Acetylation Profiles of Histone and Non-Histone Proteins in Breast Cancer

Alla Karpova

Follow this and additional works at: https://openscholarship.wustl.edu/eng_etds

Part of the Computational Biology Commons, Molecular, Cellular, and Tissue Engineering Commons, and the Systems and Integrative Engineering Commons

Recommended Citation
https://openscholarship.wustl.edu/eng_etds/433

This Thesis is brought to you for free and open access by the Engineering and Applied Science at Washington University Open Scholarship. It has been accepted for inclusion in Engineering and Applied Science Theses & Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
Acetylation Profiles of Histone and Non-Histone Proteins in Breast Cancer

by

Alla Karpova

A thesis presented to the School of Engineering of Washington University in St. Louis in partial fulfillment of the requirements for the degree of Master of Science

December 2018

Saint Louis, Missouri
Contents

Contents............................................................................................................................... ii
List of Figures............................................................................................................................ iv
List of Tables.............................................................................................................................. vii
Acknowledgments ....................................................................................................................... viii
ABSTRACT OF THE THESIS...................................................................................................... ix

1 Background........................................................................................................................... 1
  1.1 Histone Acetyltransferases and Deacetylases ................................................................. 1
  1.2 Histone Acetylation ........................................................................................................ 5
  1.3 Alterations in Acetylation in Cancer .............................................................................. 7
    1.3.1 Wrong Histone Acetyltransferases ......................................................................... 7
    1.3.2 Wrong Histone Deacetylases .................................................................................. 9
    1.3.3 Histone Acetylation in Cancer .............................................................................. 10
    1.3.4 Non-histone Proteins Acetylation: Links to Cancer .............................................. 11

2 Research Methods ............................................................................................................... 14
  2.1 Datasets Overview ......................................................................................................... 14
    2.1.1 Global Acetylome Dataset ................................................................................... 14
    2.1.2 Global Proteomics Data ....................................................................................... 15
    2.1.3 Histone Acetylation Dataset ................................................................................ 16
  2.2 Linear Regression Analysis ............................................................................................ 16
    2.2.1 Histone Acetylation Linear Model ....................................................................... 16
    2.2.2 Metabolic Enzymes Acetylation Linear Models .................................................... 17
    2.2.3 Normalization of Acetylome Data for Non-histone proteins ................................. 17
  2.3 Gene Set Enrichment Analysis ....................................................................................... 18
    2.3.1 Sample subset ...................................................................................................... 18
    2.3.2 GSEA .................................................................................................................. 18
  2.4 Statistical Analysis ......................................................................................................... 19

3 Findings .............................................................................................................................. 20
  3.1 Histone Acetylation ....................................................................................................... 20
  3.2 Metabolic Proteins Expression and Acetylation Interplay ............................................. 32
    3.2.1 Expression of metabolic enzymes characterize Basal subtype metabolism as glycolytic ................................................................................................................................. 32
    3.2.2 Differential acetylation of cytoplasmic and mitochondrial metabolic enzymes ................................................................................................................................. 37
  3.2.3 Functional role of differentially acetylated .................................................................. 45


List of Figures

Figure 1.1: Schematic representation of acetylation and deacetylation reactions (Drazic et al. 2016). A. N-terminal acetylation of polypeptides. B. Reversible acetylation of ε-amino group of lysines. C. Reaction specific for NAD⁺-dependent sirtuins. .......................................................... 2

Figure 2.1.1: Distribution of all detected acetylation values for each experiment separately. .................................................................................................................. 15

Figure 2.1.2: Overview of global acetyome dataset coverage. A. Number of acetylated peptides vs coverage. Dashed line indicates the minimal coverage for every peptide in the dataset. B. Distribution of coverage of acetylation peptides per sample. Samples with abnormally low coverage were excluded from the analysis. ............................................. 15

Figure 2.3.1: Scoring plot of PCA decomposed matrix of H2B N-terminal acetylation sites. Circle indicates a unit circle, the length of Ac-sites vectors is equal to their loading. Cluster 1 is considered as high acetylation samples and cluster 3 as low acetylation samples. ........................................... 18

Figure 3.1: Number of acetylated sites per core nucleosome histone reported in the dataset ........................................................................................................................................ 20

Figure 3.1.2: Unsupervised clustering of histone acetylation sites. Distance: Euclidean, clustering method: Ward’s. Red indicates the highest acetylation level, blue – the lowest, white – the missing data ........................................................................... 21

Figure 3.1.3: Spearman’s correlation between histone acetylation sites. A. Correlation plot for all histone sites. B. Distribution of Rho correlation values for H2A/H2B sites (yellow), or H3/H4 sites (green), or between the H2A/H2B and H3/H4 groups ........................................................................................................ 22

Figure 3.1.4: Coefficients of fitted linear models [1]. Only coefficients with p-value < 0.05 under random model are labeled .................................................................................................................. 23

Figure 3.1.5: Significantly enriched NCG sets in differentially expressed genes. A. Gene ratio of enriched sets. B. Running enrichment score for breast and kidney cancer sets ........................................................................................................ 25

Figure 3.1.6: Top ten DE genes enriched for breast cancer association. mRNA level corresponds to log₂(FPKM + 0.01), protein level corresponds to normalized relative protein abundance. Significance level of Wilcoxon test is shown ........................................................................................................... 26

Figure 3.1.7: Spearman’s correlation between H2B N-terminal Ac-sites with luminal specific DE genes. A. GATA3, mutations are labeled: M – missense mutation, F – frameshift insertion/deletion mutation, S – splice site mutation. B. FOXA1, C. ESR1. mRNA level corresponds to log₂(FPKM + 0.01), protein level corresponds to normalized relative protein abundance .................................................................................................................. 27

Figure 3.1.8: Unsupervised clustering of histone acetylation sites. Distance: Euclidean, clustering method: Ward’s. Red indicates the highest acetylation
/protein/mRNA level, blue – the lowest, white – the missing data. Mutations are labeled: M – missense mutation, F – frameshift insertion /deletion mutation, S – splice site mutation. mRNA level corresponds to log₂(FPKM + 0.01), protein level corresponds to normalized relative protein abundance .................................................................................................................. 29

Figure 3.1.9: Average Spearman’s correlation between mRNA and protein levels of differentially expressed genes and histones acetylation. A. Correlation computed for H2B K5, K11, K11_K12, K15_K16 and K20_K23. B. Correlation computed for H3 K27, K27_K36 and K36.................................................................................................................. 31

Figure 3.2.1.1: Top enriched pathways for acetylated proteins in breast cancer. The right part provides examples of acetylated proteins interactions, colorful nodes indicate that there are novel sites reported in the dataset, grey nodes corresponds to proteins with all-known Ac-sites reported in the dataset. .............. 33

Figure 3.2.1.2: Unsupervised clustering of protein expression of differentially expressed metabolic proteins and carriers (Anova, FDR < 0.05). Distance: Euclidean, clustering method: Ward’s. Red indicates the highest protein level, blue – the lowest, white – the missing data................................................................. 34

Figure 3.2.1.3: PFKP gene amplification, RNA level and protein level in breast cancer subtypes. Copy number level is considered as log₂(copy number tumor /copy number normal). The left plot has a dashed line marking 0.5, a cutoff for gene to have one additional copy. FPKM is Fragments Per Kilobase of transcript per Million mapped reads................................................................. 35

Figure 3.2.2.1: Unsupervised clustering of normalized acetylation level of differentially acetylated lysines of metabolic proteins and carriers (Anova, FDR < 0.05). Distance: Euclidean, clustering method: Ward’s. Red indicates the highest acetylation level, blue – the lowest, white – the missing data................................................................. 38

Figure 3.2.2.2: Examples of differentially acetylated sites of metabolic enzymes................................................................. 39

Figure 3.2.2.3: Coefficients of fitted linear models for HDACs [2]. Only coefficients with linear model coefficient FDR < 0.1 and p-value < 0.01 under random model are shown ......................................................................................... 41

Figure 3.2.2.4: Coefficients of fitted linear models for HATs [2]. Only coefficients with linear model coefficient FDR < 0.1 and p-value < 0.01 under random model are shown ......................................................................................... 42

Figure 3.2.2.5: Schematic representation of Acetyl-coA synthesis pathways in basal subtype. Blue arrows mean protein downregulation, pink arrows mean protein upregulation, yellow circles – elevated acetylation. Adopted from (Narita, Weinert, and Choudhary 2018). ACC1 - Acetyl-CoA Carboxylase Alpha, or ACACA; ACC2 - Acetyl-CoA Carboxylase Beta, or ACACB; ACLY - ATP Citrate Lyase, ACSS2 - Acyl-CoA Synthetase Short Chain Family Member 2; PDC – Pyruvate Dehydrogenase Complex. .................. 43

Figure 3.2.2.6: Expression of Acetyl-coA metabolism proteins. A. Protein expression of proteins involved in Acetyl-coA synthesis in cytoplasm. B. Protein expression of proteins involved in fatty acid synthesis from Acetyl-coA in cytoplasm. ................................................................................................. 44

Figure 3.2.3.1: Localization of GPI differentially acetylated sites relative to its active center. Differentially acetylated sites are marked in pink, not differentially acetylated sites are marked in grey. K519 residue involved in catalysis
reported by (Ji Hyun Lee et al. 2001) is marked in magenta. Catalytically important sites reported by (Cordeiro et al. 2004) are colored in white.

Figure 3.2.3.2: Localization of PGK1 differentially acetylated sites relative to its active center. Differentially acetylated sites are marked in cyan. Substrates 3-PG and ADP are colored in white.

Figure 3.2.3.3: Localization of MDH2 differentially acetylated sites. Differentially acetylated upregulated in basal sites are marked in pink. Differentially acetylated downregulated in basal sites are marked in blue. Not differentially acetylated cites, but previously reported as activating, are in cyan. Substrate NAD$^+$ is colored in magenta.

Figure 3.2.3.4: Localization of IDH2 differentially acetylated sites. Differentially acetylated sites are marked in pink. Protein is colored by hydrophobicity: red the most hydrophilic, white the most hydrophobic. Substrate NADP$^+$ is colored in green.
List of Tables

Table 1.1: Human major histone acetyltransferases................................................................. 2
Table 1.1.2: Human major histone deacetylases................................................................. 3
Table 3.2.2.1: Summary of protein expression and acetylation of glycolytic enzymes in breast cancer subtypes. Number of affected sites was chosen with t-test FDR < 0.1.................................................................................................................. 39
Table 3.2.2.2: Summary of protein expression and acetylation of TCA cycle enzymes in breast cancer subtypes. Number of affected sites was chosen with t-test FDR < 0.1.................................................................................................................. 40
Acknowledgments

I would like to thank my advisor Prof. Li Ding for her support and valuable advice she gave to all her students. Special thanks goes to Steven Carr lab in Broad Institute who kindly provided the data used in this Thesis. I would also like to share my gratefulness to my colleagues and friends in the lab for being the best people I am so glad to work with. This work is supported by NIH grants 5U24CA210986-03 to Steven Carr, U24CA210972 and U24CA211006 to Li Ding. I acknowledge support of computational resources from McDonnell Genome Institute and the Oncology Division of the Washington University School of Medicine. I would also like to say special thanks to the Department of Biomedical Engineering of Washington University in St. Louis and my academic advisor Prof. Dennis Barbour for giving me advice to pursue precision medicine and bioinformatics field.

