Non-redundant Clustered Gene Index
Institute for Genome Sciences, UMD School of Medicine

Author: Brandi Cantarel
Version: 2.0
Effective Date:

1 Abstract

2 Introduction

This SOP describes the creation of a non-redundant catalog of bacterial genes by body site. This was done by clustering the gene predictions from the 690 samples that passed QC, using the same identity parameter used by the Metahit project[1] to cluster their human intestinal tract data.

3 Requirements

3.1 Software requirements

USEARCH (http://www.drive5.com/usearch/)[2]

4 Procedure

Clustering was performed similar to Qin et al[1], with the exception that clustering was done here using USEARCH, rather than BLAT.

Of the 764 samples that were sequenced, 690 passed quality control screens. These 690 assemblies underwent gene prediction and annotation to generate the HMP Gene Index (available for download at http://hmpdacc.org/HMGI/). The genes for all samples were pooled by body site and clustered using usearch (http://www.drive5.com/usearch/) to generate non-redundant catalog of bacterial genes by body site. The HMP Clustered Gene Indices are available for download at http://hmpdacc.org/HMGC/.

USEARCH v.3.0.627 (64-bit) was used. Specifically, we used the UCLUST algorithm which is part of the USEARCH package. See http://drive5.com/uclust/uclust_userguide_3_0.pdf for more details.

4.1 Sorting

UCLUST requires that the sequences are sorted by decreasing length before clustering:

```
uclust --sort seqs.fasta --output seqs_sorted.fasta
```

4.2 Clustering

```
uclust --input seqs_sorted.fasta --id 0.95 --uc results.uc
```

Retrieve fasta file from uc file:

```
uclust --uc2fasta results.uc --input seqs_sorted.fasta --output results.fasta --types S
```

Non-redundant Clustered Gene Index
Institute for Genome Sciences, UMD School of Medicine

Author: Brandi Cantarel
Version: 2.0
Effective Date:

5 Implementation

6 Discussion

7 Related Documents & References

[1] Qin, J. et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature;464(7285):59-65.
[2]Edgar, RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics; 26(19):2460-1.

8 Revision History

Version	Author/Reviewer	Date	Change Made
1.0	Brandi Cantarel	11/20/2011	Establish SOP
2.0	Kemi Abolude	07/02/2013	Included more details; updated usearch version used for clustering rerun