Investigating David (Caizhi) Huang’s GSA study

By Jianying Li

In respond to David’s recent observation of an interesting but intriguing phenomenon with GSA analysis (figure 1), I have conducted a thorough investigation and am reporting what has been found.

**Fact about the dataset 1 and the analysis result**

First, I found out that the initial GSA database file missed a new line character at the end of the file (figure 2).

In my hand, I could get the results like David’s except for the order of “positively enriched” gene sets (figure 3). Adding a new line character at the end of the file does not seem to change any of the results (data not shown), and then rename the last gene list, which has duplicated “name” and add “na’ as the category. The analysis result stays the same!! (data not shown).

In conclusion, “missing the new line character” does not seem to the be problem.

**Fact about the dataset 2 and the analysis result**

Apparently, the analysis on “dataset 2” was different from the dataset 1, as reported by David. A simple test confirmed that dataset2 had problem (maybe at the data generation step).

**Test 1:**

Create a gene set 1 from 20161005.gmt by removing the last gene list genset\_1\_from\_20161005.txt, as predicted, the problem are allover!!

**Test 2:**

Create a gene set 2 from 20140627.gmt by ADDING the last gene list: genset\_1\_mod\_3.txt, as predicted, NO problem here!

In conclusion, **dataset 2** was generated differently from **dataset 1**.

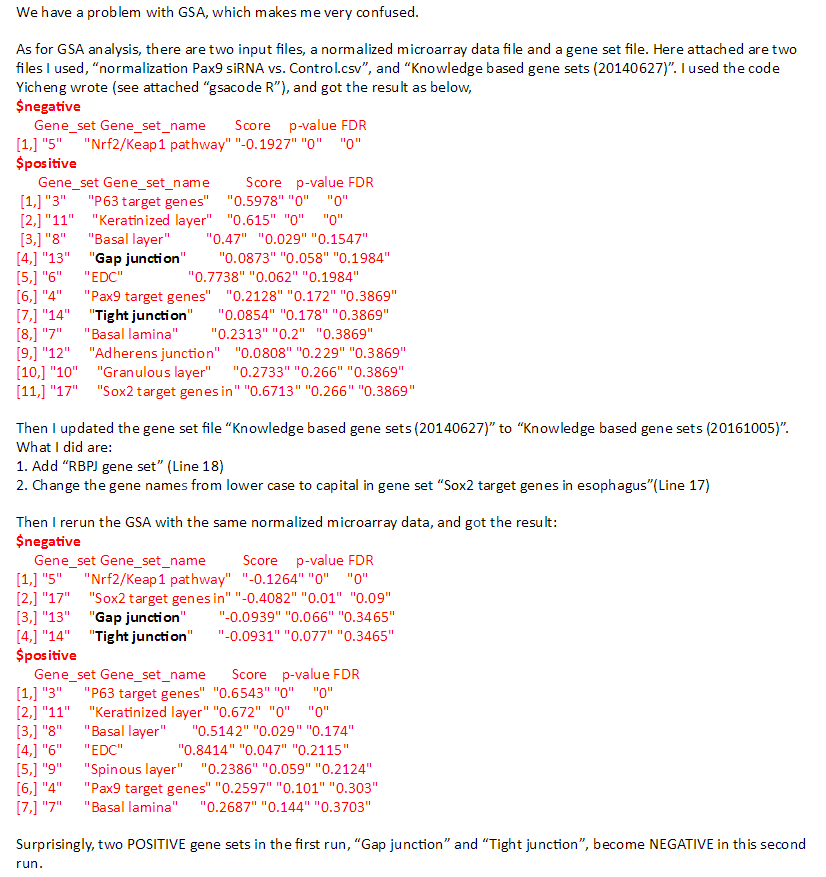


Figure 1. Different results from GSA analysis

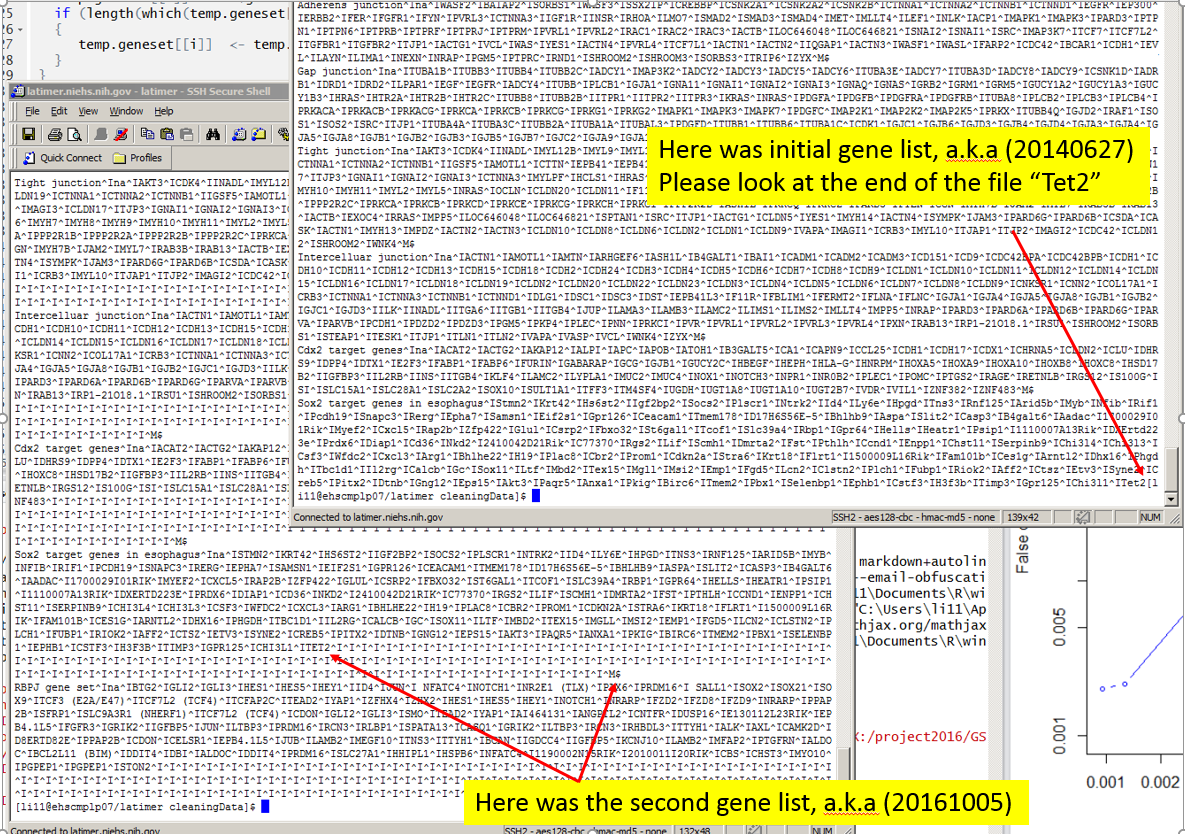


Figure 2. The first database file”xx\_20140627.gmt” misses a new line character in the end of the file.



Figure 3. Results from my analysis.