Differential expression of RNA-Seq anaylsis for killing assay of mouse-passaged *Vibrio cholerae*

Rui Liu
12/14/2018
GEN220

Key: RNA-seq pipeline was performed, and differential expression data was filtered with the small tool that I generated named diffexpression_filter.py. filtered genes were mapped to pathway using KEGG to visualize.

Introduction

Vibrio cholerae is a enteropathogen, producing cholera toxin and causing watery diarrhea. Since the year 2000 through the year 2016, countries reporting to the WHO have recorded a total of 3.4 million cholera cases and 65,600 deaths[1]. We are interested in the factors that could induce a higher colonization in host and a stronger toxicity.

Previous studies showed that host colonization could create a hyperinfectious bacterial state that is maintained after dissemination, in both human[2] and infant mouse model[3]. We want to know whether human gut microbiome would affect the colonization and gene expression of V. cholera, and whether different gut environment would affect host health, since gut microbiome plays an important role in host-microbe and microbe-microbe interaction.

Methods

Based on human gut microbiota involved in V. cholera infection[4], we designed four artificial communities, namely, simple-healthy (SH), simple-unhealthy (SunH), complex healthy (CH) and complex-unhealthy (CunH). 4-day old infant mice were treated with antibiotic and the next day they were gavaged with community+*V. cholera*. 6-day old mice were sacrificed and the whole intestine was homogenized

(Figure 1).

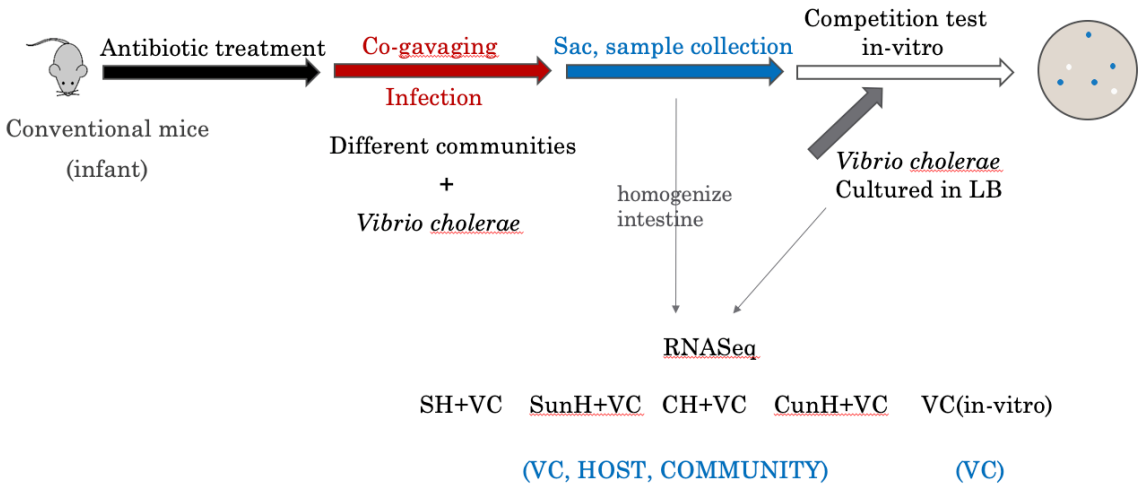

Figure 1. workflow of mice-passaged *V. cholerae* and sample collection

The killing assay showed strong phenotype through blue-white screening result. We were wondering whether the RNA-Seq data shows similar trend. We got 4 group of in-vivo RNA-seq data, including sequences from *V. cholerae*, mouse and community. To figure out why host-passaged *V. cholerae* can outcompete with the in-vitro, we set the group of LB-culture *V. cholerae* as control. For this experiment, our aim is to find some potential targets and then verify them with experiment, like gene knock out and some specific phenotype observation, eg, motility.

The data I got back from sequencing center was demultiplexed and quality-checked, therefore I started the pipeline with alignment using hisat2[5] in Linux, then transferred sam file into sorted bam file using samtools in Linux. Reads counting and gene differential expression analysis were performed in R, using summarizeOverlaps[6] and edgeR-glm model[7]. Down-stream analysis for KEGG and heatmap/plot/histogram will be performed for visualization. KEGG id for each gene was obtained using KEGGREST package in R.

For this course project, I focused on how to find out the targets. First, I had read

some related articles and have the gene lists, like genes in type 6 secretion system pathway, or chemotaxis pathway. Second, I wanted to extract the most significantly expressed genes and put them into KEGG to see if there is any pattern. Therefore, I generated a script named diffexpression_filter.py using python. Before this, I did the selection using vlookup or manual operation in excel, which is acceptable but time-consuming and it is not convenient for sample with large number of genes to open and operate the excel on normal computer or laptop. diffexpression_filter.py improves the time and accuracy.

