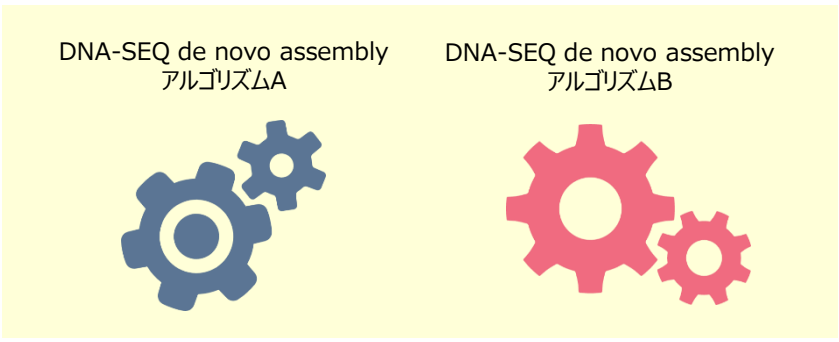
DNA-SEQ de novo assembly の品質を評価する方法

遺伝子変異研究では、以前に研究された種のサンプルを分析することがよくあります。例えば、同じ種の様々な品種、系統、または個体群のゲノムを調べることは興味深いことです。このような場合、リファレンスゲノムが利用可能であっても、配列決定された生物のゲノムを得るためにDNA-SEQ de novo assemblyを実行する必要があるケースがあります。解析の正確性と信頼性を確保するために、データセットによって異なる解析アプローチが必要となる場合があるため、様々な設定を変えたり、異なるアセンブリアルゴリズムを採用したりすることが推奨されます。

この点において、異なるde novo assemblyを簡単に比較できるQUASTツールが役に立ちます。その機能を説明するために、*Caenorhabditis elegans*(線虫)のWGSデータセットを使ったいくつかのケーススタディを紹介します。このデータセットは、*C. elegans* N2株のショートペアエンドおよびメイトペアリードで構成されており、QUASTの特徴を十分に検討することができます。

- SRA Study: DRP001005
- Reference Genome: WBcel235



正確なコンティグデータの作成には異なるアセンブリアルゴリズムや様々な設定値で評価する必要がある

▶次ページで異なるアルゴリズムでの比較ついてご紹介

biosupport@filgen.jp https://filgen.jp/ 052-624-4388 052-624-4389 フィルジェン株式会社

異なるアセンブリ アルゴリズムの比較: ABySS と SPAdes

アセンブラーによって、前提条件や手順が異なる様々なアルゴリズムを採用しています。そのため、使用するソフトウェアによって、結果として得られるアセンブリが異なります。時には、すべてのシナリオで最適なパフォーマンスを発揮する単一のアルゴリズムは存在せず、入力データによって有効性が異なる可能性もあります。ここでは、*C. elegans*データセットを用いて、ABySSとSPAdesの性能を評価します。そのために、4つの異なるアセンブリを作成しました：

- ペアエンドリードでの ABySS の使用
- ペアエンドおよびメイトペアリードでの ABySS の使用
- ペアエンドリードでの SPAdes の使用
- ペアエンドおよびメイトペアリードでの SPAdes の使用

QUASTの結果

NGxプロットは、コンティグのサイズに関する情報を提供します。例えば、NG50が10bpの場合、リファレンスゲノム全体の50%が10bpより長いコンティグでカバーされていることを示します。一般的に使用されるNx統計は、ゲノムサイズではなく、アセンブリサイズを採用しています。しかし、アセンブリごとに長さが異なるため、アセンブリ間でNxを直接比較することはできません。したがって、NGx は比較目的に適しています。

一般に、コンティグは大きい方が実際の染色体やゲノムの大きさをよりよく表しているので望ましいとされています。NGxプロットを見ると、ABySSを使って生成されたアセンブリには、より大きなコンティグが含まれていることが一目瞭然です（Figure 1）。特に、ABySSとメイトペアの結果の組み合わせは、顕著に良い結果をもたらしました。このことは、最大コンティグ長とアセンブリ全体の長さからも、さらに裏付けられます（Table1）。

アセンブリの品質を評価するためのもう一つの有用な指標は、アセンブリによってカバーされたゲノム塩基の割合を示す「Genome Fraction」です。これは、アセンブリをリファレンスゲノムにアライメントすることで測定されます。Figure 2に示すように、ABySSを用いて生成されたアセンブリは、リファレンスゲノムのより広い部分をカバーしています。これは、アセンブリのサイズが大きければ大きいほど、より多くのゲノムをカバーできる可能性があるため、アセンブリの総サイズ（Table 1）と一致します。

最後に、検出されたミスアセンブリの数から、アセンブリの精度に関する洞察を得ることができます。アセンブリのこれらの領域は、リファレンスゲノムと一致しません。このデータセットでは、ABySSとメイトペアリードを用いて作成したアセンブリにおいて、ミスアセンブリの数が比較的少なくなっています（Figure 3）。

