24/6/2020

**From:** Noelani Kamelamela <[hilo1@luria.mit.edu](mailto:hilo1@luria.mit.edu)>  
**Sent:** Friday, June 12, 2020, 5:19 PM

The data from your experiment are ready to be downloaded from the BioMicro servers.   
  
Here is a summary of the samples from this flowcell (4500T) sequenced as:  
300 + 300 bases pair-end run with 8 + 8 nucleotide indexes  
  
200219Seg/  
    D20-160001 : Axenic1 (GAACAATTCCTAGAGTTGGA)  
    D20-160002 : Axenic2 (TGTGGTCCGGAGAGCACTAG)  
    D20-160003 : Axenic3 (CTTCTAAGTCACTCTACAGG)  
    D20-160004 : Axenic4 (AATATTGCCACGGTGACACC)  
    D20-160005 : 2A1 (TCGTGCATTCGCGTTGGTAT)  
    D20-160006 : 2A2 (AAGATACACGTGTGCTAACA)  
    D20-160007 : 2C1 (TGCAATGAATCCAGAAGTAA)  
    D20-160008 : empty (CTATGAAGGACTTATACCTG)  
    D20-160009 : 2C2 (GAAGACTAGAACTAGAACTT)  
    D20-160010 : 5A1 (AGGAGTCGAGTTAGGCTTAC)  
    D20-160011 : 5A2 (TTCACTCACTTATCATGAGA)  
    D20-160012 : 5B1 (GGTCCGCTTCCTCACACAAG)  
    D20-160013 : 5B2 (CAACGAGAGCGAATTGAGTG)  
    D20-160014 : 5C1 (ATTGAGGTCCCGGATTATAT)  
    D20-160015 : 5C2 (GGAGAGACTCTTGAAGCAGA)  
    D20-160016 : 10A1 (CCGCTCCGTTTACGGCGAAG)  
    D20-160017 : 10A2 (ATACATCACATCTCCATTGA)  
    D20-160018 : 10B1 (TAGGTATGTTCGAGACCAAG)  
    D20-160019 : 10B2 (CACCTAGCACTGCTGGACAT)  
    D20-160020 : 10C1 (TTCAAGTATGGATGGTATCG)  
    D20-160021 : 10C2 (TTAAGACAAGGGCTTAATTG)

Sequences are downloaded from:

\* Use a SSH/SFTP client to connect to [bmc-150.mit.edu](http://bmc-150.mit.edu) with the following credentials  
username: segre\_ill  
password: an247bn4  
  
Download instructions are available at <https://openwetware.org/wiki/BioMicroCenter:FAQ#DOWNLOADING_DATA>  
  
The samples have been processed using the BMC/BCC 1.7 pipeline updated on 01/11/2019.  
- Upgraded cellranger from 2.0.2 to 3.0.1 for the 10X pipeline  
Details can be found at <http://openwetware.org/wiki/BioMicroCenter:Software#BMC-BCC_Pipeline>

**From:** Stuart Levine <[slevine@mit.edu](mailto:slevine@mit.edu)>  
**Sent:** Monday, June 22, 2020 9:47:16 PM  
**To:** Osborne, Melisa <[melosbor@bu.edu](mailto:melosbor@bu.edu)>

