Tutorial: Using BWA aligner to identify low-coverage genomes in metagenome sample

Umer Zeeshan Ijaz

We will use NexteraXT_even_1ng_HISEQ_AGGCAGAA-CTCTCTAT dataset to identify the list of genomes with low coverage. It is assumed that Burrow-Wheeler Aligner, Samtools, and Bedtools are installed. Here are the steps:

Step 1: Index the reference database file that comprises 59 genomes.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ bwa index Mercier_New.fasta
```

Step 2: Use BWA-MEM to align paired-end sequences. Briefly, the algorithm works by seeding alignments with maximal exact matches (MEMs) and then extending seeds with the affine-gap Smith-Waterman algorithm (SW).

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ bwa mem Mercier_New.fasta ../ID_1884/sample1/1_AGGCAGAA-CTCTCTAT_L001_R1_trim_001.fastq ../ID_1884/sample1/1_AGGCAGAA-CTCTCTAT_L001_R2_trim_001.fastq > aln-pe.sam
```

Step 3: Convert sam file to bam file.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ samtools view -h -b -S aln-pe.sam > aln-pe.bam
```

Step 4: Extract only those sequences that were mapped against the reference database. Use -F 4 switch.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ samtools view -b -F 4 aln-pe.bam > aln-pe.mapped.bam
```

Step 5: Generate a file length.genome that contains two entries per row, genome identifier and genome length.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ samtools view -H aln-pe.mapped.bam | perl -ne 'if ($_ =~ m/^\@SQ/)
{ print $_ }' | perl -ne 'if ($_ =~ m/SN:(.+)\s+LN:(\d+)/) { print $1, "\t", $2, "\n"}' > lengths.genome
```

Step 6: Sort BAM file. Many of the downstream analysis programs that use BAM files actually require a sorted BAM file. -m specifies the maximum memory to use, and can be changed to fit your system.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ samtools sort -m 1000000000 aln-pe.mapped.bam aln-pe.mapped.sorted
```

Step 7: We will now use bedtools. It is a very useful suite of programs for working with SAM/BAM, BED, VCF and GFF files, files that you will encouter many times doing NGS analysis. -ibam switch takes indexed bam file that we generated earlier, -d reports the depth at each genome position with 1-based coordinates, and -g used the genome lengths file we generated earlier.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ bedtools genomecov -ibam aln-pe.mapped.sorted.bam -d -g
lengths.genome  > aln-pe.mapped.bam.perbase.cov
```

Step 8: Look at the first few entries in the file generated above. First column is genome identifier, second column is position on genome, third column is coverage.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ head aln-pe.mapped.bam.perbase.cov
Acidobacterium_capsulatum_ATCC_51196      1     8
Acidobacterium_capsulatum_ATCC_51196      2     8
Acidobacterium_capsulatum_ATCC_51196      3     8
Acidobacterium_capsulatum_ATCC_51196      4     9
Acidobacterium_capsulatum_ATCC_51196      5     9
Acidobacterium_capsulatum_ATCC_51196      6     9
Acidobacterium_capsulatum_ATCC_51196      7     9
Acidobacterium_capsulatum_ATCC_51196      8     9
Acidobacterium_capsulatum_ATCC_51196      9     9
Acidobacterium_capsulatum_ATCC_51196      10    12
```

