1.Introduction

Endometrial carcinoma (EC) is the sixth-most-common cancer in women globally [1] with an estimated 61,880 new cases and 12,160 deaths in the United States in 2019 [2]. According to The Cancer Genome Atlas(TCGA), EC is generally divided into four categories: POLE(a rare ultramutated subtype),MSI(a hypermutated endometrioid subtype),CNV_LOW and CNV_HIGH (serous aggressive endometrioid cancers).For the CNV_HIGH subtype, its unexpected reoccurrence is usually deadly and bad in prognosis. Hence, the research on EC still has great practical medical significance and value.

Protein acetylation is an important component of protein post-translational modification. In general, protein acetylation (especially histone acetylation) plays an important role in regulating DNA synthesis, regulating DNA transcription, and participating in cell product metabolism in human body.

In our work, we combined EC and protein acetylation to explore protein acetylation levels in endometrial cancer and their possible influencing factors, while trying to screen the significant difference of modification levels of acetylation sites in different tumor types. **We believe the meaning behind it is real and huge.

Our study includes the following four parts:

1) Descriptive statistics of clinical characteristics and modification level of protein acetylation sites were performed on the tumor sample tissues (jointly completed by Chen Yumeng and Du Haoyu)

2) To explore the correlation between the level of key proteins and the acetylation modification level of histone acetylation sites in tumor samples (completed by Du Haoyu, Lasso regression analysis was used)

3) To explore the influence of key gene mutations in tumor samples on histone acetylation levels (completed by Wang Suran, t test was used)

4) To explore and screen histone acetylation sites in two classical tumor subtypes (CNV_LOW, CNV_HIGH) with significant differences in acetylation modification levels(completed by Chen Yumeng, rank sum test was used)

Here, we present statistical charts of clinical features of all 104 samples used (Figure 1). In Figure 1, clinical characteristics of each sample from samples S1-S104 were presented, including clinical stage, tumor subtype, etc.

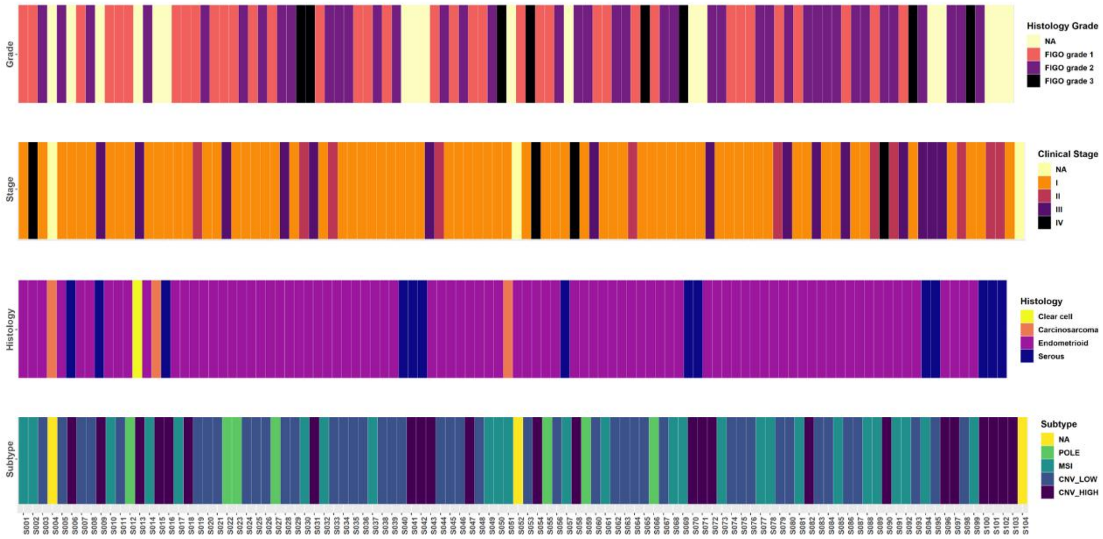

Figure 1 : Descriptive statistical charts of clinical features of each sample

We also made statistics on copy number variation (CNV) of each sample, as shown in Figure 2, which is also part of our descriptive statistical work. In fact, this is how CNV_LOW and CNV_HIGH subtypes are classified in tumor typing. We can see how each sample in Figure 1 and Figure 2 correspond.