I would like to thank my thesis committee who reviewed this thesis and gave new directions to follow in the future.
ABSTRACT OF THE THESIS

Histone and Non-Histone Proteins Acetylation Profiles in Breast Cancer

by

Alla Karpova

Master of Science in Biomedical Engineering

Washington University in St. Louis, 2018

Research Advisor: Li Ding

This study evaluates the impact of protein acetylation on breast cancer gene expression and the regulation of metabolism. Acetylation is the second abundant post-translational modification after phosphorylation, regulating protein activity and function. The alterations in acetylation of both histone and non-histone proteins is known to be related to many human diseases, including cancer. Acetylation and deacetylation of histones is closely associated with the regulation of gene expression, while acetylation of non-histone proteins may have a broad effect on major cellular processes, such as proliferation, metabolism, cell cycle and apoptosis, imbalanced regulation of which is essential for cancer development. Therefore, it’s critical to explore the role of this post-translational modification in cancer in a systematic manner. Here, utilizing a unique acetylome dataset for 120 patients with breast cancer, as well as genomic and proteomic data, I showed the impact of acetylation on gene expression and metabolic enzymes. More specifically, the association between histone H2B acetylation level and expression of FOXA1 and GATA3 transcription factors has been established. In addition, acetylation of metabolic enzymes has been demonstrated to reveal additional information on metabolism regulation in breast cancer.
1 Background

1.1 Histone Acetyltransferases and Deacetylases

Acetylation is one of the most prevalent post-translational modifications (PTMs) in eukaryotic cells (Khoury, Baliban, and Floudas 2011). Alongside with phosphorylation, ubiquitination and glycosylation, acetylation controls essential cellular processes and mediate the adjustment to changing environmental conditions. Constantly developing techniques for protein detection allow for identification of thousands of new thousands of PTM sites, leading to a deeper understanding of their function (Doll and Burlingame 2015).

Acetylation can occur in two forms: acetylation of N-terminal (Nt) amino acid in a peptide and acetylation of ε-amino group of lysines. Unlike Nt-acetylation, lysine acetylation is a reversible modification. This modification can be introduced by histone acetyltransferases (HATs) that utilize acetyl-coA as a source of acetyl group and transfer it on the ε-amino group. Since the acetylation can occur not only on histone proteins (as it was thought before), those enzymes are sometimes called lysine (K) acetyltransferases (KATs). Histone deacetylases catalyze the reverse reaction and exempt lysines from the acetyl group (Fig. 1.1.1). Similar to HATs, they are sometimes called KDACs – lysine deacetylases.
There are 17 genes encoding proteins with acetyltransferase activity annotated to be the main activity. HATs can be subdivided into three families: p300/CBP family, the GNAT family and the MYST family. Table 1.1.1 summarizes gene and protein names of major acetyltransferases and provides the examples of their substrates.

Figure 1.1.1. Schematic representation of acetylation and deacetylation reactions (Drazic et al. 2016). A. N-terminal acetylation of polypeptides. B. Reversible acetylation of ε-amino group of lysine. C. Reaction specific for NAD⁺-dependent sirtuins — class of deacetylases.

Table 1.1.1. Human major histone acetyltransferases.

<table>
<thead>
<tr>
<th>FAMILY</th>
<th>HAT</th>
<th>NEW NAME</th>
<th>HUGO GENE SYMBOL</th>
<th>SUBSTRATE EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNAT</td>
<td>HAT1</td>
<td>KAT1</td>
<td>HAT1</td>
<td>H2A, H4</td>
</tr>
<tr>
<td>GNAT</td>
<td>GCN5</td>
<td>KAT2A</td>
<td>KAT2A</td>
<td>H3</td>
</tr>
<tr>
<td>GNAT</td>
<td>PCAF</td>
<td>KAT2B</td>
<td>KAT2B</td>
<td>H3, H4, ACLY, PKM</td>
</tr>
<tr>
<td>P300/CBP</td>
<td>CBP</td>
<td>KAT3A</td>
<td>CREBBP</td>
<td>H2A, H2B, H3, H4, FOXO1</td>
</tr>
<tr>
<td>P300/CBP</td>
<td>P300</td>
<td>KAT3B</td>
<td>EP300</td>
<td>H3, FOXO1, SIRT2</td>
</tr>
<tr>
<td>_</td>
<td>TAF1</td>
<td>KAT4</td>
<td>TAF1</td>
<td>H3, H4</td>
</tr>
<tr>
<td>MYST</td>
<td>TIP60</td>
<td>KAT5</td>
<td>KAT5</td>
<td>H2A, H4, FOXP3</td>
</tr>
</tbody>
</table>
HATs can localize in different cellular locations. The majority of them functions in the nucleus, such as CBP/p300, KAT7, KAT8, HAT1 and others. However, some of the HATs can be found in both cytoplasm and the nucleus: CBP/p300, KAT2B, ELP3, ATF2 and CLOCK. In mitochondria, only one acetyltransferase (ACAT1) has been identified, modifying pyruvate dehydrogenase complex and regulating its activity (Fan et al. 2014). Histone acetyltransferases are almost always associated with other protein, defining their target and site specificity. HAT-binding proteins usually contain various domains such as bromodomain, chromodomain, WD40 repeats and PHD fingers domains aimed to recognize different modifications of histones (Lee and Workman 2007).

Histone deacetylases can be subdivided into four families: Class I, II and IV require Zn$^{2+}$ as a cofactor, while class III is NAD$^{+}$- dependent and are called sirtuins. HDACs of class I or II usually modify histones, transcription factors and chromatin remodeling complexes (Drazic et al. 2016). However, some members of those two classes are also found in cytoplasm (Table 1.1.2).

Table 1.1.2. Human major histone deacetylases.

<table>
<thead>
<tr>
<th>CLASS</th>
<th>HDAC</th>
<th>COFACTOR</th>
<th>COMPARTMENT</th>
<th>SUBSTRATE EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>HDAC1</td>
<td>Zn$^{2+}$</td>
<td>Nucleus</td>
<td>All core histones, RelA, AR</td>
</tr>
<tr>
<td>I</td>
<td>HDAC2</td>
<td>Zn$^{2+}$</td>
<td>Nucleus</td>
<td>All core histones</td>
</tr>
<tr>
<td>I</td>
<td>HDAC3</td>
<td>Zn$^{2+}$</td>
<td>Nucleus</td>
<td>All core histones, NF-kB, KAT2B, STAT1</td>
</tr>
<tr>
<td>----</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>I</td>
<td>HDAC8</td>
<td>Zn$^{2+}$</td>
<td>Nucleus/Cytoplasm</td>
<td>All core histones, p53</td>
</tr>
<tr>
<td>II</td>
<td>HDAC4</td>
<td>Zn$^{2+}$</td>
<td>Nucleus/Cytoplasm</td>
<td>All core histones, HIF1α, p53</td>
</tr>
<tr>
<td>II</td>
<td>HDAC5</td>
<td>Zn$^{2+}$</td>
<td>Nucleus/Cytoplasm</td>
<td>GATA2, GCMa</td>
</tr>
<tr>
<td>II</td>
<td>HDAC6</td>
<td>Zn$^{2+}$</td>
<td>Cytoplasm</td>
<td>α-tubulin, HSP90</td>
</tr>
<tr>
<td>II</td>
<td>HDAC7</td>
<td>Zn$^{2+}$</td>
<td>Nucleus/Cytoplasm</td>
<td>PLAG1</td>
</tr>
<tr>
<td>II</td>
<td>HDAC9</td>
<td>Zn$^{2+}$</td>
<td>Nucleus/Cytoplasm</td>
<td>ATDC</td>
</tr>
<tr>
<td>II</td>
<td>HDAC10</td>
<td>Zn$^{2+}$</td>
<td>Cytoplasm</td>
<td>HSP70, PP1</td>
</tr>
<tr>
<td>III</td>
<td>SIRT1</td>
<td>NAD$^+$</td>
<td>Nucleus</td>
<td>p53, FOXO1, HSF1, KAT7, CBP</td>
</tr>
<tr>
<td>III</td>
<td>SIRT2</td>
<td>NAD$^+$</td>
<td>Cytoplasm</td>
<td>α-tubulin</td>
</tr>
<tr>
<td>III</td>
<td>SIRT3</td>
<td>NAD$^+$</td>
<td>Mitochondria</td>
<td>GDH, TCA cycle enzymes, LCAD, ACSS2</td>
</tr>
<tr>
<td>III</td>
<td>SIRT4</td>
<td>NAD$^+$</td>
<td>Mitochondria</td>
<td>GLUD1</td>
</tr>
<tr>
<td>III</td>
<td>SIRT5</td>
<td>NAD$^+$</td>
<td>Mitochondria</td>
<td>CPS1, cytochrome c</td>
</tr>
<tr>
<td>III</td>
<td>SIRT6</td>
<td>NAD$^+$</td>
<td>Nucleus</td>
<td>H3K56ac, RBBP8</td>
</tr>
<tr>
<td>III</td>
<td>SIRT7</td>
<td>NAD$^+$</td>
<td>Nucleolus</td>
<td>H3K18ac, PAF53</td>
</tr>
<tr>
<td>IV</td>
<td>HDAC11</td>
<td>Zn$^{2+}$</td>
<td>Nucleus</td>
<td>All core histones</td>
</tr>
</tbody>
</table>

HDAC1 and HDAC2 share a lot of sequence similarity and involved in regulation of cell cycle and apoptosis (Reichert, Choukrallah, and Matthias 2012). HDAC3 is also essential for cell cycle regulation, along with DNA damage control (Reichert, Choukrallah, and Matthias 2012). Most of class I histone deacetylases are part of bigger protein complexes, such as Sin3, N-CoR/SMRT and CoREST, except for HDAC8, which has been determined to function alone (Barneda-Zahonero and Parra 2012). In general class I HDACs are expressed ubiquitously in every cell. In contrast to class I, class II HDACs are more tissue specific and play crucial role in differentiation and organism development. Along with deacetylase domain, class II HDACs have long regulatory domain, allowing for binding to tissue specific transcription factors and therefore modulating the specificity of these histone deacetylases (M. Parra and Verdin 2010). Sirtuins can be localized in various cellular location,
including mitochondria, cytoplasm and the nucleus, where they are involved in regulation of oxidative stress, aging, metabolism and DNA repair.