Step 9: Now we will count only those positions where we have >0 coverage.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ awk -F"\t" '$3>0{print $1}' aln-pe.mapped.bam.perbase.cov | sort |
uniq -c > aln-pe.mapped.bam.perbase.count
```

Step 10: To see what we have done, use the cat command

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ cat aln-pe.mapped.bam.perbase.count
4126255 Acidobacterium_capsulatum_ATCC_51196
2664070 Akkermansia_muciniphila_ATCC_BAA-835
2176909 Archaeoglobus_fulgidus_DSM_4304
6259688 Bacteroides_thetaiotaomicron_VPI-5482
5162898 Bacteroides_vulgatus_ATCC_8482
5334732 Bordetella_bronchiseptica_strain_RB50
6779108 Burkholderia1_xenovorans_LB400_chromosome_1,_complete_sequence
2968320 Caldisaccharolyticus_DSM_8903
2759841 Chlorobiumlimicola_DSM_245
3133587 Chlorobiumphaeobacteroides_DSM_266
1966851 Chlorobiumphaeovibrioides_DSM_265
2154672 Chlorobiumtepidum_TLS
5248942 Chloroflexus_aurantiacus_J-10-fl
3841892 Clostridium_thermocellum_ATCC_27405
3048801 Deinococcus_radiodurans_R1_chromosome_1,_complete_sequence
2796764 Desulfovibrio_piger_ATCC_29098
1819007 Dictyoglomus_turgidum_DSM_6724
2520796 Enterococcus_faecalis_V583
1458832 Fusobacterium_nucleatum_subsp._nucleatum_ATCC_25586
4636091 Gemmatimonas_aurantiaca_T-27_DNA
  77854 gi|220903286|ref|NC_011883.1|
2909411 gi|222528057|ref|NC_012034.1|
4920011 gi|307128764|ref|NC_014500.1|
1739058 gi|55979969|ref|NC_006461.1|
   7919 gi|83591340|ref|NC_007643.1|
6345359 Herpetosiphon_aurantiacus_ATCC_23779
1558912 Hydrogenobaculum_sp._Y04AAS1
1278529 Ignicoccus_hospitalis_KIN4/I
```

```
4901664 Leptothrix_cholodnii_SP-6
1634756 Methanocaldococcus_jannaschii_DSM_2661
1770072 Methanococcus_maripaludis_C5
1650606 Methanococcus_maripaludis_strain_S2,_complete_sequence
 459768 Nanoarchaeum_equitans_Kin4-M
2798863 Nitrosomonas_europaea_ATCC_19718
6409848 Nostoc_sp._PCC_7120_DNA
3017983 Pelodictyon_phaeoclathratiforme_BU-1
1929147 Persephonella_marina_EX-H1
2354658 Porphyromonas_gingivalis_ATCC_33277_DNA
2219120 Pyrobaculum_aerophilum_str._IM2
1998126 Pyrobaculum_calidifontis_JCM_11548
1738110 Pyrococcus_horikoshii_OT3_DNA
7145136 Rhodopirellula_baltica_SH_1_complete_genome
4106385 Ruegeria_pomeroyi_DSS-3
5762894 Salinispora_arenicola_CNS-205
5168083 Salinispora_tropica_CNB-440
5229615 Shewanella_baltica_OS185
5145817 Shewanella_baltica_OS223,
3478897 Sulfitobacter_NAS-14.1_scf_1099451320477_
3029105 Sulfitobacter_sp._EE-36_scf_1099451318008_
2666279 Sulfolobus_tokodaii
1520314 Sulfurihydrogenibium_yellowstonense_SS-5
1828687 SulfuriYO3AOP1
2361532 Thermoanaerobacter_pseudethanolicus_ATCC_33223
1884531 Thermotoga_neapolitana_DSM_4359
1823470 Thermotoga_petrophila_RKU-1
1877693 Thermotoga_sp._RQ2
2829658 Treponema_denticola_ATCC_35405
2274638 Treponema_vincentii_ATCC_35580_NZACYH00000000.1
2052189 Zymomonas_mobilis_subsp._mobilis_ZM4
```

Step 11: We will now use the above file with lengths.genome to calculate the proportions using the following one-liner. It reads lengths.genome line by line, assigns the genome identifier to myArray[0] and it's length to

myArray[1]. It then searches the identifier in aln-pe.mapped.bam.perbase.count, extracts the base count, and uses bc to calculate the proportion.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ while IFS=$'\t' read -r -a myArray; do echo -e "${myArray[0]},$(
echo "scale=5;"$(awk -v pattern="${myArray[0]}" '$2==pattern{print $1}' aln-
pe.mapped.bam.perbase.count)"/"${myArray[1]} | bc ) "; done < lengths.genome > aln-pe.mapped.bam.genomeproportion
```

Step 12: Some of the reference genomes were downloaded from NCBI. We will use a file IDs_metagenomes2.txt that contains the meaningful mapping from accession numbers to genome names.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ cat IDs_metagenomes2.txt
gi|220903286|ref|NC_011883.1|,Desulfovibrio_desulfuricans
gi|222528057|ref|NC_012034.1|,Caldicellulosiruptor_bescii
gi|307128764|ref|NC_014500.1|,Dickeya_dadantii
gi|55979969|ref|NC_006461.1|,Thermus_thermophilus
gi|83591340|ref|NC_007643.1|,Rhodospirillum_rubrum
```

Step 13: Download my annotation script (http://userweb.eng.gla.ac.uk/umer.ijaz/bioinformatics/convIDs.pl) and then use IDs_metagenomes2.txt to annotate column 1.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ perl convIDs.pl -i aln-pe.mapped.bam.genomeproportion -l
IDs_metagenomes2.txt -c 1 -t comma > aln-pe.mapped.bam.genomeproportion.annotated
```

Step 14: We have a total of 59 genomes in the reference database. To see how many genomes we recovered, we will use the following one-liner:

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ awk -F "," '{sum+=$NF} END{print "Total genomes covered:"sum}'
aln-pe.mapped.bam.genomeproportion
Total genomes covered:55.8055
```

Step 15: Now we will identify those genomes for which the proportions are less than 0.99.