For this script, we can select group, gene list, FDR (and sort), fold change (and sort). Usage:

-i: input file, example: edgeR_host.txt (table obtained from edgeR analysis)

	CH-CunH_logFC	CH-CunH_logCPM	CH-CunH_LR	CH-CunH_PValue	CH-CunH_FDR	CH-SH_logFC	CH-SH_logCPM	CH-SH_LR	CH-SH_PValue	CH-SH_FDR
Aaas	-0.396238426	2.475511296	3.283033456	0.069999438	0.806271517	-0.127272681	2.475511296	0.253095352	0.614903929	0.667878763
Aacs	0.055015731	3.559268965	0.079296281	0.778253343	1	-0.107907007	3.559268965	0.245162787	0.620501894	0.673092368
Aadac	-0.058187459	4.898146949	0.021120147	0.884452128	1	1.569210095	4.898146949	12.62136222	0.000381363	0.00094271

-g: group

-p: FDR

-f: fold change. Parameter larger than 0 means up-regulated, parameter smaller than 0 means down-regulated.

-s: sort. 1: sort; 0: non-sort

-l: gene list

-o: output file name for gene list filter

Examples:

(1) Filter by group: python diffexpression_filter.py –i edgeR_vc.txt –g “SH- SunH”
Output file: SH- SunH.txt

(2) Filter by group and FDR, and sort: python diffexpression_filter.py –i edgeR_vc.txt –g “SH- SunH” –p 0.05 –s 1
Output file: SH- SunH_s_FDR0.05.txt

(3) Filter by group, FDR and fold change: python diffexpression_filter.py –i

edgeR_vc.txt –g “SH- SunH” –p 0.05 –f -2 –s 1

Output file: FC-2_SH- SunH_s_FDR0.05.txt

(4) Filter by gene list: python diffexpression_filter.py –i edgeR_vc.txt –l gene_list_chemotaxis.txt –o de_chemotaxis.txt

Output file: de_chemotaxis.txt

Results and conclusion

Taking the output file SH- SunH_s_FDR0.05.txt for example, I got the genes that significantly expressed between SH and SunH, and mapped them into pathways using KEGG mapper. It showed all the pathways that involved in this comparison (Figure 2). And clicking on the pathway that I’m interested in gave me the map. For example, chemotaxis (Figure 3), we found that some genes were up/down regulated expressed. Based on this, we can do gene knock and test the motility phenotype in the future.

For the programming part, there are many ways to improve. First, for the gene list part, certain databases can be added, like virulence factors of pathogenic bacteria database, so people can search directly without reading paper. Also, sometimes I got gene id with different versions: maybe the gene number or maybe the gene name, increasing the difficulty for searching. Second, when adding more options for one script, because there are interactions, it also increased the logical difficulty and the codes need optimize.

Figure 2. Pathway search result

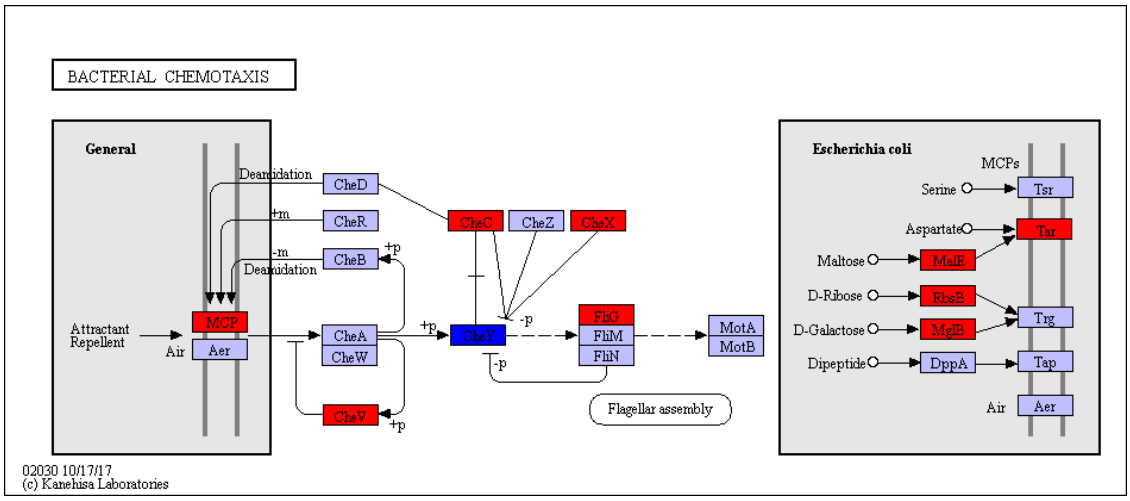

Figure 3. Bacterial chemotaxis

---

[1] https://www.who.int/gho/epidemic_diseases/cholera/epidemics/en/

[2] Merrell, D. Scott, et al. "Host-induced epidemic spread of the cholera bacterium." Nature 417.6889 (2002): 642.

[3] Alam, Ashfaqul, et al. "Hyperinfectivity of human-passaged Vibrio cholerae can be modeled by growth in the infant mouse." Infection and immunity 73.10 (2005): 6674-6679.

[4] Hsiao, Ansel, et al. "Members of the human gut microbiota involved in recovery from Vibrio cholerae infection." Nature 515.7527 (2014): 423.

[5] https://ccb.jhu.edu/software/hisat2/index.shtml

[6] http://bioconductor.org/packages/release/bioc/vignettes/GenomicAlignments/inst
/doc/summarizeOverlaps.pdf

[7] http://bioconductor.org/packages/release/bioc/html/edgeR.html