### 1.2 Histone Acetylation

Eukaryotic DNA is wrapped around and packaged with specialized protein complexes, called nucleosomes. Each nucleosome is an octamer that consists of four pairs of core histones H2A, H2B, H3 and H4. Approximately 147 bp of DNA are wrapped two times around each nucleosome with 8-114 pb of free DNA linking adjacent nucleosomes. Nucleosome is a basic unit of chromatin, and dependent on how tightly packaged the nucleosomes are, chromatin can exist in two basic states: the more relaxed euchromatin, and more condensed heterochromatin (Shahbazian and Grunstein 2007).

The openness of chromatin is thought to directly impact the expression of underlying DNA with looser chromatin being more accessible to RNA Pol and transcription factors and more actively transcribed. Chromatin structure is highly dynamic, and it plays a crucial role in epigenetic gene regulation.

All core histones have a globular domain that forms the center of the nucleosome and an N-terminal tail that protrudes away from the nucleosomes. These N-terminal tails are major sites of nucleosome regulation through post-translational modifications (PTMs) including acetylation, phosphorylation, methylation, ubiquitination and sumoylation (Shahbazian and Grunstein 2007). All four histones tails have lysines – potential sites of methylation or acetylation, and while methylation can be either repressing (H3K9me3, H3K27me3), or activating (H3K4me3) depending on the lysine position, lysine acetylation is always thought to be activating.

In vitro studies suggest that lysine acetylation reduces the electrostatic attraction between the negatively charged DNA phosphates and positively charged lysines, resulting in a less condensed
chromatin. In vivo, lysine acetylation is regulated by “writer” enzymes – histone acetyltransferases (HATs) that catalyze the transfer of acetyl group from Acetyl-CoA onto the lysine amino group and “eraser” enzymes - histone deacetylases (HDACs) that remove the acetyl group from lysines. HATs and HDACs are believed to act non-specifically genome-wide, as well as at individual gene loci when targeted by transcription activators and repressors.

Histone acetylation is dimmed to be always activation mark because it decreases the nucleosome affinity to the DNA and increases its accessibility to transcription factors. The most studied histone acetylation mark is H3K27Ac, marking active promoters and enhancers (Creyghton et al. 2010; Pradeepa 2017). H3K36Ac occurs in RNAP II promoters, while H3K36me3 is usually found within exons, marking actively transcribed genes (Morris et al. 2007). H3K9Ac marks the switch from transcription initiation to the elongation, in contrast to H3K9me3, which is a strong repressive mark of heterochromatin (Gates et al. 2017). H3K14 is critical for DNA damage checkpoint activation (Y. Wang et al. 2012), and together with H3K9ac marks bivalent promoters, enhancers and sometimes inactive promoters as well (Karmodiya et al. 2012). In one study, a consistent set of histone modification has been identified, marking promoter regions. This set contains the following histone modifications: H2A.Z, H2BK5ac, H2BK12ac, H2BK20ac, H2BK120ac, H3K4ac, H3K4me1, H3K4mc2, H3K4me3, H3K9ac, H3K9me1, H3K18ac, H3K27ac, H3K36ac, H4K5ac, H4K8ac and H4K91ac. The authors have also shown that promoters containing this marks result in higher gene expression than promoters without such modifications. They also noted that the amount of these modifications correlate genome wide (Z. Wang et al. 2008). Histone H4 lysine 16 acetylation has relatively unusual role not related to regulation of gene expression. H4K16ac controls nucleosome-level interactions, preventing the formation of evenly spaced nucleosomes, resulting in transcriptional repression (Blosser et al. 2009). However, another study has demonstrated that in mouse embryonic
stem cells loss of H4K16 acetylation does not affect high-order chromatin structure, and more surprisingly that H4K16ac is another marks of active enhancers (Taylor et al. 2013). These findings suggest that histones modifications have multifarious functions, depending on the chromatin context, developmental stage and cell type.

1.3 Alterations in Acetylation in Cancer

It has been recently demonstrated that both HATs and HDACs are required to achieve proper level of gene transcription. As it was discussed above, acetylation increases chromatin accessibility to RNAP and TFs, however, the fine tuning of acetylation marks on the promoter is required to switch from initiation to elongation. In addition, it is very important to balance the level of acetylation at transcription start sites, therefore, the activity of both HATs and HDACs is required to achieve right transcription rate.

In general, low acetylation levels can be due to either mutations, low expression, displacement or haploinsufficiency of HATs, or overexpression/aberrant recruitment of HDACs, or the combination of these events. Similarly, elevated acetylation can be achieved by high expression, mutations and aberrant recruitment of HATs or downregulation and mutations in HDACs.

1.3.1 Wrong Histone Acetyltransferases

Mutations in HATs have been reported in many cancer types. For example, CREBBP gene has many recurrent mutations in lymphoma and leukemia (Morin et al. 2011; Pasqualucci et al. 2011; Mullighan et al. 2011; Gui et al. 2011). In addition, close paralog of CREBBP, gene EP300, was shown to have both missense and truncating mutations in solid tumors (Muraoka et al. 1996; Gayther et al. 2000). In the TCGA PanCancer study, EP300 was identified as a driver gene (tumor suppressor) for bladder (together with CREBBP), endometrial and lung cancers (Bailey et al. 2018). Two functional copies of
CREBBP, but not EP300, are required to avoid defects in hematopoietic differentiation, leading to malignancies (Kung et al. 2000). Another study on mice proves even more the hypothesis that EP300 and CREBBP act as tumor suppressors. In this study embryonic stem cells deficient for p300 or CBP were inject into mice embryos. These mice were shown to have higher chances to develop hematological malignancies compared to control group (Di Cerbo and Schneider 2013). P300 mediates the functioning of a number of tumor suppressor proteins, such as TGF-β, p53 and E2F, through activation of transcription of target genes (Iyer, Özdag, and Caldas 2004). However, there are clear examples of EP300 and CREBBP being oncogenic genes. For instance, both CBP and p300 were demonstrated to form fusions with MILL (KMT2A) and MOZ (KAT6A) proteins in the mixed lineage leukemia (Di Cerbo and Schneider 2013). In addition, p300/CBP can acetylate more common fusion proteins such as AML1-ETO and positively modulate their contribution in leukomogenesis. These observations highlight the oncogenic potential of p300/CBP acetyltransferases. Not only mutation status alters the function of HATs, but the upregulation of expression level as well. Overexpression of p300 was observed in breast, liver and lung carcinomas (Di Cerbo and Schneider 2013), correlating with poor prognosis. Moreover, it was demonstrated that p300 promotes the expression of androgen receptor (AR) target genes through the acetylation of AR in ligand independent manner in prostate cancer (Debes et al. 2003). Therefore, prostate cancer can benefit from high expression of EP300. In addition, the knockdown of EP300, but not CREBBP, significantly decreases the proliferation of prostate cancer cells. In human breast cancer lines, upon BRCA1 being mutated, p300/CBP acetylates estrogen receptor α, and the ectopic expression of WT BRCA1 downregulates p300 (Di Cerbo and Schneider 2013). PCAF together with p300 acetylates p53 and increase its DNA binding affinity promoting growth arrest and apoptosis (Yamaguchi et al. 2009). Altogether, these results demonstrate
that HATs can play both oncogenic and tumor suppressor role in cancer development and their exact contribution in each particular cancer type development is still to be identified.

1.3.2 Wrong Histone Deacetylases

Similar to HATs, HDACs are also often dysregulated in cancer. The mechanism of dysregulation can be very different: from mutations to aberrant recruitment by fusion oncogenic protein. For example, class I HDACs are frequently found to be upregulated in breast, pancreatic, lung and prostate carcinomas and are almost always associated with poor prognosis (Barneda-Zahonero and Parra 2012). In one study, expression of HDAC1 and HDAC3 was correlated with estrogen and progesterone receptor expression, suggesting they could be an independent prognostic marker (Krusche et al. 2005). In osteosarcoma and breast cancer cells knockdown of HDAC1 results in cell cycle arrest and induction of apoptosis (Senese et al. 2007), while the overexpression of HDAC1, HDAC6 and HDAC8 increases cell invasion (Park et al. 2011). In addition, in breast cancer cells HDAC2 silencing enhances p53 binding ability, which correlates with cell cycle block and senescence induction (Harms and Chen 2007). Class II deacetylases were also reported to have mutations and aberrant expression in many cancer types (Barneda-Zahonero and Parra 2012). They can also affect cell proliferation rate of cancer cells. For instance, HDAC5 induces cell rapid division by regulation of p14 repression (Yarosh et al. 2008). In addition, HDAC7 together with estrogen receptor α represses a tumor suppressor Reprimo, therefore contributing to cell growth (Malik et al. 2010). HDACs might also be involved in developing of cancer chemoresistance: it was shown that DHAC7/HIF-1A complex might repress cyclin D1, contributing to chemoresistance (Wen et al. 2010). Not all HDACs function as potential oncogenes. In breast carcinomas, HDAC6 expression was demonstrated to be associated with better survival and was higher in ER and PR-positive tumors (Zhang et al. 2004). HDAC6 probably deacetylase Hsp90, preventing the hormone mediated activation and decreasing the growth
of breast cell (Barneda-Zahonero and Parra 2012). Sirtuins are more controversial in terms of their impact on cancer development. SIRT3 and SIRT7 were shown to be upregulated in breast cancer, while SIRT2 is, on contrary, downregulated in gliomas and gastric carcinomas. SIRT2 appears to act as a tumor suppressor ensuring the proper passing of mitotic checkpoint. SIRT3 overexpression opposes p53-mediated cell cycle arrest in bladder cancer cells, however, in xenograft models, SIRT3 knockdown triggers tumorogenesis (Barneda-Zahonero and Parra 2012).

These results demonstrate that HDACs play important role in cell cycle regulation and are often associated with the outcome. However, some findings are contradictory and may vary from cancer cell lines to real patients. Hence, it is very important to examine the contribution of every HDAC.

1.3.3 Histone Acetylation in Cancer

First studies on global histone acetylation changes in cancer cells revealed the overall reduction in H4K16 acetylation (Di Cerbo and Schneider 2013). Global loss of H3K18ac, H3K9ac and H4K16ac is generally associated with poor prognosis and a shorter life expectation. Moreover, global loss of methylation marks such as H3K4me2, H3K9me2 and H3K27me3 is also an indicator of poor outcome (Di Cerbo and Schneider 2013). At the same time, another group of researchers associated low level of H3K9ac and H3K18ac with a better prognosis in lung cancer (Seligson et al. 2009). The molecular mechanisms underlying such changes in global histone modifications level have not been established yet. One study has demonstrated that SIRT7 is able to deacetylate H3K18ac, resulting in inhibition of key cellular regulators, and that downregulation of SIRT7 contributes to cancer proliferation (Barber et al. 2012). Modification of core part of histones might also be dysregulated in cancer. For instance, H3K56ac involved in DNA damage response was reported to correlate with dedifferentiated state of cancer cells. So, these are not many things known about global changes in histone acetylation and even fewer of them are connected to breast cancer. Nevertheless, global
dysregulation of histone acetylation may cause global changes in gene expression patterns, which may induce and enhance tumorogenesis.

