Data Analysis Project

Yang Zhang

2019-12-04

# Summary/Abstract

# Introduction

Kientoplastids are protozoan parasites that could cause devastating cutaneous and subcutaneous diseases, such as Human African trypanosomiasis, Chagas disease, and the Leishmaniases, affecting over 23 million of people all over the world. However, there is no effective vaccine against these parasites, and increasing drug resistance challenges the efficacy of drug treatment. Novel therapeutic streatgies are therefore urgently needed, and their development relies on a better understanding to trypanosomatid biology. Trypanosoma brucei is one of the disease-causing parasites in this group. It has a complex life cycle involving different developmental life stages within the insect and mammalian host. In mammals, T. brucei exists in two forms: the slender form which is dividing in the mammalian host, and the stumpy form, which is a non-dividing form and is pre-adapted for transmission into teste fly host upon an inset bite. In teste fly, stumpy T. brucei differentiates into procyclic form that colinizes in insect midgut. Two populations of procyclic form are found in the midgut of teste fly, including early and late procyclic form. Early procyclic T. brucei is found around one week after transmisison into insect, and late procyclic form allows persistent infection of the midgut. After further differentiation, T. brucei cells colonize the salivary gland of the teste fly, and are ready to infect a mammalian host after the inset takes a blood meal.Given the complex life stages T. brucei has to go through, significant changes in its morphogy, energy source, and gene expression are required. Pertubation to the life cycle of T. brucei has been studied as a novel chemotherapeutic approach to kill the Parasites. In vitro drug induction of bloodstream form T. brucei differentiation into procyclic form has been shown to cause cell death at low concentration, indicating its medical potential. In contrast to many eurkryotes, yrypanosmatids, including T. brucei, have genes of unrelated function organized into gene arrays, which are transcribed from a single promoter into long polycistronic transcipriton units (PTUs). PTUs are then trans-spliced into monocistonic transcirption unit for protein translation. Cis- and trans-acting elements in regulating transcirption have been lacking. Regulation on individual gene expression has therefore been contributed largely to post-transcriptional regulation, with the exception of two epigenetic markers, base J and histone 3 variant (H3.V). Base J, β-D-glucosyl-hydroxymethyluracil, is a modified thymidine residue found in the genome of kientoplastids. It is mainly enriched in telomeric and subtelomeric regions. A minor subset of base J are also enriched in the regions between diverging and converging gene arrays, where transcription starts and terminates. Ablation of base J from genome leads to transcription termination defects, including transcription readthrough into the transcirption terminaiton region, and derepression of genes downstream of transcirption termination site that are silent in wild type cells. These evidence indicates that base J plays a role in regulating transcirption teriminaiton. Base J is developmentally-regulated DNA modification: it is present in blood stream form T. brucei, but it is absent from the procyclic form T. brucei. Taken together its role in regulating transcirption termination and therefore silencing genes downstream, a premature transcirption termination model is proposed, in which genes downstream of transcription termination sites are silent when base J is present, while they are expressed when base J is absent in anotehr T. brucei life stage, rendering developmental regulation on these genes. DNA is wrapped around core histones into nucleosomes. Histone variants and post-translational modificaitons have been shown to regulate transcirption from different aspects. The T. brucei Histone 3 variant shares only 45% sequence identity with the canonical histone 3. It is enriched in centromere, and it is found to colocalize with base J at transcirption terminaiton region. Its abaltion has a simlar effect on transcirtion as base J ablation, leading to increased readthrough transcirption into transcription terminaiton region and de-repression of genes downstream of transcription terminaiton sites. Furthermore, H3.V knockout has an additive effect to base J ablation on transcirption termination, suggesting that the two epigenetic markers may regulate transcirption synergistically. In this study, I will test the hypothesis that genes downstream of transcirption terminaiton sites within PTUs regaulted by premature transciprtion termination by base J and H3.V are developmentally related genes. To test this hypothesis, I will use the published RNA-seq results. Naguleswaran et al. performed RNA-seq analysis on T. brucei cells in four culturable life stages: slender form, stumpy, early and late procyclic form. This result allows me to analyze the gene expression profile across the four T. brucei life stages. The change in gene expression after base J and/or H3.V knockout was obtained from another study by Schulz et al. in which they sequenced poly(A)+ mRNA derived from T. brucei mutants lacking either base J or H3.V or both epigentic markers. The genes with upregulated expression in the mutant T. brucei cells are compared to the list of genes with highest expression level in each of the four T. brucei life stage to analyze whether base J and/or H3.V knockout has any effect on a particular set of genes.

