# Comparison of Annotation Tools Using ENCODE Project

Student: Yu Fu

Supervisor: Zhiping Weng

## Summary

The Encyclopedia of DNA Elements (ENCODE) project [1] makes it possible that many features, such as transcription factor binding sites, chromatin structures and histone modifications, can be pinpointed in the human genome. Against the long-gone term – “Junk DNA”, 80% of the human genome are now believed to be associated with biological functions, among which, most are found outside proteins coding regions. And the main goal of my rotation involves analyzing the data from the ENCODE project.

Genomic Regions Enrichment of Annotations Tool (GREAT) [2] is a tool for predicting functions of cis-regulatory regions. Previous methodology uses a hypergeometric test to obtain P values based on the fraction of all genes sharing a given ontology and the fractions of genes having the same ontology proximal to binding events. However, there are several inherent drawbacks and GREAT improved the current method by using binomial approach so that it can capture distal events accurately while keeping lower false positive rates.

Database for Annotation, Visualization, and Integrated Discovery (DAVID) [3] can annotate gene lists, displays the results in both texts and graphs, dynamically links to external resources and provides a broad and in-depth view of annotations.

My job in this rotation is mainly comparing the performance of DAVID and GREAT. To be specific, I first compared these two tools in the most intuitive way: selecting ChIP-seq peaks of certain TFs from the ENCODE project and trying annotating these transcription factor binding sites with both tools. Then, similar processes were automated to accelerate this process: combining datasets and comparing P values of binomial and hypergeometric tests in High Occupancy Target (HOT) regions. Finally, as previous method gives only a partial view, several heat maps were generated to provide a global landscape of P values.

## What I did:

1. I obtained the data (which are specialized bed files in BED6+4 format) that define called peaks and their signal strengths. As DAVID expects a list of genes as the input, genes that are proximal (less than 2 kbp) to the binding events were selected. Then the peaks from the same TFs were processes using both tools and the results were compared.

2. Those HOT regions were considered and TF binding sites from the same lab were used to generate a “HOTness gradient” of TFs, i.e., a series of files describing the genomic regions that are bound by a number of TFs that exceeds a certain threshold. Then these regions were processed by GREAT and results from different “HOTnesses” were compared.

3. In order to obtain a global graphical view of the how the distributions of binomial and hypergeometric P values differ, I used all those ChIP-seq data for TFs to query GREAT and plotted heat maps to visualize them.

## Results:

As more bed files are processed, using default settings, DAVID and GREAT consistently show similar results on their most significant annotations. Of all the eight datasets that GREAT and DAVID both annotated, they always shared more than half of all the GO terms that ranked top 20. A typical example would be Activating Transcription Factor 3 (ATF3). According previous research [4, 5], ATF3 can be induced by various signals, including the ones encountered by cancer cells and is involved in cellular stress response.

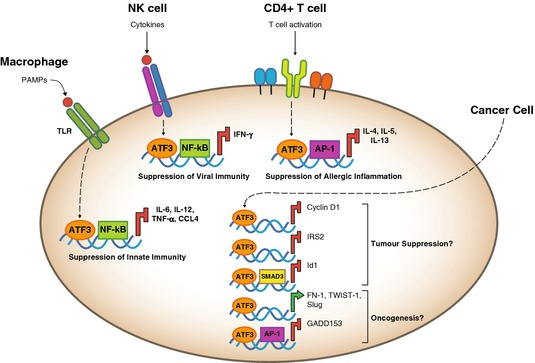


Figure 1. Roles of ATF3 in cells [4]

Previous research [5-7] also showed that ATF3 is closely related to nucleolus and acts as a hub of the adaptive-response network, so it is justifiable to observe that both DAVID and GREAT returned many terms on RNA, such as RNA processing, RNA modification, ncRNA processing, mRNA transport, rRNA processing. They also both output GO terms like lysosome and vacuole. However, in the result by GREAT, the relevant terms ranked higher, indicating that the ranking generated by GREAT contains less redundant information and shows the underestimated annotations in DAVID. All the annotations are available at <https://github.com/yfu/rotation1>.

2. Since GREAT has a better performance and it can calculate both binomial and hypergeometric P values, I used GREAT to obtain those two values under conditions of different HOTnesses. When GREAT was given a series of inputs that have different HOTnesses as the threshold, for example, peak files that contained only those regions that are covered 10, 15, 20, 25, 30, 35, 40 times by TFs, the two P values behaved differently sometimes and similarly in other cases.

