Methods for Chemical Mapping of O-GlcNAc in the *Drosophila* Genome

by

Mike Myschyshyn
BSc. Biochemistry, University of Winnipeg, 2014

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

In the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Summer 2017

Approval

Name:	**Mike Myschyshyn**
Degree:	**Master of Science**
Title:	**Methods for Chemical Mapping of O-GlcNAc in the *Drosophila* Genome**
Examining Committee:	**Chair:** **Dr. Nicholas Harden** Professor
	Dr. David Vocadlo Senior Supervisor Professor
	Dr. Ryan Morin Supervisor Assistant Professor
	Dr. Don Sinclair Supervisor Senior Lecturer, Retired
	Dr. Leonid Chindelevitch Internal Examiner Assistant Professor School of Computing Science
Date Defended/Approved:	July 10, 2017

Abstract

O-linked N-acetylglucosamine (O-GlcNAc) is an important protein modification installed onto hundreds of nucleocytoplasmic proteins by O-GlcNAc transferase (OGT). Here, I discuss the development of an antibody-free metabolic feeding approach, which enables unbiased mapping of O-GlcNAcylated proteins in a genome-wide manner. This mapping method is detailed in *Drosophila* and compared to other O-GlcNAc mapping methods related to chromatin immunoprecipitation followed by sequencing (ChIP-seq), in order to demonstrate its overall efficacy. Using a combination of experimental and bioinformatics methods, I define new genes regulated by OGT. I also report on the development of robust software used to process and analyse time course ChIP-seq data, and prove its versatility and proficiency using both simulated and published data sets. This software is then applied to the analysis of a time course O-GlcNAc chemical mapping experiment in *Drosophila* larvae, generating the first ever time course ChIP-seq experiment performed on both a protein modification and in a living organism. Using this approach I am able to distinguish between loci that are more sensitive to O-GlcNAc cycling and those that are affected more by protein turnover. These studies provide an improved understanding of the regulation of gene expression by O-GlcNAc, while providing the wider community with new computational tools for time resolved analysis of genome-wide binding by proteins.

Keywords: O-linked N-acetylglucosamine (O-GlcNAc); O-GlcNAc transferase (OGT); Chemical biology; Chromatin immunoprecipitation followed by sequencing (ChIP-seq); Time course ChIP-seq analysis software; Bioinformatics.

Acknowledgements

Below is a list of individuals that I give special thanks to for assistance in completion of my degree.

Prof. David Vocadlo: I thank Dr. Vocadlo for his superb mentorship and significant contribution to experimental design. Dr. Vocadlo always supported my pursuits of various projects while providing critical feedback. Dr. Vocadlo has provided me with a productive environment where I was able to get a strong hold of a wide range of bioinformatics techniques which I imagine will translate into other fields.

Dr. Don Sinclair: I thank Dr. Sinclair for his highly active interest in projects throughout my degree and his expertise in *Drosophila* genetics. Dr. Sinclair executed most of the *Drosophila* genetics and has provided me with great leads on gene targets to explore.

Dr. Ta-Wei Liu: I thank Dr. Liu for teaching me fundamental biochemical techniques and critical conversation of experimental methodologies. Dr. Liu is an expert in biochemical strategies and executed many of the biochemical experiments.

Prof. Ryan Morin: I thank Dr. Morin for critical feedback of experiments. Dr. Morin has given me useful advice for sequencing analysis strategies and algorithm development.

Dr Samy Cecioni: I thank Dr. Cecioni for critical feedback of experiments and chemical synthetic prowess. Dr. Cecioni synthesized many of the compounds that we used in our experiments.

Alesha MacKay: I thank Alesha MacKay for emotional support and critical reading of this thesis.

I also thank the entire Vocadlo group (2014-2017) for a positive work environment, interesting conversation, and scientific advice.

Table of Contents

List of Tables

List of Figures

List of Acronyms

[6-4]PP	(6-4)pyrimidine-pyrimidone photoproducts
3C	Chromosomal capture
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5P	Five parameter sigmoidal curve
Abf1	ARS-binding factor 1
ARS	Autonomously replicating sequence
Ac_4GalNAz	Peracetylated N-azidoacetylgalactosamine
Ac_4GlcNAz	Peracetylated N-azidoacetylglucosamine
Ac_4ManNAz	Peracetylated N-azidoacetylmannosamine
AHA	Azido homoalanine
ATP	Adenosine triphosphate
Bcl-2	B-cell lymphoma 2
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
BtOGA	Bacterial homologue of OGA
BWA	Burrows-Wheeler Aligner
CARM1	Coactivator associated arginine methyltransferase
CAS	CRISPR-associated
cDNA	Circular DNA
ChAP-MS	Chromatin affinity purification with mass spectrometry
Chec-seq	Chromatin endogenous cleavage followed by sequencing
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CpG	Cytosine and guanine rich DNA
CRISPR	Clustered regularly interspaced short palindromic repeat
CRISPR-ChAP-MS	CRISPR-based chromatin affinity purification with mass spectrometry
CS-B	Cockayne syndrome cell line
CTCF	CCCTC-Binding Factor
CuAAC	Cu(I)-catalyzed Azide-Alkyne Cycloaddition
ddNTP	Di-deoxynucleosidetriphosphate
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleosidetriphosphate
dSfmbt	*Drosophila* Scm-related gene containing four mbt domains
EMS	Ethyl methyl sulphonate
EZH2	Enhancer of Zest Homolog 2
FLASH	4',5'-bis(1,3,2-dithioarsolan-2-yl)fluorescein
FLP	Flippase
FRET	Fluorescence Resonance Energy Transfer
FRT	Flippase recognition target
GalT	Galactosyltransferase
GFAT	Glucosamine-6-phosphate transaminidase
GFP	Green fluorescent protein
GlcNAc	N-acetylglucosamine
GlcNAz	N-azidoacetylglucosamine

GPP	Glycosylation prediction program
GSK-3β	Glycogen synthase kinase 3
HBP	Hexosamine biosynthetic pathway
HCF1	Host cell factor C1
HeLa S3	Human epithelial carcinoma cell line
Hi-C	High throughput chromosomal capture
HIF1	Hypoxia-inducible factor 1
HOX	Homeotic
HR	Homologous recombination
HSF1	Heat shock factor protein 1
HyCCAPP	Hybridization capture of chromatin associated proteins for proteomics
IFN-γ	Interferon-γ
IgG	Immunoglobulin G
IKKB	Inhibitor of nuclear factor kappa-B kinase
IRI	Incorporation Rate Index
mC	Methylcytosine
MEF	Mouse embryonic fibroblasts
mRNA	Messenger RNA
Muc1	Mucin 1
N6A	N6-methyladenine
NF-kB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NGS	Next-generation sequencing
NHEJ	Nonhomologous end-joining
NHF1	Human fibroblast cell line
NRSF	Neuron-restrictive silencer factor
NTP	Nucleosidetriphosphate
O-GlcNAc	O-linked β-N-acetylglucosamine
OGA	O-linked β-N-acetylglucosamine hydrolase
ogaKO	OGA knockout
OGT	O-linked β-N-acetylglucosamine transferase
p38	Tumour protein p38
p53	Tumour protein p53
PCR	Polymerase chain reaction
PDX-1	Insulin promoter factor 1
PFK1	Phosphofructokinase-1
Ph	Polyhomeotic
Pho	Pleiohomeotic
PhoRC	Pho Repressive Complex
PICh	Proteomics of isolated chromatin segments
PRC	Polycomb repressive complexes
PTM	Post translational modification
qPCR	Quantitative polymerase chain reaction
RNA	Ribonucleic acid
RNAi	RNA interference
RT-PCR	Reverse transcription polymerase chain reaction
RT-qPCR	Reverse transcription followed by quantitative real time PCR
S2	Schneider's *Drosophila* Line 2
SMRT-Seq	Single molecule real time sequencing
STAT1	Signal transducer and activator of transcription 1
Strvn	Streptavidin
SVM	Support vector machine
SWI/SNF	SWItch/Sucrose Non-Fermentable
Sxc	Super sex combs
TAB-seq	TET assisted bisulfite sequencing

TALEN	Transcription activator-like effector nuclease
TC	Time course
TChP	Targeted chromatin purification
TDCA	Time Dependent ChIP-seq Analyser
TET	Ten-Eleven Translocation
TPR	Tetratricopeptide repeats
TTI	Turnover Time Index
UAS	Upstream activation sequence
UDP-GlcNAc	Uridine 5'-diphosphate-N-acetylglucosamine
UV	Ultraviolet
VSV	Vesicular stomatitis virus
WGA	Wheat germ agglutinin
XR-seq	eXcision repair sequencing
YY1	Ying Yang 1
ZFN	Zinc finger nuclease
φC31	PhiC

Chapter 1: **Introduction**

1.1 Basics of genetic methods

In its most reduced form, the central dogma of molecular biology states that deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA), and RNA is translated into protein[1]. Given the biological context, there are exceptions to this rule. However, discoveries on how DNA is transcribed to RNA and how RNA is translated into protein, along with innovative techniques to study RNA and protein expression levels, have fundamentally altered the way scientists understand the role of DNA in cells and multicellular organisms. This thesis largely focuses on methods to study how proteins post-translationally modified with β-N-acetylglucosamine (GlcNAc) O-linked to serine and threonine residues (O-GlcNAc) bind to DNA and influence the expression of RNA and proteins. This work has been accomplished using next-generation sequencing (NGS), quantitative polymerase chain reaction (qPCR), western blot analysis, genetic fly models, and the development of bioinformatics software. This chapter introduces these core techniques, as well as others that are pertinent to the explanation of the results.

DNA[2] and its subsequent characterization[3], has transformed the field of genetics as we know it today. Inherent in the structure of DNA were self-evident implications regarding its replication, which implicated DNA as the likely source for passing genetic information from parents to offspring[4]. These early structural studies in combination with fundamental early studies that showed DNA to be a primary component of bacterial and viral virulence[5,6], has earned DNA a central position in the field of biology.

Given that DNA is such an important part of this thesis, the next section details the fundamentals of DNA sequencing. DNA sequencing allows recovery of the primary structure of DNA, that is, the linear order of base pairs (bp). Later sections discuss how protein-DNA interactions and modifications of proteins bound to DNA influence the regulation of gene expression, which essentially underpins the field of epigenetics.

1.1.1 Sequencing technologies

The first robust method developed to sequence DNA was invented by Frederick Sanger. Sanger sequencing involves the hybridization of a single stranded DNA template to a short primer, which then undergoes templated polymerization

catalyzed by DNA polymerase[7]. The catalytic activity of DNA polymerase allows elongation of primers by consecutive addition of deoxynucleoside triphosphates (dNTPs) that are complementary to the template sequence. Elongation occurs in the 5' to 3' direction, with respect to the primer being elongated, and 3' to 5' with respect to the template strand. The key concept that permits sequencing is chain termination. This process literally terminates the growing complementary chains at different bases within the sequence. This is accomplished using dideoxynucleoside triphosphates (ddNTPs), which lack a 3' hydroxyl group compared to their dNTP counterparts, resulting in an inability to form the next phosphodiester bond, effectively terminating chain elongation[8]. When ddNTPs are added to a reaction mixture at a low concentration compared to dNTPs, chain termination can be controlled such that each elongating sequence of DNA randomly and stochastically terminates where a ddNTP substitutes for its cognate dNTP.

Initial developments of Sanger sequencing used radioactively labeled ddNTPs of the four canonical dNTPs (adenine, thymine, cytosine, and guanine). The template to be sequenced is then used in four separate reactions, each with a pool of the four canonical dNTPs and with one of the ddNTPs. The four resulting radioactive pools of DNA products each contain sequences that were terminated at various positions where a specific canonical base is normally found. These pools are then separated electrophoretically on a polyacrylamide gel in four independent lanes and the resulting bands are visualized using X-ray film. The sequence of the complement to the template sequence is then defined by ordering the bands from shortest to longest (5' to 3'). The identity of the base at any position is established by the lane in which the band is observed and the corresponding ddNTP identity. Template sequence is then inferred from complementarity.

A major advancement in Sanger sequencing was the development of fluorescent ddNTPs. These new ddNTPs contained covalently attached fluorophores that allowed visualization of DNA products based on the emission spectra of the terminal ddNTP attached[9,10]. Each of the fluorescent ddNTPs possesses a unique emission spectrum, enabling chain termination Sanger sequencing to be performed in a single reaction tube, greatly simplifying the sequencing procedure. After completion of DNA elongation, the DNA products labeled with fluorescent ddNTPs in the reaction mixture is separated by size using polyacrylamide gel or capillary electrophoresis and the identity of the base complementary to the fluorescent ddNTPs can then be determined by the colour of the fluorescence emitted by the terminal ddNTP.

Advancements in robotics and further optimizations of ddNTP fluorophore chemistry allowed for a highly robust and reliable sequencing strategy. Eventually, many genomes were sequenced using high-throughput Sanger sequencing, including the human genome, whereby several facilities used Sanger sequencing to read fragments of the human genome, which were later combined computationally. Today, Sanger sequencing is highly automated with efficient fluorescent ddNTP chain terminating reactions and separation of reaction mixtures by microfluidics, resulting in chromatograms representing DNA sequences that are then read by computers in an automated manner.

Long sequencing times and high costs, however, eventually drove the need for new DNA sequencing technologies. NGS technologies, characterized by microscopic parallelization, such that large sequencing projects could be accomplished by a single machine, gained rapid popularity due to their high-throughput nature. A popular next-generation sequencing technology developed by the company Solexa, and now part of the company Illumina, relies on an ingenious reversible dye terminator that could be removed so that consecutive rounds of chain termination and elongation can occur[11]. This technology in combination with polymerase chain reaction (PCR)[12], which allows amplification of genomic fragments, and bridge amplification greatly simplifies preparation and sequencing of DNA libraries. These technologies were ultimately used to develop the Illumina, or sequencing by synthesis, approach[13].

Illumina sequencing proceeds by the following steps: purified DNA is fragmented. Fragmented DNA is ligated to DNA sequences called adaptors which contain primer binding sites, indices, and complementary DNA to the sequences that are hybridized to the flow cell. The adaptors are annealed to the very ends of the fragmented and ligated DNA and at the very ends of the adaptors are the specific sequences required for hybridization of complementary oligomers on the sequencing by synthesis glass flow cell. DNA fragments originating from the isolated DNA are referred to as inserts. The resulting pool of DNA containing various inserts with annealed adaptors is referred to as a library. The indices have a unique sequence such that they could be used as a bar code to distinguish between libraries constructed from various sources. The library containing DNA sequences of interest are then applied to an acrylamide-coated glass flow cell that has been homogeneously coated with two different short oligonucleotide sequences that are complementary to two sequences within the adaptor sequences. The library can therefore hybridize to the flow cell to create a lawn of short oligonucleotide sequences, some of which will bear a hybridized DNA fragment for sequencing.

Next, the DNA fragments to be sequenced undergo several rounds of bridge amplification in order to generate DNA clusters. The first step of bridge amplification requires a polymerase to create a complementary strand of all hybridized DNA, effectively elongating all short oligomers that are hybridized to a DNA fragment. This results in DNA oligomers covalently bound to the flow cell containing the complementary DNA sequences of interest. The original library is then denatured and washed away. Next, the DNA fragments bound to the flow cell are sufficiently flexible so that their free adaptor end can hybridize to a nearby complementary sequence that is bound to the flow cell. A polymerase then extends the short oligo resulting in a complementary sequence to the original that is now also bound to the flow cell. The DNA is denatured and several rounds of this bridge amplification are performed to create a cluster of DNA, which is sometimes called a polony. After amplification, all reverse strands are enzymatically cleaved and washed away, resulting in clusters, or polonies, that contain only forward strands. Sequencing is then carried out.

Sequencing proceeds through the introduction of primers that hybridize to the primer binding sites. A pool of reversible fluorescent ddNTPs is then added to the flow cell, leading to elongation of the primer by the ddNTP base defined by template complementarity. Excitement of the ddNTPs results in the emission of a unique wavelength of light at each cluster that depends on the specific fluorescent ddNTP added to the growing primer. The emission of a certain wavelength of light ultimately determines the identity of the base. Chain terminators are reversed chemically[14] and another round of ddNTP addition follows. Each round is called a cycle, therefore cycle number limits the length (bp) of DNA sequenced. When the cycle threshold is reached, the index is read by introduction of an appropriate index primer. This sequence of DNA is unique for each library allowing for binning of data so that libraries from different sources can be sequenced together. If paired end sequencing is specified, index two is read as well and the sequences undergo one more round of bridge amplification followed by removal of the forward strands. The remaining reverse strands are then sequenced in a similar fashion to the forward strands. Cluster generation is essential for this sequencing strategy so that the fluorescence emitted by ddNTP addition is intense enough for detection by digital cameras.

To finalize this process of sequencing, the resulting sequences obtained are finally aligned to the reference genome from which DNA was obtained. Computational resources and sensitive machinery make the initial investment cost of the sequencing equipment high, but the sequencing price per base is comparably inexpensive once initial investments are made. Significant bioinformatics efforts have

been put into making alignment practical. Nevertheless, downstream analysis is also essential in any sequencing efforts and will be discussed in the bioinformatics section. A visual summary of the Illumina sequencing method is shown in Figure 1.1.