結論として、さまざまな指標を考慮した結果、ABySSとメイトペアリードを使用して生成されたアセンブリは、より完全で正確であると推察されます。

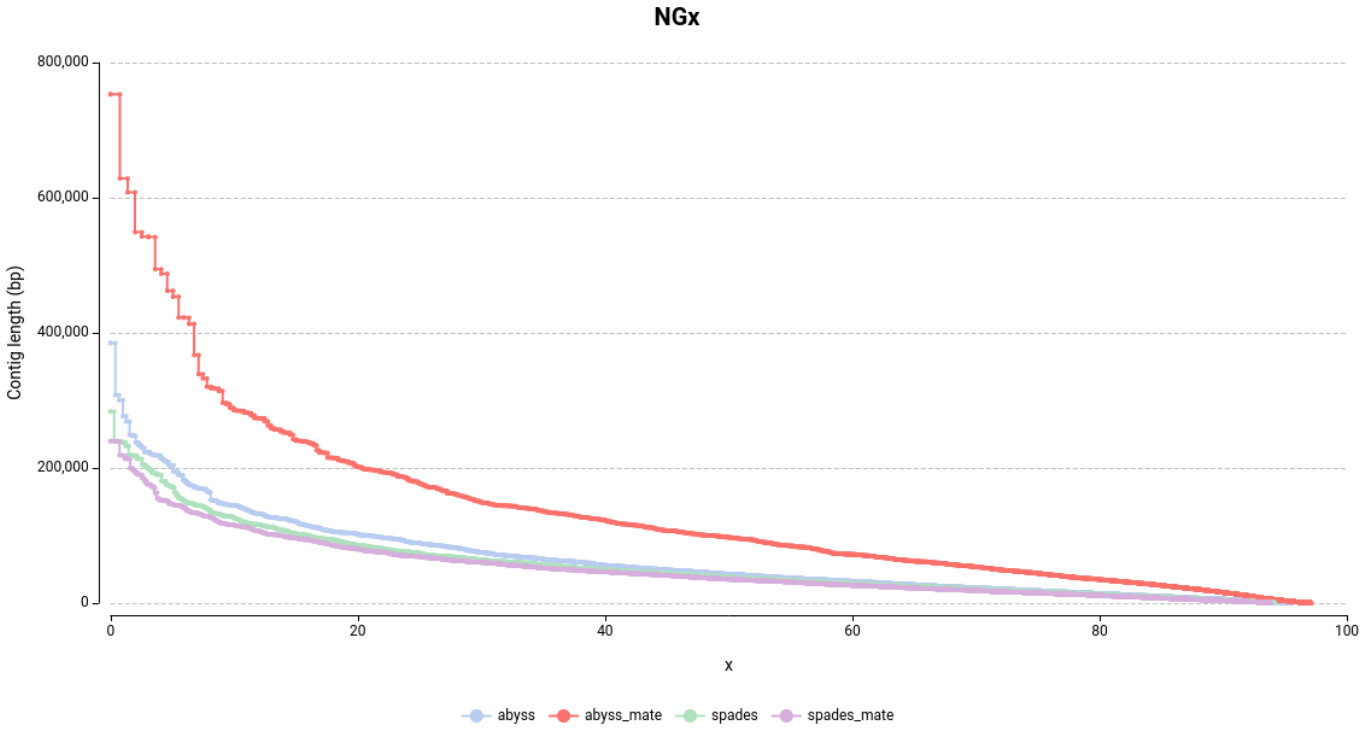


Figure 1. ABySS と SPAdes で作成されたアセンブリの NGx プロット
ABySS(青と赤)を使って生成されたアセンブリには、より大きなコンティグが含まれていることが読み取れる