One possible scenario is that the binomial P value stayed consistently high till a high HOTness point while the hypergeometric P value tended to decrease dramatically with the increase of HOTness. This demonstrates that binomial P values may remain significant even when only those most common regions shared by TFs are kept.

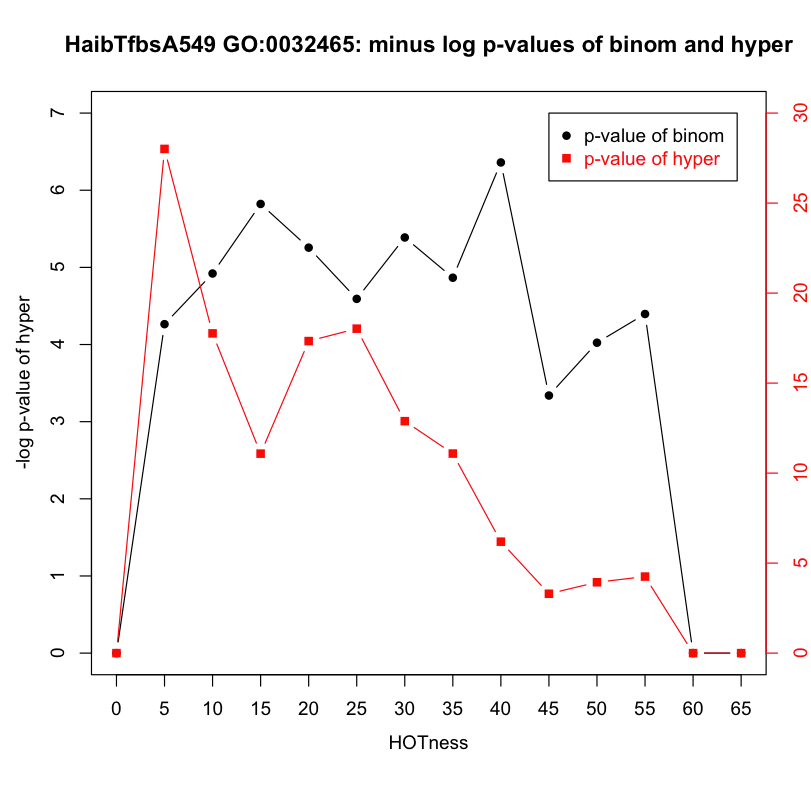


Figure 2. Two lines show the trends of binomial and hypergeometric P values.