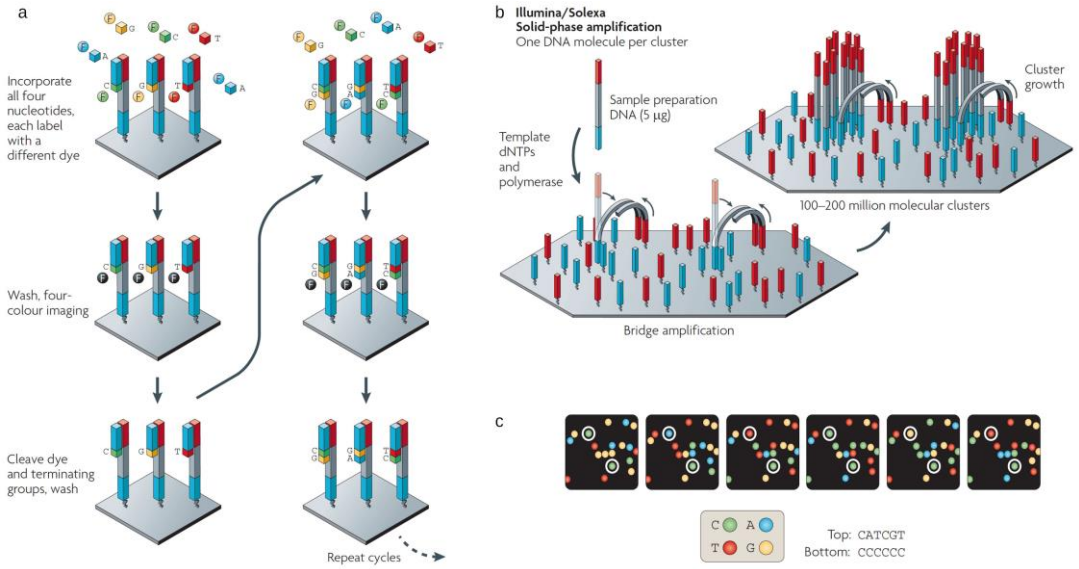


Figure 1.1. **Illumina sequencing by synthesis reversible termination method.** (a) Illumina four-colour reversible terminator ddNTPs compete for incorporation onto a primer based on complementarity to a template strand. Incorporation of a fluorescent ddNTP emits a unique wavelength. Templates are hybridized by their adaptors to oligos that are immobilized to a solid phase flow cell. After imaging, a cleavage step removes the fluorescent dyes, permitting elongation of the primer. (b) Cluster generation by bridge amplification enables the generation of a unique polony so that fluorescent signal through incorporated fluorescent ddNTPs is strong enough for cameras to detect. (c) Hypothetical images captured by cameras from the fluorescent incorporation of ddNTP onto polonies. Reprinted from Nature Reviews Genetic, vol. 11, Metzker M.L., Sequencing technologies - the next generation, p. 31-46, copyright 2010, with permission from Macmillan Publishers Ltd. [doi:10.1038/nrg2626].[15]

NGS sequencing technologies have advanced over the years to provide researchers with vast amounts of knowledge of various genomes. Some of these techniques include chromatin immunoprecipitation followed by sequencing (ChIP-seq), RNA sequencing (RNA-seq), and bisulfite sequencing to name a few. These will be discussed in the following sections in more detail. These now standard sequencing techniques can be applied to DNA isolated or prepared in a specific way in order to investigate various biological questions.

Today, new sequencing techniques continue to emerge and existing techniques continue to become more cost effective and practical. Among other sequencing techniques are pyrosequencing[16], which relies on luminescent detection of pyrophosphate release. Nanopore sequencing, which uses a membrane with an

electric potential applied from one side to the other. DNA is then passed through the pore and the change in current is monitored and the distinct signal is converted into a unique nucleotide[17], which in turn permits sequencing. The company PacBio's single molecule real time sequencing (SMRT-seq) technology uses single DNA polymerase molecules immobilized to the bottom of wells having zeptolitre volumes[18]. Incorporation of fluorescently tagged dNTPs is monitored in real time, which depends on the modification state of the template base, enabling sequencing of endogenous methylation and other DNA modifications at a single bp level[19]. This remarkable ability is an emerging need in the field of DNA sequencing as increasing attention is paid to the role of DNA modifications in regulating genome biology.

1.1.2 Quantification of gene expression

Ultimately, when researchers investigate how genes are regulated, it is important to quantify the expression of genes using reliable techniques. This section is dedicated to explaining standard strategies in quantification of gene expression.

The technique of separating a mixture of proteins in polyacrylamide gel on the basis of their size and then electrophoretically transferring the separated proteins to nitrocellulose membranes, where specific proteins can be quantified by the signal intensity stemming from binding of a fluorescently tagged antibody, was developed by Towbin and coworkers[20]. The term coined for this process is "western blot", and the process remains a staple technique to quantify expression of genes by protein levels. Western blotting requires quantification of a housekeeping protein, or proteins that should not have different expression levels in control and experimental samples. This technique also relies heavily on antibody specificity. Although popular, western blotting is low throughput. Conversely, whole proteome mass spectrometry experiments are more time consuming and costly, but allow for entire proteome comparisons[21].

An alternative to protein quantification is messenger RNA (mRNA) quantification. The semi-quantitative technique called reverse transcription polymerase chain reaction (RT-PCR) is a modified version of PCR[12], wherein RNA is reverse transcribed by the enzyme reverse transcriptase into complementary DNA (cDNA). cDNA is then amplified using standard PCR and the resulting amplified DNA can be quantified by fluorescent staining after size separation in a gel. Because PCR amplification is exponential in nature, this process is ill suited for determining the exact quantities of mRNA. A better approach is to use reverse transcription followed by quantitative real time PCR (RT-qPCR or qPCR). qPCR monitors the amplification of cDNA (obtained from mRNA by reverse transcription), in real time using

fluorescent dye that hybridizes to double stranded DNA. qPCR machines rely on primer hybridization to the cDNA template which is amplified with DNA polymerase and nucleotides. Each amplification, or cycle, results in a new double stranded cDNA copy for every template, and therefore, amplification is exponential in nature. Eventually, the fluorescent signal given off by a fluorescent dye that binds to double-stranded DNA is so intense that the detectors of fluorescence reach their limit of detection. The PCR cycle at which saturation of fluorescent signal occurs can then be compared in different samples after normalizing to housekeeping genes to enable researchers to get a precise relative quantification of mRNA in the different samples. This process has been highly automated through sophisticated machinery, allowing researchers to quickly test gene expression with very little input mRNA.

High throughput RNA analysis strategies also exist. Microarrays containing thousands of hybridized non-fluorescent DNA probes can be used to quantify gene expression[22]. Here, a population of cDNA from a sample is fluorescently labeled, which then hybridize to the modified and complementary DNA oligomers immobilized on the microarray. Coordinates of each spot on a microarray correspond to specific transcripts and the amount of cDNA binding can be quantified using fluorescence. The accuracy of this technique has been found to match RT-qPCR standards. However, a better technique is RNA-seq, whereby transcriptomes are directly sequenced using NGS platforms[23]. RNA-seq benefits from the ability to distinguish isoforms and to detect mutations, which is not directly feasible using microarrays.

1.1.3 Fly as a model organism

Model organisms are often studied to research biological processes related to humans. An attractive model organism is easy to maintain in a laboratory setting and offers particular experimental advantages. *Drosophila melanogaster* is one of the most common animals used as a model organism and has been since Morgan first chose to study fly for an evolution experiment in 1909[24]. In addition to their short life span and ability to produce large numbers of offspring, fruit flies provide an ideal choice as a developmental model due to their well-known anatomy and the variety of well-established mutations available for study. Although the anatomic divergence between fruit flies and humans is obvious, many fundamental molecular pathways are highly conserved[25]. Therefore, *D. melanogaster* are particularly advantageous for the investigation of molecular and cellular mechanisms underlying human diseases[26]. In fact, ~75% of human genes associated with disease have homologues in the *Drosophila* genome[27]. The popularity of *D. melanogaster* over other multicellular model organisms has steadily increased in recent years as new techniques have

been developed to more easily manipulate the fruit fly genome[28]. Such developments include, the ability to create molecularly designed deletions and mutations, improved genetic mapping methods, new transgenic approaches to create transgenic flies, and the ability to clone and modify large fragments of DNA[28].

Early genome engineering techniques in *Drosophila* relied on random mutagenesis followed by phenotypic screening. A popular mutagen is ethyl methyl sulphonate (EMS) due to its relatively low toxicity to humans and reported lack of sequence bias during mutagenesis[29]. Applications of transposable element mutagenesis[30] to the genetic analysis of *Drosophila* blossomed with the advent of a binary system that separated genes encoding for the transposase, necessary for transposition of the transposable elements, and the transposable element itself[31].

Targeted expression systems using the yeast transcriptional activator GAL4 has been used to enhance expression of genes engineered with an upstream activation sequence (UAS)[32]. GAL4 binds the UAS, inducing the expression of the downstream gene. Gal4 expression can be driven by developmentally specific or tissue specific promoters, enabling an array of versatile, targeted expression in a time and spatially resolved manner.

Using transposons for engineering purposes suffers from the difficulty in targeting specific genes. In particular, the transposable P element could allow pseudo-random insertions of specific sequences into certain regions of genomic DNA. In *Drosophila*, mutagenesis through gene targeting became possible with the development of homologous recombination using various recombinases[33,34]. Techniques that use site specific recombination to exchange homologous DNA with a genomic target sequence include the yeast derived Flippase (FLP) and Flippase recognition target sequence (FRT) (FLP/FRT), the bacteriophage P1 derived Cre recombinase and LoxP sequence (Cre/LoxP), and the PhiC31 (φC31) integrase. In Cre/LoxP systems, Cre targets LoxP sites and creates a double stranded break, resulting in the possibility of a deletion, insertion, inversion, or translocation depending on the orientation of the flanking sites LoxP and presence of homologous DNA. FLP recognizes FRT sites to perform similar activities as Cre[35]. FLP and Cre are examples of tyrosine integrases, with differing efficiencies and sequence recognition motifs. Serine integrases also exist, including φC31, which outperforms tyrosine integrases in their ability to unidirectionally excise and integrate DNA, making them a powerful research tool[36].

Another targeted genetic manipulation technology is RNA interference (RNAi). RNAi is a powerful approach whereby double stranded RNA or short hairpin RNA is cleaved by a protein called Dicer, to create a single-stranded RNA molecule

that hybridizes to complementary mRNA to promote its cleavage by associating with the RISC protein complex, thereby silencing the targeted gene[37]. The major benefit of RNAi is the ability to rapidly target any gene of interest in a genome-wide manner that enables screening[38]. RNAi libraries targeting almost all genes in *Drosophila* have been used to identify novel genes required for cell survival and to define phenotypes associated with gene knockdown[39]. Further, RNAi transgenes can be used to specifically silence genes *in vivo*. Limitations of RNAi include transient expression of the siRNA and incomplete knock-down of the target gene as well as off target effects that are sometimes seen.

Once the *Drosophila* genome was sequenced[40], researchers were more easily able to explore targeted genome-editing approaches, providing a targeted alternative to phenotypic screens. These editing methods include the use of meganucleases, zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and CRISPR-Cas systems (clustered regularly interspaced short palindromic repeats and CRISPR associated proteins). Genome-editing techniques enable insertions, deletions, or replacements of genomic DNA in a living organism using engineered nucleases that create site-specific double strand breaks. DNA breaks are then repaired either through non-homologous end-joining (NHEJ), resulting in deletions, or through homologous recombination (HR), resulting in insertions or replacements.

The development of ZFNs for genome editing made the execution of targeted HR easier, and eliminated the necessity of first introducing P element constructs[41]. Zinc fingers with different DNA sequence specificity can now be engineered to create sequence targeted ZFNs. However, the high risk of off target effects due to difficulty in sequence targeting is a limiting factor of this technology.

Transcription activator-like effectors (TALEs) are virulence factors first found in *Xanthomonas* plant pathogens that bind to specific genes in the host to promote pathogen spread[42]. TALEs are repetitive blocks of protein with a small portion of variability in each repeat that dictates the nucleic acid binding specificity of the repeats[43]. Specific ordering of TALE repeats enables locus specific binding and has enabled their use in biotechnology applications. TALEs fused to nucleases (TALENs) have been demonstrated to specifically cleave and knockout genes of interest via NHEJ[44]. Although TALENs show greater specificity and simplicity as compared to zinc fingers, the involved experimental design requirements have caused a shift to newer genome-editing technologies, such as clustered regularly interspaced short palindromic repeats (CRISPR) and the RNA guided DNA nuclease CRISPR associated protein 9 (Cas9), or the CRISPR-Cas system.

Many bacteria and archaea contain CRISPR DNA which is essentially a library of viral DNA signatures. Bacterial and archaeal organisms have adapted an immunity strategy that cleaves foreign DNA using Cas nucleases and incorporates the cleaved product into the CRISPR library. These can then be used to guide Cas proteins to cleave foreign DNA specifically to defend against these pathogens[45]. In the past decade, this system has been shown to generate DNA double stranded breaks and has been modified to include a dual nuclease (Cas) RNA guided approach to edit genomes[46]. CRISPR-Cas has since been expanded to generate stable cell lines and organisms with insertions, deletions, point mutations and frame shifts at targeted loci[42].

Overall, the continually exciting and intellectually stimulating field of *Drosophila* genetics exemplifies the power of continually evolving applications of genetic engineering strategies. Through the manipulation of genes within *Drosophila*, many interesting phenomenon have been discovered and translated into mammalian systems, offering a host of insights into human biology.

1.2 Epigenetics

The human genome has been sequenced and annotated[47,48] and approximately 20,000 genes have been discovered. Given that all of the somatic cells in an individual contain the same genetic code, scientists wondered how the tissue-specific patterns of gene expression could be generated. This problem drove fundamental research in the area of differential regulation of gene expression and, more recently, in an area that we now term epigenetics. Epigenetics is the study of molecular factors that change gene expression rather than alterations in the genetic code. These factors include DNA and protein modification which ultimately drive protein-DNA interactions and genome architecture.

In order to understand mechanisms of epigenetic inheritance, it is important to understand the higher order of DNA in cells. In eukaryotes, DNA is localized within the nucleus and is organised by histone and non-histone protein complexes. The functional packaging unit of DNA is composed of an octamer of histones which DNA winds around to form a nucleosome. The nucleosome can contain different isoforms of histones and various combinations of histone PTMs. This results in a potential for the nucleosome to have many chemical makeups leading to various protein-DNA dynamics which inevitably permits the cell to tailor the chromatin architectural state of individual loci. DNA itself can be modified as well, the most commonly studied base being cytosine, which further expands the potential for the cell to manifest various chromatin states at individual loci. In general, compact DNA is referred to as

heterochromatin and open DNA is referred to as euchromatin, although these terms have been superseded by more specific sub-classifications of chromatin based on their molecular makeup[49]. A general illustration of higher order chromatin is shown in Figure 1.2.

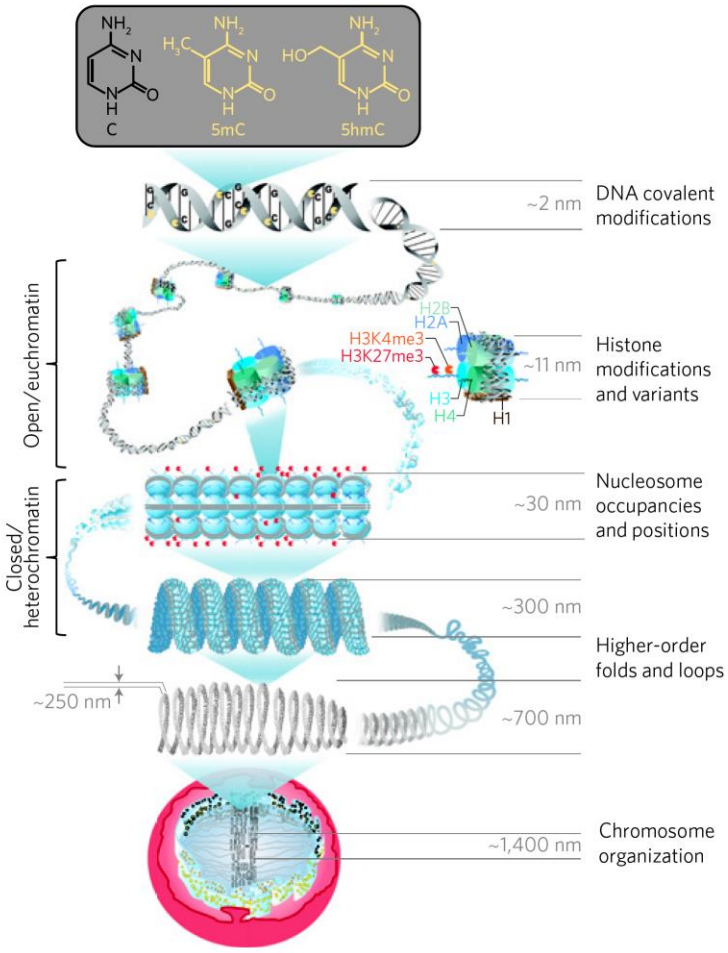


Figure 1.2. **Higher order of DNA in eukaryotes.**
Single nucleotide modifications are the smallest epigenetic code written on chromatin (top). Nucleosomes are constituted of DNA wrapped around an octamer of histone proteins. Histones can be post translationally modified (PTM), such as histone 3 lysine 27 trimethylation (H3K27me3). Nucleosomes play a fundamental role in chromatin packaging, and various histone PTMs impact chromatin structure. Heterochromatin is more tightly packaged relative to euchromatin, and is generally transcribed to a lesser extent. Reprinted from Nature Nanotechnology, vol. 8, Aguilar C.A. and H.G. Craighead, Micro- and nanoscale devices for the investigation of epigenetics and chromatin dynamics, p. 709-718, copyright 2013, with permission from Macmillan Publishers Ltd. [doi: 10.1038/nnano.2013.195].[50]

The classic Mendelian system of inheritance claims that alleles from each parent contribute to the overall phenotype of an offspring and that the nature of the phenotype is determined by the dominant relationship between the parental alleles. Upon the discovery of DNA and its essential role in replication and inheritance, it was believed that the primary genetic code was the fundamental contributor to heritable

traits. This classical genetics approach cannot explain why cells with the same genetic code are able to take on different phenotypes and cellular function. On the other hand, epigenetics explains these observed phenotypic differences by means of factors other than the base sequence of genes. Essentially these factors consist of complex molecular machinery that is able to alter chromatin structure. For example, the Polycomb group (PcG) proteins, first discovered in *Drosophila*, were found to be essential in maintaining the silent state of tissue specific genes[51,52], and most are chromatin modifying proteins that confer the silent state of chromatin. In contrast, the Trithorax group (TrxG) proteins maintain activation of tissue specific genes[51,52], and they have the opposite effect on chromatin. Another prominent example of epigenetic effects is the phenomenon of imprinting, where parental specific contribution of active versus silenced alleles occurs. This phenomenon results from DNA modifications. Ultimately, it is important to remember that DNA is a dynamic and responsive macromolecule and interpretation of gene expression requires multiple considerations. Broad categories of mechanisms to maintain regions of silent or active chromatin states depend on self-propagation of histone marks and DNA sequence specific recruitment of chromatin modifiers, illustrated in Figure 1.3. Select molecular details of these briefly mentioned epigenetic concepts will be discussed in more detail in the following sections.