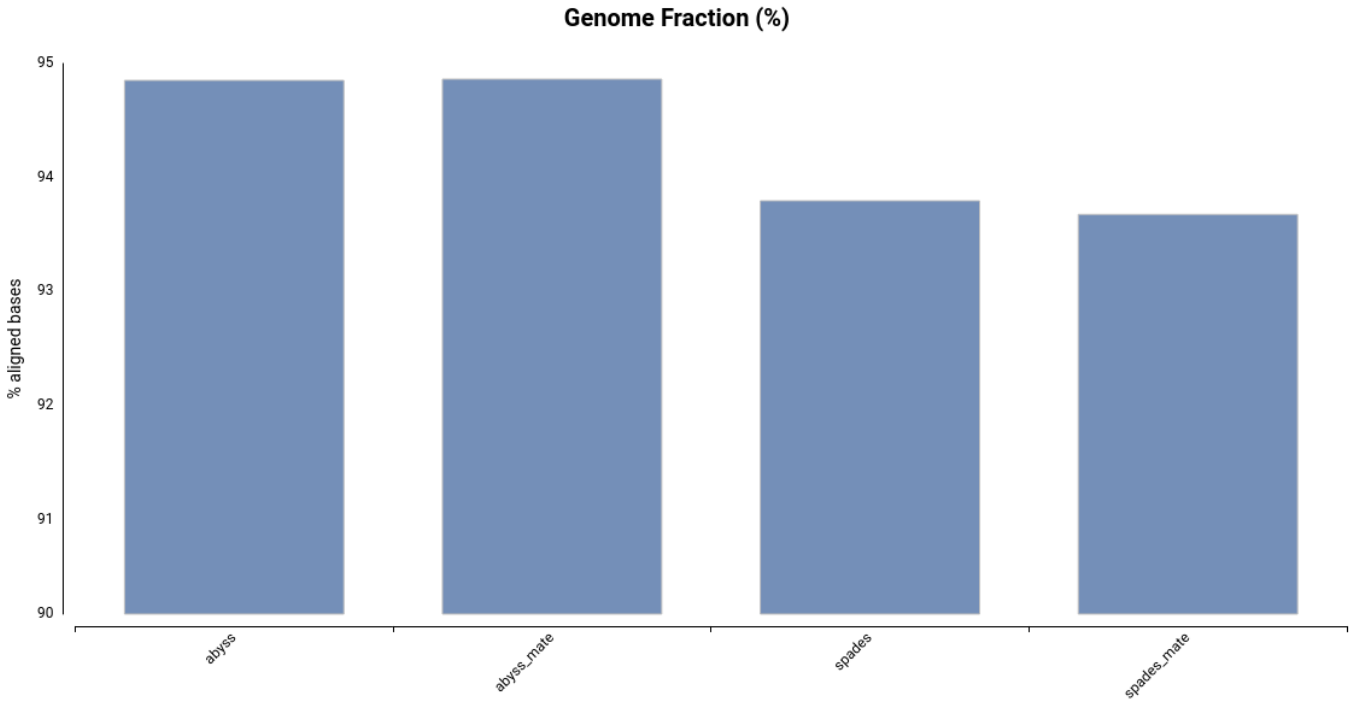


Figure 2.各アセンブリのGenome Fraction
ABySSを用いて生成されたアセンブリ(左2つ)は、リファレンスゲノムのより広い部分をカバーしていることが読み取れる

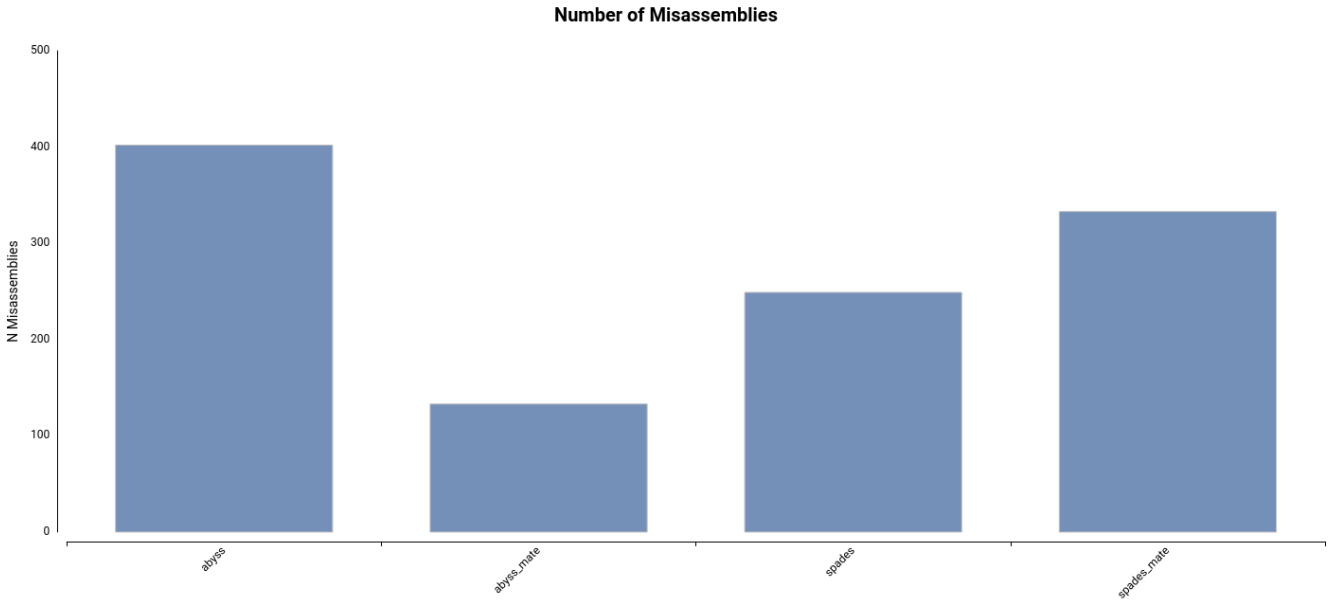


Figure 3.それぞれで検出されたミスアセンブリの数
ABySSとメイトペアリードを用いて作成したアセンブリにおいて、（左から2番目）
ミスアセンブリの数が比較的少なくなっていることが読み取れる