In other cases, the two P values had almost the same trend, shown in Figure 3

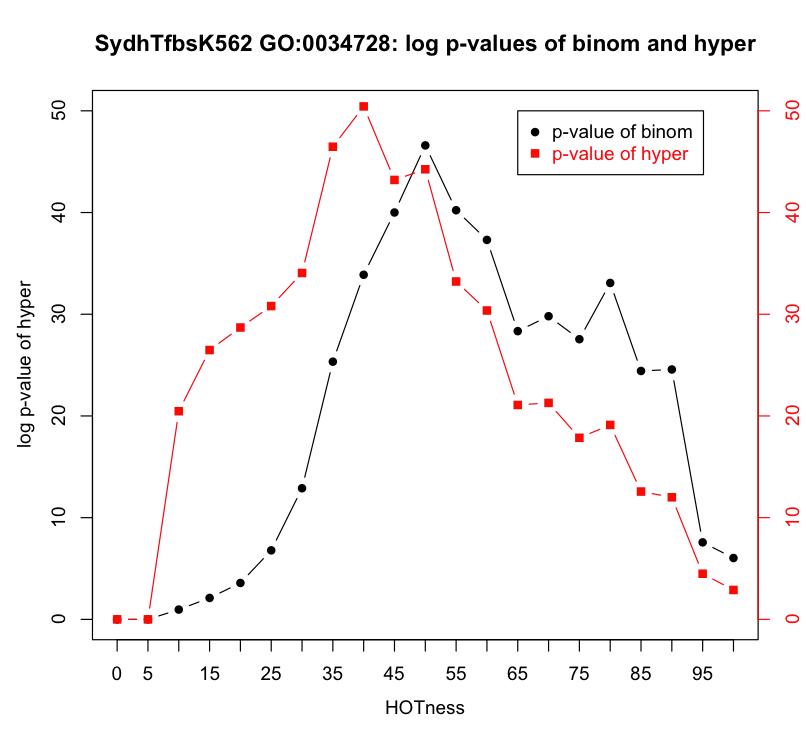
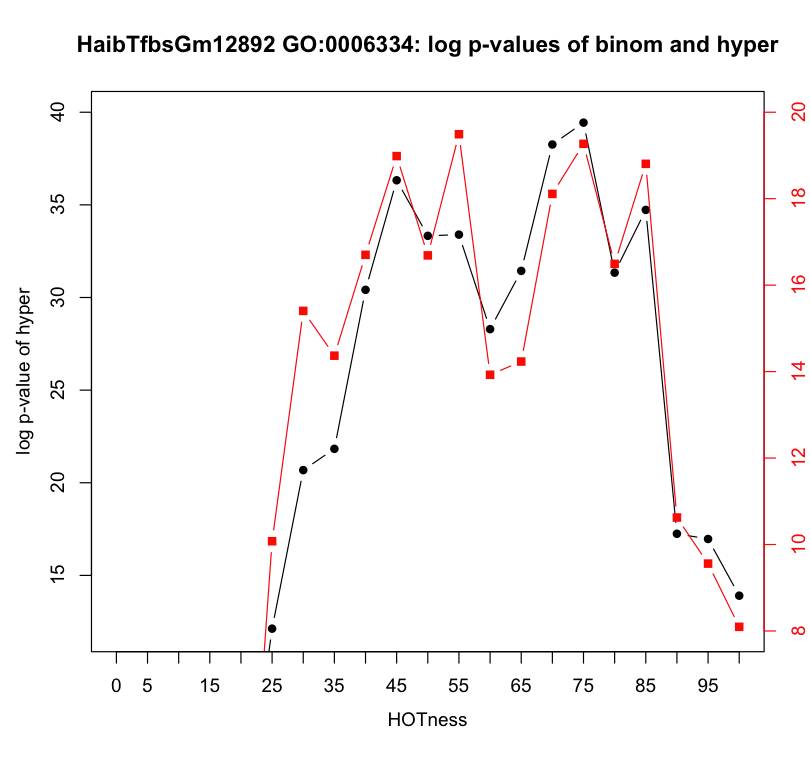


Figure 3. Two examples showing that binomial and hypergeometric P values share the same trend

The trends may be GO term specific as it is impossible to plot such graphs for all GO terms (some GO terms will only appear at certain HOTnesses and for each set of data, they have different GO terms enriched in GREAT).

3. Below are two heat maps showing the GO terms on the x-axis and TFs on the y-axis. The left figure indicates the binomial P values that are used by GREAT only. In this figure, GO terms can be approximately identified in three clusters. Those that have the most significant P values are the leftmost two, representing intracellular parts. The second cluster reports those general cellular component GO terms, meaning that cellular components are widely regulated by a large proportion of the TFs. The third cluster has no particular preferences towards a certain category of GO terms. When investigated vertically, similar cell lines often appear to remain together, suggesting that the same cell line, even under different experimental conditions, may contain TF binding sites that are closely correlated.

The figure on the right provides a visualization of hypergeometric P values, which is a criterion that both DAVID and GREAT use to filter and rank terms in results. When investigated horizontally, the GO terms can be clusters into three groups. The terms in first group reveals that terms on organelles are the most significant. The second cluster mainly contains terms on cellular component. The terms in the third cluster do not have a common theme but they largely overlap with terms in largest cluster found in the left figure.