1.3.4 **Non-histone Proteins Acetylation: Links to Cancer**

Acetylation is implicated in vital cellular processes, many of which have been linked to various diseases, including cancer. Acetylation is involved in regulation of gene expression not only in form of histone acetylation, but non-histone proteins as well. P53 is a great example of a transcription factor with activity modulated by acetylation. This key regulator of cell cycle and apoptosis is acetylation at several sites, resulting in enhanced DNA binding capacity and activation of p53-regulated genes. p53 is the most important and most frequently mutated tumor suppressor in the majority of cancer type (Narita, Weinert, and Choudhary 2018). Another major cellular process often dysregulated in cancer is cell cycle. During the cell cycle sister chromatid are grouped together in pairs, which is accomplished by cohesion complex. A key component of cohesion complex, SMC3, surrounding chromatids as a ring, is acetylated at two sites, therefore resulting in close state of this ring and tight retention of sister chromatids together. It was also reported that acetylation modifies the activity of other cell cycle regulators such as CDK1, CDK2, Aurora kinase A and B (Narita, Weinert, and Choudhary 2018).

Taking together, these results suggest that along with phosphorylation acetylation may play important role in regulation of cell cycle in both normal and cancer cells. Moreover, acetylation of DNA damage response proteins controls the choice of a pathway to repair double-strand breaks: non-homologous end-joining (NHEJ) or homology-directed repair (HDR). Acetylation interferes the recruitment of NHEJ-promoting factor TP53-binding protein 1 (53BP1) by several means: histone H4 acetylation prevents 53BP1 binding to H4K20me2 site; acetylation of H2AK15 decreases the amount of H2AK15ub sites and thus impairs 53BP1 recruitment to H2AK15ub; ATM kinase activated upon DNA damage response phosphorylates ACLY, thus increasing the production of acetyl-coA in
nucleus, resulting in higher histone acetylation and wrong chromatin localization of 53BP1; and finally, CBP acetylates 53BP1, which impairs its recruitment to double-strand breaks. Hence, acetylation of various proteins defines which pathways will be utilized to repair DSBs. If this decision making acetylation is disrupted, NHEJ may be used more frequently, resulted in higher rate of insertions and deletions - common outcome of NHEJ (Narita, Weinert, and Choudhary 2018). Finally, acetylation has been shown to be part of cell signaling process as well. For instance, phosphatase PTEN regulating the level of PIP3 can be acetylated in its catalytic and C-terminal domains. Acetylation of catalytic domain inhibits PTEN activity, while acetylation of C-tail promotes PTEN binding to proteins, enhancing its lipid phosphatase activity and recruiting it to signaling complexes (Narita, Weinert, and Choudhary 2018). Besides gene transcription, DNA damage response and cell signaling, acetylation is also associated with regulation of protein folding, cytoskeleton organization, RNA processing, metabolism, autophagy and other vital cellular processes.

Taking together, acetylation and acetylation regulating enzymes are particularly important for cancer formation and are frequently positively or negatively associated with poor outcome. Recent studies by the National Cancer Institute’s Clinical Proteomic Tumor Analysis Consortium (CPTAC) have produced proteomic, phosphoproteomic and acetylome data for infiltrating breast carcinomas. These datasets provide the opportunity to evaluate the impact of acetylation on tumor progression and subtyping using a cohort of 122 breast cancer patients. In this study, we characterized global histone acetylation profiles and evaluated the differences in metabolic enzymes expression and acetylation across breast cancer subtypes. We determined association between acetylation of N-terminal H2B histone sites and luminal transcription factors GATA3 and FOXA1. Acetylation of H2B but not other histones also significantly correlates with protein expression of many tumor suppressor chromatin modifying enzymes, such as KMT2A, KMT2C and KMT2D and this correlation evades in basal and
Her2 subtypes. Analysis of acetylation of metabolic enzymes revealed a metabolic uniqueness of basal subtype compared to all others and demonstrated that regulation of enzymatic activity by acetylation complements of that by gene and protein expression.
2 Research Methods

2.1 Datasets Overview

2.1.1 Global Acetylome Dataset

The acetylome dataset for CPTAC Breast cancer was generated by Steven Carr lab, Broad Institute, using the targeted LC-MS/MS technology with isobaric tags (TMT (tandem mass tags)-10) (Mertins et al. 2018). The peptides were enriched for acetylated lysines, using anti-lysine acetylation antibody (Svinkina et al. 2015). The dataset consists of log-ratios of intensities for experimental over common reference. The common reference is created by sampling small amount of tissue of every experiment sample and mixing them together, ensuring that peptides can be identified more consistently.

The dataset consists of 9517 detected peptides and 130 tumor samples with eight replicates. Over 9,000 detected peptides correspond to more than 10,000 unique acetylation sites for more than 300 proteins. The data were generated in 17 different experiments. Figure 2.1.1.1. shows the distribution of detected values for each experiment. Distributions are centered and scaled, so the data is normalized for batch effect.
Figure 2.1.1. Distribution of all detected acetylation values for each experiment separately.

Samples from one experiment have significantly lower coverage than all other samples (Fig. 2.1.2 B) and were excluded from the analysis. The minimum coverage for the dataset with excluded samples is 26%, the number of peptides with 100% coverage across samples is 1602, corresponding to 744 genes (Fig. 2.1.2 A).

Figure 2.1.2. Overview of global acetylome dataset coverage. A. Number of acetylated peptides vs coverage. Dashed line indicates the minimal coverage for every peptide in the dataset. B. Distribution of coverage of acetylation peptides per sample. Samples with abnormally low coverage were excluded from the analysis.

2.1.2 Global Proteomcis Data

Global proteomics data were downloaded from CPTAC Data Coordinating Center. This includes 122 tumor samples and 17 adjacent normal samples. The overlap of samples with both acetylome and proteome data is 113 tumor samples.
2.1.3 Histone Acetylation Dataset

Four core histones H2A, H2B, H3 and H4 are encoded by a large number of genes, protein products of which can differ in one amino acid in N-terminal domain. Mass spectrometry technology is able to distinguish different genes of one histone and report two functionally the same sites as separate entities. To facilitate the comprehension of the data, I averaged the values for every core histone belonging to one functional site. For example, reported peptides HIST1H2BH_K12k and HIST1H2BD_K12k were average to obtain a value for H2B_K12 site.

2.2 Linear Regression Analysis

2.2.1 Histone Acetylation Linear Model

Linear model of form $Ac_{site} \approx \beta_0 + \beta_1 * Pr_{enzyme} + \epsilon$ was fitted using lm function in R to test for association between enzyme protein abundance and acetylation level of histone site in 113 tumor samples. I used acetylation level of a histone site as a dependent variable and the protein abundance of HAT or HDAC as an independent one. The protein level of histones was not included in this model since the global proteomic dataset Only Site : Enzyme pairs represented in more than 30 samples were considered. The p-value of coefficient $\beta_1$ were adjusted to FDR using Benjamini-Hochberg procedure.

To account for the possibility that such associations may appear by chance, the similar linear models were fitted using random pairs of Protein : Ac-site. I randomly selected 50 proteins from the global proteomics dataset and 1000 Ac-sites from global acetylome dataset and performed linear regression analysis for every pair. I collected the resulted random $\beta_1$ coefficients and fitted their distribution with normal distribution. The resulted distribution was used to obtain the significance level of experimental
coefficient given the distribution of random $\beta_1$. Only pairs with coefficient p-value under random model $< 0.05$ were considered for further analysis.

### 2.2.2 Metabolic Enzymes Acetylation Linear Models

Similar to histone sites, associations between HATs and HDACs and acetylation of metabolic enzymes we tested using linear model of form: $Ac_{\text{substrate}} \approx \beta_0 + \beta_1 * Pr_{\text{enzyme}} + \beta_2 * Pr_{\text{substrate}} + \epsilon$. I used acetylation level of a metabolic enzyme site as a dependent variable and the protein abundance of HAT or HDAC and protein abundance of the metabolic enzyme as independent ones. The protein level of histones was not included in this model since the global proteomic dataset Only Site : Enzyme pairs represented in more than 30 samples were considered. The p-value of coefficient $\beta_1$ were adjusted to FDR using Benjamini-Hochberg procedure.

Significant pairs were filtered using random model p-value as described in 2.1.1.

### 2.2.3 Normalization of Acetylome Data for Non-histone proteins

To assess the difference between the amount of protein acetylation independent from protein expression level, the global acetylation data were normalized using linear regression of form: $Ac_{\text{substrate}} \approx \beta_0 + \beta_2 * Pr_{\text{substrate}} + \epsilon$. Protein abundance of any given protein was used as an independent variable and the acetylation value of an Ac-site on this protein was used as a dependent variable. The residual values $\epsilon$ of every fitted model was used as a new normalized acetylation value not explained by protein abundance.
2.3 Gene Set Enrichment Analysis

2.3.1 Sample subset

To select the samples with high- or low-level acetylation of N-terminal H2B sites, I performed Principal Component Analysis on the acetylation data of these sites, using `prcomp` function in R with centering and scaling. Then I used the returned X matrix as input for `Optimal_Clusters_KMeans` to determine the optimal number of k-means clusters and then `KMeans_rcpp` from ClusterR package. The clusters with the highest loadings of histone H2B sites vectors were treated as samples with high H2B acetylation, and clusters with the lowest negative loadings were chosen as samples with low H2B acetylation. This analysis was done independently for whole sample set and luminal only sample set.

![Figure 2.3.1.1. Scorings plot of PCA decomposed matrix of H2B N-terminal acetylation sites. Circle indicates a unit circle, the length of Ac-sites vectors is equal to their loading. Cluster 1 is considered as high acetylation samples and cluster 3 as low acetylation samples.](image)

2.3.2 GSEA

To find gene potentially regulated by histone acetylation, the gene expression data (FPKM), provided by (?) was used. Genes with zero expression level were filtered out. Baumgartner-Weiss-Schindler
(BWS) test statistic was used as a ranking metric for differentially expressed genes (Zyla et al. 2017). The latest release 6.0 of the Network of Cancer Genes database was used as gene sets for GSEA (Venkata et al. 2018; Subramanian et al. 2005). GSEA was performed in R using function gsea from ClusterProfiler package and BWS statistic as a ranking metric (G. Yu et al. 2012). Only gene sets with q-value < 0.05 were considered. Similar analysis was performed with Gene Ontology (GO) sets downloaded from MSigDB version 6.2, curated by Broad Institute (Liberzon et al. 2015).