Statistic	Abyss	Abyss_mate	Spades	Spades_unpaired
Num. contigs (>= 0 bp)	70674	68600	61459	63280
Num. contigs (>= 10000 bp)	2229	1393	2328	2377
Num. contigs (>= 50000 bp)	490	601	458	437
Num. contigs	5674	3654	5668	6112
Largest contig	384836	753101	283380	239865
Total length	95813155	97384272	94415625	239865
Reference length	100286401	100286401	100286401	100286401
auNG	62763.7	135616.3	54011.7	49442.1

Table 1. 異なるアルゴリズムで実行されたアセンブリのそれぞれについて、QUASTで得られた要約統計量

▶次ページでメイトペアデータまたはunpairedリードの追加がアセンブリにどのような影響を与えるかを調査

異なるメイトペアまたはunpairedリード をアセンブリに追加する

使用する入力データの質と量の両方が、de novo assemblyに影響を与える可能性があります。アセンブルアルゴリズムは、リファレンスゲノムに頼らず、入力されたリードのみを用いてゲノムを再構築することを目的としており、その結果、アセンブルは入力データに大きく依存することになります。一般的に、入力データが多いほど良いアセンブリができますが、提供されたデータに誤りがある場合は、必ずしもそうではありません。

このセクションでは、メイトペアデータまたはunpairedリードの追加がアセンブリにどのような影響を与えるかを調査します。unpairedリードは、FASTQ Preprocessingステップで対応するペアを失い、シングルエンドリードとなったリードです。科学者の中には、unpairedリードは短く、品質が低い可能性があるため、アセンブリにノイズが入る可能性があると指摘する人もいます。しかし、それを使わないことで貴重な情報が失われるという意見もあります。

Input Data

これを評価するために、ABySSを用いて*C.elegans*のWGSからなるデータセットを組み立てました。以下の4つの異なる入力データを用いて行われました：

- ショートペアエンドリード
- ショートペアエンドリード＋メイトペアリード
- ショートペアエンドリード+ショートunpaired (single)リード
- ショートペアエンドリード+ショートunpaired (single)リード+メイトペアリード

OmicsBoxで利用可能なQUASTを用いて、アセンブリの品質を評価・比較しました。

QUASTの結果

Figure 4を見ると、NGx プロットの線が Y 軸上で高い位置にあるため、メイトペアデータを使用して生成されたアセンブリがより大きなアセンブリを生成することがわかります。これをすばやく評価するもう 1 つの方法は、NGxカーブの下の面積であるauNG (Table 2) を調べることです。したがって、auNGが大きいほど、NGxの値も大きくなります。

このことは、サマリーレポート（Table 2）に掲載されたいくつかの統計データからも裏付けられます。例えば、大きなコンティグ（長さ>=50k bp）の数、最大コンティグ長、アセンブリの全長は、すべてメイトペアデータで生成されたアセンブリで高くなっています。

さらに、検出されたミスアセンブリの数は、メイトペアリードで生成されたアセンブリでは大幅に少なくなっています(Figure 5)。

全体として、メイトペアリードを使用する場合と使用しない場合で大きな違いがあります。これらのリードは特にアセンブリを強化するために使用されるため、これは論理的です。メイトペアリードは長距離情報を提供し、ショートリードのシーケンスデータから生成されたコンティグをスキャフォールディングにしてつなげたり、アセンブリプロセスにおけるあいまいさやエラーを解決したりするのに役立てることができます。

unpairedリードを追加するかしないかの違いについては、この例ではより大きなコンティグを構築するのに役立つことが観察されます（Table 2, Figure 4）。