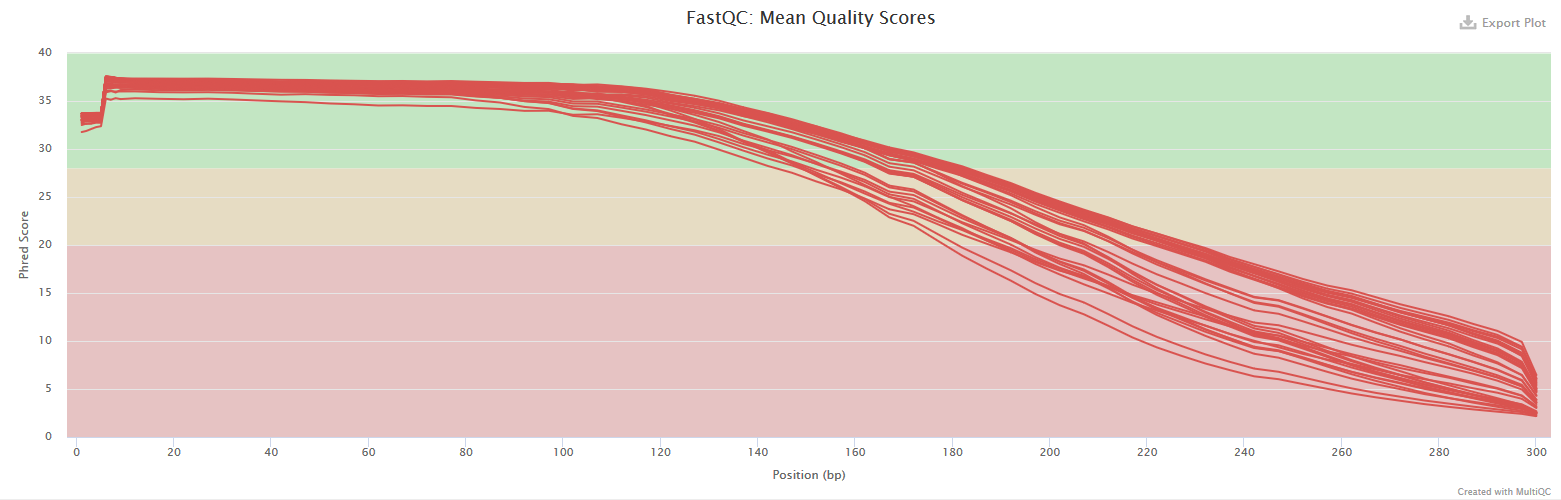
Fastqc should indicate fail as the samples are not shotgun human genome - we run it as well. Samples were prepped with Flex, which is about the best we can do for longer inserts - the solid matrix helps control insert size. I'm attaching our internal QC as a 7zip file. We do not do any trimming / pair merging of reads. This is not atypical for long miseq runs, several samples had very low input and did not perform well as Austin indicated. The run itself was robust in terms of error rate (QC plot at the top).

Samples:

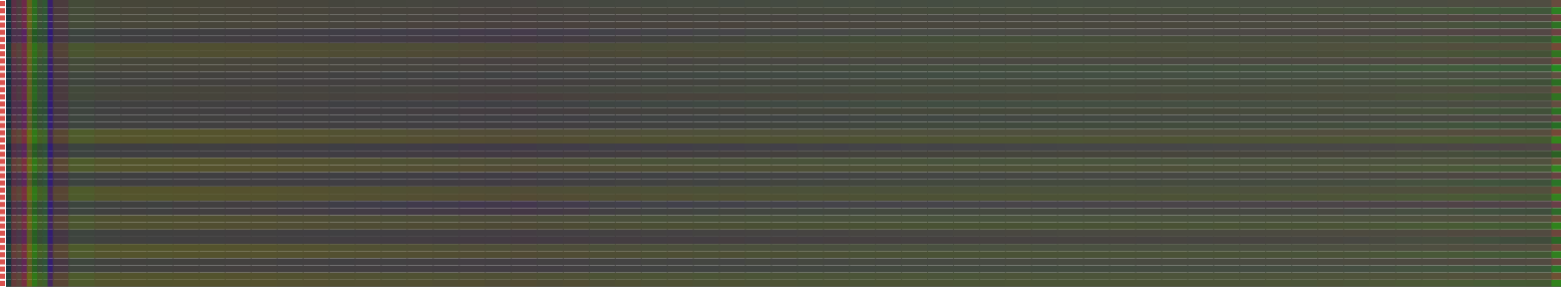
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **sample** | **PRO** | **ALT** | **exp** | **days** | **ng/ul** | **final volume** |
| 1 | Axenic | 1A3 |  |  |  | 30 | 100 |
| 2 | Axenic | DE |  |  |  | 10 | 100 |
| 3 | Axenic | 9313 |  |  |  | 0.6 | 100 |
| 4 | Axenic | MIT0604 |  |  |  | 6 | 100 |
| 5 | 2A | 9313 | 1A3 | 2 | 100 | 0.7 | 100 |
| 6 | 2A | 9313 |  | 6 | 440 | 1.2 | 100 |
| 7 | 2C | 9313 | 1A3 | 2 | 100 | 1.5 | 100 |
| 8 | empty | empty | empty | empty | empty | empty | empty |
| 9 | 2C | 9313 | 1A3 | 6 | 440 | 8 | 100 |
| 10 | 5A | MIT0604 | 1A3 | 2 | 100 | 4.28 | 100 |
| 11 | 5A | MIT0604 | 1A3 | 6 | 440 | 16 | 100 |
| 12 | 5B | MIT0604 | 1A3 | 2 | 100 | 3.75 | 100 |
| 13 | 5B | MIT0604 | 1A3 | 6 | 440 | 7.3 | 100 |
| 14 | 5C | MIT0604 | 1A3 | 2 | 100 | 9.3 | 100 |
| 15 | 5C | MIT0604 | 1A3 | 6 | 440 | 19.3 | 100 |
| 16 | 10A | MIT0604 | DE | 2 | 100 | 5.85 | 100 |
| 17 | 10A | MIT0604 | DE | 6 | 440 | 14 | 100 |
| 18 | 10B | MIT0604 | DE | 2 | 100 | 1.6 | 100 |
| 19 | 10B | MIT0604 | DE | 6 | 440 | 6.25 | 100 |
| 20 | 10C | MIT0604 | DE | 2 | 100 | 5.34 | 100 |
| 21 | 10C | MIT0604 | DE | 6 | 440 | 18 | 100 |