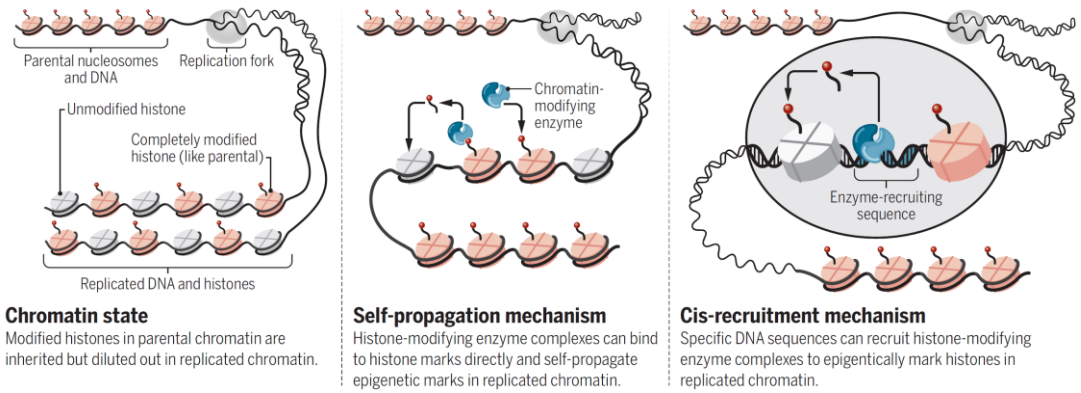


Figure 1.3. **Mechanisms to maintain regions of silent chromatin states.** During replication, silent histone marks are diluted and distributed equally among the parent and daughter chromatin (left). Silent histone marks can be re-established through self-propagation (middle) via recruitment of histone modifying enzyme complexes, or through DNA sequence specific recruitment of histone modifying enzyme complexes (right). Reprinted from Science, vol. 356, De S. and J.A. Kassis, Passing epigenetic silence to the next generation, p. 28-29, copyright 2017, with permission from AAAS [doi: 10.1126/science.aan1493].[53]

A major motivation in biological sciences is the need to pursue research projects that will ultimately benefit the well-being of humans, usually through therapeutic or industrial applications. Epigenetics has generated much interest because of its potential for providing insights into new therapeutics. However, due to the infancy of this field there are currently limited therapeutics involving epigenetic regulation of genes[54,55]. Nevertheless, many promising epigenetic therapeutics are currently under development[54]. Overall, amazing progress in the field of epigenetics has been made in recent decades, which will be highlighted in the next few sections. Key techniques that permit scientists to study such phenomenon will also be discussed.

1.2.1 Histone modifications

The fundamental structural protein complex that is responsible for DNA organization is the nucleosome. The nucleosome is composed of an octamer containing two copies of each of the histones H2A, H2B, H3, and H4[56]. A significant amount of research has focused on defining the exact physical structure of nucleosomes bound to DNA. Of note, crystallographic and biophysical studies have found that 146 bp of DNA are organized in a superhelix around the nucleosome with specific histone-histone and histone-DNA interactions coordinating important structural features[57]. Besides nucleosomes, protamines have also been shown to organize chromosomes in certain cell types[58]. Interestingly, the core nucleosome complex has been shown in some cases to contain a range of histone variants[59], which are speculated to play cell type and cell cycle specific roles.

Much research has been devoted towards defining how nucleosomes interact with DNA, how and why they are positioned at specific regions of DNA and why nucleosome positioning varies based on cell type. In yeast, the adenosine triphosphate (ATP) driven SWItch/Sucrose Non-Fermentable (SWI/SNF) complex was discovered to influence repositioning of nucleosomes[60]. Later, it was found that CCCTC-Binding Factor (CTCF) domains played a role in anchoring nucleosome positions[61]. A time course Fluorescence Resonance Energy Transfer (FRET) experiment revealed rapid DNA winding and unwinding at nucleosomes that occurs at the millisecond time scale[62]. These studies demonstrate processes that govern chromatin remodelling.

Additional complexity has been uncovered through fundamental biochemical experiments that have identified histone PTM "writers", "readers", and "erasers"[63–71]. As part of the histone code, "writers" serve to modify histones with various PTMs, "readers" function as proteins that read histone PTMs and carry out downstream

processes, and "erasers" serve to remove various histone PTMs. These proteins influence active or silent states of chromatin. Of note are the Polycomb Repressive Complexes (PRC) 1 and 2. PRC1 and PRC2 compact DNA in part by ubiquitination of H2A and di- and tri-methylation of H3K27, respectively[72,73]. Polyhomeotic (Ph) is a core component of PRC1. The PRC2 protein Enhancer of Zest Homolog 2 (EZH2) catalyzes the trimethylation of H3K27, which is then recognized by PRC1, resulting in ubiquitination of H2AK119 (118 in *Drosophila*) by the E3 ubiquitin ligase Sex Combs Extra (Sce)[74]. Importantly, a plethora of specific types of mutations in histone-modifying genes have been discovered in cancers[75]. In *Drosophila*, PRC1 and 2 complexes are recruited to DNA motifs called Polycomb Response Elements (PRE). Recent reports suggest that PREs and similar sequences in yeast are required for maintenance of histone PTMs[76–78]. It has also recently been shown that the repressive mark H3K27me3 can self-propagate by allosteric activation of PRC2[73]. This observation has led to the hypothesis that cells can balance the regulation of gene expression by modulating factors involved in maintenance and propagation of histone marks[53].

Many other histone PTMs have been discovered that are linked with activating or repressing capacity[72,79]. Because histone PTMs can have combinatorial effects, novel techniques are being developed to investigate the effect of nucleosomes with multiple PTMs. A recent strategy using a single-cell technique in which nucleosomes were hybridized to glass slides and imaged using fluorescent antibodies specific to histone PTMs, enabled researchers to resolve truly bivalent modified nucleosomes, which were found to vary by cell type[80]. Furthermore, the genomic coordinates of these nucleosomes were determined by sequencing. Synthetic standards of nucleosomes with particular modifications are also being developed to help clarify how proteins are recruited to their binding sites[81]. In this regard, an important question to ask is how PcG proteins are recruited to nucleosomes to carry out their functions. This is fundamental to the understanding of development and is an intense field of study. It is known that many different transcription factors influence PcG recruitment. Not surprisingly, many transcription factors involved in recruitment are also mutated in various cancers[82]. Non-coding RNA has also recently been found to mediate PcG recruitment[83]. Further, DNA modifications have emerged as mediators of PcG recruitment, and increasingly are attracting interest due to the environmentally sensitive nature of these modifications. The next section will briefly discuss DNA modifications.

1.2.2 DNA modifications

Modification of DNA bases is an essential strategy for the epigenetic regulation of gene expression in mammals. The Ten-Eleven Translocation (TET)-catalyzed oxidation of methylcytosine (mC) leads to the formation of 5-hydroxymethylcytosine (5hmC)[84,85], 5-formylcytosine (5fC)[86,87], and 5-carboxylcytosine (5caC)[87,88] bases. These modified bases are intermediates in the active DNA demethylation pathway[89,90] and also serve as stable[91,92] epigenetic marks with unique functions[93–95]. An overview of cytosine modification is illustrated in Figure 1.4. Cytosine is not the only DNA base modified in eukaryotes. Recently, N6-methyladenine (N6A) has been discovered in *Drosophila* and has been shown to be important in development[96]. Accordingly, particular interest has been devoted to developing methods to identify DNA modifications using DNA sequencing technology and to discover their roles in biology. Single base sequencing techniques using sequencing by synthesis methods rely on chemical treatment of DNA to modified or unmodified bases such that they are sequenced as another base. For example, bisulfite-sequencing requires treatment of DNA with bisulfite, which converts cytosine, 5fC, and 5caC to uracil, while mC and 5hmC remain as cytosines[97]. Oxidative bisulfite sequencing, which precedes bisulfite sequencing with potassium perruthenate treatment, results in conversion of 5hmC to uracil along with endogenous 5fC and 5caC, allowing differentiation of mC and 5hmC bases[98]. TET-assisted bisulfite sequencing (TAB-seq) is a strategy to map single base positions of 5hmC[99]. Here, endogenous 5hmC is blocked with glucose by ß-glucosyl transferase. Remaining cytosines are oxidized using TET and then converted to uracil through bisulfite treatment. Non-single base approaches also exist akin to chromatin immunoprecipitation (ChIP) followed by sequencing (ChIP-seq) procedures where antibodies for specific modified bases are developed[100,101]. An *in vitro* chemical labeling technique to purify 5hMeC DNA has also been developed, relying on bioorthogonal chemistry that specifically ligates a biotin handle to the hydroxyl group on 5hmC[102]. The single base sequencing techniques described above all suffer from the requirement for high sequencing depth, but generally supersede the non-single base resolution strategies [97]. Overall, DNA modification mapping strategies are proving to be essential tools in epigenetics research.

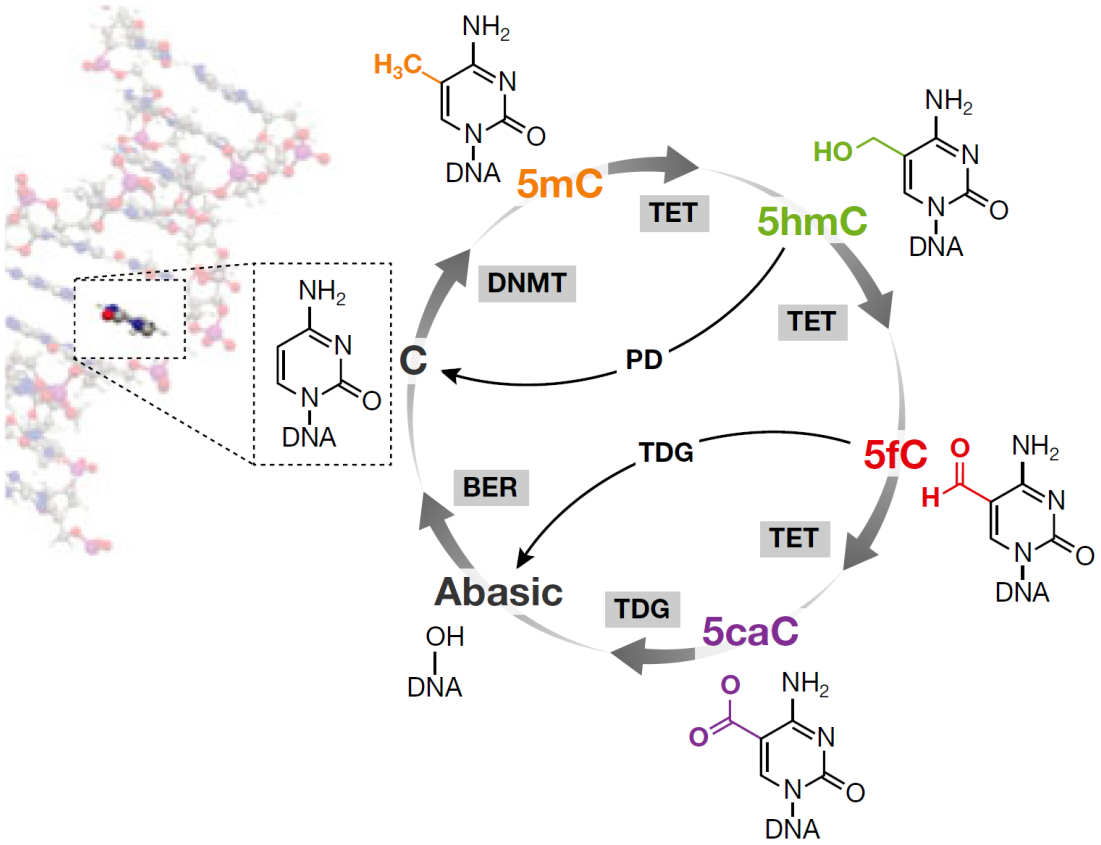


Figure 1.4. **Cytosine modifications in mammals.**
Unmodified cytosine can be methylated by DNA methyl transferases (DNMTs) and further oxidized by TET enzymes. 5fC or 5caC can be recovered back to unmodified cytosine through the action of thymine-DNA glycosylase (TDG) leading to an abasic site which is repaired through the base excision repair (BER) pathway. Passive dilution (PD) indicates replication based loss modified cytosines. Reprinted from The EMBO Journal, vol. 33, Tuesta L.M. and Y. Zhang, Mechanisms of epigenetic memory and addiction, p. 1091-1103, copyright 2014, with permission from John Wiley and Sons [doi:10.1002/embj.201488106].[103]

Since most modifications on cytosine occur in cytosine and guanine (CpG) rich areas of the genome, methods have been developed to enrich CpG rich DNA prior to chemical treatment and sequencing of DNA in order to reduce sequencing costs[104]. This is a partial solution to the problem. Other superior techniques are being developed to sequence-identify endogenous DNA modifications. For example, SMRT-seq can be used to identify endogenous modification status of bases by direct sequencing[18].

Experimental evidence suggests that DNA modifications may have an array of fascinating cellular roles. Historically, imprinting was discovered to be a major role of DNA methylation, which caused parental specific contribution of active versus silent alleles. Imprinted alleles are referred to genes that are silenced in a specific parent, thus the imprinted gene is contributed by a single parent. Of note, the *Igf2* and *H19* loci are well known to be imprinted in the maternally contributed allele, and

therefore the paternal allele is necessary for proper offspring development[105]. Recently, various environmental treatments have been found to impact DNA methylation and these have therefore gained increasing attention in the research community. In particular, exposure to harmful toxins has been shown to alter the DNA methylome[106]. Researchers have also shown that 5hmC levels vary at genes in a part of the brain associated with addiction in mice treated with cocaine[107]. Changes in CpG methylomes have also been observed in alcoholics[108], in embryos of alcohol treated mice[109], in placental DNA of smoking mothers[110], and in type 1[111] and type 2[112] diabetes sufferers. In addition, cytosine 5hmC has been shown to be essential for development[113]. Enzymes responsible for methylation of cytosine and mC itself were shown to contribute actively to neuronal plasticity[114]. mC was also shown to inhibit CTCF binding at exons of certain genes, resulting in alternatively spliced transcripts[115]. There are many other diseases and epigenetic phenomena in which DNA modifications are implicated[106,116]. To summarize, DNA modifications play an important role in epigenetic regulation of genes, most likely through recruitment of proteins[117].

1.2.3 Techniques in epigenetics

The study of epigenetics largely relies on investigations of the DNA interactome to infer links between protein-DNA and gene expression. The binding of proteins to DNA can be detected by ChIP, where bound proteins are crosslinked to DNA at loci to which they bind and interact with DNA. DNA is then sheared and the regions of DNA bound to proteins of interest are enriched using an antibody. The immuno-purified DNA is released from bound proteins by reversing the crosslinks and by protein digestion after which the liberated DNA can be interrogated with primers for loci of interest via PCR and analysed by gel electrophoresis. More accurately, one could ChIP DNA and amplify loci by qPCR. Both these strategies require specific primer design and DNA amplification for each locus and do not easily facilitate the discovery of novel loci at which specific proteins are bound, but rather investigate or validate loci at which interactions are suspected. Efforts to explore DNA interactomes in a faster and in a high throughput manner led to the development of DNA microarrays that allow hybridization of immuno-purified DNA. This strategy is called ChIP on chip, and was first implemented to investigate cohesin binding to chromosome 3 of yeast[118]. Soon after, microarrays were developed that contained all open reading frames and intergenic yeast DNA[119,120]. Today, whole genome microarrays are available for a variety of genomes with relatively high resolution probes of around 50 bp in length.

Clever techniques to investigate DNA interactomes that incorporate the high throughput nature of NGS exist. ChIP-seq was introduced in 2007, and quickly developed into a concrete DNA interactome investigation technology through repeated demonstration of its consistency and concordance with older studies as well as its obvious utility for studies in a genome-wide manner[121–124]. This technique relies on sequencing ChIP enriched DNA using NGS, thereby eliminating the need for ChIP on chip. ChIP-seq eliminates many of the limitations inherent in ChIP on chip experiments such as the difficulty and cost of constructing chips with highly variable regions as well as their limited resolution.

Even the very first ChIP-seq experiments revealed the obvious potential of this technology. The distribution of over 20 methyl histone marks in human CD4+T cells were determined by ChIP-seq and compared with ChIP-seq data for RNA polymerase II, which was used to pinpoint active genes. The locations of many of these histone marks were consistent with previously assigned repressing or activating roles[122]. ChIP-seq investigations on the distribution of H3K27me3 and H3K4me3 in three mouse cell lines revealed that these two histone marks regulate repression and expression, respectively, of genes, and these ChIP-seq experiments were supported by RT-PCR experiments[124]. Furthermore, occupancy of both the repressive mark H3K27me3 and activating mark H3K4me3 at certain promoters, called bivalent promoters, were referred to as poised genes. In human, ChIP-seq of neuron-restrictive silencer factor (NRSF) lead to the discovery of NRSF specific binding motifs and enrichment at a specific set of genes opened new lines of research into how this protein regulates targeted gene expression[121]. Another pioneering ChIP-seq study targeting Signal transducer and activator of transcription 1 (STAT1) in interferon-γ (IFN-γ) stimulated and unstimulated human epithelial carcinoma (HeLa) S3 cells discovered differential gene targeting of STAT1 dependent on IFN-γ stimulation[123].

These early ChIP-seq experiments represent the hallmarks of ChIP-seq data acquisition that many researchers desire when performing such experiments. Mainly, these genome-wide interactome analyses generate impressive amounts of high resolution data that can not only validate hypotheses but also promote discovery of unexpected genome wide distributions that enable new insight of epigenetic phenomenon.

The development of chromosomal conformation capture (3C) allowed researchers to study long distance DNA-DNA interactions of known loci[125]. This strategy was expanded for use in a high throughput chromosomal capture (Hi-C) using NGS, allowing genome wide observations of DNA-DNA interactions[126].

Combining Hi-C and ChIP-seq in a technique called Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) enabled researchers to observe proteins that bind to multiple loci, separated by long distances, simultaneously[127]. Methods to study long distance DNA-DNA interactions will inevitably assist our understanding of fundamental molecular processes. A recent example is a time course study investigating cohesin binding during the mitotic phase to certain loci in budding yeast[128]. This study found that cohesin moves along DNA, thereby providing evidence supporting a loop extrusion hypothesis, in which DNA loop origins extrude from a single locus in order to avoid DNA knotting.