2.4 Statistical Analysis

Protein expression and normalized acetylation means in breast cancer subtypes were compared by Student’s t-test. The t-test p-values were adjusted to FDR using Benjamini-Hochberg procedure. The differentially expressed metabolic proteins (Fig. 3.2.1.2) and normalized acetylation of metabolic proteins sites (Fig. 3.2.2.1) for hierarchical clustering were chosen based on ANOVA test p-value with p-value FDR adjusted cutoff 0.05.
3  Findings

3.1  Histone Acetylation

Acetylome dataset reports ~80 acetylated peptides corresponding to various genes of four canonical histones H2A, H2B, H3 and H4. The histone acetylation data have been preprocessed as described in section 2.1.3. to facilitate results interpretation. The resulting protein histone acetylation dataset contains mostly sites from H2B histones (Fig 3.1).

![Bar graph of acetylated sites per core nucleosome histone](image)

Figure 3.1.1. Number of acetylated sites per core nucleosome histone reported in the dataset.

To evaluate general trends in histone acetylation marks across samples and identify general patterns in the acetylation of histones, unsupervised clustering of acetylation values has been performed. Figure 3.1.2 demonstrates the result of the clustering in combination with PAM50 subtype, stage and other clinical annotations. As it seen in Figure 3.1.2, based on the histone acetylation, samples can be subdivided into 3 groups: with overall low, average and high acetylation of histones. There is a group
of samples, sharing high acetylation of all four core histones; another group shares H3/H4 elevated acetylation, while H2A/H2B sites are average; and there are also patients with increased H2B sites acetylation, but low H3/H4 acetylation.

![Figure 3.1.2. Unsupervised clustering of histone acetylation sites. Distance: Euclidean, clustering method: Ward’s. Red indicates the highest acetylation level, blue – the lowest, white – the missing data.](image)

Additionally, H3 and H4 acetylation sites correlate with each other more than with H2B and H2A sites. Indeed, the average Spearman’s correlation value between sites within H3/H4 group is much higher than that between them and H2A/H2B group (Fig. 3.1.3). The observed correlation between histone sites can be explained by nucleosome structure. Nucleosome assembly is a sequential process,
starting from the formation of H2A/H2B and H3/H4 dimers, followed by tetramer and then octamer formation.

Figure 3.1.3. Spearman’s correlation between histone acetylation sites. A. Correlation plot for all histone sites. B. Distribution of Rho correlation values for H2A/H2B sites (yellow), or H3/H4 sites (green), or between the H2A/H2B and H3/H4 groups.

Many histone sites are known to be involved in transcription regulation and chromatin structure maintenance. The acetylation level of some sites may be particularly important for cancer cells and therefore tightly regulated. To determine whether there are sites regulated majorly by one of known histone acetyltransferases, I used linear regression approach and found histone Ac-sites that can be regulated by major acetyltransferases in breast cancer. In this model, Ac-site is an independent variable, whereas the protein abundance of an acetyltransferase is a dependent variable:

\[ Ac_{site} \approx \beta_0 + \beta_1 \times Pr_{enzyme} + \epsilon \]  

[1]

The bigger \( \beta_1 \), the stronger one unit of enzyme protein abundance affects one unit of the acetylation level of an Ac-site. To account for random associations that might appear, the similar regression models were fitted with random Ac-sites and random protein abundances. The distribution of resulted
linear model coefficients were used to determine the p-value of experimental linear model coefficients. Figure 3.1.3 shows the breakdown of linear model coefficients $\beta_1$ per HAT. As seen from Fig. 3.1.4, CREBBP protein product (CBP) demonstrates the strongest association with N-terminal H2B histone Ac-sites: H2B K5, K11, K12, K15, K16, K20 and K23. This observation is concordant with a recent paper findings, stating that p300/CBP are responsible for acetylation of N-terminal H2B Ac-sites (Weinert et al. 2018).

![Figure 3.1.4. Coefficients of fitted linear models [1]. Only coefficients with p-value < 0.05 under random model are labeled.](image)

EP300 (p300) also shows association with N-terminal H2B Ac-sites, but the coefficients are smaller and less significant under the random associations model. In addition, CLOCK, KAT5 and KAT7 exhibit significant associations with H2AK124, H3K23 and H2BK57 sites, respectively, however the physical interaction between these proteins and Ac-sites has not been reported so far. High association between KAT7 and KAT8 protein and H2BK5, K11 sites can be accounted to co-expression of these acetyltransferases (Spearman’s Rho$_{EP300/KAT7}$ = 0.6, Rho$_{CREBBP/KAT8}$ = 0.45).
I have decided to focus on CBP : H2B association because H2B sites have better coverage than other significant sites, providing more power for statistical tests, and because this histone has never been explored in cancer studies. Since acetylation of histones in general decreases the nucleosome positive charge, leading to the impaired interaction with DNA phosphate groups and therefore, resulting in a more relaxed and accessible chromatin state, global differential acetylation of histone H2B may affect the expression of some. One possible mechanism of action on gene expression level might be associated with H2A-H2B dimers removal from nucleosomes upon acetylation that may help maintain the open state of certain genome regions (Ito et al. 2000). Additionally, at least for yeast, H2B N-terminal lysines have been shown to be involved in upregulation of genes involved in NAD$^+$ and vitamins synthesis (M. A. Parra et al. 2006), suggesting that changes in global histone acetylation can influence the expression of specific genes. To determine if differential acetylation of H2B N-terminal sites is connected to gene expression, gene set enrichment analysis (GSEA) has been performed using RNA-seq gene expression data. I compared gene expression in two groups of samples with either high or low acetylation level of H2B N-terminal sites. Using Network of Cancer Genes (NCG) curated sets, we identified that ‘Breast cancer’ set is enriched for upregulated genes (high H2B samples compared to low H2B ones), along with kidney, bladder and liver cancer types. On the contrary, the downregulated genes were found to be enriched for melanoma associated genes (data not shown). Since many cancer types share cancer associated protein, such as p53, c-Myc, NRAS and other well-known oncogenes and tumor suppressors, it is not surprising to see other cancer types to be enriched as well. Regardless of other cancer types, breast cancer set shows the biggest number of genes contributing to maximum enrichment score (34 genes versus 23 in kidney and 24 in bladder cancer).
Figure 3.1.5. Significantly enriched NCG sets in differentially expressed genes. A. Gene ratio of enriched sets. B. Running enrichment score for breast and kidney cancer sets.

mRNA and protein expression level of top ten breast cancer associated genes, contributing to the maximum enrichment score, are shown in Fig 3.1.6. For these genes the difference in gene expression level is translated into the protein level as well.
Among the top ten breast cancer DE genes, there are three genes involved in the luminal development vector establishment in breast tissue: GATA3, FOXA1 and ESR1. Their expression remains high in breast cancer subtypes derived from luminal cell types, and positively associated with good outcome (Shou et al. 2016; Yoon et al. 2010). Due to subtype specificity, it is possible that the observed difference in expression of those genes can be driven by subtype, but not H2B acetylation. To see if H2B acetylation shows similar association with GATA3, FOXA1 and ESR1 expression in luminal samples only, I tested the mean difference between H2B high and low samples, defined for luminal subtypes only, using similar approach (see 2.3.1). It turned out that there is no significant difference in GATA3, FOXA1 or ESR1 mRNA level between H2B high and low luminal samples, while GATA3 and FOXA1 protein level is on contrary significant (p-value <0.05) (See Appendix A Fig. A.1). In addition, correlation analysis between H2B Ac-sites and GATA3, FOXA1 and ESR1 mRNA or
protein in luminal samples revealed that protein has stronger association with H2B histone than mRNA for all three genes (Fig. 3.1.7). However, in all samples together both mRNA and protein correlate well with H2B acetylation (Fig A.2). Such dramatic difference in correlation between mRNA and H2B acetylation in all samples and luminal samples separately tells us that correlation in all samples together is driven by differences between subtypes and not H2B. Protein correlation with H2B remains similar for all samples and luminal samples only, suggesting that this correlation is real.
Figure 3.1.7. Spearman’s correlation between H2B N-terminal Ac-sites with luminal specific DE genes. A. GATA3, mutations are labeled: M – missense mutation, F – frameshift insertion/deletion mutation, S – splice site mutation. B. FOXA1, C. ESR1. mRNA level corresponds to log₂(FPKM + 0.01), protein level corresponds to normalized relative protein abundance.

So, we see very low correlation between H2B acetylation and GATA3/FOXA1 mRNA level. However, the protein is highly correlated with H2B acetylation. This may indicate that these gene products might be regulated on protein level by H2B acetylation. It is known that in the multiprotein complexes subunits that are not incorporated in the complex get degraded very soon (Mueller et al. 2015). GATA3 has been demonstrated to act upstream of FOXA1 and mediate ESR1 binding ability.
(Theodorou et al. 2013), therefore these proteins may form complexes at gene regulatory regions. As a support for this statement, our data indicates that protein expression of GATA3, FOXA1 and ESR1 is highly concordant in luminal breast cancer (Fig 3.1.8) - average Spearman’s correlation value between protein abundances of these three proteins is 0.41. Hence, these findings suggest us that GATA3, FOXA1 and ESR1 form a complex that somehow is linked to H2B acetylation.

Even though three important transcription factors correlate with H2B acetylation, it is unlikely that they directly interact since none of these proteins have a bromodomain able to recognize acetylated proteins. Hence, there must be a bromodomain containing mediator able to bind H2B histone sites that serves as a linker between the TF and the histone modification. There are at least five proteins with a bromodomain that were shown to interact with H2B acetylated lysines: BRD2, BRD3, BRD4, CBP (CREBBP gene product) and p300 (EP300 gene product). In luminal breast cancer, all five proteins demonstrate good correlation with GATA3 protein and histone H2B acetylation (Fig. 3.1.8 and Fig. A.3), suggesting that they interact with them and this interaction involves H2B histone.
Figure 3.1.8. Unsupervised clustering of histone acetylation sites. Distance: Euclidean, clustering method: Ward’s. Red indicates the highest acetylation/protein/mRNA level, blue – the lowest, white – the missing data. Mutations are labeled: M – missense mutation, F – frameshift insertion/deletion mutation, S – splice site mutation. mRNA level corresponds to log2(FPKM + 0.01), protein level corresponds to normalized relative protein abundance.

In fig. 3.1.8, we can clearly see a group of samples with high acetylation of H2B, high protein expression of GATA3, FOXA1, ESR1 and bromodomain containing proteins (on the right), and a group of samples with low expression of these proteins and low histone acetylation. However, there is also a group with elevated acetylation of H2B, but average or sometimes decreased expression of transcription factors. One possible explanation to this observation is that other proteins can interact with H2B as well, contributing to the longevity of acetylation marks. Such protein can be found by similar approach, searching for high protein correlation with H2B acetylation, but low mRNA correlation. Such analysis can be done in the future.

