Ecoli strain identification

Chunyu Zhao

December 29, 2018

Refseq and ATCC E.coli genomes

There are 24 E coli genomes in ATCC database.

- read in assembly_summary.txt

metadata

In this report, we looked at the `meconium` samples from IGRAM run1, and identified the strain of E.coli (Escherichia coli) from the annotation of de novo assembled contigs.

Conclusion: the most abundance ATCC E. coli strain found in the samples is `ATCC 11775`, and the second candidate is `ATCC 25922`.

read in Sunbeam results

read in metaphlan2 results

Annotation of De novo assembled contigs

Since there are some assemblies dataset in the refseq genome, thus we filtered out contigs (annotated as E. coli) shorter than 10 kbps.

- read in ecoli_annotation_data.rda

Then we mapped the reads back to each contigs (kepp all multiple alignment) by bowtie2 and collected the number of mapped reads for each contig.

Heatmap for number of mapped reads for each ATCC E. coli strain

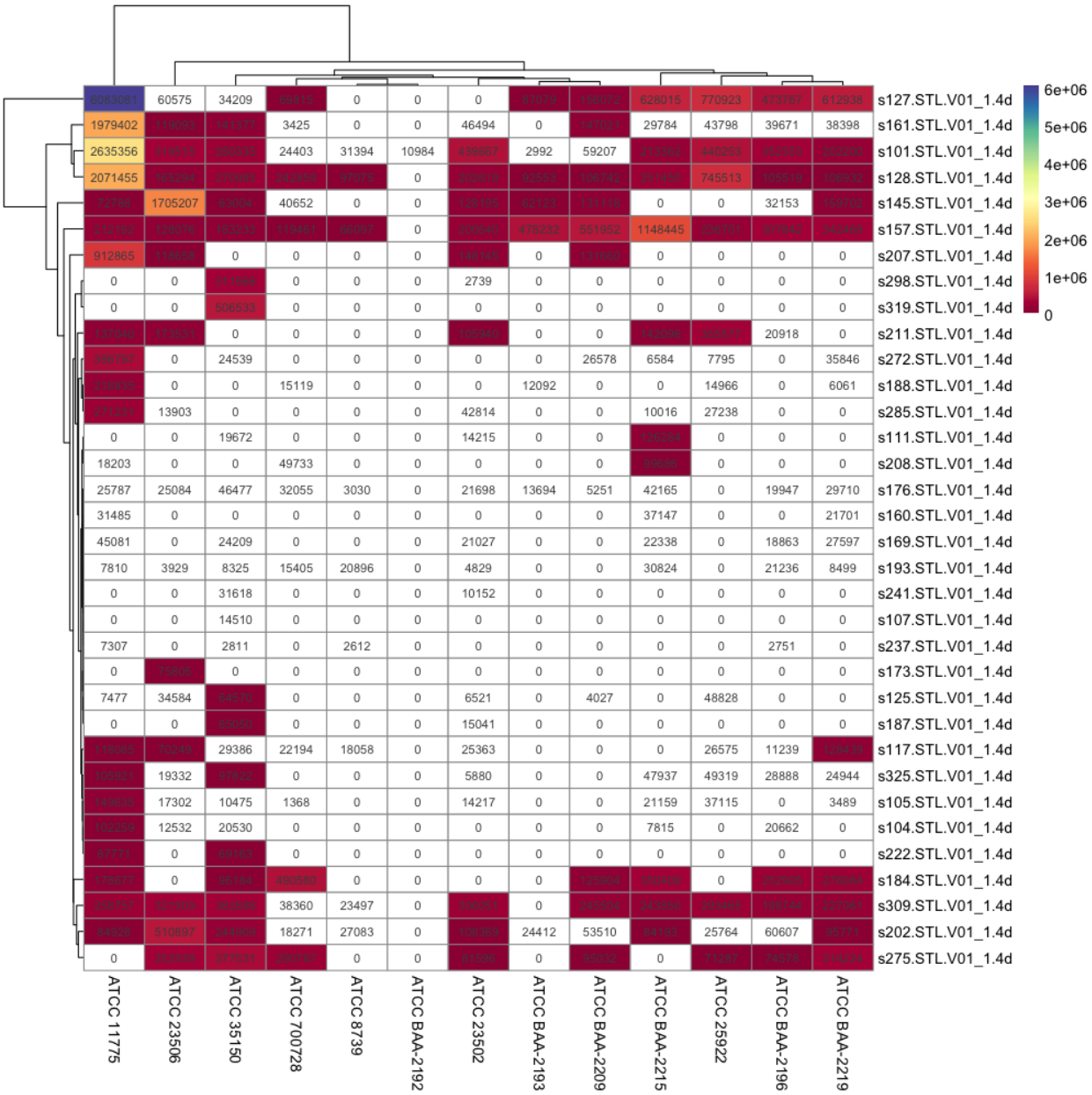
6e+06
5e+06
4e+06
3e+06
2e+06
1e+06
0
s127.STL.V01_1.4d
s161.STL.V01_1.4d
s101.STL.V01_1.4d
s128.STL.V01_1.4d
s145.STL.V01_1.4d
s157.STL.V01_1.4d
s207.STL.V01_1.4d
s298.STL.V01_1.4d
s319.STL.V01_1.4d
s211.STL.V01_1.4d
s272.STL.V01_1.4d
s188.STL.V01_1.4d
s285.STL.V01_1.4d
s111.STL.V01_1.4d
s208.STL.V01_1.4d
s176.STL.V01_1.4d
s160.STL.V01_1.4d
s169.STL.V01_1.4d
s193.STL.V01_1.4d
s241.STL.V01_1.4d
s107.STL.V01_1.4d
s237.STL.V01_1.4d
s173.STL.V01_1.4d
s125.STL.V01_1.4d
s187.STL.V01_1.4d
s117.STL.V01_1.4d
s325.STL.V01_1.4d
s105.STL.V01_1.4d
s104.STL.V01_1.4d
s222.STL.V01_1.4d
s184.STL.V01_1.4d
s309.STL.V01_1.4d
s202.STL.V01_1.4d
s275.STL.V01_1.4d
ATCC 11775
ATCC 23506
ATCC 35150
ATCC 700728
ATCC 8739
ATCC BAA-2192
ATCC 23502
ATCC BAA-2193
ATCC BAA-2209
ATCC BAA-2215
ATCC 25922
ATCC BAA-2196
ATCC BAA-2219