1.2.4 Processing big data

In this era of scientific data collection, it is becoming standard to perform whole genome high throughput experiments, especially in the field of epigenetics. Due to the nature of high throughput experiments, it has become necessary to develop efficient and robust processing and analysis software. Constant development of hardware in the technology sector enhances the ease of data processing, which allows greater data collection which in turn stimulates hardware development. In addition, there is a need for software that enables biologically and statistically relevant analyses. Among the most important software in genetics have been aligners. The basic concept of alignment algorithms is to identify the location of a query DNA sequence to a reference genome. The well-known Basic Local Alignment Search Tool (BLAST) algorithm uses a heuristic approach where fragments of a query sequence are searched for in a database of short oligomers in a process known as seeding[129]. Algorithms for NGS data analysis require increased speed and efficiency, which BLAST cannot provide. Some popular software such as Burrows-Wheeler Aligner (BWA)[130] and Bowtie[131] use an algorithm originally designed for file compression, called the Burrows-Wheeler Transform. This approach resulted in the first practical tools for use in NGS alignment efforts. Many other types of algorithms exist for statistical calculations, data pre-processing, and data analysis. These tools have emerged as critical components for aiding experimental studies and greatly aid clarity in data reporting.

1.3 Carbohydrates in biology

Living systems require tightly regulated molecular processes to facilitate development and sustain homeostasis. Proteins are essential in this process as they provide many structural components and factors that execute specific tasks. In many cases, however, a properly folded protein in a suitable complex with its binding

partners is not sufficient to enable the complex to carry out its normal functions. PTMs are often necessary additions that are required for many proteins to become fully functional and/or stable. Various amino acids in a protein are subject to many types of PTMs, including phosphorylation, acetylation, methylation, glycosylation, and others[132]. These modifications are critical for the proper functioning of biological systems[133].

Of relevance to this work is the glycosylation of serine and threonine residues of nuclear and cytoplasmic proteins with O-linked ß-N-acetylglucosamine (O-GlcNAc). The enzyme O-GlcNAc transferase (OGT) uses uridine 5'-diphosphate-N-acetylglucosamine (UDP-GlcNAc) as a substrate, which is generated from glucose in a six step process by the enzymes of the hexosamine biosynthetic pathway (HBP). O-GlcNAc is an interesting and important PTM in eukaryotes because it is abundant and nutrient responsive[134] and has emerged as being important in many intracellular functions[135]. However, there are many other glycans that have important functions in biology. This subsection provides a brief overview of protein glycosylation modifications and their general importance, which is then followed by a section on O-GlcNAc and O-GlcNAc processing enzymes.

Key features of glycosylation include: 1) bond and linkage type, 2) glycan type, 3) glycan branching, and 4) glycan length. The combination of these four features, as well as the identity of the protein being modified, dictate the nature of the downstream functions of glycosylation. In general, glycosylation alters the physicochemical attributes of the modified protein, thereby altering its function directly or by altering its interactome. Most glycosylation occurs within the endoplasmic reticulum and Golgi apparatus where an array of glycosyltransferases and glycoside hydrolases exist. Among other functions, these glycosyltransferases and glycoside hydrolases in combination with lectins, which bind glycosylated proteins, promote the proper folding of proteins through a complex cycle of glycosylation and attempted folding, occasional mis-folding, which may be followed by further glycosylation, refolding, and deglycosylation[136].

Glycans also play a prominent role in immunity. Mammalian cells express unique antigens at the cell surface that are important for recognizing self and distinguishing from pathogens. Interestingly, some pathogens have evolved means to present host-like glycans so that they can evade our immune system[137,138]. Antibodies such as immunoglobulin G (IgG) subtypes have their specificity affected by modification with sialic acid, and this has been exploited in order to increase specificity of recombinant antibodies[139]. In another example, a reported 80% of

cancer cells have abnormal expression of the glycopeptide Mucin 1 (MUC1), to which therapies have been and are currently being developed[140].

Consistent with the varied and important roles of glycosylation, over 40 congenital glycosylation disorders have been reported that stem from aberrant N- and/or O- linked glycosylation[141]. Interestingly, many of these disorders are neurological in nature. Although these diseases are rare, they underscore the essential nature of glycosylation in human health and disease. Overall, it is clear that glycosylation is a complex phenomenon. It is imperative that researchers design tools to effectively study distinct glycans in order to facilitate understanding of their molecular roles. There are many other roles glycans play, such as in signalling, trafficking, and gene expression[135,142]. Fundamentally, glycans are emerging to be of greater interest because they are being uncovered as new targets for therapeutic intervention and play diverse and fundamental roles in many processes involved in homeostasis[143].

1.3.1 O-linked β-N-acetylglucosamine transferase

When scientists choose to study the molecular mechanisms of proteins and their function, it is logical to choose fundable topics. Generally, in biological sciences, funding can be more readily obtained if the protein has clear targetable roles in human disease or health. From this perspective, OGT is an important enzyme to study since it has been implicated in many aspects of human health, development and disease.

OGT catalyzes the transfer of a single GlcNAc moiety from UDP-GlcNAc onto select hydroxyl groups of serine and threonine residues of nucleocytoplasmic proteins (Figure 1.5, a). This PTM can be removed from proteins through the action of the glycoside hydrolase O-GlcNAcase (OGA). Curiously, S-linked GlcNAcylation has been discovered in mammals, which apparently cannot be hydrolyzed by OGA[144]. Standard O-linked GlcNAc is thought to repeatedly cycle through addition by OGT and removal by OGA. This cycling permits a dynamic and responsive nutrient sensing mark. O-GlcNAcylation was discovered in 1984[145], and since then, OGT and O-GlcNAc has been found in all metazoans and plants studied[135], and is essential for development and survival in mammals.

Initial characterization using standard biochemical techniques identified OGT as a heterotrimer with two 110 and one 78 kDa subunits[146]. OGT was also found to have a high (545 nM) affinity for UDP-GlcNAc[146]. Sequencing analysis has revealed that OGT is highly conserved from humans to worm and that a series of tetratricopeptide repeats (TPR) are an important feature of the enzyme[147]. Crystal

structure studies of OGT in complex with a peptide substrate revealed more detailed features of the structure of OGT, as well as an ordered sequential bi-bi kinetic mechanism for this bisubstrate enzyme[148]. OGT is essential for mouse embryogenesis[149,150], and via an interesting experiment whereby mouse fibroblast cells containing a loxP flanked *Ogt* allele was transduced with natural vesicular stomatitis virus (VSV-G) or VSV containing Cre protein[150], this enzyme was shown to be essential for cell viability. RT-PCR analysis of OGT and actin four days postinfection showed far less OGT mRNA in cells transduced with VSV-Cre and less O-GlcNAcylated proteins seen by immunoblot when using a pan-specific antibody for O-GlcNAc. Twelve days post-transduction, fibroblast cells containing a loxP flanked *Ogt* allele treated with VSV-Cre were dramatically less viable compared to its VSV treated counterpart.

Of the virtually thousands of O-GlcNAc modified proteins, more than 700 have been site mapped for their O-GlcNAc modification, some of which contain multiple O-GlcNAc sites (Figure 1.5, b)[151]. Interesting substrates of OGT from the perspective of epigenetics includes Host Cell Factor C1 (HCF1)[152], Ph[153], and all three of the TET proteins[151,154–156]. Remarkably, mammalian OGT was found to cleave HCF1[152]. Recent studies determined that the proteolysis occurs in the same active site but through a different mechanism[157].

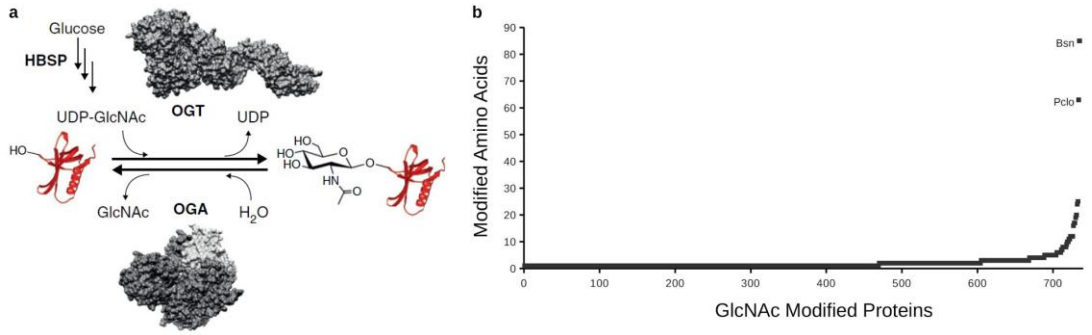


Figure 1.5 **O-GlcNAc installation and number of O-GlcNAc residues in O-GlcNAc modified proteins.**
(a) Dynamic cycling of UDP-GlcNAc onto serine and threonine residues of target proteins by OGT and OGA. UDP-GlcNAc is enzymatically produced from glucose in a six step process via the hexosamine biosynthetic pathway (HBSP). (b) Number of O-GlcNAc residues in O-GlcNAcylated proteins from PhosphoSitePlus[151]. Over 700 proteins have been O-GlcNAc site mapped. Most proteins have one known modified residue. Some have many, though in these cases it is not known how many sites are modified on any one protein molecule. Bassoon (Bsn) and Piccolo (Pclo) are among proteins with the most potential O-GlcNAc modified residues. Figure (a) was reprinted from Current Opinion in Chemical Biology, vol. 16, Vocadlo D.J., O-GlcNAc processing enzymes: catalytic mechanisms, substrate specificity, and enzyme regulation, p. 488-497, copyright 2012, with permission from Elsevier [doi: 10.1016/j.cbpa.2012.10.021].)[158].

Varying levels of O-GlcNAc have been found in different organelles[134,159]. In line with this, different OGT isoforms have also been discovered[159]. The rate limiting enzyme in the HBP, which leads to the synthesis of UDP-GlcNAc, is glucosamine-6-phosphate transaminidase (GFAT). It is interesting that OGT, OGA, and GFAT isoforms 1 and 2 have been shown to be differentially expressed in diabetic and non-diabetic individuals[160]. These observations suggest the importance of tight regulation of genes in the HBP as well as OGT.

As noted above, OGT and O-GlcNAc are essential for basic cellular processes and maintenance of proper health. In *Arabidopsis*, an OGT homologue was found to affect gibberellin signal transduction[161]. O-GlcNAc has also been found to stabilize nascent polypeptide chains[162]. On the other hand, studies have shown that glycosylation of Ph[163] and Tau[164] prevent aggregation and promote proper functioning of these substrate proteins. Several transcription factors, including insulin promoter factor 1 (PDX-1) in pancreatic ß-cells, were found to be modified by O-GlcNAc[165] and subsequently, in a gel shift assay, it was shown that PDX-1 bound more insulin promoter DNA when these cells were fed with high glucose compared to low glucose. Many heat shock proteins have been found to be OGT substrates. Interestingly, the transcription factor Heat Shock Factor protein 1 (HSF1) mediates the transcription of heat shock proteins. Glycogen synthase kinase 3 (GSK-3β) inactivates HSF1 by phosphorylation, which is increased in OGT knockout cells[166]. Moreover, reduced expression of several heat shock proteins was observed in OGT knockout cells, suggesting a mechanism by which O-GlcNAc participates in a stress response mechanism[166].

Recently, interest in O-GlcNAc and its role in cancers has grown. Several genes involved in the HBP were found to be overexpressed in human prostate cancers[167]. Many papers have also been published showing increased OGT and O-GlcNAc levels in different cancer lines[168]. It is well known that many types of tumour cells have increased glucose uptake. Phosphofructokinase-1 (PFK-1) was shown to be O-GlcNAc modified at serine 529 and the S529A mutation was shown to reduce cell proliferation and lung tumour growth in live mice[169]. Furthermore, O-GlcNAc was shown to stabilize hypoxia-inducible factor 1 (HIF1), thereby increasing the expression of a glucose transporter which was shown to increase glycolysis and tumour growth[170]. The study also found that human breast cancers with high levels of HIF1 had high levels of OGT[170]. The phenomenon of high glycolytic rates in many cancers, termed the Warburg effect, in combination with these and other implications of OGT in metabolism, has stimulated research on OGT and O-GlcNAc as a biomarker for some tumour types[168]. Interestingly, tumour suppressor protein p53

(p53), one of the most frequently mutated proteins associated with numerous cancers, is a substrate for OGT.

OGT, OGA, and O-GlcNAc have also been linked to cardioprotection. In a study that treated rat cells with glucosamine, a precursor to UDP-GlcNAc, or PUGNAc, an inhibitor of OGA, researchers observed increased O-GlcNAc levels that correlated with improved cell viability after simulated ischemia and reperfusion[171]. Multiple OGA inhibitors also increased mouse viability after induced ischemia and reperfusion, *in vivo*[172,173]. Studies suggest that increases in O-GlcNAcylation mediate their cardioprotective effects through tumour protein p38 (p38) activation and B-cell lymphoma 2 (Bcl-2) translocation[171,174].

As mentioned, Tau is a substrate for OGT and it is stabilized against aggregation when glycosylated[164]. Tau is important for microtubule stabilization and proper neuronal function. Hyperphosphorylation of Tau causes its aggregation which contributes to neurodegeneration in Alzheimer's disease. OGA inhibitors, which increase global O-GlcNAc levels, have proven efficient at reducing phosphorylation of Tau[175] and hinder progression of AD-like tauopathy in transgenic mouse models[164].

1.3.2 Tools to study O-GlcNAc

There have been impressive advances in the development of tools used to study OGT and O-GlcNAcylation over the past few years. Inhibitors, chemical probes, genetically modified organisms, and computer programs have all increased our understanding of this important protein and its catalytic activity. This section is committed to detailing some of the notable research tools in the field of O-GlcNAc.

1.3.2.1 *In silico* O-GlcNAc tools

Many *in silico* analyses have been carried out on the substrate specificity of OGT. From data pertaining to specific O-GlcNAc modification of proteins that has been generated by mass spectrometry, researchers have been able to conduct amino acid motif analysis of substrates. The representation of a proline nearby O-GlcNAc modified sites in OGT substrates was observed more than 25 years ago[176]. However, over 25 years of collecting and analysing O-GlcNAc substrates has yielded little progress in this direction. Below is a brief discussion of currently available tools for *in silico* O-GlcNAc substrate identification.

Early O-GlcNAc prediction studies relying on sequence inferences discovered differences in both the number of residues predicted to be O-GlcNAcylated and the position of the modification depending on the functional category of the protein[177].

More recent prediction tools have supported this claim by investigating sequence preferences in transmembrane containing versus non-transmembrane containing O-GlcNAc substrates[178]. A support vector machine (SVM) algorithm based program called EnsembleGly[179] is trained from a database containing known N-, O-, and C-linked glycosylated proteins to enable predictions of unknown, yet putative glycoproteins. However, EnsembleGly is not specific for O-GlcNAcylated protein prediction. Glycosylation prediction program (GPP)[180] uses a random forests algorithm to determine N- or O- linked sites of glycosylation within a peptide query. GlycoMine[181] is essentially an updated version of GPP (random forests) with an option to specify the type of glycan linkage being queried. OGlcNAcScan is primarily based on amino acid sequence stretches flanking the site of O-GlcNAcylation[182]. O-GlcNAcPRED analyses amino acid composition and physicochemical subgrouping of amino acids flanking O-GlcNAc sites[183]. Another tool called OGTSite takes amino acid composition consideration one step further by binning OGT substrate by motifs, enabling increased O-GlcNAc prediction accuracy[184].

These informatics prediction tools essentially function in a similar manner: researchers use a curated O-GlcNAc database and examine sequence prediction effectiveness based on different variables. A position-specific scoring matrix (PSSM), where amino acid identity relative to the site of O-GlcNAcylation was investigated[185], and was found to be the most sensitive, specific, and accurate way to predict O-GlcNAc sites compared to other sequence-based analysis strategies. Interestingly, one of the least predictive methods employed is based on surface area accessibility. This coincides with finding that O-GlcNAcylation can occur co-translationally[162], potentially giving rise to O-GlcNAc residues within folded domains. The PSSM analysis examined non-O-GlcNAcylated residues to gain insight into motifs that are unlikely to harbour O-GlcNAc modifications. Enriched amino acids were sub grouped based on a statistical model of positional clustering of residues surrounding the O-GlcNAc modification site. Furthermore, the prediction model was shown to vary in power depending on the motif subgroup to which the query protein belongs. This observation supports claims that OGT specifically interacts with substrates through protein partner complexes[186–188], and suggests that prediction software that considers secondary and tertiary protein information might be more powerful for O-GlcNAc substrate predictions, as shown in other PTM predicting software[189].

Overall, bioinformatics prediction and analysis of O-GlcNAc substrates have provided some useful knowledge, and will probably improve as methods are refined. Ultimately however, meaningful research requires the validation of O-GlcNAc substrates and mapping of modified residues. Recent developments in substrate

identification are receiving considerable attention. This will be discussed next, along with other O-GlcNAc research tools.

1.3.2.2 O-GlcNAc biochemical tools

Antibodies that recognize O-GlcNAc[190,191] have been raised, but tend to be sequence specific[192], and this limits their utility[193]. Developments in antibody technologies has allowed for more selective O-GlcNAc antibodies[193], but these are still limited in scope. Lectins are a class of proteins that specifically bind to glycans depending on the sugar modification. Lectins are important in biological contexts for glycan binding. A lectin found in wheat germ, called wheat germ agglutinin (WGA) was characterized by Burger and Goldberg[194], and has shown to bind ß-linked O-GlcNAc[195]. Other lectins also bind O-GlcNAc[196], however WGA seems to be the lectin of choice in research settings. Lectins, however, including WGA, have relatively weak interactions to their ligands[197], and lack the ability to specifically bind to single glycan modifications[195,198]. Chemical probes developed to capture O-GlcNAc modified proteins have therefore become one of the favoured techniques of scientists in the field.

Two major chemical labeling approaches are used to study O-GlcNAc modified proteins: 1) metabolic feeding with sugar analogues that are incorporated onto O-GlcNAc modification sites, and 2) chemoenzymatic labeling of O-GlcNAc modified substrates. Both of these techniques use bioorthogonal chemistry and will be reviewed below.

Bioorthogonal chemistry enables chemical reactions to take place in complex environments such as cell lysates and living systems, which in turn enables researchers to monitor biomolecules and cellular processes. Generally, the goal is to design chemical probes that specifically and unobtrusively react with or serve as a surrogate for one compound in living systems. Once the chemical probe is incorporated into the living system, researchers can monitor the incorporated probe directly, or through reactions that specifically react with the incorporated probe. Observations are then treated as though they were on the target of the chemical probe.