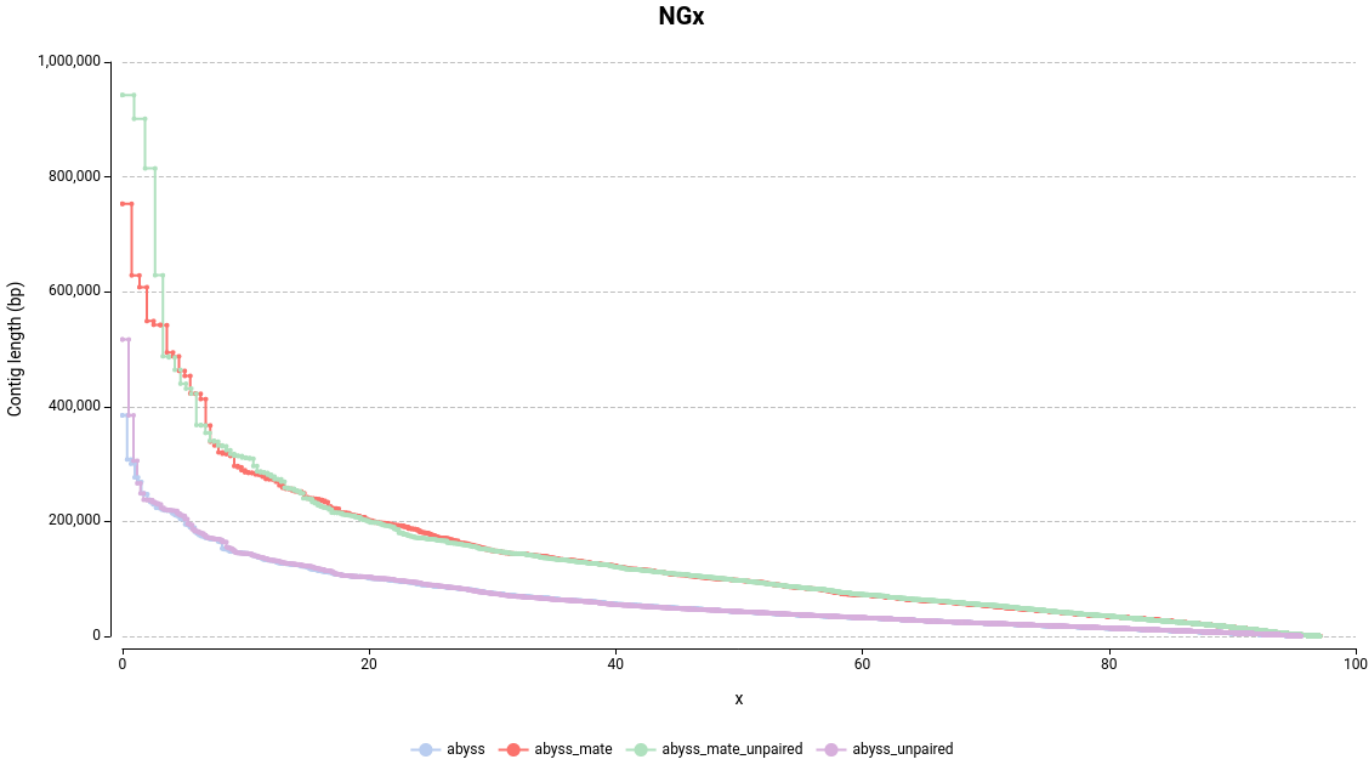


Figure 4.異なる入力データで取得されたアセンブリの NGx プロット
メイトペアデータのアセンブリ（赤と緑）がより大きなアセンブリを生成することが読み取れる
またunpairedリードの有無（紫と青/赤と緑）からこのリードの追加により大きなコンティグを構築できることが読み取れる

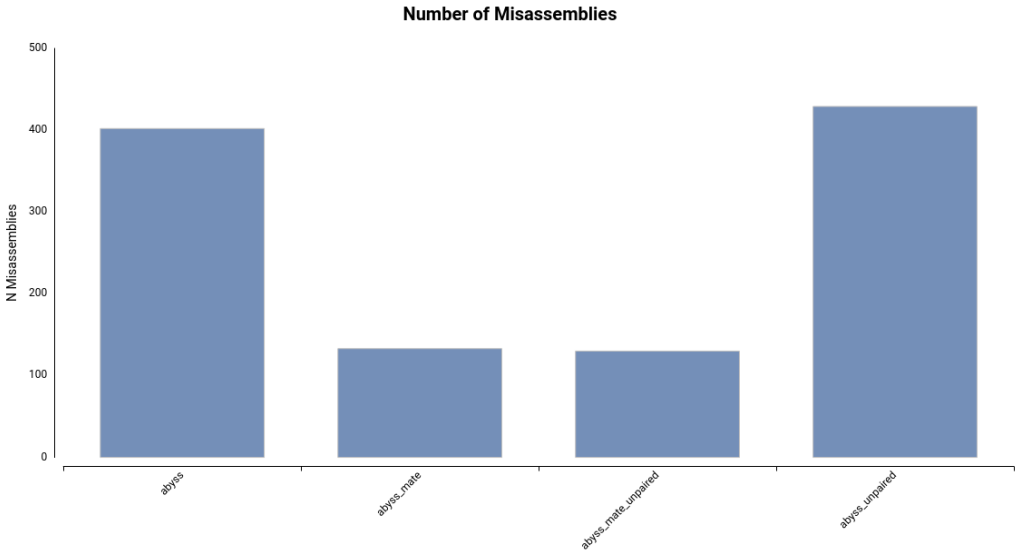


Figure 5.各アセンブリで検出されたミスアセンブリの数
メイトペアリードで生成されたアセンブリ（左から2，3番目）では大幅に少なくなることが読み取れる

Statistic	Abyss	Abyss_mate	Abyss_mate_unpaired	Abyss_unpaired
Num. contigs (>= 0 bp)	70674	68600	85717	87640
Num. contigs (>= 10000 bp)	2229	1393	1396	2235
Num. contigs (>= 50000 bp)	490	601	613	478
Num. contigs	5674	3654	3644	5532
Largest contig	384836	753101	942534	516982
Total length	95813155	97384272	97379516	95812895
Reference length	100286401	100286401	100286401	100286401
auNG	62763.7	135616.3	141949.4	64254.4