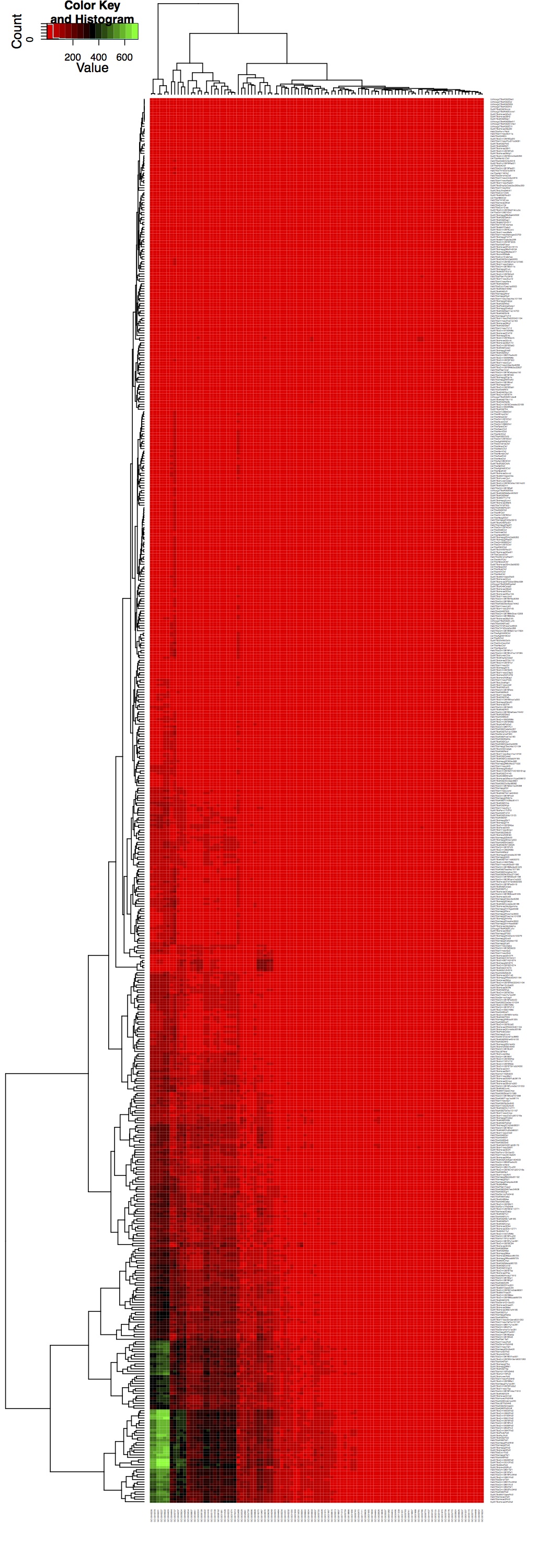
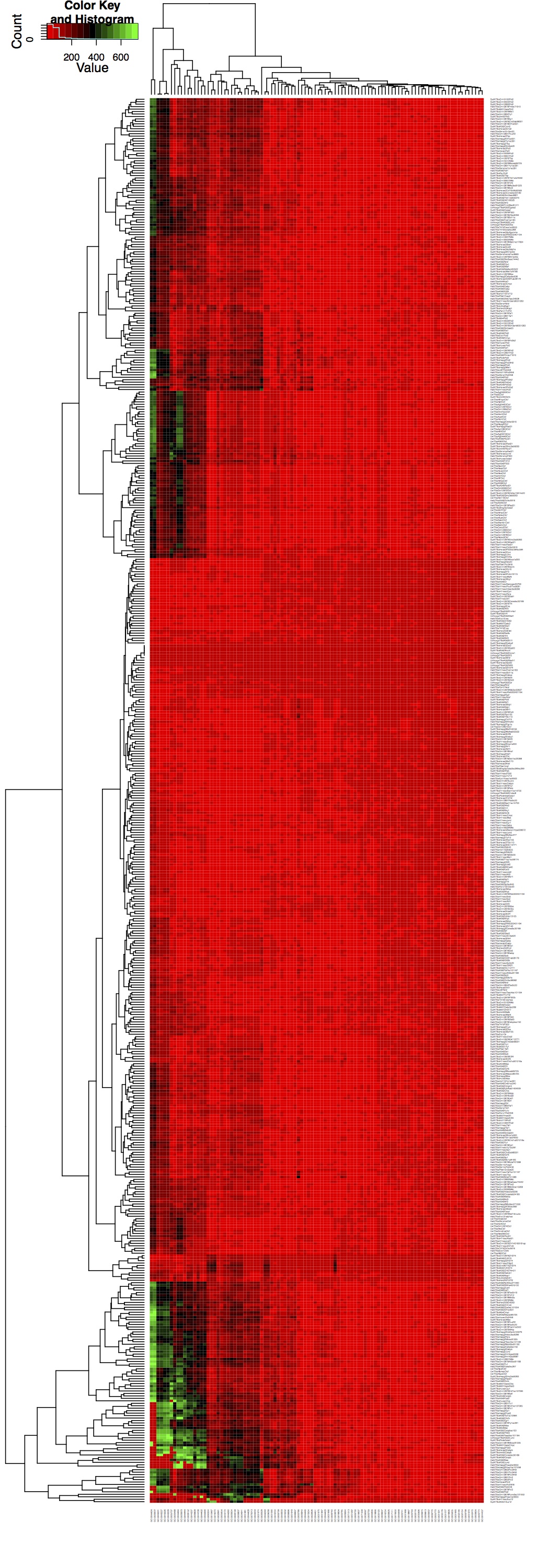


Figure 4. Heat map showing the binomial (left) and hypergeometric (right) P values of different enriched GO terms from different TFs.