One of the first examples of a bioorthogonal reaction was reported in 1998 when researchers discovered that a dye termed FLASH (4',5'-bis(1,3,2-dithioarsolan-2-yl)fluorescein) could bind to a specific protein peptide and become fluorescent, akin to tagging a protein with green fluorescent protein (GFP)[199]. This strategy also required genetic engineering of the target protein but afforded a more realistic way to image proteins due to the small size of the engineered peptide motif relative to GFP.

Another strategy to tag proteins is to feed cells with an azide containing methionine analogue called azidohomoalanine (AHA), which is incorporated onto newly translated proteins[200]. AHA is assimilated by the cellular machinery and incorporated into proteins in the same way as methionine. Although AHA is not specific for particular proteins since it is incorporated into all proteins, it has great value in studying newly translated proteins.

Not all chemical probes that exploit bioorthogonal chemistry can be metabolically incorporated into regular cultured cells. Incorporation of some probes requires genetically engineered enzymes that enable cellular assimilation of the chemical probes[199,201]. Many different chemical probes have been synthesized for a variety of purposes, including, for example, to monitor protein glycosylation [202], protein methylation[201], and protein lipidation[203]. There are also different bioorthoganol chemistries that can be applied to a biological sample containing a chemical probe, each with their own reported advantages and limitations.

When a chemical probe is incorporated into living systems, researchers must then perform highly targeted and efficient chemical reactions to capture or monitor the probe modified biomolecule. For example, in the case of FLASH, reaction with a tetracysteine peptide motif leads to fluorescence. For chemical probes such as AHA, reactions must be performed that specifically tag the biomolecule with a fluorophore or capture reagent such as biotin. Azide groups are absent from biological systems and are well tolerated, which makes them excellent candidates for bioorthogonal chemistry[204,205]. A reaction such as the Staudinger-Bertozzi ligation allows azide-specific chemical ligation with a range of different phosphine probes[204,205]. The Staudinger-Bertozzi ligation probe positions a methyl ester ortho to the triarylphosphine, which upon reaction with an amide results in a stable amide linkage[206]. In fact, glycan labeling with azido sugars followed by ligation to phosphine probes via the Staudinger-Bertozzi ligation is a well-documented and highly reliable methodology[207].

The Staudinger-Bertozzi ligation works well and is highly specific but is slow and phosphine probes oxidize rapidly. To combat this, the copper catalysed [3+2] cycloaddition of alkynes, or click chemistry, was introduced to the field of bioorthogonal chemistry[208]. This reaction affords greater speeds and more stable probes but suffers from the cytotoxicity of copper. The strain promoted [3+2] cycloaddition permitted copper free click chemistry through the release of energy associated with ring strain of the cyclooctyne[209]. Ultimately, the bioorthogonal chemicals used should be amenable to the specific experiments being performed.

Of interest to this thesis are probes that are used to modify endogenous O-GlcNAc proteins. In 2003, Vocadlo *et al*.[210] synthesized peracetylated N-azidoacetylglucosamine (Ac_4GlcNAz), and showed that the GlcNAc salvage pathway could effectively process this probe to generate UDP-GlcNAz which was shown to label Nup62, a well-known O-GlcNAcylated nuclear pore protein. Following on this concept, there have been OGT inhibitors created that are activated by assimilation through the HBP[211]. Interestingly, azide modified GlcNAc analogues were found to have higher HBP processing efficiency compared to their alkyne counterparts[212]. Per-O-acetylated N-azidoacetylgalactosamine (Ac_4GalNAz) was found to be effectively converted to UDP-GalNAz through the galactosamine salvage pathway and later used as a donor for N-acetyl-α-galactosaminyltransferases to label O-linked mucins[213]. It was later found that metabolic feeding with Ac_4GalNAz results in intracellular UDP-GlcNAz and therefore could be used by OGT to label O-GlcNAcylated proteins[214]. Furthermore, labeling of O-GlcNAcylated chromatin bound proteins, was found to be more efficient using Ac_4GalNAz compared to Ac_4GlcNAz[215]. Interestingly, Ac_4GalNAz metabolically labeled zebrafish treated *in vivo* with a difluorinated cyclooctyne provided the first live imaging experiment using bioorthogonal chemistry[216]. A similar study using different metabolic incorporated sugars was done in *C. elegans* soon after, and highlighted significant changes in the locations of GlcNAc, GalNAc, and ManNAc during the developmental stages of worm[216].

More recently, other O-GlcNAc targeting probes have been developed with advantageous qualities. For example, recent developments have created an OGA-resistant probe that enters the GlcNAc salvage pathway more selectively[217]. Alternatively, using a genetically engineered UDP-GlcNAc pyrophosphorylase, researchers were able to treat cells with a diazirine-modified partially protected GlcNAc-1-phosphate analogue, which resulted in endogenous O-GlcNAc proteins labeled with a diazirine analogue of GlcNAc (GlcNDAz). This permitted ultraviolet (UV) inducible cross linking of proteins binding to O-GlcNAc substrate partners, which could then be identified by mass spectrometry[218].

An alternative approach to the use of metabolic feeding for labeling O-GlcNAc modified proteins is a chemoenzymatic strategy that can modify native O-GlcNAc residues. O-GlcNAc exists as a terminal residue. Galactosyltransferase (GalT) is an enzyme that can modify terminal O-GlcNAc with galactose[145]. Earlier studies took advantage of this by synthesizing [^{3}H] radiolabeled galactose, which could then be enzymatically transferred to O-GlcNAc residues using recombinant GalT to enable quantification and imaging of these proteins by autoradiography[219]. A more

sophisticated technology has been demonstrated by Khidekel, *et al.*[220], where a mutated GalT (Y289L) was used for O-GlcNAc labeling with a UDP-galactose substrate analogue containing a ketone group. This mutated GalT accepted the UDP-galactose analogue and allowed downstream bioorthogonal chemistry to capture O-GlcNAc modified proteins. More specific azide chemical probes developed for GalT chemoenzymatic labeling[221] have proven efficient at capturing novel O-GlcNAcylated substrates[222].

Other strategies have been pursued to identify O-GlcNAc modified proteins. For example, development of a biotinylated peptide microarray with 13 amino acid peptides allowed the detection of peptides that were preferentially O-GlcNAc modified[223]. Subsequent analysis of crystal structures of OGT with these synthetic peptide substrates provided insight on the interaction of OGT with its substrates. Importantly, similar sequence motifs were observed here as found in native O-GlcNAc substrates. Others have used a peptide array in a similar fashion to gauge OGT substrate preference[224]. The major issue with these methods is the short length of the peptides, which may not realistically represent interactions that occur *in vivo*, particularly given the extensive TPR domain that is known to bind to OGT substrates. Clearly, there are many available tools to label and capture O-GlcNAcylated proteins and identify OGT substrates. The various methods have been developed in attempts to create reliable and unbiased approaches. These probes have different qualities and researchers should choose an appropriate strategy depending on the application being pursued.

Notably, OGT plays important roles beyond its catalytic function. Recent studies suggests it has important roles in complexes[225]. In fact, earlier studies have shown, using a yeast two-hybrid system, that OGT is capable of maintaining stable interactions with other proteins[226]. Furthermore, it has been suggested that TPRs are important interaction domains. Recent development in microarray technology has allowed OGT interactome assays to be conducted in a high throughput format, and have identified several potential binding partners[227].

1.3.3 OGT in epigenetics

In 1984, Ingham identified the *super sex combs* (*sxc*) locus and characterized it as a member of the Polycomb group (PcG) of regulatory genes in *Drosophila*. Using ethyl methanesulfonate-induced *sxc* alleles, Ingham showed that this gene is necessary for the correct repression of homeotic genes. $sxc^{-/-}$ pharate adults showed inappropriate expression of the *ANT-C* and *BX-C* genes, as evidenced by homeotic phenotypes of antenna to leg transformation, transformation of anterior

compartments of the mesothoracic and metathoracic legs to prothoracic identity, as well as wing to haltere transformation, etc.[228].

Interestingly, unlike mutants of many other PcG members, which fail to survive embryogenesis when homozygous, *sxc*$^{-/-}$ mutants die as pharate adults; however, Ingham used the technique of pole-cell transplantation to show that such mutants fail to survive embryogenesis when maternally contributed *sxc* is eliminated and the dead embryos showed extensive homeosis. Ingham also showed that the *sxc*$^{+}$ product is required during the larval period for normal development. Inghams data support the idea that the *sxc* function is required throughout development, but that the maternal contribution of *sxc*$^{+}$ allows development to the pharate adult stage. As mentioned, because *sxc* is required for repression of Hox genes, it has been categorized as a PcG gene. In 2009 it was discovered that *Drosophila sxc* encodes *Ogt*[153,229]; O-GlcNAcylation levels were shown to be drastically reduced in *sxc* knockout flies and through use of an antibody for human OGT it was shown that fly OGT was absent in *sxc* alleles, indicating high structural conservation of the proteins[153,229]. Moreover, the human OGT transgene was able to rescue *sxc* lethality[229], indicating striking functional conservation for OGT between the two species. Interestingly, it was shown that Ph colocalizes with O-GlcNAc on polytene chromosomes[229] and that this protein was O-GlcNAc-modified[153], indicating at least one initial putative function of OGT as a PcG protein. In fact, the localization of Ph at PREs was shown to be reduced in imaginal discs of *sxc*$^{-/-}$ larvae compared to wild type[153]. Subsequent research found this reduced binding was likely due to the non-productive aggregation of Ph induced by loss of glycosylation[163], similar to Tau protein, but without any known implications of phosphorylation.

The analysis of Ph and sxc in flies has provided a striking illustration of how the O-GlcNAcylation of an important epigenetic protein is essential for its proper functioning. This work also underscores the advantages of fly as a model organism. However, the exclusive emphasis of OGT effects on Ph could obscure in the possibility of other important biological functions of sxc/OGT in flies and OGT in humans[230]. Furthermore, not all Ph targets are affected by loss of sxc. Clearly there are many other O-GlcNAc substrates in fly[215], including many proteins involved in gene regulation, and recent advances in metabolic feeding based approach for O-GlcNAc ChIP-seq may expand this perspective considerably[215]. As mentioned, many transcription factors are known to be O-GlcNAc modified; indeed, it has been reported that approximately 25% of O-GlcNAc substrates are transcription factors[231]. This implies that O-GlcNAc has a fairly broad role in regulation of gene expression. Relatively few of these putative substrates have been studied in detail, including the

previously mentioned PDX-1[165], p53[232], and Ph[153]. Sp1 is modified and stabilized by O-GlcNAc[233]. O-GlcNAc modification of FoxO1 increases in response to glucose, which increases FoxO1 transcriptional activity[234,235].

Other non-transcription factor nuclear proteins are modified by OGT including RNA polymerase II[236]. Interestingly, the polycomb repressive deubiquitinase complex (PR-DUB) composed of BAP1 and ASXL1 and 2 in mammals has been shown to interact with OGT[237]. Earlier genetic studies showed a genetic interaction between *Asx* and *sxc* heterozygotes in *Drosophila*[238].

Widespread O-GlcNAcylation of histones has been reported[239,240], and further it has been reported that some of these modifications are mediated by TET[187]. However, another study has found no support for O-GlcNAcylation of histones[241]. In mitosis synchronized HeLa cells it was shown that overexpressed OGT modulated H3 Ser10 phosphorylation, H3K9 acetylation, and H3K27 trimethylation[239]. In addition, OGT also influences histone modifiers. For example, histone-arginine methyltransferase (CARM1) associates with OGT and shows reduced activity in OGT overexpressed cells[242]. As mentioned HCF1 is cleaved by OGT and O-GlcNAcylation has been shown to stabilize the HCF1 methyltransferase complex SET1/COMPASS[243]. This process is likely regulated upstream by OGT interactions with TET 2 and 3[244]. It has also been reported that OGT modifies EZH2 and stabilizes methylation of H3K27[245].

Overall, it is apparent that OGT and O-GlcNAcylation play key roles in epigenetic regulation of a large number of target genes, in mammals to flies. Therefore, it is reasonable to suggest that there are many mechanisms of OGT gene regulation yet to be fully elucidated and that there is a compelling need for the continued development and implementation of novel and effective methodologies to study the panorama of OGT roles in gene regulation.

1.4 Thesis overview

The preceding discussion highlights key genetic and epigenetic concepts, connecting OGT and O-GlcNAcylation to many phenomena including regulation of gene expression. The next chapters detail research undertaken to increase understanding of how OGT regulates gene expression. In this work, novel methodologies were developed to study O-GlcNAcylated proteins bound to the *Drosophila* genome, but these approaches are amenable to studying such phenomena in any organism. In collaboration with others, I have developed a novel antibody free metabolic feeding approach to map O-GlcNAc proteins genome wide, developed robust software to analyse time course ChIP-seq data, and generated

time course O-GlcNAc metabolic feeding data in live *Drosophila*, providing the first ever time course ChIP-seq data set in a live organism and on a PTM.

Chapter 2: **Genome-wide chemical mapping of O-GlcNAcylated proteins in *Drosophila melanogaster***

Note: The work presented in this chapter has been published in *Nature Chemical Biology,* vol. 13, p. 161-167, 2017.

2.1 Background

As mentioned above, N-acetylglucosamine -O-linked to nucleocytoplasmic proteins (O-GlcNAc) is implicated in the regulation of gene expression in organisms from humans to *Drosophila melanogaster*. Within *Drosophila,* O-GlcNAc transferase (OGT) is one of the Polycomb Group proteins (PcGs) that act through Polycomb Group Response Elements (PREs) to silence homeotic (*HOX*) and other PcG target genes. PcG proteins act to impose silencing of gene expression through their interactions with genomic sequences known as Polycomb Group Response Elements (PREs) that are found within PcG target genes including the homeotic (*HOX*) genes[246]. Many PcG proteins are recruited to PREs within multiprotein complexes including the Pho Repressive Complex (PhoRC), which in turn recruits PRC1 and PRC2[247,248]. Specific post-translational modifications of constituent PcG proteins impact their stability and interacting partners and thereby contribute to the proper functioning of these PcG complexes[163,249]. Accordingly, the localization of PcG proteins to the genome has proven important for understanding the mechanisms leading to appropriate silencing of gene expression. The mapping of PcG and other chromatin associated proteins that bear post-translational modifications is, however, challenging because high quality antibodies for this purpose are not generally available[250]. Indeed, given the recognized limitations of using antibodies ChIP-seq[250,251], the development of new methods to purify chromatin bound proteins with biologically important modifications to enable ChIP-seq is of wide current interest[102,252].

One protein modification that regulates PcG function in *Drosophila* is the modification of nuclear and cytoplasmic proteins with N-acetylglucosamine (GlcNAc) residues O-linked to serine and threonine residues of target proteins (O-GlcNAc). The O-GlcNAc modification is conserved among all metazoans[135] and the *Drosophila* glycosyltransferase, OGT, which catalyzes transfer of GlcNAc from uridine diphosphate GlcNAc (UDP-GlcNAc) to acceptor proteins[135] is encoded by the PcG gene *sxc*[153,229]. Early studies show that many chromatin-bound proteins within

Drosophila[253] are O-GlcNAc modified. To date, however, the only PcG protein shown to be O-GlcNAcylated in *Drosophila* is Polyhomeotic (Ph). Notably, O-GlcNAcylation of Ph is required to control *HOX* gene silencing, as well as the silencing of certain other PcG target genes[153,163]. Given the varied physiological roles played by *O*-GlcNAc on diverse proteins and its presence on many chromatin associated proteins[253], coupled with the difficulty in generating pan-selective antibodies directed toward O-GlcNAc[193], we believed that the development of antibody and lectin free methods to map O-GlcNAc to the genome would be useful for interrogating the functional roles of O-GlcNAcylated proteins within the genome.

Here we describe a method for genome-wide mapping of *O*-GlcNAc that circumvents the use of antibodies and lectins. This method exploits emerging metabolic feeding strategies to assimilate sugar analogues bearing bio-orthogonally reactive chemical moieties onto proteins[254]. Subsequent chemoselective ligation[208] of metabolically labeled O-GlcNAcylated proteins enabled us to map O-GlcNAc across the genome in cultured cells and *D. melanogaster*. We compare this method to the use of the lectin wheat germ agglutinin (WGA)[153,253], and to a chemoenzymatic approach[220] we implement here for a ChIP-seq-like strategy. Using this metabolic labeling approach we identify non-PRE-containing genomic loci bearing O-GlcNAc, and show that OGT regulates expression of genes at various loci including several within heterochromatin.

2.2 Results

To develop a chemical strategy to map O-GlcNAcylated proteins to the genome we used a metabolic engineering approach that exploits the ability of OGT to transfer a close analogue of GlcNAc known as N-azidoacetylglucosamine (GlcNAz) onto proteins[210]. Using downstream bio-orthogonal chemistry, the pendent azide group of O-GlcNAz-modified proteins can be modified with biotin (Figure 2.1, a)[210,255]. We speculated we could then use streptavidin to enrich for genomic DNA fragments to enable a genome-wide ChIP-seq-like strategy and thereby identify genomic loci bearing O-GlcNAc in a manner that would not depend on lectin or antibody specificity.

