

細菌ゲノムのde novo assemblyと マニュアルキュレーションによる高品質ゲノムの作成

フィルジェン株式会社 バイオインフォマティクス部
(support@filgen.jp)



断片化してシーケンス

ATGCGCGTAGCGATGACGACA



GC GCGTAGC
ATGACGACA
TAGCGATGA
ATGCGCGT
CGATGACGT
CGTAGCGATG

1 ATGCGCGTAGC
2 AGCGATGACG
3 GATGACGACA

de novo assembly

完全に元の配列を復元できる場合はほとんどない
→いくつかの断片（コンティグ）に分かれるのが普通

ATGCGCGTAGCGATGACGACAATCGTGTGACGTGA

アセンブリ1

- ・コンティグ3本
- ・1コンティグあたり平均15bp

ATGCGCGTAGCGATG
TGACGACAATCGTGT
AATCGTGTGACGTGA

アセンブリ2

- ・コンティグ6本
- ・1コンティグあたり平均8.8bp

ATGCGCGT
GTAGCGAT
GATGACGACA
CAATCGTG
TCGTGTGAC
TGTGACGTGA

アセンブリ1の方がもとのゲノムに近い
→アセンブリ1の方が高品質

ATGCGCGTAGCGATGACGACAATCGTGTGACGTGA

アセンブリ1

- ・コンティグ3本
- ・1コンティグあたり平均15bp

ATGCGCGTAGCGATG
TGACGACAATCGTGT
AATCGTGTGACGTGA

アセンブリ2

- ・コンティグ6本
- ・1コンティグあたり平均8.8bp

ATGCGCGT
GTAGCGAT
GATGACGACA
CAATCGTG
TCGTGTGAC
TGTGACGTGA

コンティグの本数が少なく、各コンティグの長さが長いほど高品質
