシーケンスカバレッジが必要であることが推定される。

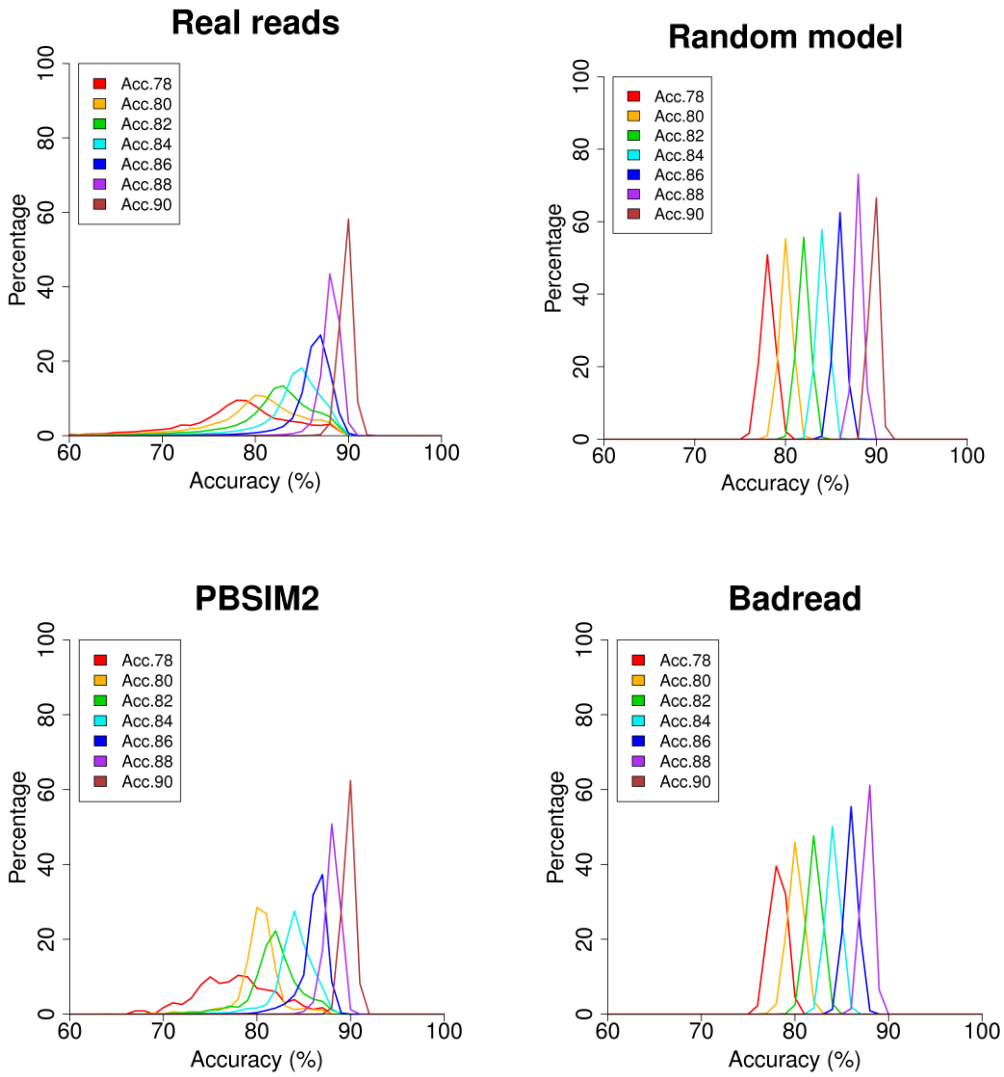

図２　ロングリードにおけるエラーの不均一性

線虫の PacBio リードにおけるエラーの不均一性を可視化した。まず、ロングリードを精度別のグループに分ける。次に、各ロングリードを 800nt のインターバルに分割し、各インターバルの精度を計算する。各グラフはインターバルの精度の分布を表しており、グラフの色は精度別のグループを示している。"Real reads"は実データから作成した分布であり、 "Random Model"はエラーのランダム性を前提にして作成した分布である。この２つの比較から実データには明らかにエラーの不均一性があることが分かる。 "PBSIM2"と”Badread”はそれぞれ PBSIM2 と Badread（https://github.com/rrwick/Badread　既存のロングリードシミュレーターの１つ）のシミュレーションデータから作成した分布である。PBSIM2 が実データの分布を正確に再現していることが分かる。