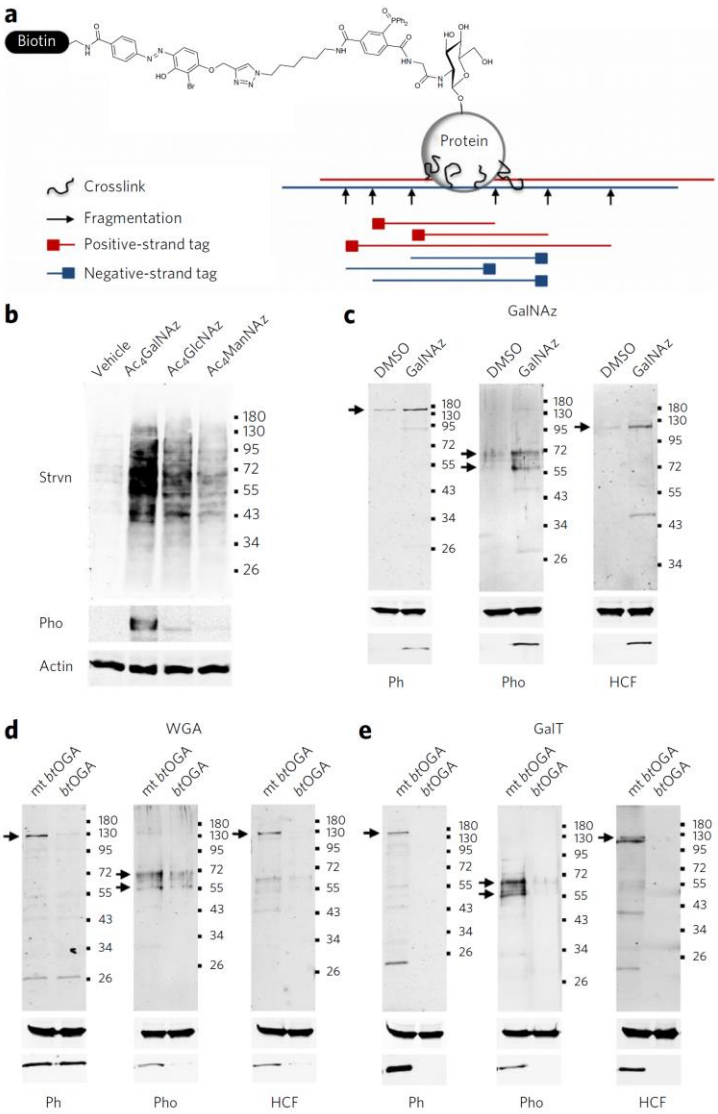


Figure 2.1. **A combined metabolic feeding-chemoselective ligation strategy enables labeling of chromatin associated proteins from *Drosophila* S2 cells.** (a) Antibody-free method to enable mapping O-GlcNAcylated proteins to the genome. DNA is cross-linked to O-GlcNAcylated proteins that are then ligated to biotin using the Staudinger Ligation. Enriched and fragmented DNA is then sequenced. (b) Nuclear proteins from S2 cells in the presence and absence of metabolic labeling with Ac_4GalNAz (GalNAz), Ac_4GlcNAz, and Ac_4ManNAz (16 h) were conjugated to a biotinylated phosphine cleavable capture reagent (Biotin-*azo*-phosphine). Biotinylated proteins were detected using Streptavidin (Strvn) blot analysis (upper panel) and Pho antibody (middle panel). The lower panel shows actin loading control before streptavidin purification. (c) Nuclear proteins from S2 cells treated for 16 hours with Ac_4GalNAz or vehicle were conjugated to Biotin-*azo*-phosphine. Biotinylated proteins were detected after Streptavidin (Strvn) immunoprecipitation by immunoblot. Ph, Pho, and dHCF Biotin-azo-phosphine conjugated proteins are recovered after $Na_2S_2O_4$-mediated release. Arrows in upper panel indicate the size of protein being probed. Middle panel shows actin as an input control. Lower panel shows immunoprecipitated actin. (d) Nuclear proteins as in (c) from S2 cells precipitated using WGA followed by blot analysis. Lysates were treated with WT and D243A (mt) *Bt*OGA. (e) Nuclear proteins from S2 cells precipitated with streptavidin after UDP-Ac_4GalNAz and GalT treatment and chemical

conjugation to biotin. Lysates were treated with WT and D243A (mt) *Bt*OGA All blots were reproducible using two biological replicates.

2.2.1 Metabolic labeling of *Drosophila* S2 cells

We first tested the efficiency of several per-O-acetylated azido-sugars in labeling *Drosophila* proteins. *Drosophila* S2 cells were incubated with per-O-acetyl N-azidoacetylglucosamine (Ac_4GlcNAz), N-azidoacetylgalactosamine (Ac_4GalNAz), N-azidoacetylmannosamine (Ac_4ManNAz), or vehicle alone for 16-24 h, after which we isolated nuclear proteins (Figure 2.1, b and Figure 2.2). In our hands neither the copper-free nor the Cu(I)-catalyzed Azide-Alkyne Cycloaddition (CuAAC) showed the high selectivity in *Drosophila* lysates that is essential for use in downstream DNA sequencing experiments. The Staudinger-Bertozzi ligation, however, showed excellent selectivity[256] (Figure 2.3) that we believed would enable accurate ChIP-seq studies. Given that sodium dithionite ($Na_2S_2O_4$)-cleavable azobenzene linkers are reliable[257], we used a phosphine probe containing this linker (Biotin-*azo*-phosphine)[162]. Capture of the biotinylated proteins followed by cleavage of the linker and analysis by streptavidin blotting revealed to us that labeling of proteins within S2 cells by Ac_4GalNAz was the most robust approach, as reported previously for mammalian cells[214], and was most efficient at ~16 hours (Figure 2.3). Other sugar analogues are more specific than Ac_4GalNAz[258] in labeling O-GlcNAcylated proteins over glycoproteins within the secretory pathway. However, since O-GlcNAc is the only chromatin associated glycan modification, we felt that the higher labeling efficiency of Ac_4GalNAz relative to other sugar analogues[259], would be more suitable for ChIP-seq experiments.

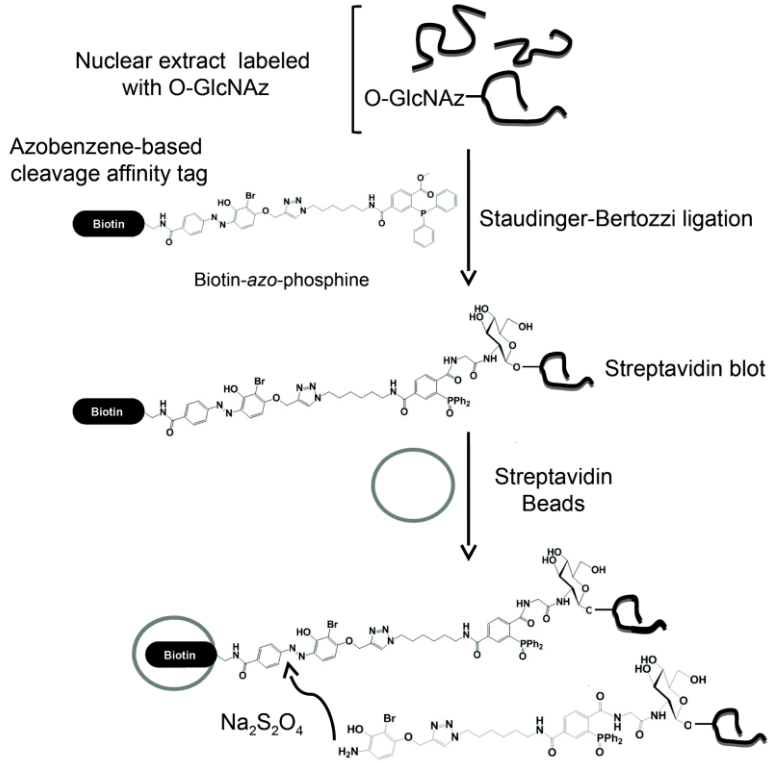


Figure 2.2. Schematic illustration of an antibody free method to enable ChIP-seq analysis of O-GlcNAc.

Cells incorporate Ac_4GalNAz and through the GalNAc salvage pathway is converted to UDP-GlcNAz, which is subsequently used as a substrate by OGT to modify PcG and other proteins (represented as dark curly line). DNA is cross-linked and fragmented to the O-GlcNAzylated proteins which is then ligated with a Biotin-azo-phosphine probe using the Staudinger Ligation. Fragments of DNA cross-linked O-GlcNAc-biotin probe can then be purified via streptavidin beads and later released with $Na_2S_2O_4$ treatment. Without DNA cross-linking, this method can be fully utilized to attain protein ID via western using a streptavidin immunoprecipitation and blot with antibody of choice.

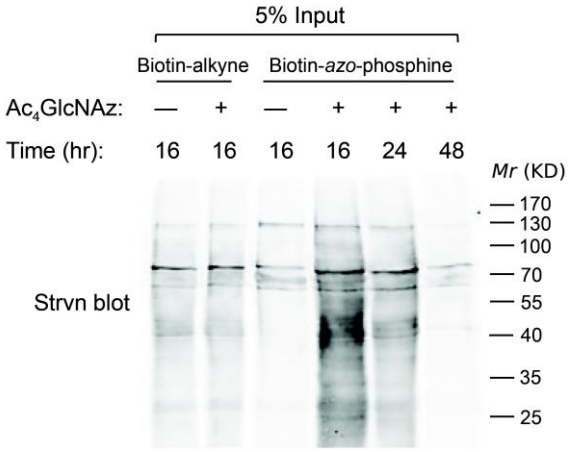


Figure 2.3. Biotinylated tags Biotin-alkyne and Biotin-azo-phosphine for CuAAC- and Staudinger-Bertozzi ligation dependent detection of O-GlcNAz. S2 cells were treated with 200 μM of Ac_4GalNAz for 16-48 hours and analyzed by Streptavidin (Strvn) blot using Odyssey (LI-COR Biosciences). Data was reproducible using two replicates.

2.2.2 Identification of Pho and dHCF-1 as O-GlcNAcylated proteins

Ph was previously identified as an OGT substrate by WGA pull down[153], however, earlier data suggests several yet to be identified chromatin-bound proteins may also be O-GlcNAcylated in *Drosophila*[253]. Given the role of OGT in silencing and the observation that Ying Yang 1 (YY1), the mammalian homologue of *Drosophila* Pleiohomeotic (Pho), is O-GlcNAcylated[260], we speculated that Pho and other PcG proteins may also be modified and contribute to the observed distribution of O-GlcNAc throughout the *Drosophila* genome[253]. Because Pho is involved in recruitment of PRC1 and PRC2 to the genome, the presence of Pho at genomic loci is often considered a marker for identification of PREs. Knowledge of its O-GlcNAcylation state might therefore help rationalize any observed distribution of O-GlcNAc across the genome. We therefore used Ac_4GalNAz-labeled S2 nuclear lysates, coupled the azide moieties with Biotin-*azo*-phosphine, and then used streptavidin resin to purify O-GlcNAz-modified proteins. After eluting proteins using $Na_2S_2O_4$ we probed the eluate with antibodies against several proteins known to bind to PREs including the PRC1 subunit Ph, Pho, as well as the PcG and Trithorax group (TrxG) associated epigenetic regulator host cell factor 1 (HCF-1), which is extensively modified in mammals[243] (Figure 2.1, c). We find these three proteins are O-GlcNAz-modified and verified this observation using both WGA pull down and an alternative O-GlcNAc proteomic technology involving use of a mutant β-1,4-galactosyltransferase (Y289L GalT)[220]. Importantly, when using the WGA and GalT methods, we find that O-GlcNAc-based reactivity was diminished after treatment with a bacterial homologue of OGA (*Bt*OGA) that is able to cleave O-GlcNAc (Figure 2.1,

d-e)[261,262]. Notably, we observed that, in agreement with actin being a reported substrate for OGT in mammals[263], this protein is pulled down after metabolic labeling with Ac_4GalNAz, WGA, and chemoselective labeling with GalT. Further, we observed actin immunoreactivity after WGA precipitation regardless of *Bt*OGA treatment, suggesting O-GlcNAc-independent binding of this abundant protein by WGA (Figure 2.1, c-e). We validated the identification of these proteins as being O-GlcNAcylated by also examining the O-GlcNAcylation of Pho and three control proteins including Drpr (CG2086), Crb (CG6383), and Pak (CG10295) from both wild type and *sxc*$^{-/-}$ pupae. Based on the low level of predicted disordered regions present in these control proteins, we anticipated they would not be O-GlcNAcylated since O-GlcNAc is generally found within disordered regions of proteins. Metabolic feeding of wild type and *sxc*$^{-/-}$ pupae, followed by Staudinger ligation and streptavidin precipitation, enabled detection of Pho but none of the three control proteins. (Figure 2.4, a-c). Collectively, these data showed that multiple PcG or PcG-associated proteins are O-GlcNAcylated and detectable using this metabolic labeling approach.

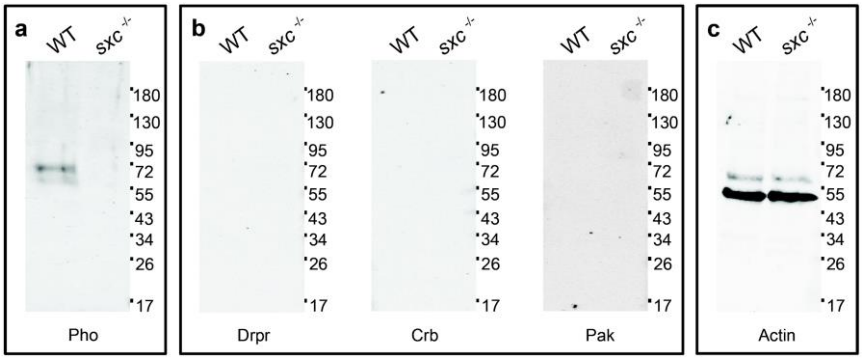


Figure 2.4. **Ac_4GalNAz fed wild type and *sxc*$^{-/-}$ *Drosophila* followed by Biotin-alkyne conjugation and streptavidin purification enriches specific proteins.** *Drosophila* white pre-pupae were fed with 100 μM Ac_4GalNAz from embryogenesis and their proteins were immunoprecipitated with phosphine probe followed by western blot with Pho antibody (a), and Drpr, Crb, and Pak antibodies (b). Actin levels are shown before immunoprecipitated with phosphine probe as a loading control (c).

2.2.3 Chemical mapping of O-GlcNAc across the *Drosophila* genome

We next tested whether this metabolic labeling approach, used in conjunction with chemoselective ligation and next-generation sequencing of protein-bound DNA fragments, could reveal the genomic locations to which O-GlcNAcylated proteins bind in *Drosophila* S2 cells. As a positive control we performed ChIP to test for the presence of Pho at loci known to possess PREs. We focused on regions from the characterized PcG target genes *Abd-B*, *Dll*, *en*, *pnr*, *Scr*, *tsh*, and *Ubx* (Figure 2.5) and compared these to genes lacking a PRE, including *dpr12* and *CG11665* (Figure

2.5). The pattern of bands we observed for these targeted genes during polymerase chain reaction (PCR) analysis of DNA fragments obtained during O-GlcNAz-mediated precipitation corresponds to the pattern of ChIP-isolated DNA fragments we observed for Pho. Importantly, we also found that biotin capture enriches DNA from O-GlcNAz-labeled chromatin but not DNA from vehicle treated cells, indicating that we detected only DNA sequences at which O-GlcNAz modified proteins are bound within cells. We find that the level of enrichment of DNA observed using metabolic labeling with Ac_4GalNAz in tandem with the Biotin-*azo*-phosphine probe was at least 380-fold, which was superior to the use of CuAAC reagents were 140-fold enrichment was observed. These data indicate that this chemical method enables efficient isolation of DNA from these genes known to contain PREs, which bear O-GlcNAcylated proteins.

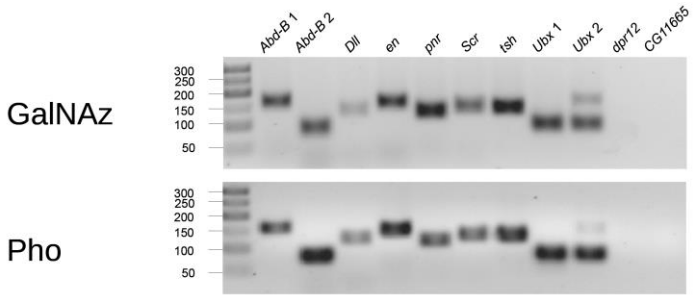


Figure 2.5. DNA purified from metabolic feeding-chemoselective ligation followed by ChIP-PCR mimics patterns of DNA pulled down by Pho antibody at discrete loci in *Drosophila*.
O-GlcNAz (upper panel) co-localizes with Pho (lower panel) on HOX gene PcG responsive elements (PREs) in S2 cells. DNA recovered by ChIP- was amplified using the primers for PRE regions of each HOX gene [lane 1, *Abd-B* (+2.1 kb); lane 2, *Abd-B* (+72.3 kb); lane 3, *Dll* (-1 kb); lane 4, *en* (-0.3 kb); lane 5, *pnr* (+3.9 kb); lane 6, *Scr* (+0.2 kb); lane 7, *tsh* (+19.4 kb); lane 8, *Ubx* (-29.6 kb); lane 9, *Ubx* (-29.4 kb)] in the input fraction of the ChIP assays. Select euchromatic [lane 10, *dpr12* (-3.2 kb)] and heterochromatic [lane 11, *CG11665* (+12.5 kb)] regions were included as negative controls to which PcG proteins do not bind.

Given that these targeted ChIP results were qualitatively positive and DNA enrichment was high, we next examined the potential of this strategy in conjunction with next-generation DNA sequencing to perform genome-wide mapping of O-GlcNAc. Because Pho is O-GlcNAcylated and is predominantly found at PREs at which silencing occurs[264], we compared the O-GlcNAz ChIP-seq profile of chromatin we obtained from analysis of S2 cells with the profile we obtained for the distribution of Pho. We also performed ChIP-seq using the lectin WGA and chemoenzymatic GalT labeling and compared these data to the results obtained for O-GlcNAz and Pho (Figure 2.6, a-b). We find that signals from the O-GlcNAz and WGA ChIP-seq

data were consistent with our targeted study (Figure 2.5) and showed similar localization at discrete genomic loci, including those encoding *HOX* genes (Figure 2.6, a-b). These loci overlapped to a great extent with well characterized PREs bound by Pho, including for example, *Ubx-Abd-B*, *inv-en*, *lab-Antp*, *InR-E2f*, *srp-pnr*, *hh-pnt* as well as the TrxG-response-element (TRE)-associated locus, *roX1* (Figure 2.6, a-b and Figure 2.7, a-e). Control cells that were treated with vehicle instead of Ac_4GalNAz, yielded extremely low levels of DNA, consistent with the high enrichment we observed for metabolic labeling. Therefore, for subsequent ChIP-seq experiments we used whole genomic DNA as a control to enable analysis of the distribution of sequencing reads as previously recommended[265]. These data support the specificity of this metabolic feeding ChIP-seq strategy. Inspection of the data obtained using GalT, however, revealed comparatively poor enrichment at these loci (Figure 2.6, a-b and Figure 2.7, a-e), suggesting this method as implemented here lacks sufficient specificity for use in ChIP-seq studies, perhaps due to the different ligation methods and linkers used for these two approaches.