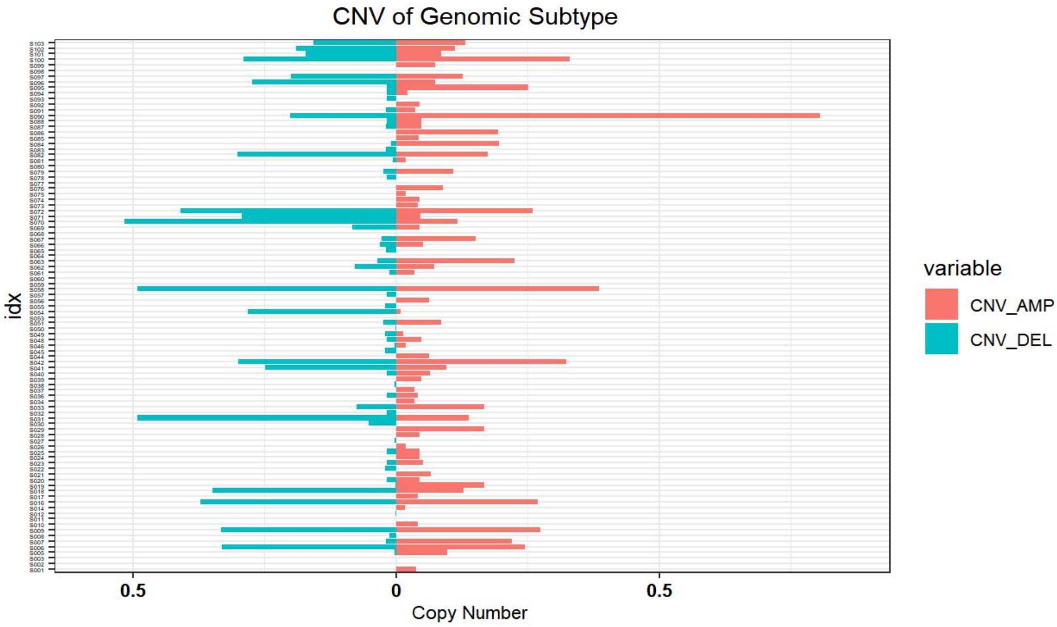

Figure 2 : Presentation of each sample’ s copy number variation

In addition, we also plotted the acetylation levels of each sample protein in Figure 3, from which it can be seen that there is a general up-regulation of the acetylation levels of the sample proteins.

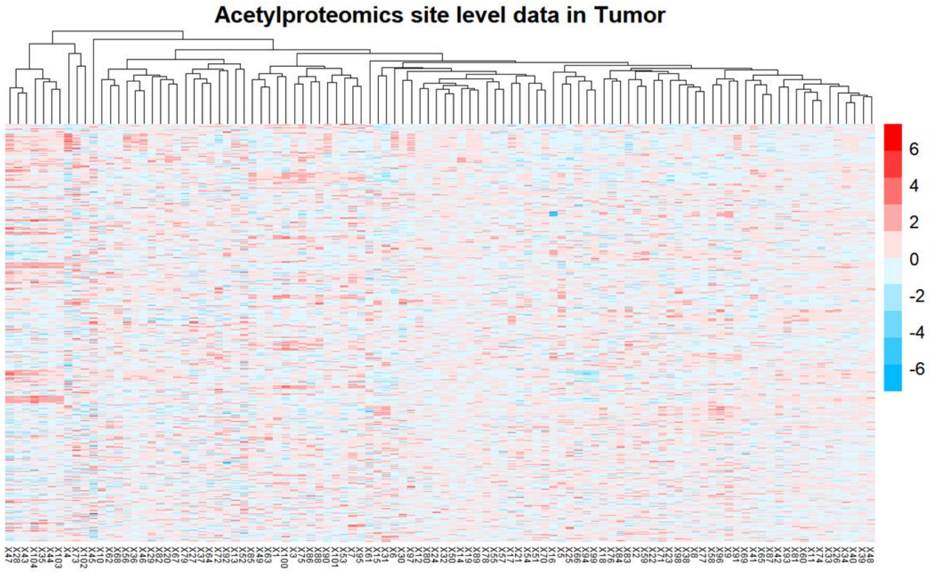

Figure 3: Diagram of protein acetylation levels in each sample

2.The relationship between histone acetylation levels and the expression levels of certain key acetylation-process enzymes and histones in tissues

From the paper and above results, we have found that the level of protein acetylation is significantly up-regulated in tumor samples Endometrial carcinoma. So what factors might be related to protein acetylation levels in endometrial cancer? We mainly considered two possible factors: the expression level of key acetylation enzymes and the mutation of key genes. It must be noted that we mainly focus on the acetylation modification of histones. Because there are many acetylation sites of histones, we believe that they can be universally representative. At the same time, the acetylation of histones has its special biological functions, including participating in DNA synthesis and promoting DNA transcription, which is of great significance.