Table 2.異なる入力データセットで得られた各アセンブリについて、QUASTで得られた要約統計量

さまざまな k-mer サイズを試す

同じアルゴリズム、同じ入力データであっても、パラメータが異なれば、出来上がるアセンブリは全く異なるものになります。特に、アセンブリに大きく影響するパラメーターの 1つは、k-mer サイズです。ゲノムde novo genome assemblyにおいて、k-merはシーケンシングリードから抽出される長さ（通常20-150ヌクレオチド）の短い連続した配列です。k-merサイズは、オーバーラップを検出しコンティグを組み立てる際の感度と特異性のレベルを決定するため、アセンブリプロセスにおいて重要なパラメータです。

Input Data

*C. elegans*データセットのリードの長さは110bpです。このセクションでは、3つの異なるk-merサイズを使用した場合の性能を評価します。

- 55bpのk-merサイズ、これはリード長の50%です。
- 77bpのk-merサイズ、これはリード長の70%です。
- 25bpのk-merサイズ、これはリード長の~23%にあたります。

QUASTの結果

他のセクションと同様に、NGxプロット（Figure 6）を見ると、k-merサイズ77で得られたアセンブリは、より大きなコンティグを含んでおり、望ましいことが容易にわかります。意外なことに、デフォルトの構成（k-mer55）は、より小さなコンティグで構成されているため、最もパフォーマンスが悪いアセンブリであるようです。

しかしながら、Genome Fraction（Figure 7）を見ると、k-mer27で得られたアセンブリは、リファレンスゲノムのより少ない割合をカバーしていることがわかります。言い方を変えると、コンティグが大きくなっても、k-mer55のアセンブリと比較すると、完成度が低いです。

コンティグが大きい場合でも、k-mer55で得られたアセンブリと比較して、アセンブリはそれほど完全ではありません。また、アセンブリの総長からも明らかで（Table 3）、小さなアセンブリは、k-mer27で得られます。これは、アセンブリを評価するために、さまざまな統計を調べることがいかに重要であるかを示しています。

さらに、abyss_kmer27アセンブリでは、コンティグの総数（Num. contigs (>= 0 bp), Table 3 ）が非常に高くなっています。したがって、他の2つのアセンブリと比較して、アルゴリズムがより大きなシーケンスに結合できなかった小さなコンティグが多数あるようです。