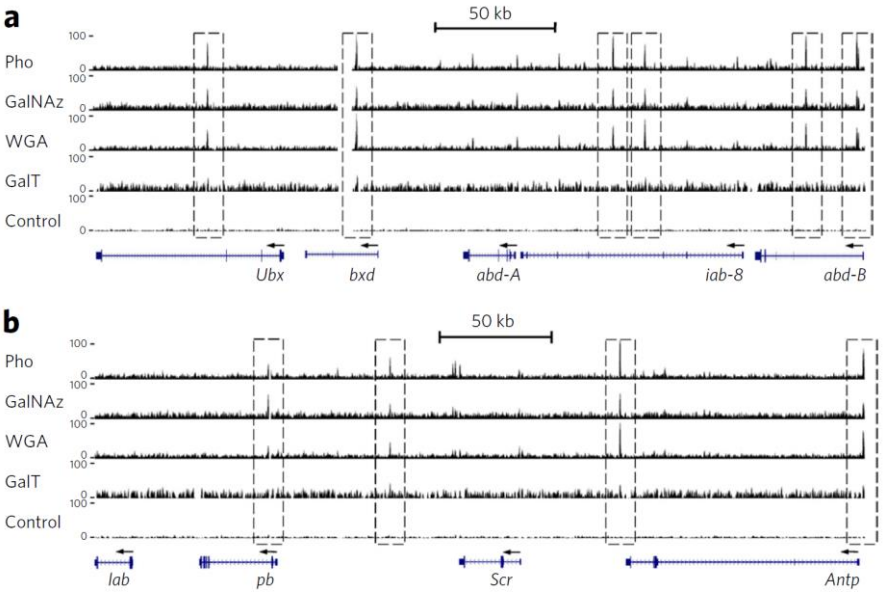


Figure 2.6. Metabolic feeding combined with antibody-free genome wide chromatin precipitation and sequencing reveals O-GlcNAcylated proteins at discrete loci in *Drosophila* S2 cells.
ChIP-seq tracks of normalized read density shown in descending order were obtained using anti-Pho antibody (Pho - top track), Ac_4GalNAz (GalNAz) feeding, WGA pull down (WGA), GalT labeling (GalT), and non-enriched genomic DNA (Control - bottom track). Genes are labeled and drawn in blue under tracks and show exons as darker boxes, introns as thin lines and the transcriptional start site is indicated by an arrow. Tracks show (a) *Ubx* - *Abd-B* and (b) *lab* - *Antp* loci. Peaks are highlighted in dashed boxes. Peaks called using MACS 1.4 (p=0.005).

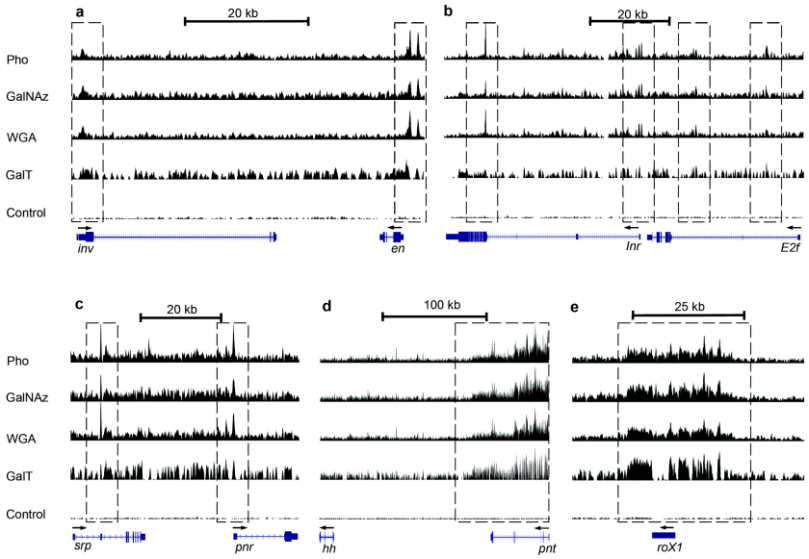


Figure 2.7. ChIP-seq tracks at various HOX loci in S2 cells.
ChIP-seq tracks of normalized read density of Pho (top track), Ac_4GalNAz, WGA, GalT labeling and genomic DNA control (bottom track) in S2 cell. Bioinformatics analysis (MACS version 1.4.2, p-value = 0.005) identified the HOX gene regions *inv* to *en* (scale = 0-100) (a), *lnR* to *E2f* (scale = 0-150) (b), *srp* to *pnr* (scale = 0-150) (c), cnc cluster (scale = 0-150) (*hh* to *pnt*, d) and TrxG response element (TRE), *roX1* locus (scale = 0-100) (e), to contain O-GlcNAc protein(s). Blue boxes represent exons and thin lines represent introns. Peaks are highlighted with dashed boxes.

2.2.4 Informatics analysis of the genomic distribution of O-GlcNAc

Given the abundance of O-GlcNAc at PREs observed when using metabolic feeding, we wanted to determine the extent of overlap between our O-GlcNAz datasets and Pho ChIP-seq experiments at PRE-containing genes, with reference to external datasets of known PREs in *Drosophila*. A previous analysis of PREs within S2 cells, using antibodies to various PcG proteins and H3K27me3, used ChIP in conjunction with microarrays (ChIP-on-chip)[266]. More recently a computational approach, EpiPredictor[267], has provided the most comprehensive prediction of PREs to date[264,266–268]. Within the S2 ChIP-on-chip data set[266] we used the high confidence set of genes associated with PREs and from the EpiPredictor data set we selected the set of high confidence set of genes associated with predicted PREs. We thus obtained two lists consisting of 179 (S2 ChIP-on-Chip) and 314 (EpiPredictor) predicted PRE-containing genes. We next analyzed our ChIP-seq data using different analysis software including MACS 1.4.2, MACS 2.1, and HOMER 4.7 with different parameter settings (Figure 2.8, a-c). We found that MACS 1.4.2 (p-value = 0.005) gave the most robust determination of peaks, where each peak represents a large number of overlapping DNA sequencing reads at a specific genomic locus. Using MACS 1.4.2 (p-value = 0.005) we identified 973 Pho peaks, 784 Ac_4GalNAz peaks, 2233 WGA peaks, and 5720 GalT peaks (Table 2.1). We then analyzed the

relationship between peaks observed in our Pho, O-GlcNAz, WGA, and GalT ChIP-seq data sets with the S2 ChIP-on-chip and EpiPredictor-identified PREs, for a total of eight analyses. We found that data from our Pho ChIP-seq and all our O-GlcNAc ChIP-seq experiments significantly correlate with the annotated PRE-containing genes (Table 2.2). Notably, we find that our O-GlcNAz dataset correlates most closely with the largest PRE dataset provided by EpiPredictor and shows better overlap with this PRE dataset than our WGA and GalT datasets (Table 2.2).

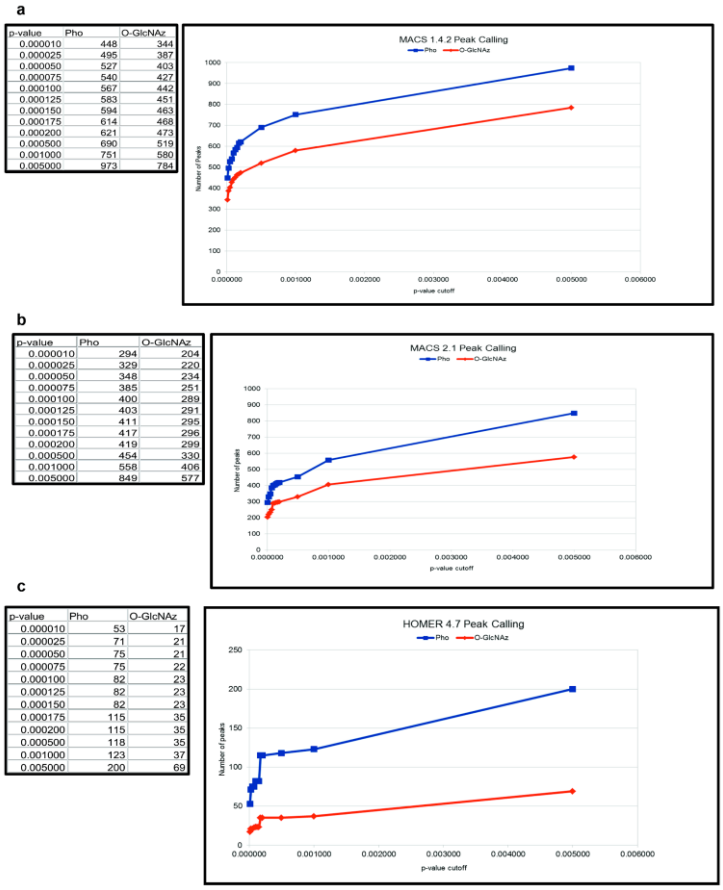


Figure 2.8. **Effect of ChIP-seq peak calling algorithms and parameter settings on the number of peaks called.**
Number of peaks called using MACS 1.4.2 (a), MACS 2.1 (b) and HOMER 4.7 (c) with different p-value cutoff on ChIP-peak calling.

Table 2.1. Bioinformatics summary of ChIP-seq data from Ac_4GalNAz fed S2 cells, WGA purified loci and GalT labeled loci.

Basic characteristics of MACS peaks from O-GlcNAc ChIP-seq experiments in S2 cells by Ac_4GalNAz (GalNAz), WGA and GalT and Pho ChIP-seq. Number, length, total genome coverage and overlapped RefSeq genes of MACS peaks from each ChIP-seq experiment.

MACS Peak Stats	GalNAz	WGA	GalT	Pho
MACS Peaks Called	784	2233	3554	973
Average Peak Length (bp)	3040	1377	2020	2712
Total Genome Coverage (%)	1.41	1.82	4.25	1.56
Non-Redundant Genes Overlapped	545	1124	2081	733

Table 2.2. GenometriCorr analysis of MACS peaks in S2 cells. Correlation of Pho, Ac_4GalNAz, WGA and GalT ChIP-seq peaks with Shwartz and Zeng PRE genes.

If the relative ecdf deviation area is positive then there is a positive correlation between the datasets. The higher the value of relative ecdf area correlation indicates a higher correlation and the value of the relative ecdf deviation area p-value indicates if the correlation between the datasets is significant. Ac_4GalNAz is most correlated with the Zeng dataset.

Query Population	Pho		GalNAz		WGA		GalT	
Reference Population	Shwartz	Zeng	Shwartz	Zeng	Shwartz	Zeng	Shwartz	Zeng
Relative ecdf deviation area	0.011	0.018	0.01	0.02	0.019	0.012	0.009	0.01
Relative ecdf area correlation	0.045	0.074	0.04	0.081	0.078	0.046	0.038	0.04
Relative ecdf deviation area p−value	0.036	<0.001	0.085	<0.001	<0.001	0.001	<0.001	<0.001

We next performed a more detailed analysis of our S2 cell ChIP-seq datasets to compare the different O-GlcNAc ChIP-seq strategies. Examining the peaks called for each O-GlcNAc ChIP-seq strategy we found that the percent of total genome coverage for Ac_4GalNAz, WGA, and GalT was 1.41%, 1.82% and 3.08%, respectively (Table 2.1 and Figure 2.9, a). Interestingly, the number of peaks called using the WGA (2233) and GalT (5720) methods was from 3- to 7-fold higher than for

Ac_4GalNAz (784 peaks), while the genome coverage obtained using WGA and GalT are only 1.3 and 2.2-fold higher than when using Ac_4GalNAz. This difference likely stems from larger peaks being called for the Ac_4GalNAz dataset as compared to the WGA and GalT datasets (Table 2.1 and Figure 2.9, b). When analyzing the O-GlcNAz strategy, we find it had the highest fraction of shared peaks between all three O-GlcNAc ChIP-seq strategies (58.7%) followed by WGA (47.3%) and lastly GalT (18.5%) (Figure 2.10). Furthermore, when we explored the overlap of each O-GlcNAc ChIP-seq strategy with the distribution of Pho at genes, we found that the Ac_4GalNAz and Pho datasets are the most similar (Figure 2.11 and Figure 2.12), which is consistent with our comparisons using genome wide peaks (Figure 2.13, a-d).

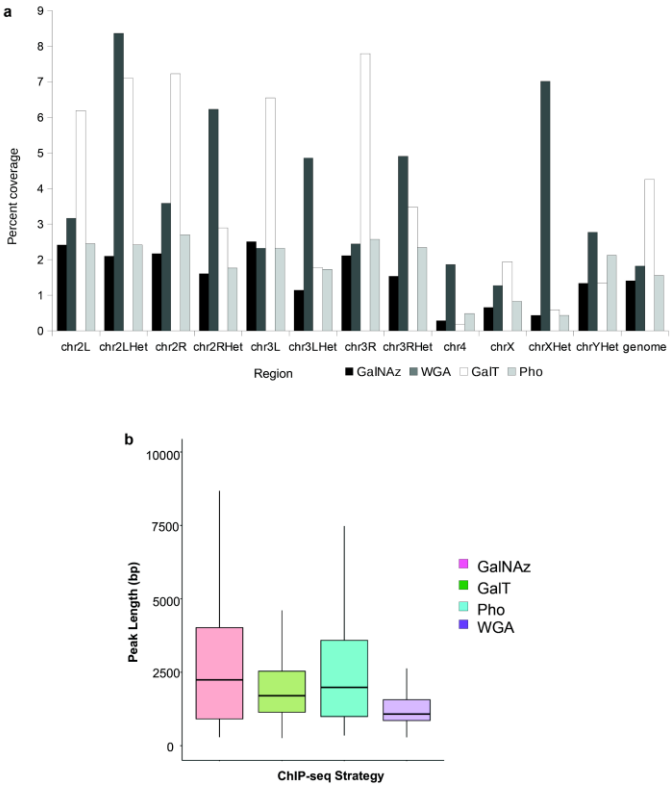


Figure 2.9. Basic MACS peak characteristics.
(a) Nucleotide coverage represented as a percent at across the genome. (b) Whisker plot of peak lengths of Ac_4GalNAz (red), GalT(green), Pho (blue) and WGA (purple) ChIP-seq experiments. Each box shows the median (mid line) and the lower and upper quartiles (25% and 75%).

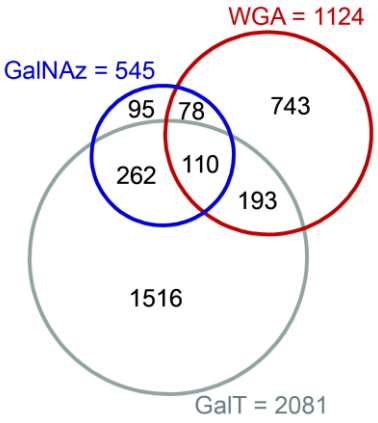


Figure 2.10. **Overlap of Ac_4GalNAz, WGA, and GalT ChIP-seq genes.**
Venn diagram of RefSeq genes that overlap with peaks from each O-GlcNAc ChIPseq experiment. Peaks from Ac_4GalNAz, WGA and GalT overlap with 545, 1124 and 2081 RefSeq genes respectively.

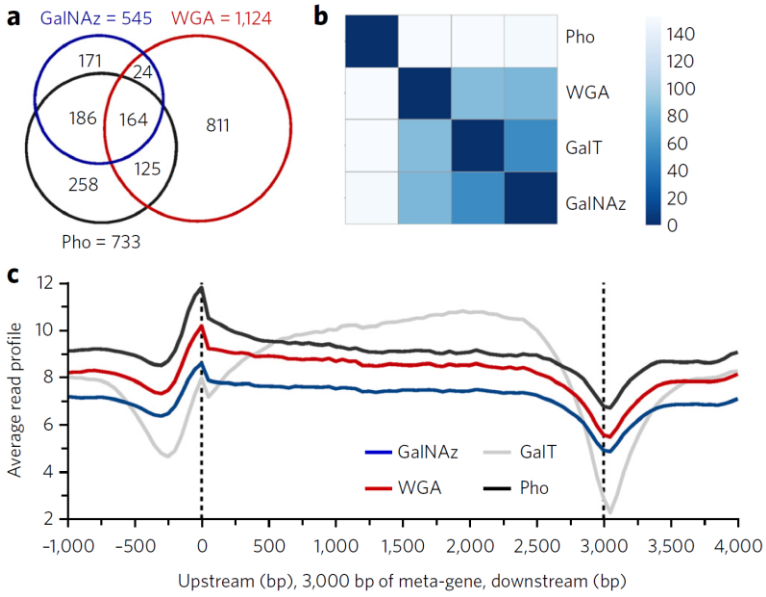


Figure 2.11. **Comparative bioinformatics analysis of next-generation sequencing data from S2 cells using Ac_4GalNAz feeding, WGA precipitation, and GalT labeling.**
(a) Venn diagram of RefSeq genes that overlap with MACS peaks from Pho, Ac_4GalNAz and WGA ChIP-seq experiment. Peaks from Pho, Ac_4GalNAz and WGA overlap with 733, 545 and 1124 RefSeq genes respectively. (b) DESeq2 differential expression analysis using all aligned reads from BAM files. Different O-GlcNAc ChIP-seq strategies were analyzed genome wide at 100 bp bins. Scale is shown in log2. (c) Average gene profile image of ChIP-seq experiments obtained with CEAS from BAM file read density at genes with peaks for each experiment. Y-axis indicates average read signal. The prefix "meta" indicates that genes have been normalized to the same length. Data was normalized to account for differences in sequencing depth.

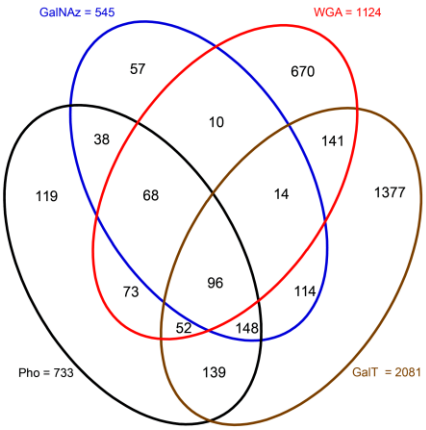


Figure 2.12. Venn diagram of RefSeq genes that overlap with MACS peaks from each O-GlcNAc ChIP-seq experiment and Pho ChIP-seq experiment. MACS peaks from Pho, Ac_4GalNAz, WGA and GalT overlap with 733, 545, 1124 and 2081 RefSeq genes respectively. WGA and GalT have many peaks that are not shared between other ChIP-seq strategies.