We first focused on the possible relationship between the expression levels of key acetylation process enzymes such as acetyltransferase (HAT, etc.), deacetylase (HDAC, etc.) and histones(H2A,H2B,H3,H4) in tumor samples and the acetylation levels of histones in the samples. We therefore ask this biological question :

Whether there is a correlation between the expression level of (partial) acetylation process enzymes in tumor samples and the acetylation modification levels of (partial) histone in tumor samples.

Since this problem involves the expression levels of various acetylation enzymes and the acetylation modification levels of various histones, it is not in line with the actual situation to use simple linear models (Pearson coefficient, Spearman coefficient, etc.) for analysis. We complied with the selection in the paper and decided to use the Lasso model for regression analysis. Lasso regression is an optimized multiple linear regression model. Let y_i be the outcome and $x_i:=(x_1,x_2......,x_p)^T$ be the covariate vector for the ith case,its main objective is to make:[3]

$$\min_{\beta_0,\beta}\{\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\}\ subject\ to \sum_{j=1}^{p}|\beta_j|\leq t$$

$$\beta\in(\beta_0,\beta_1......\beta_i): Regression\ Coefficients;$$

$$t: constraint$$

Meanwhile, note that since the P-value of regression coefficients of some independent variables has been taken into account in the regression analysis process (variables whose p-values do not meet the requirements will be deleted from the resulting regression equation, and the regression coefficient will be 0), we do not give the P-value of each regression equation here.

In the regression analysis, we took the level of acetylation modification of a specific histone as the dependent variables, and the level of acetylation enzymes and histone expression as independent variables. Finally, each regression coefficient in corr.xlsx is obtained and shown in Figure 4.

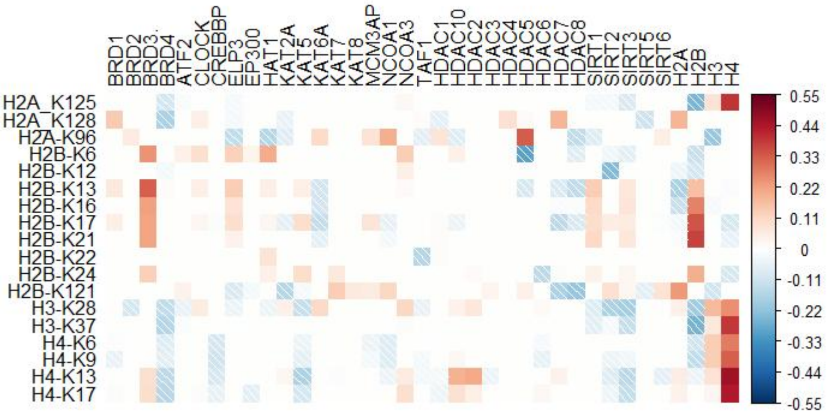

Figure 4: Regression coefficients between modification levels at each acetylation site and levels of key proteins

From the figure, we can find that the two variables with the most significant positive correlation are H4 level and

H3 acetylation level, followed by BRD3 level and H2B and H4 acetylation level. The variable pairs with significant negative correlation include H2B level and H3 acetylation level, SIRT3, BRD4, KAT5 expression level and H3 and H4 acetylation level. The correlation between histone expression level and acetylation level indicates the possible regulatory relationship between histones. Because BRD3 belongs to the bromine domain(BRD) protein, it can bind to the amino acid at the acetylation site, which may protect it from acetylation. SIRT3 is deacetylase, and the correlation between their expression level and histone acetylation level is consistent with their physiological role, which indicates that they may participate in and regulate the histone acetylation process.

Based on these results and inferences, we conclude that there is indeed a correlation between (some) acetylation enzyme expression levels and (some) sample histone acetylation levels.

3. Correlation between the levels of key proteins and acetylation modification of histone acetylation sites

The characterization of acetyl groups in cancer tissues from patients is limited. Similar to previous work in cell lines, studies have observed the enrichment of acetylated proteins in EC tumors, involving splicing, RNA transport, synthesis and degradation of precursor proteins and metabolic pathways. Histone acetylation patterns were observed to be highly heterogeneous in tumor samples, but not strongly associated with discrete genomic subtypes or clinical features. We evaluated how histone acetylation was affected by EC mutation, and found that H3 site was up-regulated in ARID1A and KRAS mutation samples, including K27 and K36. Our observation highlights the heterogeneity of acetyl groups in EC and the potential impact of SMG mutations on the acetylation level of histones.These relationships have a general impact on the specificity of EC or some gene mutations on the acetylation of histones. A similar comprehensive study of acetyl groups in other cancers is needed.[4]

