パイプラインの概要、および、新手法（RoomD, RoomE）の要約

NGSの劇的な進化によってゲノム配列決定はより身近なものになってきた。一方で、正確なゲノムアノテーションはすべてのゲノム解析研究を円滑に進めるために重要だが、真核生物に対して正確なゲノムアノテーションを簡潔に構築するパッケージの配備は不十分である。

- **NGSによって決定されたゲノム配列の精度に関する課題**
  出力配列に含まれる特有のエラー（欠失や挿入）への対処
- **正確なアノテーションの決定に考慮するべき問題**
  遺伝子クラスター(やCNV)の正確なアノテーション

我々が運用している**「真核生物用アノテーションパイプライン」**、関連する新規開発中のソフトウェア：
「kmer-based ポリッシュプログラム」「リピートマスク手法」「RNA-seqアセンブラ」について紹介する。

【ゲノムアノテーションパイプラインの概要、および関連するソフトウェア】

新規ゲノム配列決定（PacBio, Nanopore）
ロングリードを用いて決定された配列に含まれるエラー修正（ポリッシュ；kpolish）

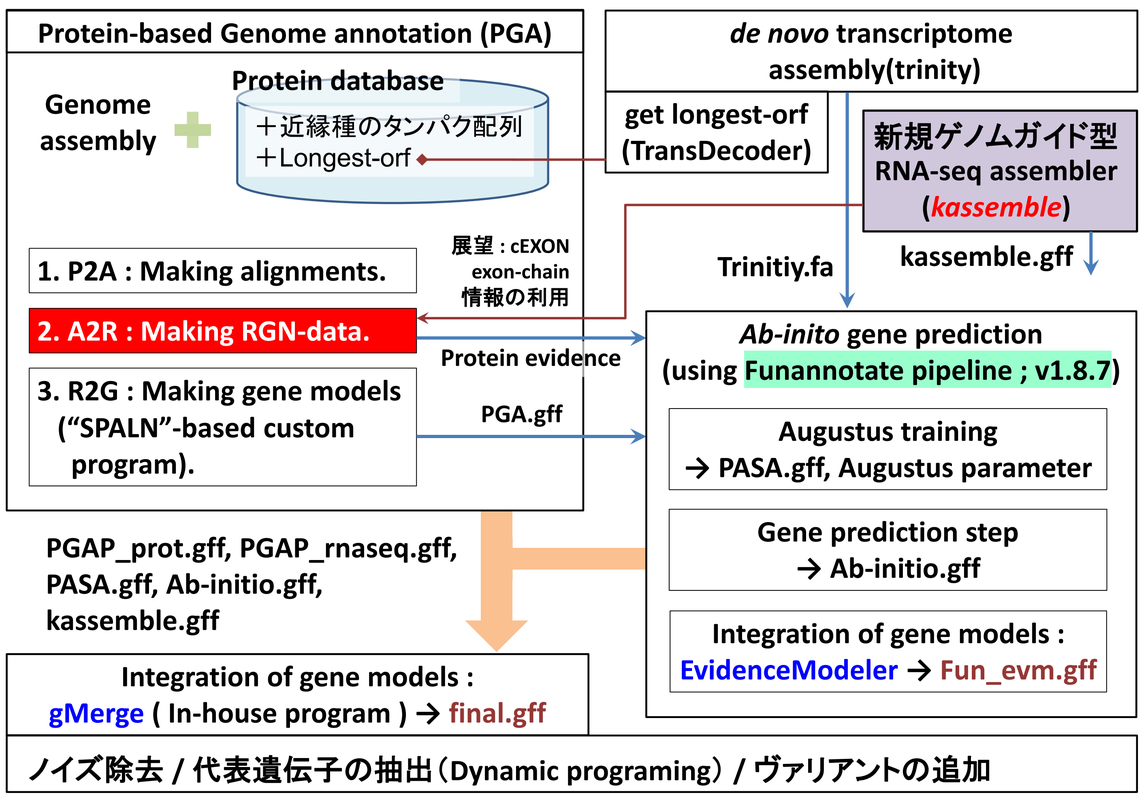
カスタムリピートライブラリの作成（RepeatModeler & RepeatMasker）
⇒ マスク済みゲノム配列におけるCDS領域の誤マスク修正 (RM+F)