#Methods and Results  
##Data qcquisition RNA-seq results were obtained from Naguleswaran et al. and Schulz et al. Both are in readable excel format and can be easily exported into R studio for further analysis.

##Data import and cleaning Find the Data cleaning code in ./code/processing\_code. Run Yang\_data\_processing.Rmd to obtain the data processing results.

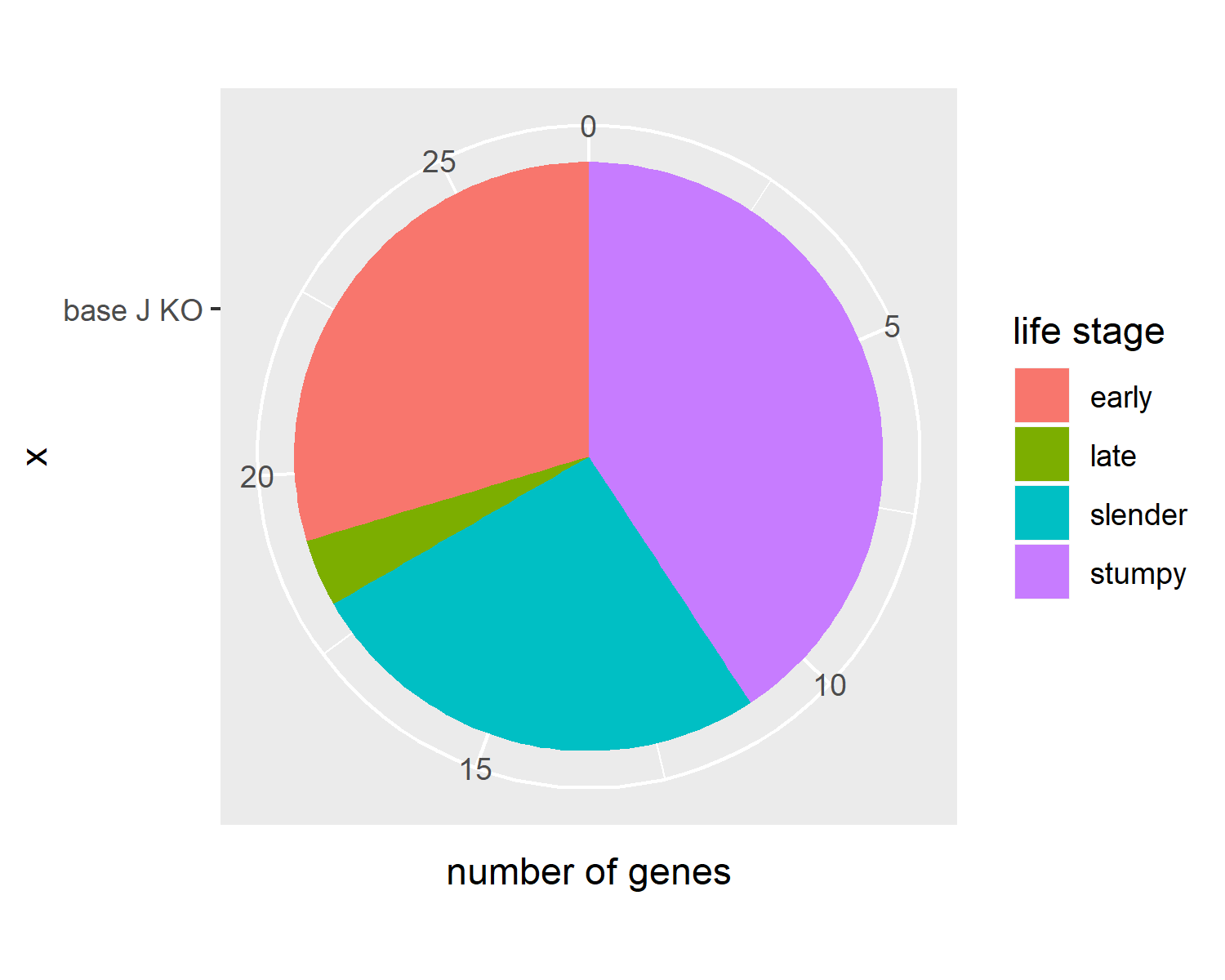
##Data analysis First, I want to know which genes have up-regulated expression in each of four T. brucei life stages. To answer this question, if a certain gene has the highest transcirpts read in one of the life stage, then I regard the gene to be over-represented in that life stage. For exmaple, for gene Tb927.4.4520, the transcripts reads for slender form, stumpy form, Early and late procyclic form are 3.77, 4.17, 3.99, and 2.06, respectivelt. This gene is upregulated in the stumpy stage according to the definition. There are 2612, 2853, 2819, and 1420 genes that have the max expression levels in slender, stumpy form, early and late procyclic forms, respectively (supplementary tables: Gene\_upregulated\_Early, Gene\_upregulated\_Late, Gene\_upregulated\_slender, \_Gene\_upregulated\_stumpy).