When reproducing the data in the article, I first showed the quantitative characteristics of ARID1A and KRAS gene mutations and wild types in the form of pie chart.Then the acetylation levels of some protein sites of the mutant and wild type of these two genes were reproduced in the form of box diagram. Among them, ARID1A genotype shows its wild type and mutant H2A_ K95， H3_ K27 and H3_ Acetylation level of K36 protein, KRAS genotype showed its wild type and mutant H3_ K27，H3_ K36 and H4_ Acetylation level of K12 protein. After that, I used T test to test each pair of data, and finally found that it was necessary to reject the original hypothesis, that is, there were significant differences in the acetylation levels of these proteins of the wild type and mutant of these two key genes.

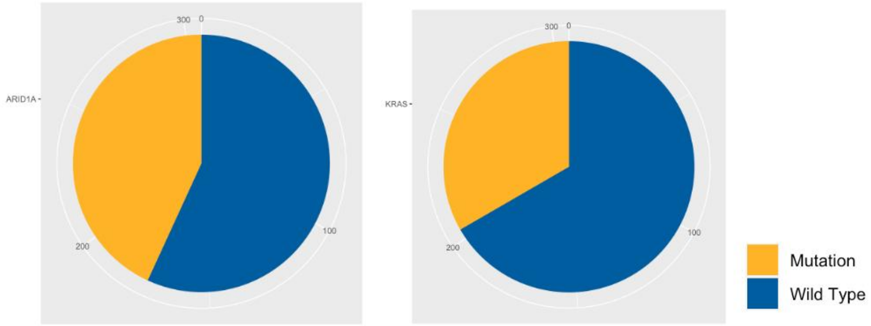

Figure 5:The quantitative characteristics of ARID1A and KRAS gene mutations and wild type.

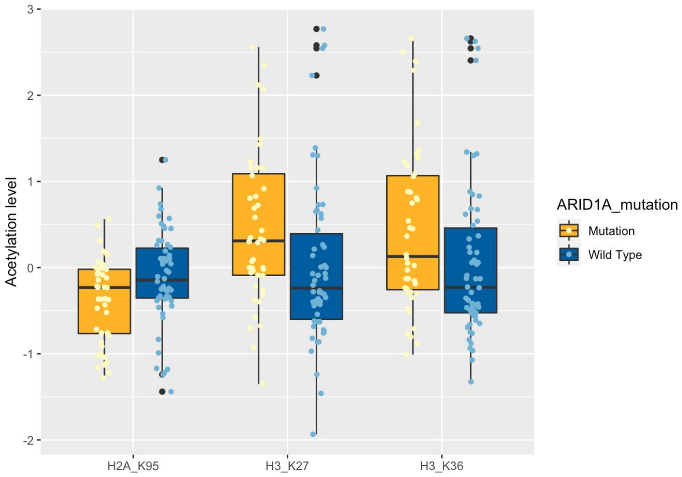

Figure 6:Acetylation level of ARID1A gene mutation and wild type

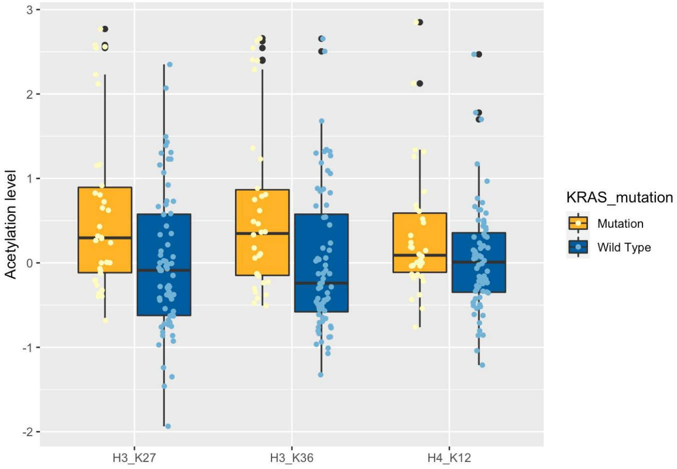

Figure 7:Acetylation level of KRAS gene mutation and wild type

4. Screen out significant histone acetylation sites

(a) Descriptive statistical analysis

According to Acetylproteomics site level data,we can draw the heatmap(Figure 3). According to the overall sample information obtained from the descriptive statistics, we can infer the basic situation of the sample from the whole that,compared to enriched normal samples, some sites in the tumor samples are significantly up or downregulated.