FastQC/multiQC on all samples:

### Sequence Quality Histograms



**Per Base Sequence Content**



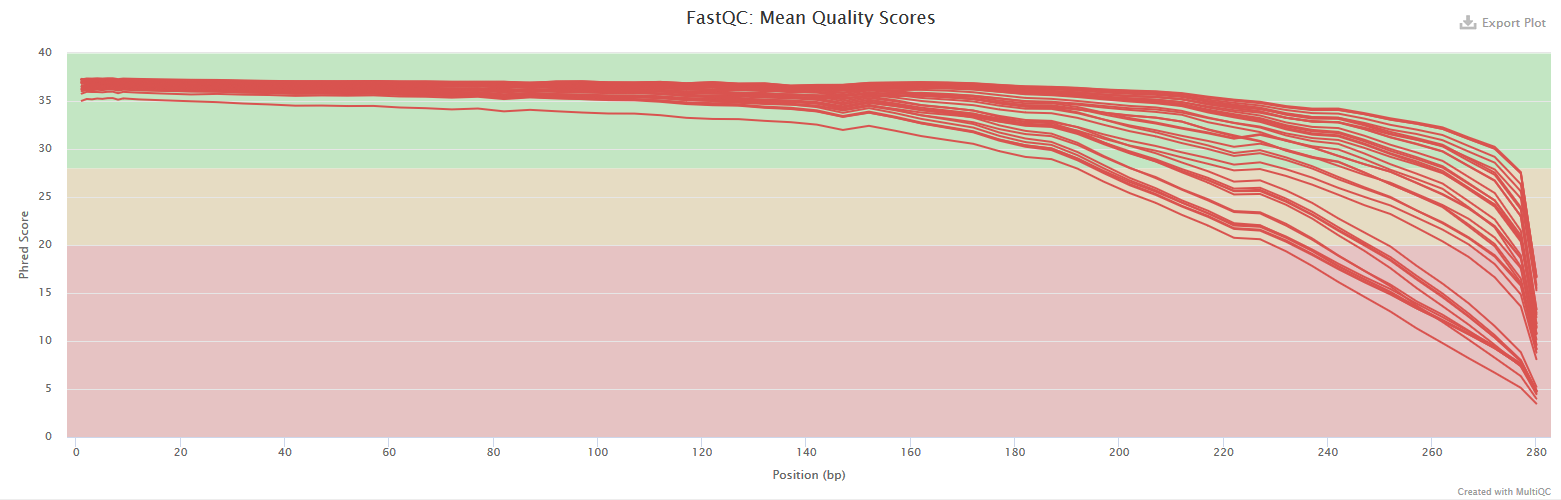
Decided to trim the first 20 bases. And not to trim the low quality at the end as advice from bwa mem developer is not to trim low quality and let bwa do soft clipping during run time.