最後に、ミスアセンブリの数が多いアセンブリは、デフォルトの設定であるk-mer55で得られたものであることが判明しました（Figure 8）。

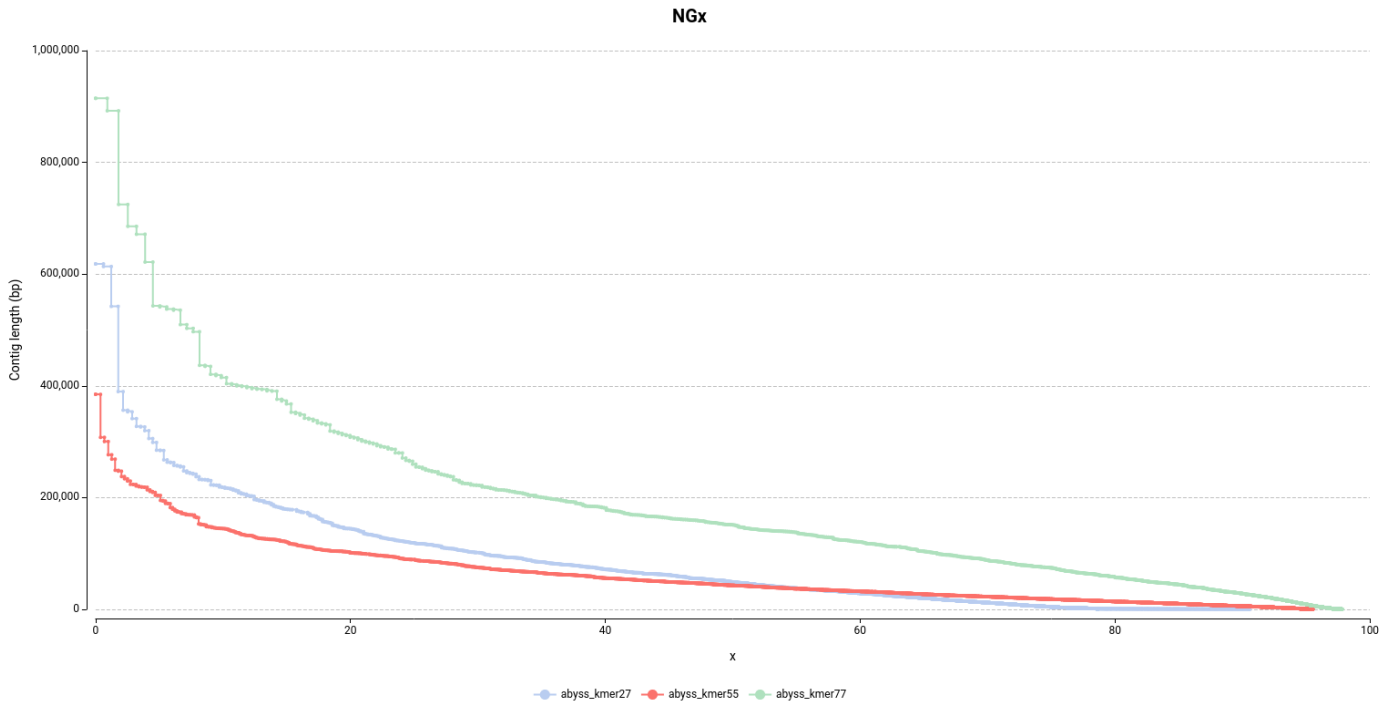


Figure 6.異なる k-merサイズで取得されたアセンブリの NGx プロット
k-mer77(緑)が大きなコンティグを含み、 k-mer55(赤)がパフォーマンスの悪いアセンブリであることが読み取れる

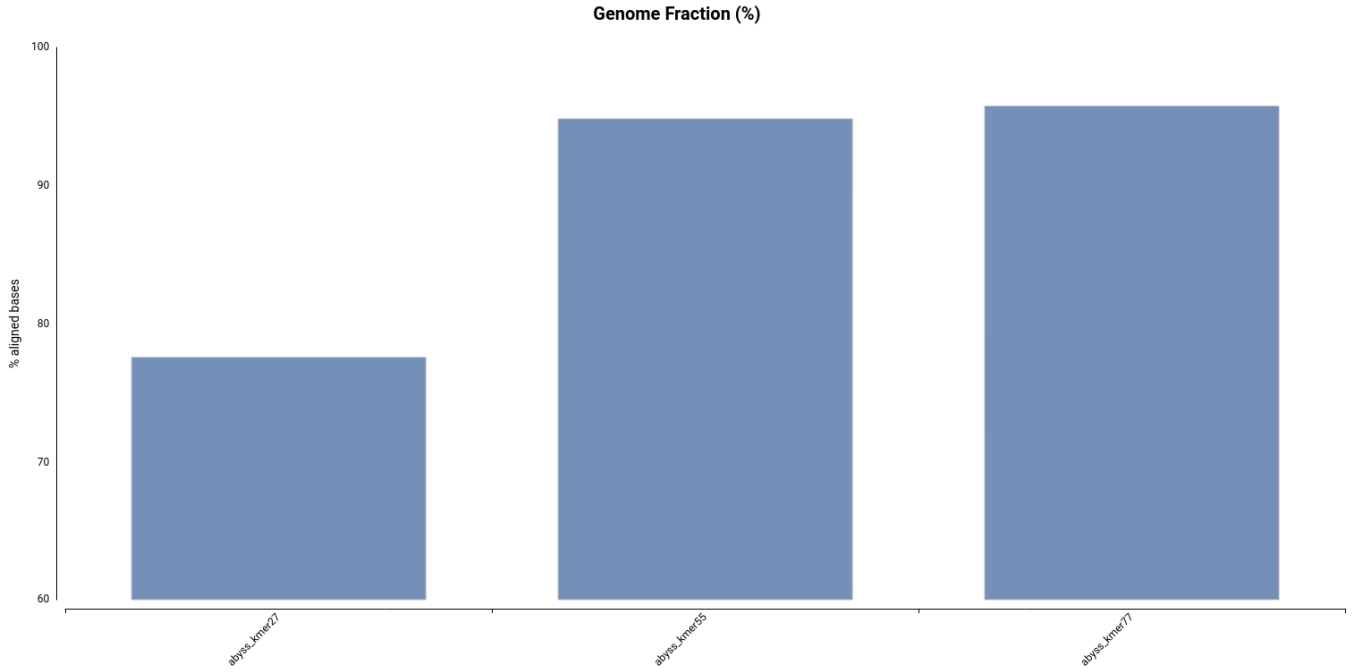


Figure 7.各アセンブリーがカバーするゲノムの割合
Fig6と異なりk-mer27(左)で得られたアセンブリはk-mer55(中央)のアセンブリと比較すると、完成度が低いことが読み取れる