Protein database の選定

1. ribosomal rna の予測 (using SILVA)
2. ターゲット種の分類階層を同定
3. 使用するDB の選択

ex) Eukaryota (n = 33,931,931)
Metazoa (n = 15,049,121)
Insecta (n= 3,203,208)

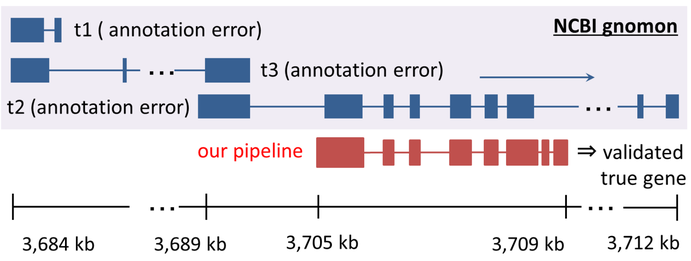
UniRef90 / Pre-selected DB / User-customized DB (特定の分類階層（taxid）を指定)

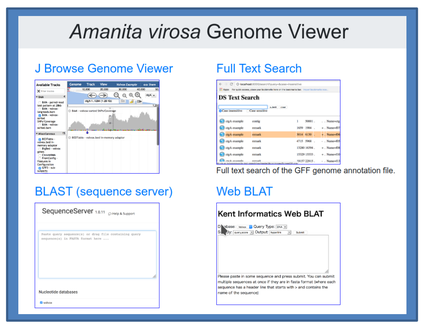


Required: Genome assembly, maximum intron size only.
・For Funannotate pipeline : Busco-DB name and/or Augustus pre-traning parameter, if RNA-seq is not entered.

昆虫	Total [Repeat%]	#gene (BUSCO=C+F%)	爬虫類	Total [Repeat%]	#gene (BUSCO=C+F%)
Target1	= 667M [32.6]	#15,373 (99.0%)	Target2	= 1.404G [39.9]	#20,356 (96.2%)
			Target3	= 1.467G [40.1]	#20,047 (96.3%)
近縁種A	= 908M [42.1]	#14,284 (92.9%)			
近縁種B	= 471M [32.7]	#15,773 (95.7%)	近縁種C	= 1.413G [41.3]	#25,134 (88.0%)
近縁種C	= 843M [33.7]	#27,558 (96.5%)	近縁種D	= 1.591G [34.6]	#18,445 (69.9%)

On the *Ipomoea nil* genome annotation test





C. elegans ゲノムを用いたベンチマークによる Precision の評価

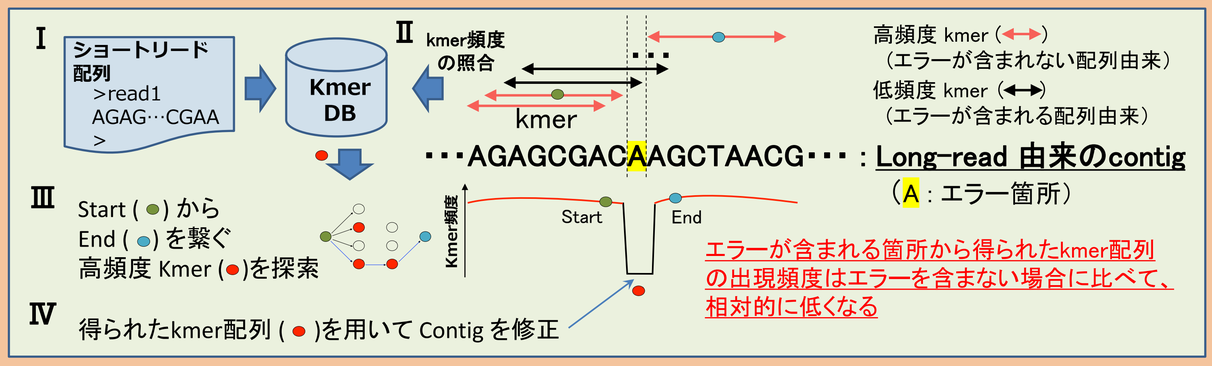
90%程度の相同性を持つ配列がDBにある時、予測されたエクソンの 96% 程度が正しい