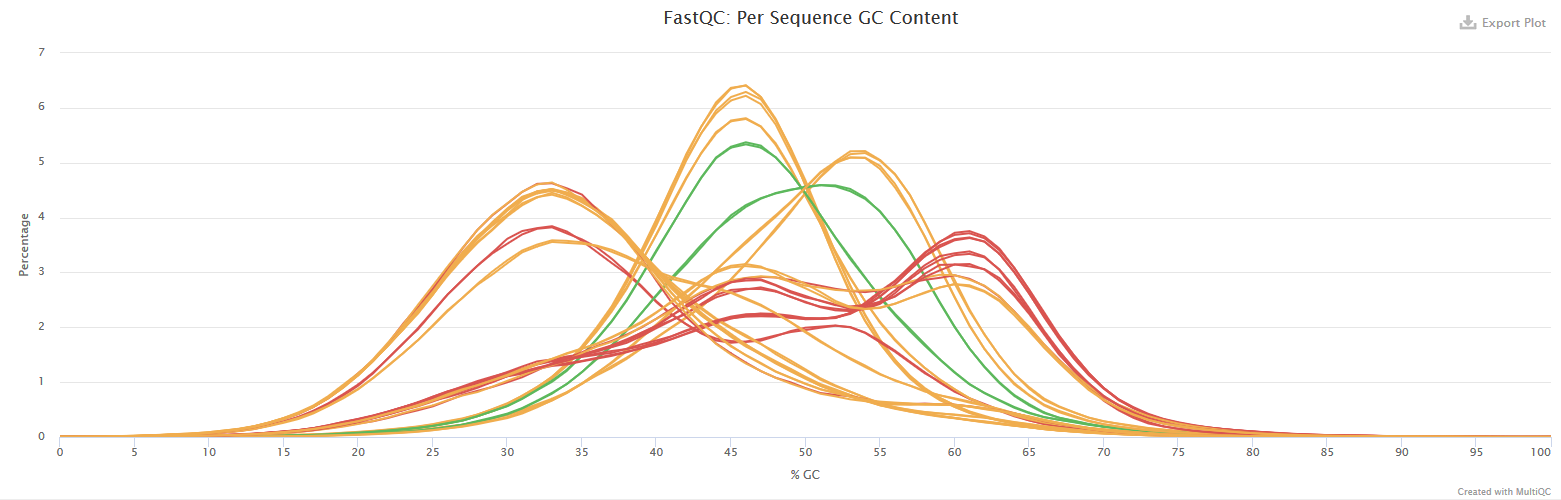
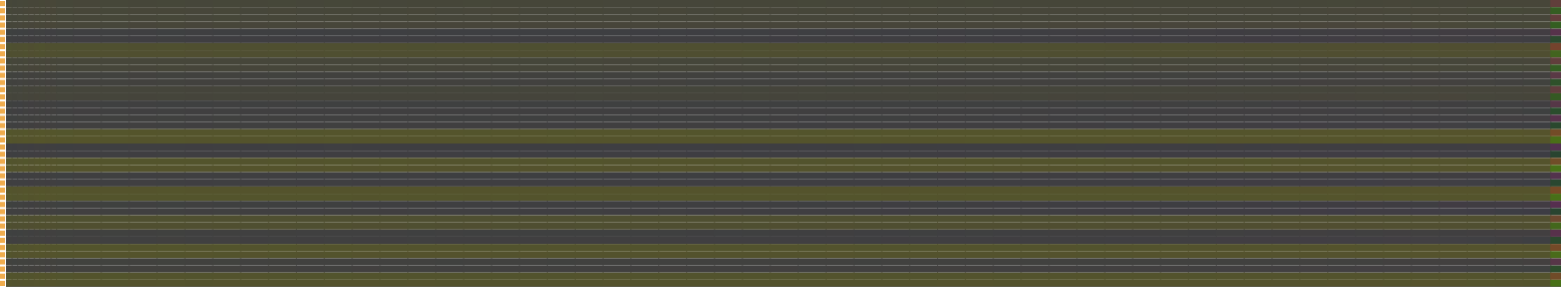
Ran the following command to trim adapters and barcodes from all fastq files:

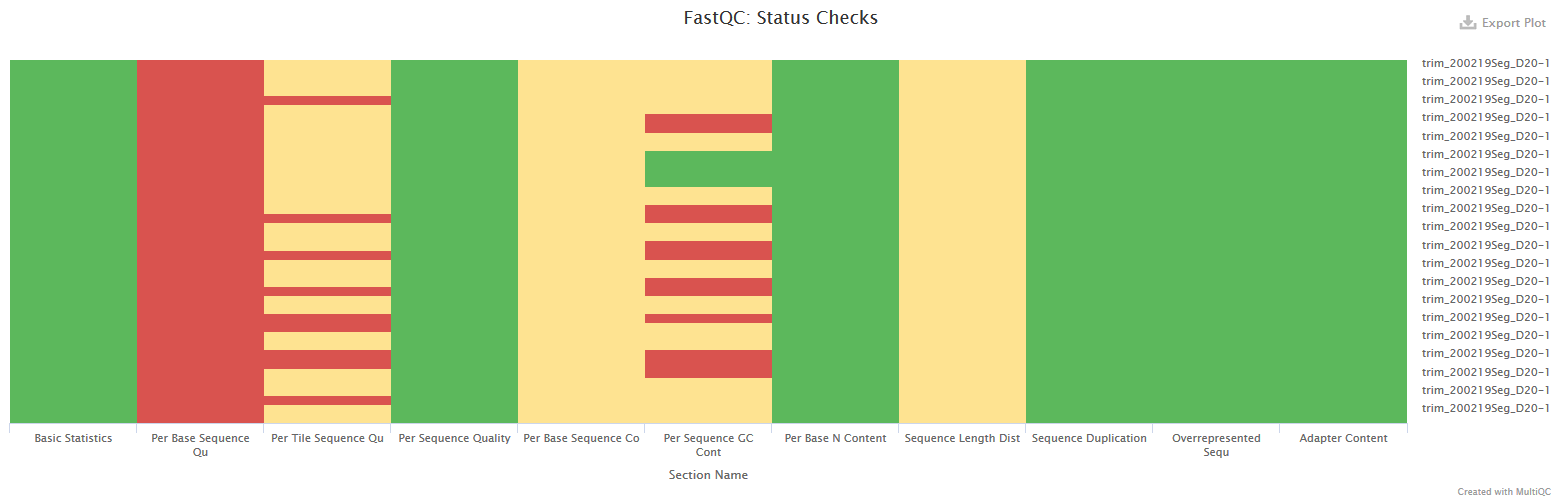
adapter=~/anaconda3/pkgs/trimmomatic-0.39-1/share/trimmomatic/adapters/**NexteraPE-PE.fa**

trimmomatic PE -threads 10 -trimlog ${outdpath}/trimmomatic.log -summary ${outdpath}/trimmomatic.log -validatePairs ${fastq1path} ${fastq2path} ${pout1} ${uout1} ${pout2} ${uout2} **ILLUMINACLIP:${adapter}:2:30:10:1:TRUE HEADCROP:20**

After trimming:







Download genomes:

Genomes were downloaded from NCBI on 25/6/2020:

|  |  |  |  |
| --- | --- | --- | --- |
| Species | Page | Accession | URL |
| DE | Alteromonas mediterranea DE, complete sequence | NC\_011138.3 | <https://www.ncbi.nlm.nih.gov/nuccore/NC_011138.3> |
| 1A3 | Alteromonas macleodii strain HOT1A3 plasmid pAM1A3, complete sequence | NZ\_CP012203.1 | <https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP012203.1> |
| 1A3 | Alteromonas macleodii strain HOT1A3 chromosome, complete genome | NZ\_CP012202.1 | <https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP012202.1> |
| MIT0604 | Prochlorococcus sp. MIT 0604, complete genome | NZ\_CP007753.1 | <https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP007753.1> |
| MIT9313 | Prochlorococcus marinus str. MIT 9313, complete genome | NC\_005071.1 | <https://www.ncbi.nlm.nih.gov/nuccore/NC_005071.1> |

## variant calling using snippy:

All samples were mapped to the reference genome and variant called using snippy (based on bwa mem mapping , freebayes variant call, snpEff annotation.