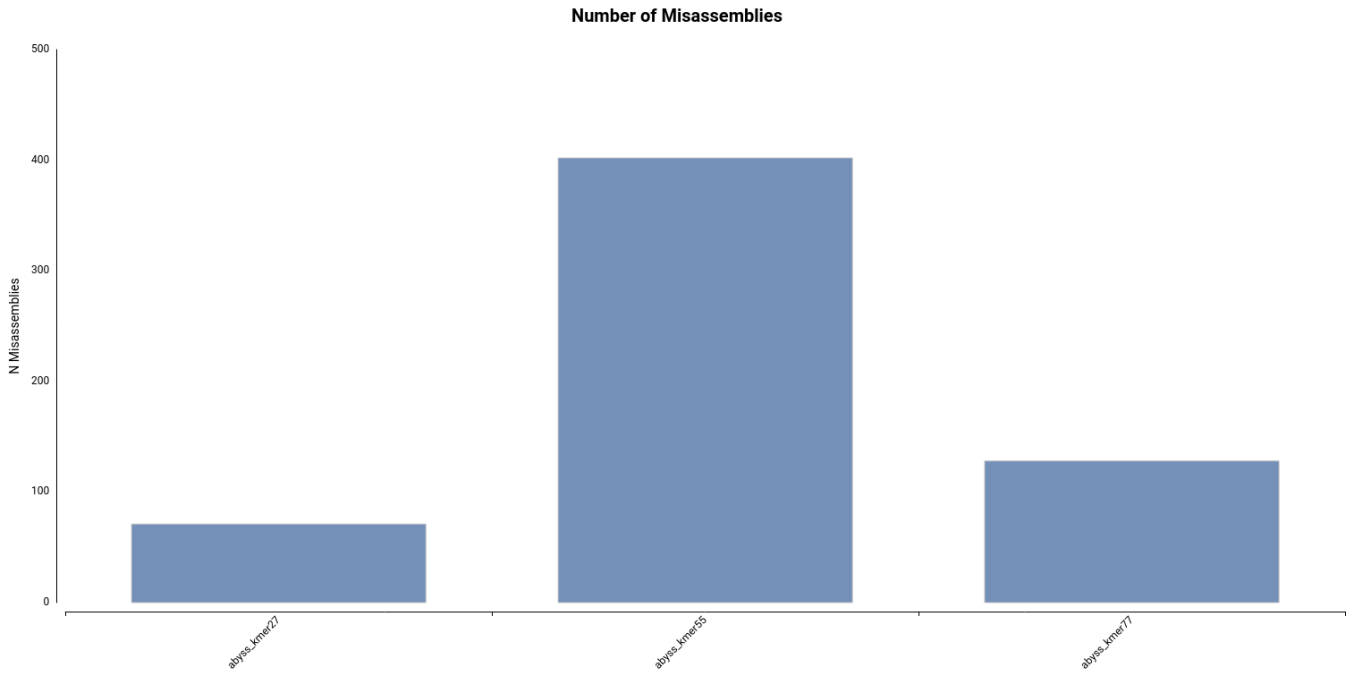


Table 8.各アセンブリで検出されたミスアセンブリの数
k-mer55(中央)が最も多いことが読み取れる

Statistic	Abyss_kmer27	Abyss_kmer255	Abyss_kmer77
Num. contigs (>= 0 bp)	479663	70674	53279
Num. contigs (>= 10000 bp)	1375	2229	1009
Num. contigs (>= 50000 bp)	452	490	553
Num. contigs	20530	5674	2641
Largest contig	618279	284836	914631
Total length	90902996	95813155	98116581
Reference length	100286401	100286401	100286401
auNG	84684.1	62763.7	197579.9

Table 3.異なるk-merサイズで得られた各アセンブリについて、QUASTによって取得された要約統計

References
Bolger, A.M., Lohse, M. and Usadel, B. (2014) “Trimmomatic: A flexible trimmer for Illumina sequence data,” *Bioinformatics*, 30(15), pp. 2114–2120. Available at: Trimmomatic: a flexible trimmer for Illumina sequence data .

Gurevich, A. *et al.* (2013) “Quast: Quality Assessment Tool for Genome Assemblies,” *Bioinformatics*, 29(8), pp. 1072–1075. Available at: QUAST: quality assessment tool for genome assemblies .

Jackman, S.D. *et al.* (2017) “Abyss 2.0: Resource-Efficient Assembly of large genomes using a bloom filter,” *Genome Research*, 27(5), pp. 768–777. Available at: ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter .

Li, H. 2020 *Aun: A new metric to measure assembly contiguity*, *Sitewide ATOM*. Available at: auN: a new metric to measure assembly contiguity (Accessed: February 27, 2023).

Prjibelski, A. *et al.* (2020) “Using Spades de Novo Assembler,” *Current Protocols in Bioinformatics*, 70(1). Available at: https://doi.org/10.1002/cpbi.102.