シングルセルゲノムデータ解析手順

1. 取得したfastqに対し、bbduk.sh [version 39.01]を用いてアダプター配列のトリミング、低品質リードの除去を行いました。指定しているオプションは、ktrim=r ref=adapters k=23 mink=11 hdist=1 tpe tbo qtrim=r rimq=10 minlength=40 maxns=1 minavgquality=15 となります。

2. bbmap.sh [version 39.01]を用いて、マスクされたヒトゲノムに対しリード配列をマッピングし、マップされたリードをヒトコンタミリードとして除去しました。指定しているオプションは、 quickmatch fast untrim minid=0.95 maxindel=3 bwr=0.16 bw=12 minhits=2 path=human_masked_index(*) qtrim=rl trimq=10 となります。

3. SPAdes genome assembler [v3.15.5]を用いて、(2)に対しシングルセルモードでアセンブルを行いました。指定しているオプションは、--sc --careful --disable-rr --disable-gzip-output –k 21,33,55,77,99,127 となります。

4. seqkit [v2.6.1]を用いて、(3)で取得したアセンブリ配列から200塩基未満のコンティグを削除しました。

5. Prokka [version 1.14.6]を用いて、(4)で取得したアセンブリ配列から遺伝子領域を推定しました。指定しているオプションは、 --rawproduct --mincontiglen 200 となります。

6. QUAST [v5.2.0]を用いて、(4)で取得したアセンブリ配列のコンティグ数、全長、GC含量などを評価しました。オプションはデフォルトとなります。

7. CheckM [v1.1.3]を用いて、(4)で取得したアセンブリ配列の完全性および汚染度を評価しました。指定しているオプションは、lineage_wfにおいて -r --nt、taxonomy_wfにおいて taxonomy_wf --nt domain Bacteria となります。

8. GTDBTk [version 2.3.2]を用いて、(4)で取得したアセンブリ配列の生物系統情報を推定しました。オプションはデフォルトとなります。生物系統推定に使用したデータベースのバージョンは、release214となります。

(*) hg19_main_mask_ribo_animal_allplant_allfungus.fa.gz from https://zenodo.org/record/1208052#.X1hBFWf7SdY

bitBiome