Snippy command:

snippy --ref ${refgb} --pe1 ${pout1} --pe2 ${pout2} --cpus 20 --outdir ${snoutdir} --prefix snp\_1A3\_all --unmapped --rgid D20-${idx} –report

Following this call, snippy was run again on the co-culture using the unmapped reads output from the second strain in the co-culture.

The % of unmapped reads in MIT0604 samples was large. (axenic: ~30%, experiment 2: ~70%, experiment 6: 10-30%)

|  |
| --- |
|  |
|  |
|  |

|  |
| --- |
|  |

I ran qualimap on the bam files (post filtering of reads mapped to the other strain).

qualimap command:

qualimap bamqc -nt 20 -bam ${i}/\*.bam -outdir ${i}/qualimap\_out

Results of qualimap + multiQC are at the excel file: ccpa/unmapped\_reads.xlsx

|  |  |
| --- | --- |
|  |  |

ALT coverage in most co-cultures is very low (1A3: 2-9x for most samples, DE: 0-2x), which is in sufficient for variant calling (we are looking for at least 10x coverage). In PRO strains, coverage of experiment 2 (100 days) is too low for variant calling, but coverage for the axenic and 440 days is good (30-36 for axenic, 38-63 for 440 days, 5-12 for 100 days).

## Unmapped reads

Upto 70% of MIT0604 reads are unmapped.

Several possibilities: MIT0604 was not axenic, there is additional contaminating bacteria, MIT0604 reference genome does not match the strain we are using (not likely because the %unmapped is low on MIT0604 samples on day 440, ~5% for MIT0604 + DE samples).

To try and understand what the contaminating bacteria is I tried several things:

#### Kraken

Kraken + bracken on all reads per sample (kraken DB built using default parameters on 9/7/2020):

kraken2 --db /data/home/dsher/oweissber/kraken2db --confidence 0.05 --threads 20 --output ${krakendpath} --minimum-base-quality 3 --report ${krakenreport} --paired --use-names ${pout1} ${pout2}

bracken -d /data/home/dsher/oweissber/kraken2db -i ${krakenreport} -o ${brakendpath} -l S

|  |
| --- |
|  |

Most of the reads are assigned to the PRO and ALT we are growing. A small number of reads is assigned to *Maircaulis maris*, but most of the reads receive *Unclassified* status, meaning that kraken failed to classify them ☹.

#### spades + blastn

Assembled all MIT0604 unmapped reads from all samples using spades and run blast on the resulting contigs.

20 Taxa with longest total alignment length:

|  |  |  |  |
| --- | --- | --- | --- |
| **Subject Sci Names** | **Total Alignment Length** | **Number of Contigs** | **Number of hits** |
| **Alteromonas macleodii** | 7460367 | 2066 | 6248 |
| **Alteromonas macleodii ATCC 27126** | 3022314 | 2002 | 2091 |
| **Alteromonas macleodii str. 'English Channel 673'** | 2611950 | 1954 | 2056 |
| **Alteromonas mediterranea** | 2469162 | 922 | 2750 |
| **Alteromonas macleodii str. 'Balearic Sea AD45'** | 2458683 | 1900 | 2067 |
| **Alteromonas macleodii str. 'Black Sea 11'** | 1860123 | 1514 | 1677 |
| **Maricaulis maris MCS10** | 955015 | 14 | 584 |
| **Alteromonas sp. BL110** | 918890 | 852 | 946 |
| **Methylophaga nitratireducenticrescens** | 743174 | 38 | 488 |
| **Alteromonas mediterranea UM7** | 573632 | 703 | 769 |
| **Alteromonas sp. Mex14** | 555758 | 514 | 551 |
| **Alteromonas mediterranea DE1** | 534202 | 662 | 728 |
| **Alteromonas mediterranea UM4b** | 519872 | 666 | 739 |
| **Alteromonas mediterranea U8** | 310676 | 333 | 356 |
| **Alteromonas mediterranea U7** | 281618 | 303 | 352 |
| **Alteromonas mediterranea 615** | 250452 | 242 | 258 |
| **Glycocaulis alkaliphilus** | 233940 | 6 | 210 |
| **Alteromonas mediterranea U4** | 214386 | 259 | 281 |
| **Methylophaga frappieri** | 214225 | 24 | 136 |
| **Alteromonas mediterranea MED64** | 206153 | 244 | 265 |