相同性	90≦%<95	80≦%<90	70≦%<80	60≦%<70	50≦%<60	40≦%<50
RGN	11,897	3,142	3,481	3,907	3,802	3,087
NT-level	99.5	98.7	97.9	97.5	97.1	95.9
Exon-level	96.1	78.6	72.2	63.5	58.4	62.6

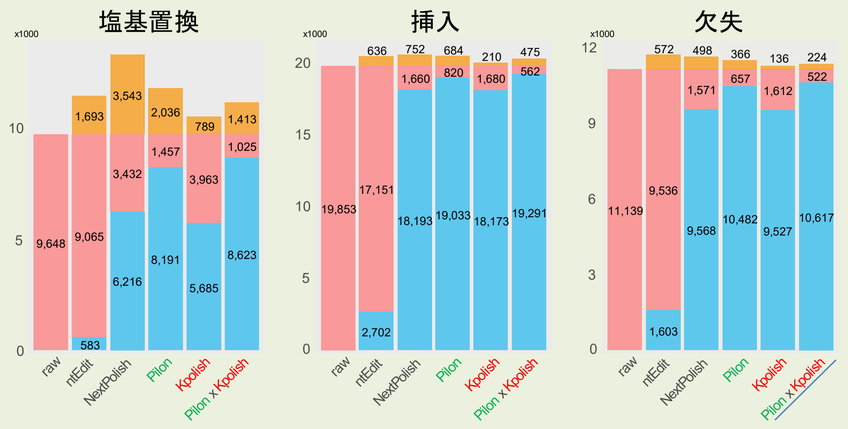
RoomD

【Kmer-based 新規ポリッシュ手法の開発; kpolish 】

- ロングリードを出力するNGS (PacBioやNanopore) を用いた配列決定が一般化
- これらNGSには、特有のシーケンスエラーがある
  ⇒ アセンブル後のコンティグに含まれるエラーをショートリード配列により修正
  = **Polish (ポリッシュ) は高精度な配列決定に重要**
- 既存のプログラムには、その精度や計算時間において課題が残されている



酵母のPacBio配列 (x68) ⇒ wtdbg2 → Arrow [polish (long-read)]
⇒ Polish (short-read ; MiSeq ≒ x83) : ntEdit, NextPolish, Pilon (x4), kpolish



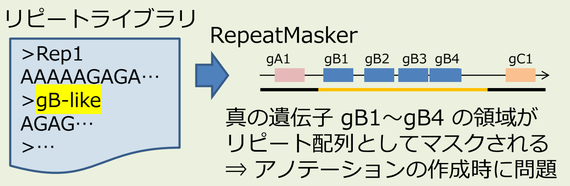
: kpolish < 他手法
誤修正は最も少ない

: kpolish < pilon

kpolish（kmer-based）
+ pilon（mapping-based）
⇒ 性質の異なる手法を組み合わせ、精度の向上が可能

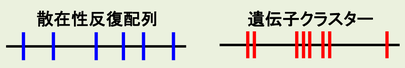
【新規リピートマスク手法（RepeatModler&Maskerによるリピートマスク修正; RM+F）】

- RepeatModeler (RM) はゲノム配列をマスクする際の 反復配列ライブラリ の作成に広く用いられている
- しかしながら、遺伝子に類似した配列が同ライブラリに混入する可能性がある

