Script R16: CRISPR Analysis

HANNIGAN GD, GRICE EA, ET AL.

Abstract

This protocol outlines how to generate tables to analyze CRISPR interactions. Based on methods from the following publication:

Hannigan, Geoffrey D., et al. "The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome." *mBio* 6.5 (2015): e01578-15.

Citation: HANNIGAN GD, GRICE EA, ET AL. Script R16: CRISPR Analysis. **protocols.io** dx.doi.org/10.17504/protocols.io.ejqbcmw
Published: 10 Mar 2016

Guidelines

sessionInfo()

```
## R version 3.2.0 (2015-04-16)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.4 (Yosemite)
## ## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats graphics grDevices utils datasets methods base
##
## loaded via a namespace (and not attached):
## [1] magrittr_1.5    formatR_1.2    tools_3.2.0    htmltools_0.2.6
## [5] yaml_2.1.13    stringi_0.4-1    rmarkdown_0.7    knitr_1.10.5
## [9] stringr_1.0.0    digest_0.6.8    evaluate_0.7
```

Before start

Supplemental information available at:

https://figshare.com/articles/The_Human_Skin_dsDNA_Virome_Topographical_and_Temporal_Diversity_Genetic_Enrichment_and_Dynamic_Associations_with_the_Host_Microbiome/1281248

Protocol

Step 1.
Load the required libraries.

cmd COMMAND
```
library(plyr)
packageVersion("plyr")
```

EXPECTED RESULTS

```
## [1] '1.8.2'
```

Step 2.
Read in metadata for skinmet and virome samples.

cmd COMMAND
```
metadata<-
read.delim("../../IntermediateOutput/Mapping_files/SkinMet_and_Virome_001_metadata.tsv")
meta<-
metadata[,c("NexteraXT_SampleID","NexteraXT_Virome_SampleID","SubjectID","Site_Symbol","Sit
e_Categories","TimePoint")]
colnames(meta)[1]<-"SkinMetSampleID"
colnames(meta)[2]<-"ViralSampleID"
```

Step 3.
Next, we want to match spacers from the metagenome samples that have hits to our viral contigs. First read in information about which spacer match which viral contigs.

cmd COMMAND
```
viral_contig_hits<-
read.delim("../../IntermediateOutput/CRISPRs/all_viral_hits.txt",header=FALSE)
colnames(viral_contig_hits)<-c("Spacer", "ViralContig")
viral_contig_hits$ViralSampleID<-gsub(viral_contig_hits$ViralContig, pattern="contig-
[0-9]*_[0-9]*_nucleotides_MG",replacement="MG")
```

Step 4.
Read in the TaxonID for the viral contigs that contain spacers.

cmd COMMAND
```
viral_hits_taxonomy<-
read.delim("../../IntermediateOutput/CRISPRs/viral_hit_seqs_blastn_formatted.txt",header=FA
LSE)
colnames(viral_hits_taxonomy)<-c("ViralContig", "TaxonID")
viral_contig_hits2<-merge(viral_contig_hits, viral_hits_taxonomy, by="ViralContig")
```

Step 5.
Merge viral contig hits with metadata.

cmd COMMAND
```
viral_contig_hits_data<-merge(viral_contig_hits2, meta, by="ViralSampleID", all=FALSE)
viral_contig_hits_data$SkinMetSampleID<-NULL
colnames(viral_contig_hits_data)<-
c("ViralSampleID","ViralContig","Spacer","ViralTaxonID","ViralSubjectID","ViralSite_Symbol"
,"ViralSite_Categories","ViralTimePoint")
```

Step 6.
Remove nose samples that we do not use in this study.

cmd COMMAND
```
viral_contig_hits_data2<-
subset(viral_contig_hits_data, viral_contig_hits_data$ViralSite_Symbol != "No")
```

Step 7.
Then, we determine which metagenome contigs the spacers came from and categorize them taxonomically. Read in skinmet read taxonomy assignments.

cmd COMMAND
```
read_assignment<-
read.delim("../../IntermediateOutput/CRISPRs/read_assignment_blastn.txt",header=FALSE)
colnames(read_assignment)<-c("Repeat","SkinMetContig","TaxonID")
read_assignment$Repeat<-
gsub(read_assignment$Repeat, pattern="_blastn_formatted.txt",replacement="")
read_assignment$SkinMetSampleID<-gsub(read_assignment$SkinMetContig, pattern="contig-
[0-9]*_[0-9]*_nucleotides_MG",replacement="MG")
```
Step 8.
Merge skinmet taxonomy assignments with metadata.

cmd COMMAND
```
read_assignment_data<-merge(read_assignment, meta, by="SkinMetSampleID")
read_assignment_data$ViralSampleID<-NULL
```
Step 9.
Read in file telling which repeat/spacers are found in which skinmet contigs.

cmd COMMAND
```
read_spacers<-
read.delim("../../IntermediateOutput/CRISPRs/read_spacer_assignment.txt",header=FALSE)
colnames(read_spacers)<-c("Repeat","SkinMetContig")
read_spacers$Spacer<-gsub(read_spacers$Repeat, pattern="repeat_[0-9]*_",replacement="")
read_spacers$Repeat<-gsub(read_spacers$Repeat, pattern="_spacer_[0-9]*",replacement="")
read_spacers$SkinMetSampleID<-gsub(read_spacers$SkinMetContig, pattern="contig-
[0-9]*_[0-9]*_nucleotides_MG",replacement="MG")
```
Step 10.
Merge taxonomy assignments with skinmet contig information.

cmd COMMAND
```
skinmet_data<-
merge(read_spacers, read_assignment_data, by=c("Repeat","SkinMetContig","SkinMetSampleID"))
colnames(skinmet_data)<-
c("Repeat","SkinMetContig","SkinMetSampleID","Spacer","SkinMetTaxonID","SkinMetSubjectID","
SkinMetSite_Symbol","SkinMetSite_Categories","SkinMetTimePoint")
```
Step 11.
Lastly, we merge the metagenome and the viral hit information to get a final informative table.

cmd COMMAND
```
summary<-merge(skinmet_data, viral_contig_hits_data2, by=c("Spacer"),all=FALSE)
write.table(summary,"../../IntermediateOutput/CRISPRs/crispr_hits.txt",row.names=FALSE, quo
te=FALSE, eol="\r\n", sep="\t")
```