Most of the taxa the unmapped reads map to are Alteromonas which seems to indicate that our mapping to Alteromonas strains may be improved (the reads contain unmapped reads from MIT0604 axenic, MIT0604+DE, MIT0604+1A3).

Strains on the list that are not Alteromonas are highlighted in yellow.

The most common strain is *Maricaulis maris MCS10*. I downloaded it’s gb file and tried to map the unmapped reads to it. the number of reads mapped was very low (a few thousands). It seems that this is not the source of the unmapped reads.

## Unmapped reads for MIT0604 axenic

Next, I’ve assembled the unmapped reads for the axenic MIT0604 (sample 4) and eun blastn on the result vs NCBI nt db (downloaded on 6/7/2020).

blast command:

blastn -db nt -query contigs.fasta -evalue 1e-06 -num\_threads 20 -num\_alignments 10 -num\_descriptions 10 -out unmapped.axenic.nt.blastn -outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore staxids sscinames"

Kraken + Bracken results on axenic MIT0604 (sample 4):

# Unmapped second strategy

Following the suggestion from Green, Stefan:

Any chance it could be a bacteriophage? I would recommend performing a de novo assembly both on all the data and also on the unmapped data. It would be interesting to look at how the data plot (we plot summed assembly length vs average coverage when sorted from highest coverage to lowest).

|  |  |
| --- | --- |
|  |  |

Blasting the top coverage clusters vs nr:

selected the top coverage contigs:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Max coverage | Min coverage | Number of contigs |
| Unmapped reads | 7881 | 50 | 8 |
| Axenic unmapped reads | 553 | 20 | 5 |

./select\_top\_coverage\_contigs.py --inf dnaseq\_adapter\_only/spades\_unmapped\_MIT0604/contigs.fasta --mincov 50

./select\_top\_coverage\_contigs.py --in ../dnaseq\_adapter\_only/spades\_unmapped\_MIT0604\_axenic/contigs.fasta --mincov 20

And run blastx on them:

blastx -db nr -query ${contigfa} -evalue 1e-06 -num\_threads 20 -num\_alignments 10 -num\_descriptions 10 -out ${contigfa}.nr.txt -outfmt '6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore staxids sscinames'

Results:

All MIT0604 unmapped reads

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Node** | **Length** | **Coverage** | **Subject** | **Accession** | **Qcov** | **Ident** |
| 2792 | 128 | 7881 |  |  |  |  |
| 2781 | 405 | 209.223 | transposase [Palleronia rufa] | WP\_161785528 | 0.25 | 0.21 |
| 483 | 1652 | 71.37639 | IS1380 family transposase [Marinobacter sp. B9-2] | PTB93145 | 0.28 | 0.28 |
| 2783 | 349 | 66.88739 |  |  |  |  |
| 2793 | 128 | 64 |  |  |  |  |
| 647 | 1283 | 54.32959 | CBS domain-containing protein [Cyanothece sp. SIO1E1] | NET30525 | 0.04 | 0.03 |
| 2790 | 200 | 53.53425 |  |  |  |  |
| 2006 | 538 | 51.72993 | hypothetical protein A9Q78\_09185, partial [Methylophaga sp. 41\_12\_T18] | OUR71561 | 0.18 | 0.14 |