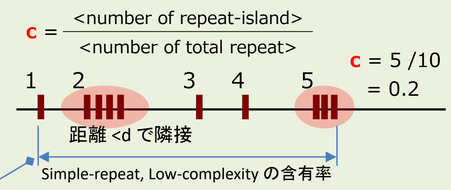


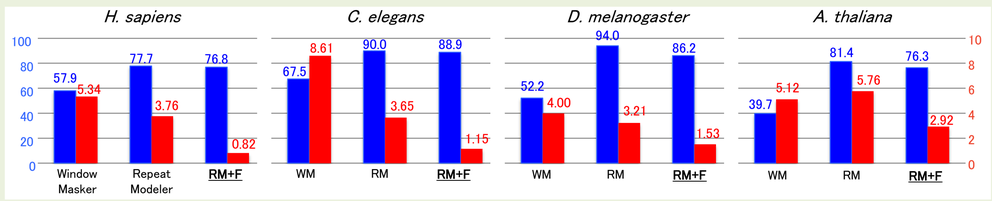
Ⅰ. 単純反復配列 ⇒ マスク
Ⅱ. 出現数 f の評価
<10 → アンマスク
≥ f → マスク
f = GenomeSize / 1Mb
(ex: ヒト = 3,088)

Ⅲ. 散在率 “c” の評価
c ≥ 0.5 → マスク (散在性リピート)
c < 0.5 ⇒ step.Ⅳ



Ⅳ. 単純反復配列の含有率 s の評価
if. 末端にあるリピート配列間領域: s < 95% ⇒ アンマスク（遺伝子クラスター）





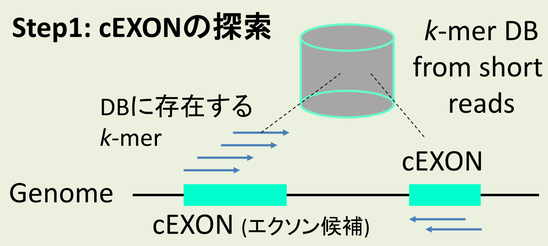
RoomE

【マッピング非依存的なゲノムガイド型RNA-seqアセンブラの開発; kassemble 】

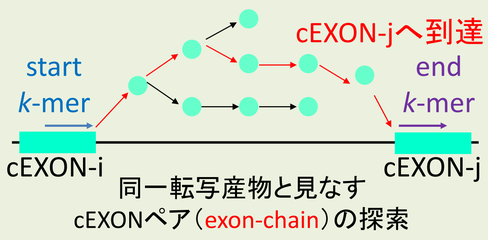
	必要データ	精度	主な処理	例
***de novo*型**	リード配列	▲	グラフ探索	trinity
ゲノムガイド型	リード配列＋参照配列	○（参照配列の精度に依存）	マッピング → グラフ探索	stringtie

両タイプの利点を有する新規ゲノムガイド型RNA-seqアセンブラ: ***kassemble***
⇒ ・従来の**ゲノムガイド型より高速**な動作
・**参照配列を利用することによる高精度化**
・一方で、*de novo*型のような**参照配列への低依存性**

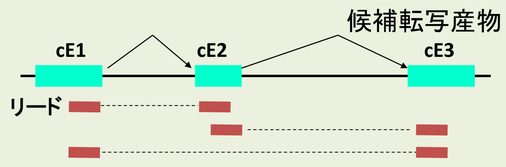
Step1: cEXONの探索



Step2: パス探索/アラインメント



Step3: 後処理（開発中）/出力

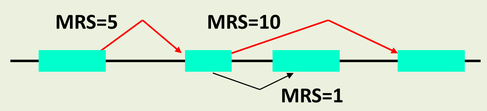


Paired-end 配列を利用し、cEXONペアを厳選する

ベンチマーク
（マウスのRNA-seqデータを使用）

発現量を基準にした transcript配列 のデータセット	*kssemble* (k=25) Sensitivity	*stringtie* Sensitivity
低発現（≥1；n=3,688）	97.3%	96.7%
中発現（≥5；n=626）	99.4%	98.2%
高発現（≥10；n=624）	99.7%	98.9%

・大部分の exon-chain を捕捉
・Stringtie (ゲノムガイド型)よりも高感度

Min read support (MRS)	Sensitivity (exon chain)	Precision (exon chain)
1	98.6%	90.4%
3	98.5%	92.3%
10	98.1%	94.3%

・Paired-end リードを用いて、exon-chainを評価することで、偽のcEXONへの exon-chain を削減
・Transcriptの精度向上に活用できる