Table: Genes affected by Base J and/or H3.V mutants

(#tab:} {r resulttable)Result Table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | base J alone | H3V alone | either |
| number of genes | 34 | 6 | 2 |

It shows here that base J and H3.V knockout leads to upregulation of gene expression for 36 and 8 genes, respecively, with 2 genes shared between them.

I then cross-compared the genes affected by base J and/or H3.V knockout to the list of genes that have the highest expression level in each of the four stages. Figure: The distribution of genes affected by base J knockout in slender, stumpy form, early and late procyclic T. brucei.  


It shows that Base J ablation leads to upregulation of genes with a maximum of gene expression primarily in early procyclic form, slender and stump blood stream form. The majority of the genes affected have a maximum expression level in the blood stream form, which fits into our expectation. As reported before, base J regulates expression of genes downstream of transcription termination with PTU in blood stream form T. brucei, leading to the hypothesis of regulation of gene expression by premature transcription termination. After the ablaiton of base J, it is expected that genes normally downstream of chromosomal-internal base J will be de-repressed in the blood stream form.

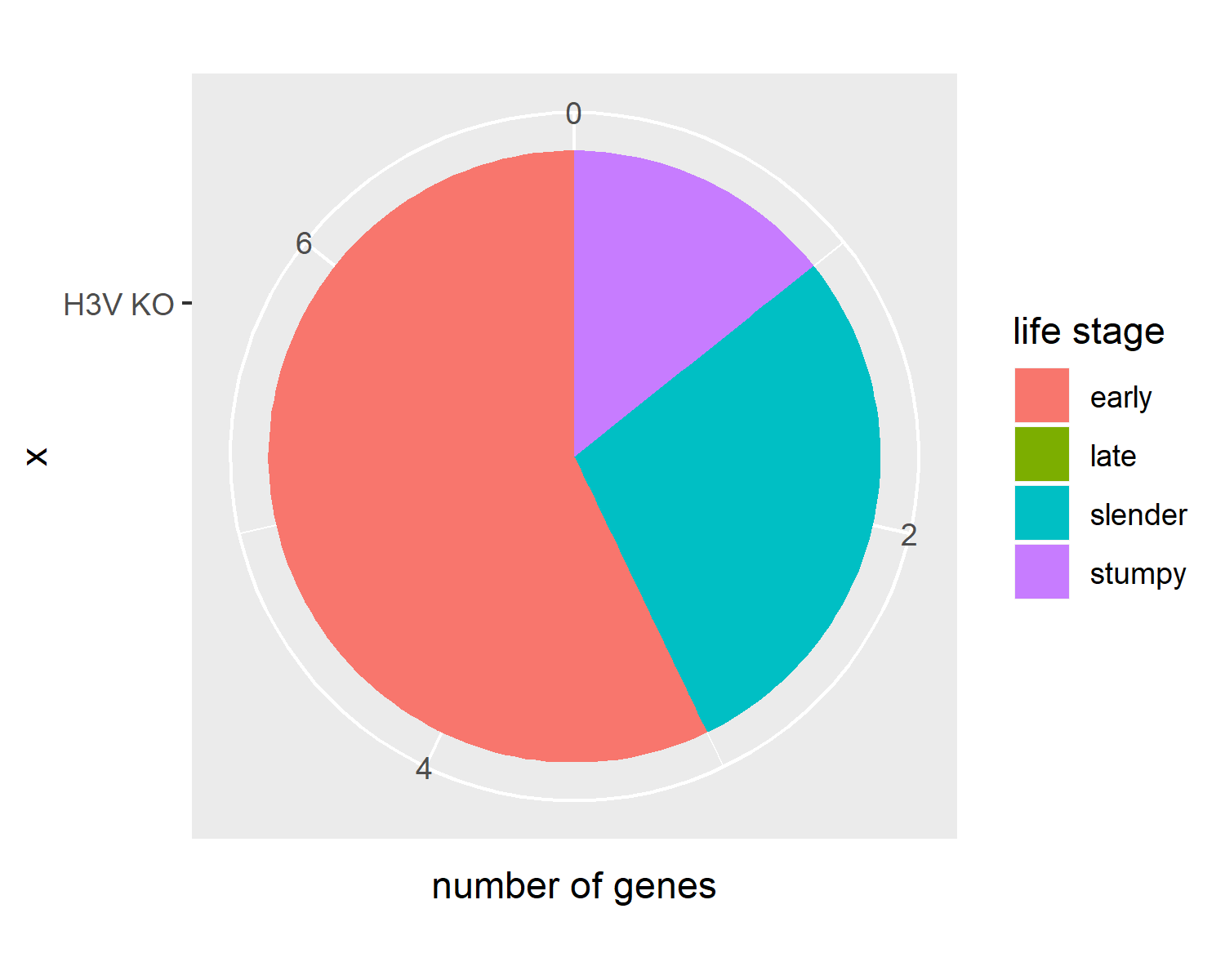
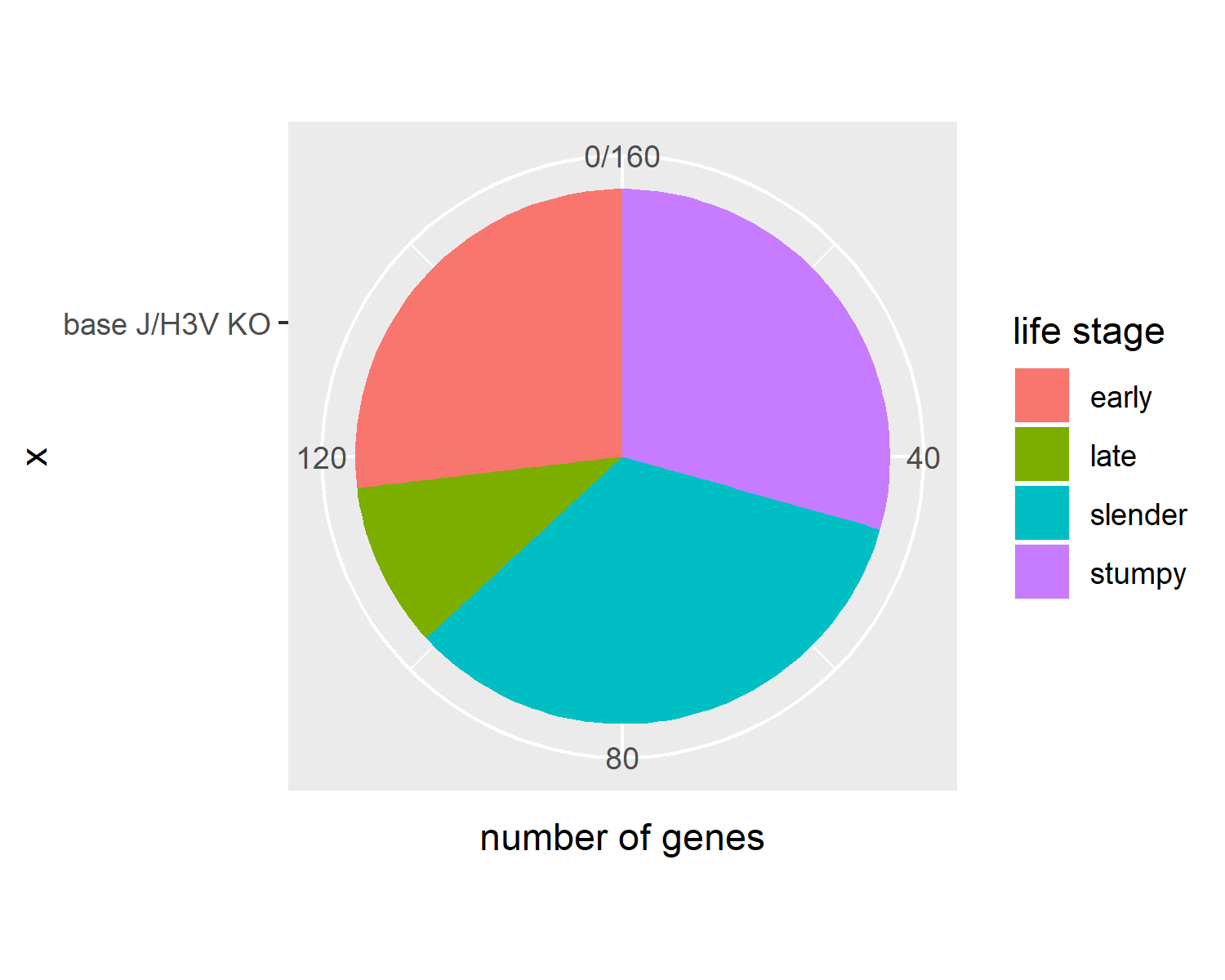
Figure: the distribution of genes affected by H3.V knockout to the list of genes that have the highest expression level in each of the four stages.  H3.V has been established to be an epigenetic marker for regulating transcirption termination. Its ablation leads to increased transcirption readthrough into the transcirption terminaiton region, and de-repression of genes downstream of base J/H3.V peaks. As shown in the figure, H3.V affected the genes primarily in early procyclic form, and blood stream form, with no gene in late procyclic form. Unlike base J, H3.V is not a developmentally regulated epigenetic marker, and the parasite will not naturally remove it in certain life stage. Therefore, it is expected that genes affected by H3.V knockout do not fall into a specific category.

Figure: the distribution of genes affected by basee J/H3.V double knockout to the list of genes that have the highest expression level in each of the four stages.  Base J and H3.V double knockout affected much more genes than either knockout alone. This is consitent with the idea that base J and H3.V act synergistically to regulate transcirption termination, and thereby silence genes dowonstream of base J/H3.V peak within PTUs. As shown in the figure, the majority of the affected genes belongs to the slender and stumpy form.

# Discussion

## Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

# References