Table 1: E coli reference genome information summary

assembly_accession	species_taxid	taxid	organism_name	atcc	NumSequences	MaxSeqLen	MinSeqLen	MedianSeqLen
GCF_000019385.1	562	481805	Escherichia coli ATCC 8739	ATCC 8739	1	4746218	4746218	4746218
GCF_000690815.1	562	866789	Escherichia coli DSM 30083 = JCM 1649 = ATCC 11775	ATCC 11775	2	4906844	131289	2519066
GCF_000743255.1	562	1322345	Escherichia coli ATCC 25922	ATCC 25922	3	5130767	24185	48488
GCF_000401755.1	562	1322345	Escherichia coli ATCC 25922	ATCC 25922	116	592765	213	6950
GCF_000335055.2	562	1051353	Escherichia coli ATCC 700728	ATCC 700728	131	351360	295	9728
GCF_000734955.1	562	866789	Escherichia coli DSM 30083 = JCM 1649 = ATCC 11775	ATCC 11775	135	382114	207	3820
GCF_000935075.1	562	562	Escherichia coli	ATCC 35150	143	375090	1007	3636
GCF_000333195.1	562	1268238	Escherichia coli O5:K4(L):H4 str. ATCC 23502	ATCC 23502	160	329648	203	4637
GCF_000333175.1	562	1246623	Escherichia coli O10:K5(L):H4 str. ATCC 23506	ATCC 23506	224	427054	200	1610
GCF_000934865.1	562	562	Escherichia coli	ATCC 35150	242	202085	1007	6802
GCF_000496345.2	562	1404261	Escherichia coli ATCC BAA-2192	ATCC BAA-2192	330	362226	200	1111

assembly_accession	species_taxid	taxid	organism_name	atcc	NumSequences	MaxSeqLen	MinSeqLen	MedianSeqLen
GCF_000473725.2	562	1389955	Escherichia coli ATCC 35150	ATCC 35150	358	342071	207	1212
GCF_000506845.1	562	1405294	Escherichia coli ATCC BAA-2215	ATCC BAA-2215	373	260086	203	1005
GCF_000496365.2	562	1404262	Escherichia coli ATCC BAA-2193	ATCC BAA-2193	451	286305	204	969
GCF_000508385.1	562	1405295	Escherichia coli ATCC BAA-2219	ATCC BAA-2219	495	143583	201	1128
GCF_000508365.1	562	1405292	Escherichia coli ATCC BAA-2196	ATCC BAA-2196	503	275398	201	882
GCF_000506825.1	562	1405293	Escherichia coli ATCC BAA-2209	ATCC BAA-2209	1763	256727	200	365

name conversion for the second time point

```
## # A tibble: 3 x 36
## # Groups:   SubjectID [1]
##   SampleID SampleType Secondary.Sampl~ HostSpecies SubjectID study_day
##   <chr>    <chr>      <chr>            <chr>       <chr>     <chr>
## 1 s158.ST~ STL        SWB              Human       158-2     V01_1.4d
## 2 s158.ST~ STL        SWB              Human       158-2     V02_1mo
## 3 s158.ST~ STL        SWB              Human       158-2     V02_1mo
## # ... with 30 more variables: Box.Column <int>, Box.Row <int>,
## #   dna_plate_original <int>, dna_location_original <chr>,
## #   fungal_looking_growth <lgl>, dna_concentration_original <dbl>,
## #   dna_concentration_re.extract <dbl>, reextract <lgl>,
## #   dna_location_re.extract <chr>, dna_plate_source <int>,
## #   dna_source_location_for_epmotion <chr>,
## #   dna_destination_location_for_epmotion <chr>, barcode_index_set <chr>,
## #   barcode_index_fwd <chr>, barcode_index_rev <chr>,
## #   library_concentration_ng_ul <dbl>, avg_fragment_size_bp <int>,
## #   library_nM <dbl>, pooled_vol_ul <dbl>, pool <int>, both_kept <int>,
## #   dropped <int>, false <int>, fwd_only <int>, input <int>,
## #   rev_only <int>, true <int>, low_quality <dbl>, human <dbl>,
## #   non_human <dbl>
```

- now lets check the name convention

```
##            SampleID SubjectID SampleType study_day SampleIDpart1
## 1 s281.STL.V01_1-4d       281        STL  V01_1.4d          s281
##          SampleIDre isEqual
## 1 s281.STL.V01_1.4d       0
```

name conversionn

- read in 20171120_Ecoli.txt, 20171122_Bvulgatus.txt and 20171117_Efaecalis.txt
- write samples_1month_20180110.txt
- write sample_conversion_1month_20180110.txt