de novo アセンブリの品質指標の一つであるN50もこの条件で良くなる

ATGCGCGTAGCGATGACGACAATCGTGTGACGTGA

アセンブリ1

- ・コンティグ3本
- ・1コンティグあたり平均15bp

ATGCGCGTAGCGATG
TGACGAC AATCGTGT
AATCGTGT GACGTGA



アセンブリ1.1

- ・コンティグ2本
- ・1コンティグあたり平均**18.5bp**

ATGCGCGTAGCGATG
TGACGACAATCGTGTGACGTGA

コンティグを適切にアライメントして2つのコンティグをマージできれば、
コンティグの本数は減り、コンティグの平均長さを増やすことができる。
→より高品質になる

ATGCGCGTAGCGATGACGACAATCGTGTGACGTGA

アセンブリ1

- ・コンティグ3本
- ・1コンティグあたり平均15bp

ATGCGCGTAGCGATG
TGACGAC AATCGTGT
AATCGTGTGACGTGA

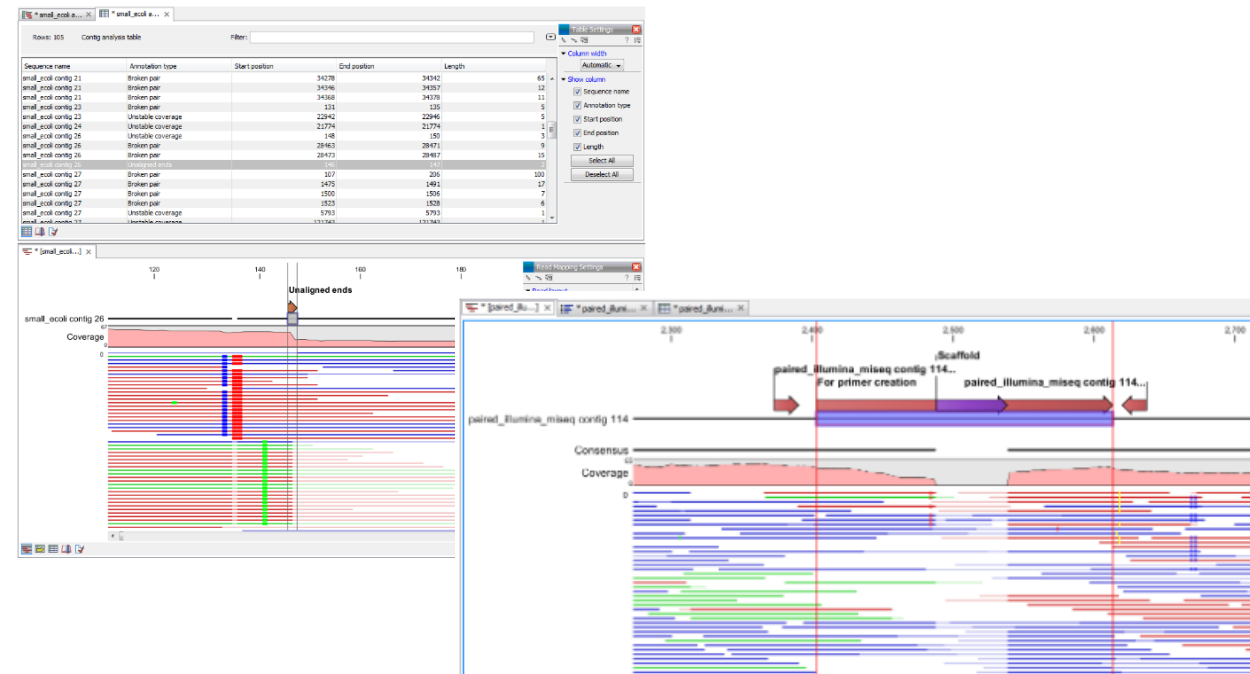
アセンブリ1.1

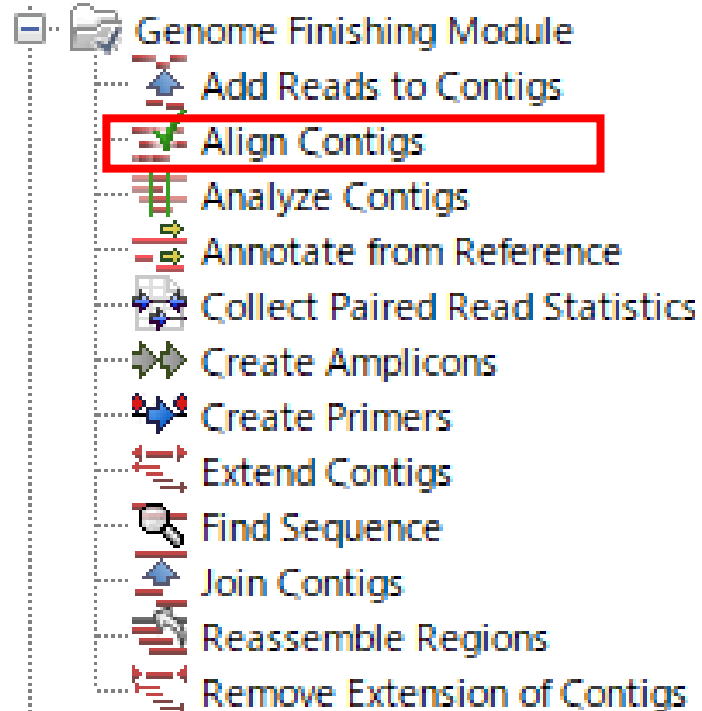
- ・コンティグ2本
- ・1コンティグあたり平均**18.5bp**

ATGCGCGTAGCGATG
TGACGACAATCGTGTGACGTGA

- コンティグをアライメントする方法としては、
- ①BLASTを利用して、コンティグ間のオーバーラップを見つける
 - ②ゲノムが既知の近縁種にマッピングする
- などの手法が考えられる。

- ✓ CLC Genomics Workbench Premiumに含まれるモジュールの一つ。
- ✓ コンティグのアライメントを行ったり、コンティグ間のオーバーラップに基づいたコンティグをマージしたり、コンティグのマニュアルキュレーションを行うためのモジュール