I also noticed that genes that were previously found as differentially expressed between samples with high and low H2B acetylation have dramatic differences between their protein and mRNA correlations with H2B acetylation. It turned out that in luminal samples these genes correlate with H2B acetylation on a protein, but not on a mRNA level. Many of these are chromatin-interacting proteins. This indicates the possibility that they are more likely to be involved in interaction with H2B, than to be regulated by H2B on the mRNA level (Fig. 3.1.9 A). Such correlation is not observed for H3 histone sites marking active chromatin (Fig. 3.1.9 B), or H2A, or H4 (data not shown). Among proteins with high correlation with H2B there are transcription factors: ARID1A, ARID1B, ARID2, FOXA1, GATA3; methyltransferases KMT2A, KMT2C and SETD2; EP300 and other chromatin interacting and modifying enzymes.
Figure 3.1.9. Average Spearman’s correlation between mRNA and protein levels of differentially expressed genes and histones acetylation. A. Correlation computed for H2B K5, K11, K11_K12, K15_K16 and K20_K23. B. Correlation computed for H3 K27, K27_K36 and K36.

To sum up, we showed that acetylation of histone H2B is significantly associated with CBP/p300 protein level across subtypes, suggesting that CBP/p300 may acetylate H2B N-terminal sites. We also demonstrated that in luminal subtypes protein level of various TFs, including luminal specific TFs, correlates with H2B acetylation better than that of their mRNA level, suggesting the direct or indirect interaction between those factors and histone H2B.
3.2 Metabolic Proteins Expression and Acetylation Interplay

3.2.1 Expression of metabolic enzymes characterize Basal subtype metabolism as glycolytic

Acetylation is the second most common post-translational modification in eukaryotic cells. Along with histone proteins, many non-histone proteins have been found to be acetylated as well. Among acetylated proteins, there are ones involved in chromatin remodeling, metabolism, translation, splicing and other major vital cellular processes (Choudhary et al. 2009). Acetylation of non-histone proteins has been shown to play essential role in cancer development, therefore it is important to see how the acetylation of those proteins can contribute to breast cancer subtyping and development.

To find out what cellular pathways are enriched for acetylated proteins in the existing acetylome dataset for breast cancer, the KEGG pathways enrichment analysis has been performed with acetylated proteins reported in the dataset as foreground and the whole list of genes from RNA-seq data as a background. The top ten overrepresented KEGG pathways are shown in Fig. 3.2.1.1. Among the enriched acetylated pathways there are spliceosome, ribosome, proteasome subunits along with enzymes involved in carbon metabolism, amino acid synthesis etc. (Fig. 3.2.1.1). All these gene products are housekeeping and relatively abundant in every cell, hence it is not surprising to see them enriched for acetylated proteins. This observation points out to one limitation of targeted mass-spectrometry measurement, which we have to keep in mind, while analyzing the data: less abundant proteins may not be captured by antibodies due to their saturation by excessive amount of housekeeping proteins.
I have decided to focus on acetylation of cancer metabolism and look for subtype differences in this process. Cancer is known to have a rewired metabolism to support constant needs in energy and monomers for building new cells. One possible way to reprogram the metabolism is to change the expression of metabolic genes. Another way to regulate protein activity without changing the expression is to use post-translational modifications. And acetylation is known to be a major regulatory modification of metabolic enzymes.

First, I evaluated how protein expression of metabolic genes is different across subtypes (Fig. 3.2.1.2). On figure 3.2.1.2 basal cancer subtype forms an isolated cluster, supporting the fact that basal triple negative subtype has a distinctive form of metabolism (Lanning et al. 2017).
Figure 3.2.1.2. Unsupervised clustering of protein expression of differentially expressed metabolic proteins and carriers (Anova, FDR < 0.05). Distance: Euclidean, clustering method: Ward’s. Red indicates the highest protein level, blue – the lowest, white – the missing data.

Many solute outer membrane carriers are upregulated in basal subtype along with glycolysis enzymes. Luminal subtypes and basal subtype tend to have their metabolic proteins regulated in opposite manner: highly expressed genes in one subtype are lowly expressed in another one. Her2 cells show mixed behavior, but the expression is more concordant with basal subtype, than luminal A/B. 7 out of 26 upregulated proteins in basal subtype are involved in glycolysis: glucose importers GLUT1/GLUT3, hexokinase HK2/HK3, phosphofructokinase PFKP and enolase ENO1 (t-test FDR < 0.05 compared to LumA/B). In addition, lactate dehydrogenase LDHB is also overexpressed in basal compared to all other subtypes (t-test FDR < 0.001), suggesting that the final compound of glycolysis pyruvate (Pyr) is further transformed into lactate and then exported from the cell by Monocarboxylic Acid Transporter 1 (MCT1), which is also upregulated (t-test FDR < 0.01 compared to LumA and Her2). Interesting that two out of three enzymes controlling glycolysis rate limiting reactions are upregulated (HK and PFKP) in basal, meaning that basal subtype probably has an elevated glycolysis flux compared to luminal A and B. I would also like to point out that PFKP gene
in basal subtype has higher copy number variation (CNV) compared to other subtypes, which probably causes its upregulated expression (Fig 3.2.1.3). Even though Her2 subtype shows high levels of PFKP protein product, the CNV of this gene is not altered in Her2, therefore, another mechanism might be involved in the upregulation of protein expression.

Figure 3.2.1.3. PFKP gene amplification, RNA level and protein level in breast cancer subtypes. Copy number level is considered as \( \log_2(\text{copy number tumor} / \text{copy number normal}) \). The left plot has a dashed line marking 0.5, a cutoff for gene to have one additional copy. FPKM is Fragments Per Kilobase of transcript per Million mapped reads.

The same association between CNV, RNA and protein level is observed for GLUT1 and LDHB (Fig. A.4 and A.5), but not for HK2/3 (data not shown). Noteworthy, along with upregulation of key glycolytic enzymes, basal subtype has significantly reduced protein expression of a key gluconeogenesis pathway enzyme - Fructose-1,6-bisphosphatase 1 (FBP1) (t-test FDR < 0.05 compared to LumA/B), strongly supporting the importance of glycolysis flux for basal subtype.

Glycolysis is tightly connected to serine synthesis pathway, which starts from glycolysis intermediate compound 3-phosphoglycerate (3-PG). Serine synthesis appears to be upregulated in basal subtype as well since enzymes controlling first two reactions, phosphoglycerate dehydrogenase (PHGDH) and phosphoserine aminotransferase 1 (PSAT1), are upregulated on both RNA and protein levels in basal
subtype (t-test FDR<0.05). Next, serine can be used for protein synthesis, or for synthesis of another amino acid – glycine. The enzyme catalyzing serine into glycine transformation in cytoplasm – Serine Hydroxymethyltransferase 1 (SHMT1) - is slightly downregulated in basal compared to luminal (t-test FDR<0.2 compared to LumA/B), while on contrary similar mitochondrial protein is upregulated (t-test FDR<0.005 compared to LumA/B). Such dysregulation of SHMT1/2 indicates that basal subtype may utilize mitochondrial transformation of serine to the glycine more actively as opposed to luminal subtypes, which rely on cytoplasmic reaction more. In mitochondria, glycine might be further decarboxylated by glycine decarboxylase (GLDC), which is also upregulated in basal breast cancer subtype (t-test FDR < 0.05).

Tumors also often use glutamine as an alternative source of energy to fuel TCA cycle and make citric acid required for amino acid synthesis (Scalise et al. 2017). Glutamine can be used for many transaminase reactions as a source of amino group. For example, glutamine is utilized by Phosphoribosyl Pyrophosphate Amidotransferase (PPAT), catalyzing the first step of purine synthesis; Guanine Monophosphate Synthase (GMPS), member of de novo guanine synthesis pathway; Glutamine--Fructose-6-Phosphate Transaminase 1/2 (GFPT1/2), controlling the flux of glucose into the hexosamine pathway; Asparagine Synthetase (ASNS), transforming aspartate into asparagine. All these proteins are upregulated in basal subtype (t-test FDR < 0.05, compared to LumA for PPAT; t-test FDR < 0.2, compared to LumA for GFPT1; t-test FDR < 0.05, compared to LumA/LumB for GFPT2, GMPS and ASNS). In order for all these reactions to be supported, tumor cell should have enough glutamine coming from extracellular space. However, none of known glutamine transporters are upregulated in any subtype (SLC1A4, SLC1A5, SLC6A19, SLC38A1, SLC38A2), except for SLC6A14, upregulated in Basal and Her2 subtypes on mRNA level (t-test FDR < 0.05, compared to LumA/LumB) (Scalise et al. 2017). In addition to being source of amino group
for other chemical compounds, glutamine can be hydrolyzed by Glutaminase (GLS), protein expression of which is significantly elevated in Basal subtype (t-test FDR < $10^{-4}$). Many tumors rely on glutamine because it can fuel TCA cycle in mitochondria: first Gln is deaminized into Glu, which is further transformed into α-ketoglutarate through reaction catalyzed by Glutamate Dehydrogenase 1 (GLUD1). As opposed to all other enzymes we have discussed so far, protein level of GLUD1 is significantly decreased in Basal subtype (t-test FDR < 0.05), suggesting fueling TCA cycle is not a preferable option for this breast cancer subtype.

### 3.2.2 Differential acetylation of cytoplasmic and mitochondrial metabolic enzymes

As we see from analysis of gene and protein expression of metabolic enzymes, Basal subtype exhibits aerobic glycolytic form of metabolism. In the frame of glycolytic metabolism, TCA cycle and oxidative respiration are usually suppressed (Vander Heiden, Cantley, and Thompson 2009). However, protein expression analysis has not revealed any changes in expression of TCA cycle enzymes in breast cancer subtypes. Hence, other mechanisms such as post-translational modifications might be involved in balancing mitochondrial reactions.

Therefore, I evaluated differentially acetylated sites across subtypes. I found 36 metabolic protein with differentially acetylated sites, 8 of which are involved in glycolysis and 9 in TCA cycle. In figure 3.2.2.1 clustering of normalized differentially acetylated Ac-sites is shown. Mostly all basal samples have high acetylation of mitochondrial enzymes, while cytoplasmic enzymes are hypoacetylated. A subset of luminal A samples has notably elevated acetylation of glycolytic enzymes such as Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH), ENO1, Glucose-6-Phosphate Isomerase (GPI), Triosephosphate Isomerase 1 (TPI1), Phosphoglycerate Kinase 1 (PGK1) and Pyruvate Kinase (PKM), as well as cytoplasmic paralogs of TCA cycle enzymes: malate dehydrogenase 1 (MDH1) and
Isocitrate Dehydrogenase (NADP⁺) (IDH1).