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ awk -F "," '$NF<=0.99{print $0}' aln-
pe.mapped.bam.genomeproportion.annotated
Burkholderia1_xenovorans_LB400_chromosome_1,_complete_sequence,.69664
Desulfovibrio_desulfuricans,.02709
Desulfovibrio_piger_ATCC_29098,.98358
Dictyoglomus_turgidum_DSM_6724,.98030
Enterococcus_faecalis_V583,.78333
Fusobacterium_nucleatum_subsp._nucleatum_ATCC_25586,.67088
Ignicoccus_hospitalis_KIN4/I,.98534
Methanocaldococcus_jannaschii_DSM_2661,.98185
Nanoarchaeum_equitans_Kin4-M,.93661
Rhodospirillum_rubrum,.00181
Sulfolobus_tokodaii,.98943
Thermus_thermophilus,.94016
Treponema_vincentii_ATCC_35580_NZACYH00000000.1,.90457
```

Step 16: The complete list is as follows:

```
[uzi@quince-srv2 ~/TSB_METAGENOMES/prob_genomes]$ cat aln-pe.mapped.bam.genomeproportion.annotated
Acidobacterium_capsulatum_ATCC_51196,.99973
Akkermansia_muciniphila_ATCC_BAA-835,.99998
Archaeoglobus_fulgidus_DSM_4304,.99931
Bacteroides_thetaiotaomicron_VPI-5482,.99989
Bacteroides_vulgatus_ATCC_8482,.99994
Bordetella_bronchiseptica_strain_RB50,.99916
Burkholderia1_xenovorans_LB400_chromosome_1,_complete_sequence,.69664
Caldicellulosiruptor_bescii,.99646
Caldisaccharolyticus_DSM_8903,.99934
Chlorobiumlimicola_DSM_245,.99879
Chlorobiumphaeobacteroides_DSM_266,.99989
Chlorobiumphaeovibrioides_DSM_265,.99999
Chlorobiumtepidum_TLS,.99987
Chloroflexus_aurantiacus_J-10-fl,.99817
Clostridium_thermocellum_ATCC_27405,.99963
```

Deinococcus_radiodurans_R1_chromosome_1,_complete_sequence,.99601
Desulfovibrio_desulfuricans,.02709
Desulfovibrio_piger_ATCC_29098,.98358
Dickeya_dadantii,.99943
Dictyoglomus_turgidum_DSM_6724,.98030
Enterococcus_faecalis_V583,.78333
Fusobacterium_nucleatum_subsp._nucleatum_ATCC_25586,.67088
Gemmatimonas_aurantiaca_T-27_DNA,.99981
Herpetosiphon_aurantiacus_ATCC_23779,.99980
Hydrogenobaculum_sp._Y04AAS1,.99961
Ignicoccus_hospitalis_KIN4/I,.98534
Leptothrix_cholodnii_SP-6,.99842
Methanocaldococcus_jannaschii_DSM_2661,.98185
Methanococcus_maripaludis_C5,.99399
Methanococcus_maripaludis_strain_S2,_complete_sequence,.99366
Nanoarchaeum_equitans_Kin4-M,.93661
Nitrosomonas_europaea_ATCC_19718,.99529
Nostoc_sp._PCC_7120_DNA,.99938
Pelodictyon_phaeoclathratiforme_BU-1,.99991
Persephonella_marina_EX-H1,.99941
Porphyromonas_gingivalis_ATCC_33277_DNA,.99990
Pyrobaculum_aerophilum_str._IM2,.99851
Pyrobaculum_calidifontis_JCM_11548,.99443
Pyrococcus_horikoshii_OT3_DNA,.99977
Rhodopirellula_baltica_SH_1_complete_genome,.99993
Rhodospirillum_rubrum,.00181
Ruegeria_pomeroyi_DSS-3,.99925
Salinispora_arenicola_CNS-205,.99594
Salinispora_tropica_CNB-440,.99705
Shewanella_baltica_OS185,.99998
Shewanella_baltica_OS223,,.99998
Sulfitobacter_NAS-14.1_scf_1099451320477_,.99697
Sulfitobacter_sp._EE-36_scf_1099451318008_,.99921
Sulfolobus_tokodaii,.98943
Sulfurihydrogenibium_yellowstonense_SS-5,.99087
SulfuriYO3AOP1,.99469

Thermoanaerobacter_pseudethanolicus_ATCC_33223,.99945
Thermotoga_neapolitana_DSM_4359,.99998
Thermotoga_petrophila_RKU-1,.99997
Thermotoga_sp._RQ2,1.00000
Thermus_thermophilus,.94016
Treponema_denticola_ATCC_35405,.99523
Treponema_vincentii_ATCC_35580_NZACYH00000000.1,.90457
Zymomonas_mobilis_subsp._mobilis_ZM4,.99797