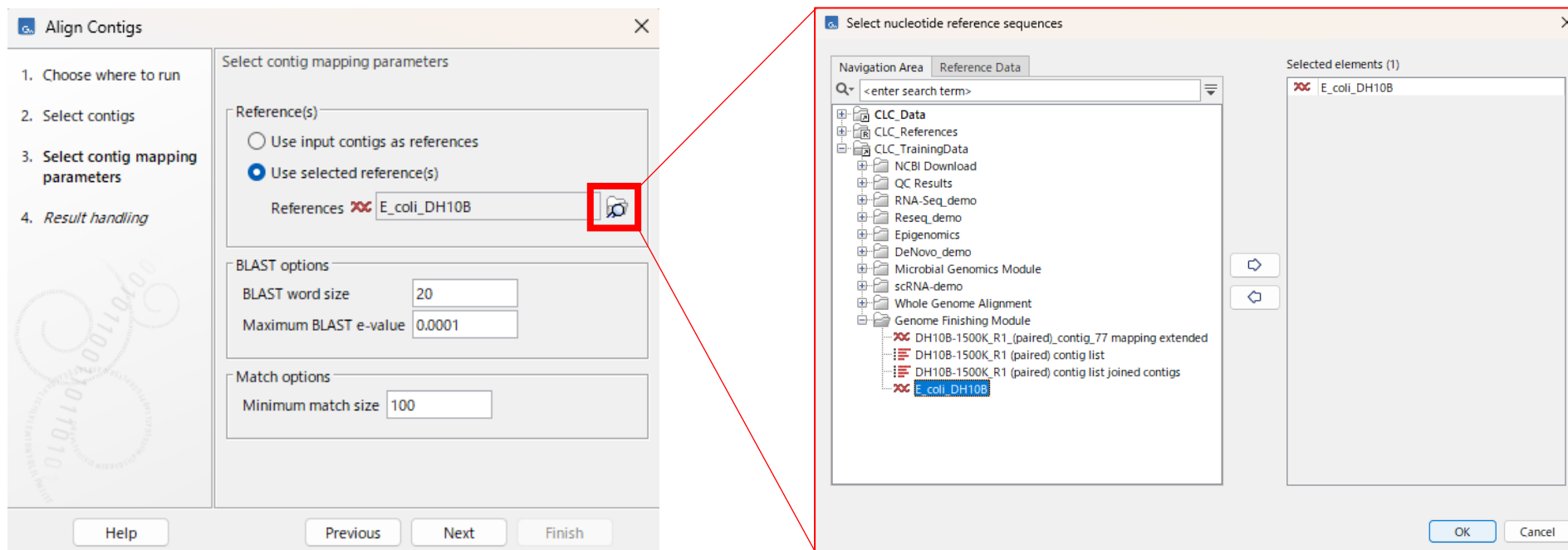




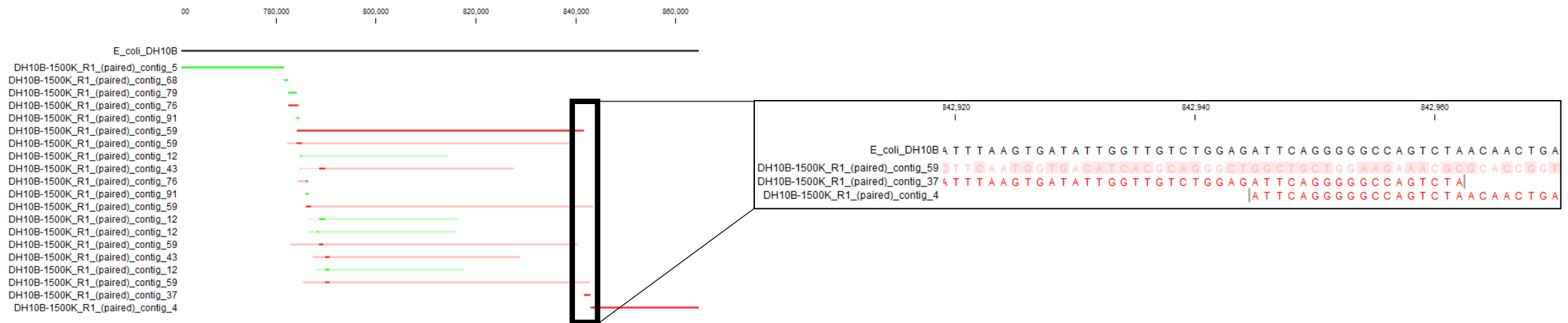
Align Contigs

de novo assemblyで得られたコンティグのアライメントを行う

- ✓ 近縁種のゲノムを指定し、それをリファレンスとしてコンティグをマッピングするか、リファレンスを介さずに、直接コンティグどうしのオーバーラップを調べるかを選択できる。