Figure 3.2.2.1. Unsupervised clustering of normalized acetylation level of differentially acetylated lysines of metabolic proteins and carriers (Anova, FDR < 0.05). Distance: Euclidean, clustering method: Ward’s. Red indicates the highest acetylation level, blue – the lowest, white – the missing data.

Interesting that there is a little overlap between differentially expressed and differentially acetylated glycolysis enzymes. The summary of observed changes in glycolytic enzymes across subtypes is shown in Table 3.2.2.1.
Table 3.2.2.1 Summary of protein expression and acetylation of glycolytic enzymes in breast cancer subtypes. Number of affected sites was chosen with t-test FDR < 0.1.

<table>
<thead>
<tr>
<th>GENE</th>
<th>PROTEIN EXPRESSION</th>
<th>ACETYLATION</th>
<th>NUMBER OF AC-SITES AFFECTED</th>
<th>NUMBER OF AC-SITES DETECTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCL2A1</td>
<td>Higher in basal</td>
<td>NA</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>SCL2A3</td>
<td>Higher in basal</td>
<td>NA</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>HK1</td>
<td>No change</td>
<td>No change</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>HK2, 3</td>
<td>Higher in basal</td>
<td>NA</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>GPI</td>
<td>No change</td>
<td>Lower in basal</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>ALDOA</td>
<td>No change</td>
<td>Higher in basal and her2</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>ALDOB, C</td>
<td>No change</td>
<td>NA</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>TPI</td>
<td>No change</td>
<td>Lower in basal</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>GAPDH</td>
<td>No change</td>
<td>Lower in basal</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>PGK1</td>
<td>No change</td>
<td>Lower in basal</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>PGAM1</td>
<td>No change</td>
<td>No change</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>ENO1</td>
<td>Higher in basal</td>
<td>Lower in basal</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>PKM</td>
<td>No change</td>
<td>Lower in basal</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>LDHA</td>
<td>No change</td>
<td>Lower in basal</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>LDHB</td>
<td>Higher in basal</td>
<td>No change</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

All differentially acetylated sites of glycolytic enzymes, except for those of aldolase A (ALDOA), are hypoacetylated in Basal subtype. Worth noting that basal subtype differs the most in acetylation of metabolic enzymes, but not others (Fig. 3.2.2.2).
In Table 3.2.2.2 the summary of changes in acetylation of TCA cycle enzymes and pyruvate dehydrogenase complex (PDC) is provided. As it seen from this table none of the TCA cycle proteins is differentially expressed, only one subunit DLAT of PDC. However, I found nine mitochondrial enzymes with upregulated acetylation in basal subtype.

Table 3.2.2.2 Summary of protein expression and acetylation of TCA cycle enzymes in breast cancer subtypes. Number of affected sites was chosen with t-test FDR < 0.1.

<table>
<thead>
<tr>
<th>GENE</th>
<th>PROTEIN EXPRESSION</th>
<th>ACETYLATION</th>
<th>NUMBER OF AC-SITES AFFECTED</th>
<th>NUMBER OF AC-SITES DETECTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>No change</td>
<td>No change</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>ACO2</td>
<td>No change</td>
<td>Higher in basal</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>IDH2</td>
<td>No change</td>
<td>Higher in basal</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>IDH3A,G</td>
<td>No change</td>
<td>No change</td>
<td>0,0</td>
<td>7,2</td>
</tr>
<tr>
<td>IDH3B</td>
<td>No change</td>
<td>Higher in basal</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>OGDH</td>
<td>No change</td>
<td>No change</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>SUCLG1,2</td>
<td>No change</td>
<td>Higher in basal</td>
<td>1,1</td>
<td>5,9</td>
</tr>
<tr>
<td>SUCLA2</td>
<td>No change</td>
<td>No change</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>SDHA</td>
<td>No change</td>
<td>Higher in basal</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>SDHB, C</td>
<td>No change</td>
<td>No change</td>
<td>0</td>
<td>1,0</td>
</tr>
<tr>
<td>FH</td>
<td>No change</td>
<td>Higher in basal</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>MDH2</td>
<td>No change</td>
<td>Higher in basal / lower in basal</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>PDHA1</td>
<td>No change</td>
<td>No change</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>PDHB</td>
<td>No change</td>
<td>Higher in basal</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>DLD</td>
<td>No change</td>
<td>Higher in basal</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>DLAT</td>
<td>Higher in basal</td>
<td>No change</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

We identified that glycolysis enzymes are upregulated and hypoacetylated in basal subtype, while TCA cycle enzymes are hyperacetylated. Next, I investigated if there are HATs or HDACs that cause such difference between subtypes.

Hence, I studied the association of expression of HATs and HDACs with acetylation of metabolic enzymes. The association was tested using linear model [2]:

$$Ac_{substrate site} \approx \beta_0 + \beta_1 * Pr_{enzyme} + \beta_2 * Pr_{substrate} + \epsilon$$

[2]
Deacetylase SIRT3 was found to have strong association with a number of mitochondrial enzymes acetylation sites. SIRT3 is a major deacetylase in mitochondria, for which many TCA cycle enzymes were reported to be a substrate (Cimen et al. 2010; W. Yu, Dittenhafer-Reed, and Denu 2012; Ozden et al. 2014). SIRT3 shows negative association (as expected for a deacetylase) with MDH2, Acyl-CoA Dehydrogenase Very Long Chain (ACADVL), Succinate-CoA Ligase Alpha Subunit (SUCLG1), 3-Hydroxymethyl-3-Methylglutaryl-CoA Lyase (HMGCL), catalyzing the final step of leucine degradation in mitochondria (Fig. 3.2.2.3). These proteins seem to be under control of SIRT3 deacetylase in mitochondria.

![Figure 3.2.2.3](image)

Figure 3.2.2.3. Coefficients of fitted linear models for HDACs [2]. Only coefficients with linear model coefficient FDR < 0.1 and p-value < 0.01 under random model are shown.

Linear models for another mitochondrial deacetylase SIRT5 were not significant under FDR cutoff and random model p-value cutoff. All other significant associations with negative $\beta_1$ seem to be not functional because all significant sites are located in mitochondria, while HDACs are either nuclear, or cytoplasmic. A number of positive associations can be explained by high protein coexpression of a
group of HATs and HDACs (Fig. A.5). At least, HDAC1, HDAC2 and HDAC3 protein correlates with the expression of various HATs.

Similar to HDACs plot, Fig. 3.2.2.4 is showing significant linear model coefficients for HATs.

Figure 3.2.2.4. Coefficients of fitted linear models for HATs [2]. Only coefficients with linear model coefficient FDR < 0.1 and p-value < 0.01 under random model are shown.

Again, various HATs display strong correlation with acetylation of cytoplasmic enzymes such as PGAM1, PGK1, GAPDH, G6PD, GPI, IDH1, Phosphofructokinase Liver Type (PFKL), MDH1, ATP Citrate Lyase (ACLY). Interesting that mostly all significantly associated with HATs Ac-sites are not differentially acetylated in basal subtype, and therefore HATs expression cannot explain difference in acetylation between subtypes. However, for one protein PGK1 differentially acetylated sites were found significantly associated with HATs. In PGK1 two out of five acetylated lysine found in linear model regression analysis (K15, K141, K146, K323, K361) are differentially acetylated across subtypes. Two Ac-sites K141 and K361 demonstrate association with nuclear TAF1 and KAT5 that might be explained by the fact that PGK1 has an alternative function in the nucleus, regulating DNA replication
and repair (J. Wang et al. 2007). Similar to PGK1, three out four significant MDH1 Ac-sites are differentially acetylated in breast cancer subtypes (K107, K118, K236, K239). As a result, linear regression analysis can only partially explain differential acetylation of metabolic enzymes in basal subtype compared to LumA/B.

Not only enzyme concentration can affect the amount of acetylated protein in cell, but also the concentration of Acetyl-coA, the essential compound for acetyltransferase reaction. Therefore, we searched the Acetyl-coA synthesis pathways for any alterations across subtypes (Fig. 3.2.2.5).

Figure 3.2.2.5. Schematic representation of Acetyl-coA synthesis pathways in basal subtype. Blue arrows mean protein downregulation, pink arrows mean protein upregulation, yellow circles – elevated acetylation. Adopted from (Narita, Weinert, and Choudhary 2018). ACC1 - Acetyl-CoA Carboxylase Alpha, or ACACA; ACC2 - Acetyl-CoA Carboxylase Beta, or ACACB; ACLY - ATP Citrate Lyase, ACSS2 - Acyl-CoA Synthetase Short Chain Family Member 2; PDC – Pyruvate Dehydrogenase Complex.

There are two pathways for acetyl-coA synthesis in cytoplasm. First one involves ACLY enzyme and citric acid, the second one involves ACSS2 enzyme and acetate. Both cytoplasmic acetyl-coA synthesis pathways turned out to be downregulated in basal subtype. The main acetyl-coA synthesis enzyme ACLY is decreased compared to Her2 (t-test FDR < 0.01), LumB (t-test FDR < 0.08) and LumA (t-test FDR < 0.2). Similar to ACLY, ACSS2 is slightly diminished in basal compared to all subtypes (t-
test FDR < 0.2), as well as mitochondrial ACSS3 (t-test FDR < 0.05), but not ACSS1 (t-test FDR < 0.3) (Fig.3.2.2.6A). In cytoplasm, acetyl-coA serves as a main source of carbon for fatty acid synthesis. However, in basal fatty acid synthesis appears to be downregulated as well. Acetyl-CoA Carboxylase Alpha (ACACA) and Fatty Acid Synthase (FASN), catalyzing the first two steps of fatty acid synthesis in cytoplasm, are downregulated on the protein level (ACACA t-test FDR < 0.03, FASN t-test FDR < 0.001) (Fig.3.2.2.6B). Regulatory beta subunit of Acetyl-CoA Carboxylase Complex (ACACB) is decreased only compared to LumA subtype (t-test FDR < 0.01).

Figure 3.2.2.6. Expression of Acetyl-coA metabolism proteins. A. Protein expression of proteins involved in Acetyl-coA synthesis in cytoplasm. B. Protein expression of proteins involved in fatty acid synthesis from Acetyl-coA in cytoplasm.
Such a noticeable downregulation of Acetyl-coA synthesis and metabolism in cytoplasm can be an indicator of a low concentration of Acetyl-coA in this compartment, which can in turn explain reduced acetylation of many Ac-sites in cytoplasm. It has been reported that expression of ACLY can predict the level of histone H2B, H3 and H4 acetylation level (Wellen et al. 2009; Carrer et al. 2017), indicating that expression level of this protein mirrors the concentration of its reaction product – Ac-coA, which in turns affect the acetylation of other proteins. In mitochondria and nucleus, another source of acetyl-coA is active - PDC, catalyzing pyruvate decarboxylation to Ac-coA. All subunits of PDC have similar expression level across subtypes, except for DLAT, being significantly upregulated in basal (t-test FDR < 0.001 compared to LumB, FDR < 0.15 compared to LumA). Another subunit of PDC, PDHB, was found to be differentially acetylated in basal subtype relatively to LumA/B (Table 3.2.2.2), but the function of these sites is unknown. Hence, the local concentration of Ac-coA in mitochondria and nucleus can be maintained by PDC in Basal subtype.