When compared with each other, the first heat map apparently is “darker” than the second one, which implies that the binomial P value tends to be more significant. Also, binomial test is able to better differentiate the GO terms and cell lines. Hence the ranking by binomial test is more reliable and meaningful.

## Tools and Method:

peak2gene.py is a tool that, given a peak file, can search upstream and downstream genes according to a user-specified distance. In this rotation, a distance of 2kb is set. Peaks were processed by this tool to generate gene lists.

DAVID [3] can annotate gene lists using binomial test and mine the common biological themes shared by the genes. Gene lists generated by the above tools are fed into DAVID to interrogate its performance.

GREAT [2] is a tool used to annotate genomic regions using hypergeometric test. Corresponding peaks were input to GREAT to obtain the hypergeometric P values.

BEDTools [8] is a toolbox for operations of bed (narrowPeak) files. In this rotation, intersectBed (which is used to find overlapping features and report them appropriately), mergeBed (which is used to merge features while throwing away redundant features), sortedBed (which is used to sort the features by genomic locations) functions were used to process peak files and generate appropriate formats for downstream analysis.

great.pl and great.py are scripts that can call GREAT web API to process peak files and then parse the returned results.

gplots [9] is an R package that provides the function to plot the heat maps using hierarchical clustering.

All necessary scripts can be found online at <https://github.com/yfu/rotation1>

## What I Learned:

For the first time, I was able to follow one of the recent trends of bioinformatics, in my case, the ENCODE project. This invaluable piece of experience gave me a general idea of what research might look like from the perspective of a graduate student.

I also learned to communicate more efficiently with other people in the lab. As bioinformatics develops, it is impossible to handle all the tasks by oneself. More communication means more prior knowledge and less repeated work. Guidance from the supervisor instilled me with new ideas and corrected errors. Others’ questions and comments force me to consider problems more thoroughly.

Technically, this rotation trained me to solve more complex biological problems using computational methods. I gained a better understanding of Python, Perl, R, shell scripts and other languages, which made me more confident when faced with more programming in the future. Besides, I was more familiar with using clusters and manipulating APIs of some web services. Also, I had the chance go deep into the nitty-gritty of those annotation tools.

Reference:

[1] Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247.

An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247.

An integrated encyclopedia of DNA elements in the human genome.

[2] Nat. Biotechnol. 28(5):495-501, 2010. PMID 20436461

"GREAT improves functional interpretation of cis-regulatory regions". Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano.

[3] Genome Biol. 2003;4(5):P3. Epub 2003 Apr 3. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA.

[4] J Mol Med (Berl). Author manuscript; available in PMC 2010 November 1. ATF3 transcription factor and its emerging roles in immunity and cancer. Matthew Thompson, Dakang Xu, and Bryan R.G. Williams\*

[5] Mol Cell Proteomics. 2008 Mar;7(3):499-508. Epub 2007 Nov 19.

Toward a confocal subcellular atlas of the human proteome.

Barbe L, Lundberg E, Oksvold P, Stenius A, Lewin E, Björling E, Asplund A, Pontén F, Brismar H, Uhlén M, Andersson-Svahn H.

[6] Scand J Plast Reconstr Surg Hand Surg. 2005;39(2):65-72. Induction of activating transcription factor 3 (ATF3) by peripheral nerve compression. Isacsson A, Kanje M, Dahlin LB.

[7] Gene Expr. 2010;15(1):1-11. ATF3, a hub of the cellular adaptive-response network, in the pathogenesis of diseases: is modulation of inflammation a unifying component? Hai T, Wolford CC, Chang YS.

[8] Bioinformatics. 2011 Dec 15;27(24):3423-4. doi: 10.1093/bioinformatics/btr539. Epub 2011 Sep 23. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations.

[9] gplots: Various R programming tools for plotting data, Gregory R. Warnes. Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables, 2012, http://CRAN.R-project.org/package=gplots