- ✓ 各コンティグがどのようにリファレンスにマッピングされたかを確認できる。
- ✓ コンティグ末端にズームインすると、隣のコンティグとのオーバーラップの有無も確認できる。



- ✓ 十分なオーバーラップがあった場合は、Join Two Contigsによりそれらをマージできる。

The screenshot displays a genomic alignment view with three tracks. The top track, labeled 'E_coli_DH10B', shows a reference sequence with a blue highlight over the region 'ATTCAGGGGGGCCAGTCT'. The middle track, 'DH10B-1500K_R1_(paired)_contig_59', shows a red sequence 'ATCACGCAGGGCTGGCTGCTGGAAGAAAC'. The bottom track, 'DH10B-1500K_R1_(paired)_contig_37', shows a red sequence 'TGGTTGTCTGGAGATTCAGGGGGGCCAGTCT'. A third track, 'DH10B-1500K_R1_(paired)_contig_4', shows a red sequence 'ATTCAGGGGGGCCAGTCT'. A context menu is open over the alignment, listing various actions. The 'Join Two Contigs' option is highlighted in blue. The menu also includes options like 'Delete Selection', 'Edit Selection', 'Transfer Selection to All Reads', 'Transfer Selection to Contig Sequence', 'Extract from Selection...', 'Find Broken Pair Mates...', 'Extract New Consensus Sequence...', 'Copy', 'Open Selection in New View', 'BLAST Selection against NCBI...', 'BLAST Selection against Local Data...', 'Set Numbers Relative to This Selection', 'Add Annotation...', 'Reassemble', and 'Split Contig...'.

842,940 | 842,960 | 842,980

E_coli_DH10B T G G T T G T C T G G A G A T T C A G G G G G C C A G T C T A C A A C T G A T C A C G G T T T G A

DH10B-1500K_R1_(paired)_contig_59 A T C A C G C A G G G C T G G C T G C T G G A A G A A A C T G T T

DH10B-1500K_R1_(paired)_contig_37 T G G T T G T C T G G A G A T T C A G G G G G C C A G T C T T T G A

DH10B-1500K_R1_(paired)_contig_4 A T T C A G G G G G C C A G T C

- Delete Selection
- Edit Selection
- Transfer Selection to All Reads
- Transfer Selection to Contig Sequence
- Extract from Selection...
- Find Broken Pair Mates...
- Extract New Consensus Sequence...
- Copy
- Open Selection in New View
- BLAST Selection against NCBI...
- BLAST Selection against Local Data...
- Set Numbers Relative to This Selection
- Add Annotation...
- Join Two Contigs**
- Reassemble
- Split Contig...