Taking together we identified possible reasons for such large differences between metabolic enzymes acetylation in subtypes. Hypoacetylated state of cytoplasmic proteins in basal subtype can be partially explained by expression of HATs and supported by possibly decreased level of cytoplasmic acetyl-coA. Hyperacetylation of mitochondrial enzymes in basal is observed probably due to downregulation of mitochondrial SIRT3 deacetylase.

### 3.2.3 Functional role of differentially acetylated

In previous sections, the differences in protein and protein acetylation levels between subtypes have been described. If upregulated expression of a protein is almost always can be treated as an increased rate of the reaction catalyzed by this protein, post-translational modifications can be as activating, as
inhibiting. Hence, there is a need in identification of functional role of these sites determined from literature.

The first glycolysis enzyme found with differentially acetylated sites is GPI. The fact that acetylation affects its activity has not been reported so far. However, in 3-dimensional space six out of fourteen detected acetylation sites are differentially acetylated and are located together on one side of the protein surface not interacting with other subunits of GPI. Four of them (K252, K447, K454, K524) frame a site believed to be important for F-6P binding (Cordeiro et al. 2004; Ji Hyun Lee et al. 2001). I hypothesize that acetylation of lysines surrounding the active site makes it more hydrophobic and not preferable for hydrophilic substrate binding. Thus, low acetylation may maintain high enzymatic activity.

Figure 3.2.3.1. Localization of GPI differentially acetylated sites relative to its active center. Differentially acetylated sites are marked in pink, not differentially acetylated sites are marked in grey.
K519 residue involved in catalysis reported by (Ji Hyun Lee et al. 2001) is marked in magenta. Catalytically important sites reported by (Cordeiro et al. 2004) are colored in white.

Similarly, PGK1 has two differentially acetylated sites (K216 and K220) located in proximal distance to the substrate binding site. The acetylation of K220 was determined to decrease enzymatic activity by disrupting the binding of ADP (S. Wang et al. 2015). The 3D conformation of native PGK1 active site can be seen in fig. 3.2.3.2

![Figure 3.2.3.2. Localization of PGK1 differentially acetylated sites relative to its active center. Differentially acetylated sites are marked in cyan. Substrates 3-PG and ADP are colored in white.](image)

Acetylated sites can interfere not only substrate binding, but also the binding of allosteric regulator as it happens in PKM. PKM becomes allosterically activated by fructose-1,6-bisphosphate (FBP) (Lv et al. 2013), and K433 acetylation prevents it from binding to PKM, and therefore, reduces the activity
of the enzyme. Moreover acetylation of K433 promotes nuclear location of PKM and triggers its kinase activity, as opposed to its canonical phosphatase activity in glycolysis (Lv et al. 2013). K433 site is differentially lower acetylated in basal subtype, allowing for binding the allosteric regulator and keeping glycolytic function of PKM on the required level.

To sum up, acetylation generally inhibits glycolytic enzymes in different ways. Given the diminished acetylation of Ac-sites in basal subtype, we can conclude that acetylation agrees with protein expression in maintaining glycolytic flux on the high level.

TCA cycle enzymes appear to be regulated by acetylation in a similar fashion. Almost all TCA cycle enzymes are subjects to deacetylation by SIRT3 (Sol et al. 2012). For instance, deacetylation of four lysines in SDH complex by SIRT3 was shown to increase its activity (Cimen et al. 2010). Two out of four these regulated lysines are differentially acetylated in breast cancer subtypes. Elevated acetylation of K179 and K538 of SDHA in basal subtype suggests that the function of this complex is decreased. Since SDH is the only unique enzyme acting as part of both TCA cycle and respiratory chain, we can suppose that respiratory chain is also affected by K179 and K538 acetylation. One possible explanation why K179 acetylation can inhibit the activity of SDHA was proposed by Peter Chhoy (Chhoy et al. 2016). He postulates that acetylation of K179 may prevent the binding of SDHE subunit of SDH complex, which is necessary for loading FAD inside the SDHA subunit.

Another enzyme MDH2 was shown to have Ac-sites, regulated by SIRT3, as well. One study has shown that increased acetylation of K185, K301, K307 and K314 upon inhibition of SIRT3 leads to the increased activity of MDH2 towards malate formation (Zhao et al. 2010). On the other hand, another study has demonstrated that upon SIRT3 inhibition and caloric restriction, K239 is significantly hyperacetylated, leading to the decreased enzymatic activity of MDH2 (Hebert et al. 2013). Such change in activity can be explained by K239 residue localization in MDH2 complex. It
locates nearby NAD$^+$ binding site and may impair its binding (Figure 3.2.3.3). Interesting that four residues reported by Zhao et al. are located together on the outer surface of MDH2 complex

Figure 3.2.3.3. Localization of MDH2 differentially acetylated sites. Differentially acetylated upregulated in basal sites are marked in pink. Differentially acetylated downregulated in basal sites are marked in blue. Not differentially acetylated cites, but previously reported as activating, are in cyan. Substrate NAD$^+$ is colored in magenta.

Interesting, K239 was found to be hyperacetylated in basal subtype, but not other four lysines. However, it is essential to notice that under normal conditions in mitochondria and low NADH level MDH2 catalyzes the reduction of malate to oxaloacetate, even though the reverse reaction is thermodynamically more favorable. In in vitro studies of MDH2 activity the rate of reverse reaction is always measured, therefore, it is hard to conclude the exact effect of acetylation on the malate
reduction reaction happening in mitochondria. But if acetylation has the same effect on both forward and reverse reactions and K239 acetylation is truly inhibiting, then we can say that MDH2 is less active in basal subtype that in others.

IDH2 has also been reported as SIRT3 substrate. The specific K413 position controlled by SIRT3 level was stated to be important for IDH2 dimerization required for proper functioning (Zou et al. 2017). In another study, acetylation mimicking mutagenesis demonstrated that acetylated K180, K251, K256, K272, K275 and K413 decrease the activity of IDH2 (Xu et al. 2017). Among these sites, only one site K272 was found to be differentially acetylated in the current dataset, not K180, K256, K275 or K413 (site K251 is not detected). This fact suggests that IDH2 activity may not be decreased in basal subtype compared to others. However, there is one site K193, which was not tested in previous studies, located in clasp-domain of IDH2 that is involved in protein dimerization and possible tetramerization (Fig. 3.2.3.4). K193 site locates in highly hydrophilic region of the clasp-domain, and acetylation might disrupt its structure since it makes lysines even more hydrophobic. Therefore, acetylation of K193 residue may interfere IDH2 dimer formation that is critical for its proper functioning. If it is so, then basal subtype has IDH2 activity decreased compared to other subtypes.
Figure 3.2.3.4. Localization of IDH2 differentially acetylated sites. Differentially acetylated sites are marked in pink. Protein is colored by hydrophobicity: red the most hydrophilic, white the most hydrophobic. Substrate NADP+ is colored in green.

Finally, Aconitase ACO2 has been also proved to be SIRT3 substrate along with other TCA cycle enzymes. However, the effect of acetylation has been poorly investigated. One study demonstrates that upon treatment with acetylating agent, the activity of ACO2 is increasing at low concentration of the agent and is pluming at its high concentrations (Fernandes et al. 2015). This controversial behavior indicates that either the degree of acetylation is important, or that acetylation of essential for catalysis sites is highly not favorable under low concentrations of acetic anhydride. Unfortunately, all four differentially acetylated sites are located on the outer surface of ACO2, relatively far from Fe-S cluster and substrate binding site, so it is hard to predict if acetylation of those site influences the activity.

To sum, we identified that the activity of glycolytic and TCA cycle enzymes is usually diminished upon acetylation, therefore, basal breast cancer subtype can be characterized by the increased level of glycolysis and decreased level of TCA cycle.
4 Conclusions

In this work we explored the associations of histone and non-histone proteins acetylation with breast cancer subtyping. In luminal A and B subtypes acetylation of N-terminal sites of histone H2B significantly correlates with expression level of luminal specific transcription factors FOXA1 and GATA3. Moreover, the correlation is higher with protein expression, than with mRNA expression, suggesting these factors might be part of complexes interacting with H2B acetylated sites. In addition, H2B acetylation correlates with protein expression of a number of tumor suppressor genes, indicating that H2B acetylation may be a new prognostic factor in luminal breast cancer. Acetylation of non-histone proteins can additionally characterize the metabolism of basal subtype. We showed that in addition to upregulated expression of glycolysis genes, basal cancer subtype maintains lowered acetylation level of cytoplasmic enzymes, but elevated acetylation of mitochondrial enzymes, which altogether help cancer favor glycolysis over the TCA cycle. Such differences in acetylation can be accounted for differential expression of mitochondrial HDACs and downregulated Ac-coA synthesis pathways in basal subtype compared to LumA/B. As a result, hypoacetylated cytoplasmic enzymes function more efficiently, while the activity of hyperacetylated mitochondrial enzymes appears to be decreased, supporting the established by gene expression vector of aerobic glycolytic metabolism in basal subtype of breast cancer.
Appendix A

Figure A.1. Gene and protein expression of luminal specific DE genes. Significance level of Wilcoxon test is shown.
Figure A.2. Spearman’s correlation between H2B N-terminal Ac-sites and luminal specific DE genes across all subtypes. A. GATA3, mutations are labeled: M – missense mutation, F – frameshift insertion/deletion mutation, S – splice site mutation. B. FOXA1. C. ESR1. mRNA level corresponds to log(FPKM + 0.01), protein level corresponds to normalized relative protein abundance.

Figure A.4. SCL2A1 gene amplification, RNA level and protein (GLUT1) level in breast cancer subtypes. Copy number level is considered as log2(copy number tumor/ copy number normal). The left plot has a dashed line marking 0.5, a cutoff for gene to have one additional copy. FPKM is Fragments Per Kilobase of transcript per Million mapped reads.

Figure A.5. LDHB gene amplification, RNA level and protein (GLUT1) level in breast cancer subtypes. Copy number level is considered as log2(copy number tumor/ copy number normal). The
left plot has a dashed line marking 0.5, a cutoff for gene to have one additional copy. FPKM is Fragments Per Kilobase of transcript per Million mapped reads.

Figure A.6. Self-correlation of protein expression of HATs and HDACs.


Parra, Michael A, David Kerr, Deirdre Fahy, Derek J Pouchnik, and John J Wyrick. 2006.


62


Xu, Yuqun, Lingwen Liu, Akira Nakamura, Shinichi Someya, Takuya Miyakawa, and Masaru


Protein acetylation in breast cancer, Karpova, M.S., 2018