However, we can also see that compared to the overall number of sites, only a small number of sites have significant variation, from which we use a hypothesis testing method in order to screen out significantly differentiated site.

In the following analysis, we mainly focus on these changing sites.

(b) Biological background knowledge

According to the biological acknowledgement, tumors were classified into the four genomic subtypes outlined in the TCGA EC landmark study (Kandoth et al., 2013):POLE, MSI, CNV-low (also called endometrioid-like), or CNVhigh (also called serous-like)

Among them, CNV is one of the important criteria for classification. Copy number variation (copy number variation, CNV) refers to the increase or decrease of copy number of some large segments of the genome, which can be divided into two types: deletion (deletion) and repeat (duplication). CNV is a kind of genomic structural variation that can regulate organism plasticity by changing gene dosage and transcriptional structure, and it is one of the main genetic bases of individual phenotypic diversity and population adaptive evolution. Abnormal DNA copy number change (CNV) is an important molecular mechanism in many human diseases (e.g., cancer, hereditary diseases, cardiovascular diseases).

Among the four types of tumors, the CNV-high subtype is more more aggressive. Its unexpected reoccurrence is usually deadly and bad in prognosis. According to the research, Histone acetylation has a close relationship with the course and prognosis of cancer.

Therefore, we wished to screen out important histone acetylation sites that were significantly up or down regulated in the two subtype: CNV-High and CNV-low.

(c) Hypothesis Test Analysis

Due to the large gap in gene expression between different samples, there is no guarantee that the samples are from the normal population. So we use Wilcoxon Rank Sum Test (Benjamini Hochberg (BH) correction, FDR <0.1) between two samples (CNV-High, CNV-Low). For each set of samples:

H_0: the acetylprotein site level data of CNV-High and CNV-Low subtype have Identical Distribution.

log2 (FC) (Fold Change) is used as the abscissa. We identified 20 downregulated sites and 7 upregulated sites in the CNV-low subtype compared to the CNV-high subtype (Figure 8).

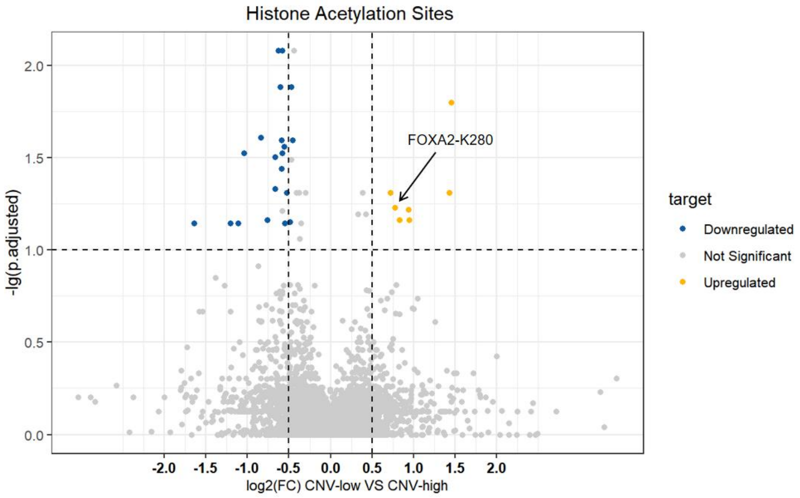

Figure 8:Wilcox Sum Test between CNV-High and CNV-Low(BH,FDR<0.1)

(d) Discussion

One upregulated site, FOXA2-K274, has been found(Figure). FOXA2 itself has been linked to increased cell proliferation and invasion in colon cancer (Wang et al.,2018). Hence, the increased FOXA2 acetylation could indicate improved stability and activity of the protein.Thus, reduced levels of FOXA2 acetylation may be responsible for the high aggressiveness of CNV-high EC tumors. Therefore, we were able to statistically screen important loci that could be further investigated by our hypothesis testing. So further work is required to define the role of FOXA2 acetylation in EC.

According to Figure, we can further identify other sites with large differences in acetylin expression in CNV-high, CNV-Low, which can be further investigated for the proteins compiled at this site, and we can use analytical methods to further explore the role of FOXA2 on cancer progression.We can use FOXA2 as an important drug target for future treatment using Survival analysis studies.

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 68, 394–424.

[2] Siegel, R.L., Miller, K.D., and Jemal, A. (2019). Cancer statistics, 2019. CA Cancer J. Clin. 69, 7–34.

[3] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

[4]Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V., and Mann, M. (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science 325, 834–840