ATGCGCGTAGCGATGACGACAATCGTGTGACGTGA

contig1
contig2

ATGCGCGTAGCGATGA

ACAATCGTGTGACGTGA

隣接するコンティグの間に数塩基のギャップ
→どちらかがあと十数塩基長ければマージできそう



コンティグにリードをマッピングしたときの○部分のリードを配列を使って、
コンティグの端よりも先の配列を知ることが出来る場合がある。

- ✓ Align Contigsで出力されるテーブルから、リードをコンティグにマッピングすることができる。

* DH10B-1500K_R1 (paired) conti... X

<enter search term> Filter Selection... Show All Rows: 98

Contig	Contig length	Total read count	Average coverage	Contig matches length	Contig matches count
DH10B-1500K_R1_(paired)_contig_62	1455	3417	345.34	5381	4
DH10B-1500K_R1_(paired)_contig_63	3182	13846	634.67	21331	7
DH10B-1500K_R1_(paired)_contig_64	13008	8634	99.46	13008	1
DH10B-1500K_R1_(paired)_contig_65	9157	5897	96.24	9157	1
DH10B-1500K_R1_(paired)_contig_66	12768	8259	96.69	12953	2
DH10B-1500K_R1_(paired)_contig_67	8733	5740	92.57	8619	1
DH10B-1500K_R1_(paired)_contig_68	818	1577	281.33	3014	4
DH10B-1500K_R1_(paired)_contig_69	1657	1115	99.48	1609	1
DH10B-1500K_R1_(paired)_contig_70	8554	5345	93.28	8554	1
DH10B-1500K_R1_(paired)_contig_71	224	213	119.05	672	3
DH10B-1500K_R1_(paired)_contig_72	856	5475	757.64	8389	11
DH10B-1500K_R1_(paired)_contig_73	5368	6994	195.08	12823	5
DH10B-1500K_R1_(paired)_contig_74	858	3371	396.00	8399	11
DH10B-1500K_R1_(paired)_contig_75	1031	1232	173.55	2062	2
DH10B-1500K_R1_(paired)_contig_76	2153	2666	183.82	9469	6
DH10B-1500K_R1_(paired)_contig_77	6590	4443	100.68	6590	1
DH10B-1500K_R1_(paired)_contig_78	11167	7296	97.28	11167	1
DH10B-1500K_R1_(paired)_contig_79	1749	863	73.07	6551	5
DH10B-1500K_R1_(paired)_contig_80	252	324	172.27	838	4
DH10B-1500K_R1_(paired)_contig_81	1045	2064	284.86	3135	3
DH10B-1500K_R1_(paired)_contig_82	2233	1433	95.31	2542	2
DH10B-1500K_R1_(paired)_contig_83	3102	2082	99.52	3246	2

Show Contigs Add/Extract... Copy Contig... Map Reads... Join Contigs... Remove Contigs

Map Reads Again

* [E_coli_DH10B] X * [E_coli_DH10B] X

[E_coli_DH10B] X

513,140 513,160 513,180 513,200

E_coli_DH10B :TGCTTTAGGGAAGGTGCGAATAAGCGGGGAAATTCTTCTCGGCTGACTCAGTCATTTTCATTTCTT

DH10B-1500K_R1_(paired)_contig_77 :TGCTTTAGGGAAGGTGCGAATAA

DH10B-1500K_R1_(paired)_contig_39 :GGGAAATTCTTCTCGGCTGACTCAGTCATTTTCATTTCTT

DH10B-1500K_R1_(paired)_contig_29 :ACAACAAGGGAAGGTGCGAATAAGCAGGTCATTTCTTCCCAAGCTGACTCGGTGATTAAATTTTC

DH10B-1500K_R1_(paired)_contig_61 :ACATGATCTCTTTTGTCTTGCCTTTTGTGCAGAAAGCATCACAACGAGCTGCATACGGATACAGTGG

DH10B-1500K_R1_(paired)_contig_50 :ACACCCCTGCGCGCCAGCCCTGAACATATGCCAACGAGATGTACATCGATACTGGTGCGGTGTTTC

- ## コンティグ

マッピングされたリード

リードの配列に基づいて延長された部分

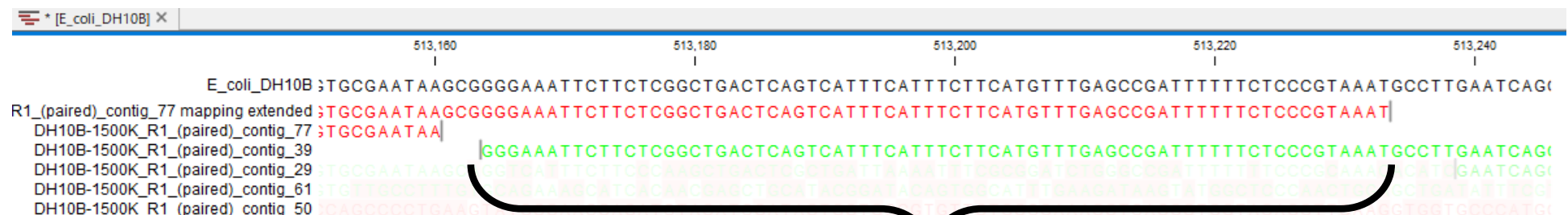


- ✓ コンティグを延長し、隣接コンティグとのオーバーラップをつくることできれば、それを介してさらにコンティグをマージすることができる。

延長前



延長後



オーバーラップが出現

お問い合わせ先：フィルジェン株式会社

TEL: 052-624-4388 (9:00～17 : 00)

FAX: 052-624-4389

E-mail: support@filgen.jp