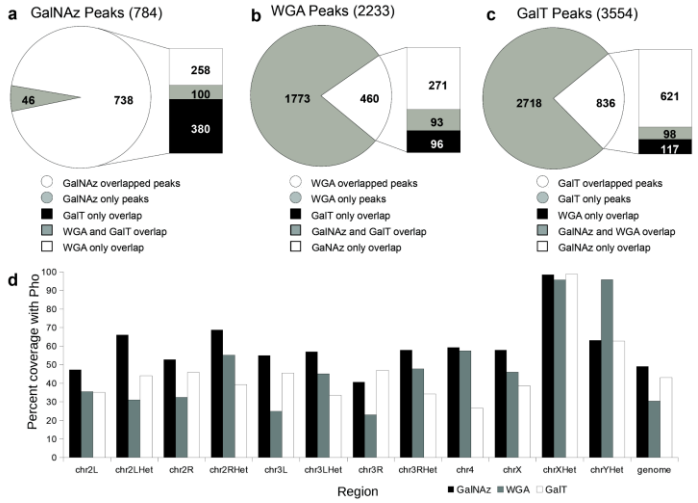
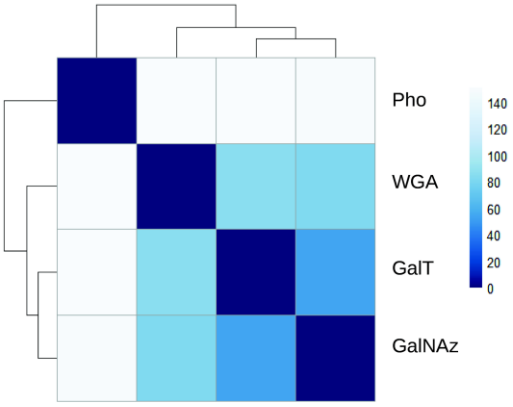


Figure 2.13. Analysis of total MACS peaks called for each GlcNAc ChIP-seq strategy and with Pho ChIP-seq in S2 cells. Bar of pie charts showing total MACS peaks (inter- and intragenic) from (a) Ac_4GalNAz, (b) WGA and (c) GalT that are unique or are shared (overlap) with other O-GlcNAc ChIP-seq strategies. The bar shows a detailed breakdown of how the peaks are shared. (d) Genome wide coverage of Pho peaks represented as a percent of nucleotides within Pho peaks are shown as a histogram for each O-GlcNAc ChIP-seq. Pho coverage on Pho would be 100% at every region. Ac_4GalNAz has the highest coverage of Pho.

As a complementary method to interrogate our datasets we used a DNA enrichment analysis of all sequences aligned to the genome[269]. We divided the genome into a series of bins each 100 bp long and then tallied the number of DNA sequencing reads we obtained using each strategy into each bin. We then compared the number of sequences we obtained within each bin for each ChIP-seq method across the genome. Consistent with our peak overlapping analyses, we found that of the three O-GlcNAc ChIP-seq datasets the distribution of Pho was most similar to the

data obtained using metabolic feeding with Ac_4GalNAz. We also saw that the other O-GlcNAc ChIP-seq experiments were more similar to the Ac_4GalNAz strategy than to each other (Figure 2.11, b and Figure 2.14). Furthermore, Ac_4GalNAz, WGA, and Pho show similar distributions across the genome, with enrichment at transcriptional start sites and depletion at transcriptional end sites, whereas GalT shows a different pattern (Figure 2.11, c), again suggesting that the GalT strategy is not optimal for ChIP-seq using the current conditions and reagents.



	GalT	GalNAz	WGA
GalNAz	53.67200		
WGA	85.49858	81.78772	
Pho	150.74088	148.73746	150.40125

Figure 2.14. All ChIP-seq experiments performed are most similar to Ac_4GalNAz when compared against each other in DESeq2.
DESeq2 analysis using raw reads from BAM files. Strategies were analyzed genome wide at 100 bp bins. Scale is shown in Log2 and the values of dissimilarity are shown on the bottom chart.

2.2.5 Genomic mapping of O-GlcNAc *in vivo* within *Drosophila*

We next examined whether this metabolic labeling strategy could be used *in vivo* to map O-GlcNAc to the genome in *Drosophila*. We raised flies on media containing different concentrations of Ac_4GalNAz and found, upon analysis of whole organism lysates, that pupae grown on 100 μM Ac_4GalNAz were optimally labeled (Figure 2.15 and Figure 2.16). We observed that different stages of *Drosophila* could be labeled (Figure 2.15, a and Figure 2.17) and that *sxc*$^{-/-}$ pupae showed a great decrease in labeling compared to wild type pupae (Figure 2.15, b). Metabolic ChIP-seq analysis of wild type and *sxc*$^{-/-}$ Ac_4GalNAz-fed pupae revealed O-GlcNAc at PRE target genes in wild type but not in *sxc*$^{-/-}$ pupae (Figure 2.15, c and Figure 2.18, a-h). This specificity and high enrichment is consistent with our data from S2 cells. Notably, we found that the genomic distribution of O-GlcNAc within pupae resembled that seen for S2 cells at major PREs and is, again, not widely distributed across the

genome (Figure 2.6, a-b and Figure 2.18, a-d). MACS analysis of O-GlcNAz ChIP-seq data in pupae revealed 2432 peaks which overlap 1651 genes. We found that these genes at which O-GlcNAc is bound include most of those we detected in S2 cells, as well as a large number of genes not detected in S2 cells (Figure 2.15, d). We also observed that *Drosophila* pupae had greater numbers of peaks for the Ac_4GalNAz ChIP-seq data set at exons as compared to S2 cells (Figure 2.19). These differences in genomic distributions are consistent with expected differences between cells and flies and highlight the utility of Ac_4GalNAz feeding *in vivo*.

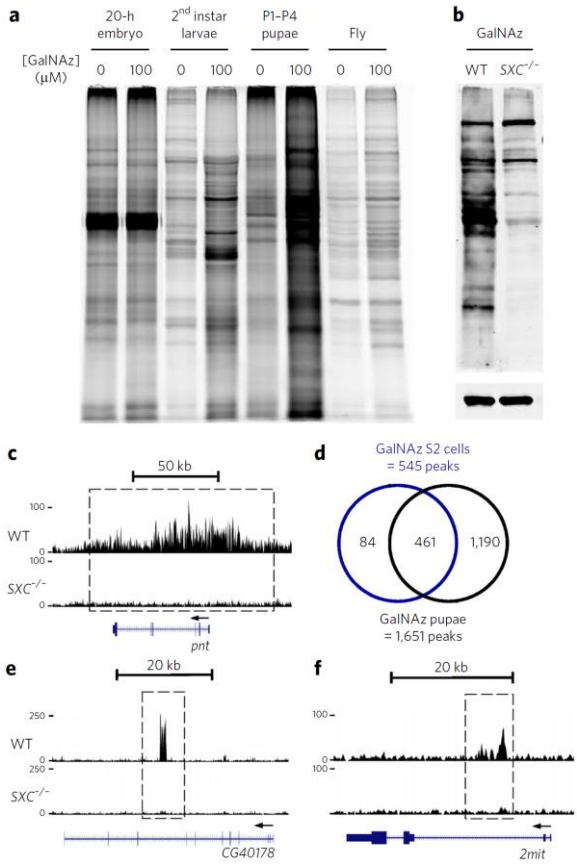


Figure 2.15. Ac_4GalNAz feeding enables *in vivo* labeling of *Drosophila* at larval, pupal, and fly stages.

(a) Strvn blot of different stages of *Drosophila* that have been fed 0 or 100μM Ac_4GalNAz since birth. Cu catalyzed Huisgen [3+2] cycloaddition of a biotin containing alkyne probe was used to label O-GlcNAzylated proteins in this case rather than the phosphine probe. (b) Strvn blot of wild type and *sxc*$^{-/-}$ white pre-pupae *Drosophila* that have been fed 100μM Ac_4GalNAz (GalNAz) since birth on top panel. Actin levels on bottom panel. All blots were reproducible using at least two biological replicates. (c) ChIP-seq tracks of normalized read density for Ac_4GalNAz fed wildtype pupae (top tracks) and *sxc*$^{-/-}$ pupae (bottom tracks) shows O-GlcNAc signal only in wild type pupae at loci containing the known PRE *pnt.* Genes are labeled and drawn in blue with exons depicted as darker boxes, introns as thin lines, and the transcriptional start site marked by an arrow. Peaks are highlighted in dashed boxes. (d) Venn diagram showing overlap between Ac_4GalNAz ChIP-seq peaks in S2 cells and pupae. ChIP-seq tracks as in (c) for the previously

undetermined GlcNAc regulated loci *2mit* (e) and heterochromatic gene *CG40178* (f).

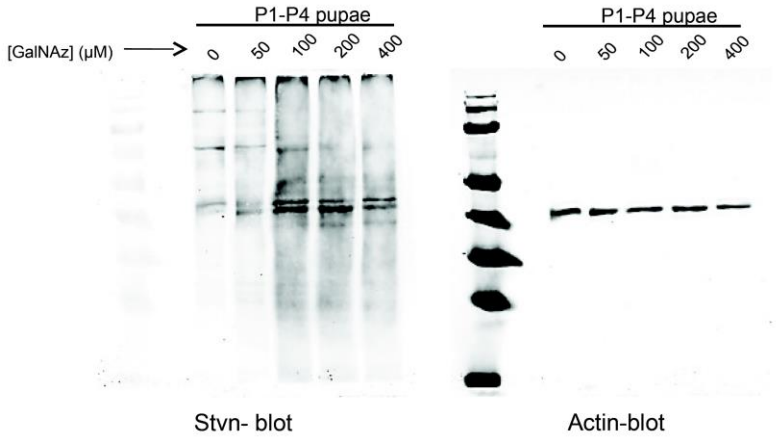


Figure 2.16. 100 μM Ac_4GalNAz is an optimal concentration for *in vivo Drosophila* labeling.
Drosophila pupae were fed different concentrations of Ac_4GalNAz from embryogenesis and their proteins were immunoprecipitated with phosphine probe followed by Strvn blot (left). Actin blot to show protein loading is on right. Data was reproducible using two replicates.

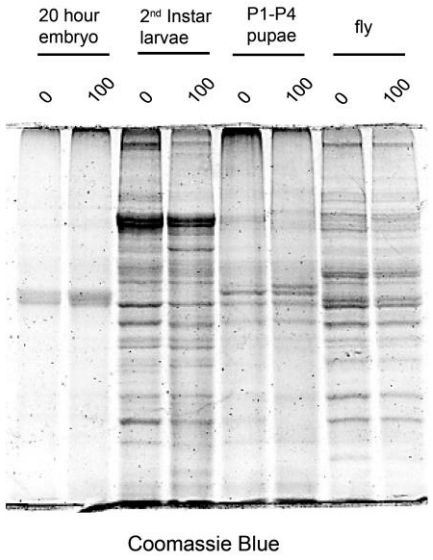


Figure 2.17. Loading control for Ac_4GalNAz fed *Drosophila*.
Different stages of *Drosophila* were fed with 100 μM Ac_4GalNAz from embryogenesis and their proteins were immunoprecipitated with phosphine probe followed by streptavidin blot (Figure 2.15, a). A Coomassie blue stained gel shows the protein loading. Data was reproducible using two replicates.

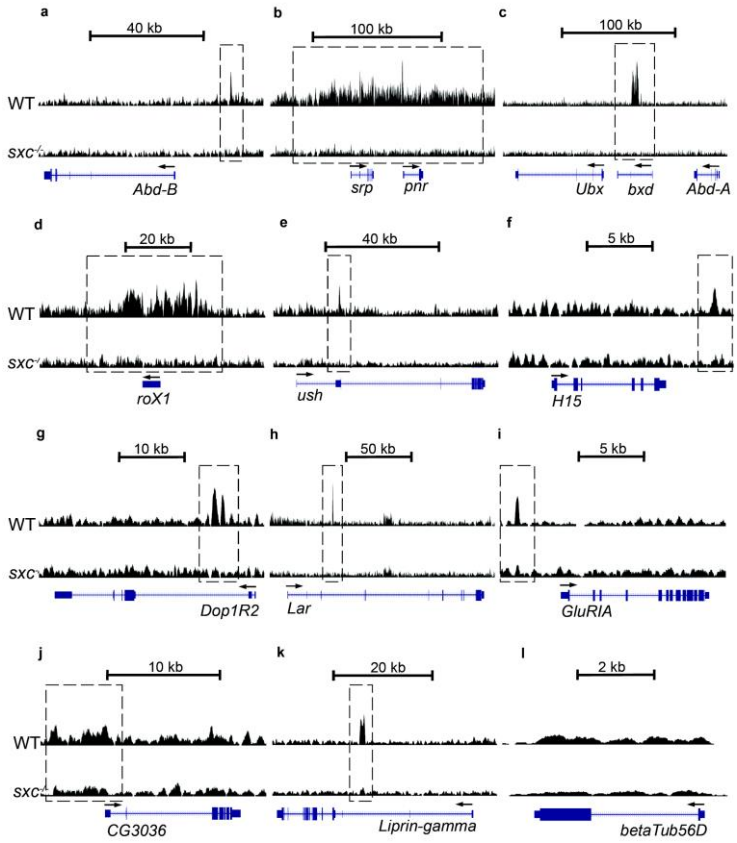


Figure 2.18. O-GlcNAz ChIP-seq tracks in wild type and *sxc*$^{-/-}$ *Drosophila* pupae at HOX genes and several other O-GlcNAcylated loci.
ChIP-seq tracks of normalized read density of wild type (top tracks) and *sxc*$^{-/-}$ (bottom tracks) 100 μM Ac_4GalNAz fed pupae. *Abd-B* (scale = 0-100) (a), *srp-pnr* (scale = 0- 100) (b), *Ubx-Abd-A* (scale = 0-100) (c), *roX1* (scale = 0-50) (d), *ush* (scale = 0-100) (e), *H15* (scale = 0-35) (f), *Dop1R2* (scale = 0-50) (g), *Lar* (scale = 0-100) (h), *GluRIA* (scale = 0-50) (i), *CG3036* (scale = 0-50) (j), *Liprin-gamma* (scale = 0-100) (k), and *betaTub56D* (scale = 0-100) (l). Exons are represented as blue boxes and introns as lines. Peaks are highlighted with dashed boxes.

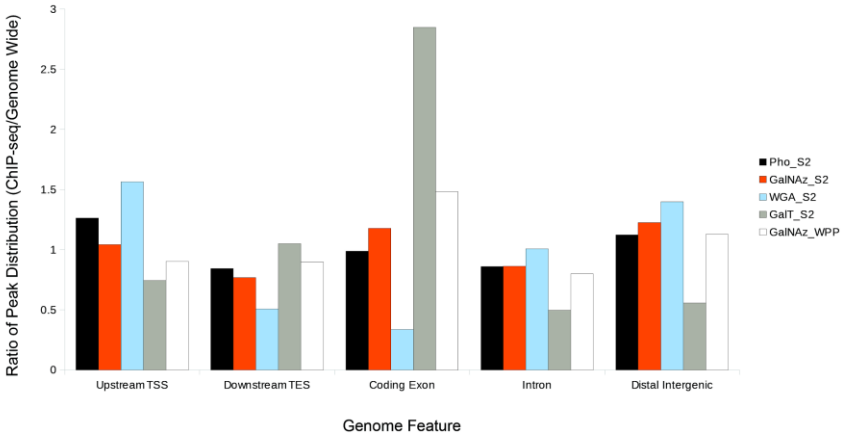


Figure 2.19. Summary of genomic features found at ChIP-seq peaks.
ChIP-seq peaks of Ac_4GalNAz in S2 cells, WGA in S2 cells, GalT in S2 cells, Pho in S2 cells, and Ac_4GalNAz in pupae were analyzed for distribution around genomic features. Upstream TSS includes 5'UTR and 3000bp upstream of transcriptional start site (TSS). Downstream TES includes 3'UTR

and 3000bp downstream of transcriptional end site (TES). Values are reported as a ratio to each feature seen in *Drosophila* dm3 genome.

2.2.6 OGT regulates expression from diverse O-GlcNAcylated loci

Interestingly, among our datasets from both S2 and pupae we observed some loci at which both O-GlcNAc and Pho are found (Figure 2.20, a-c and Figure 2.21, a-f) as well as sites at which O-GlcNAc was present alone (Figure 2.15, e-f, Figure 2.18, i-k). Interestingly, none of these loci are either annotated[264,268] or predicted[266,267] to be associated with a PRE. We therefore analyzed genes at which we observed O-GlcNAc binding in pupae in the context of previous datasets describing the genomic distribution of Ph[153], WGA[153], *Drosophila* Scm-related gene containing four mbt domains (dSfmbt)[270], and Pho[270] from imaginal discs, and H3K27me3[271] in pupae[215]. Surprisingly, we found that among these O-GlcNAc bound genes, there are 1492 that lacked binding of any of these PcG proteins and/or H3K27me3 marks and are not annotated[264,268] or predicted[266,267] as containing PREs. This observation suggests that OGT influences gene expression via mechanisms independent of PcG regulation. Notably, present among these loci were many heterochromatic genes; this is of interest, since, although various studies have described more general effects of perturbations in proteins important for assembly and/or maintenance of heterochromatin on expression of genes located therein, little is known about the molecular regulation of specific heterochromatic genes[272]. We therefore wondered whether OGT might play a role in regulating expression of genes found at these various non-PRE-containing sites, including heterochromatic genes. To address this question we performed quantitative PCR (qPCR) analysis using wild-type and *sxc*$^{-/-}$ mutant pupae[229] of 15 genes from three classes of genes that bound O-GlcNAc modified proteins in our study (Figure 2.20, a-c, Figure 2.18, e-k. Five of these genes, for which we expected differential expression, are annotated as possessing PREs (*pnt* and *H15*) or are associated with one or more of Ph, Pho, dSfmbt, or H3K27me3 marks and are predicted to be PREs in the EpiPredictor dataset (*Dop1R2*, *Lar*, and *ush*). A further five are euchromatic genes not previously identified or predicted as containing PREs and which bear none of these PcG proteins or H3K27me3 marks (*chn*, *2mit*, *CG3036*, *GluRIA*, *Lipirin-gamma*). The final five are selected from heterochromatic genes, expression of which has not been linked to PcG function (*Parp*, *CG17514*, *Set1*, *CG40178*, and *scro*). We normalized expression ratios against the housekeeping gene, *betaTub56D*, which has never been linked to PcG function nor speculated to contain a PRE (Figure 2.18, l) and which shows very similar expression between wild type and *sxc*$^{-/-}$ mutant pupae in

qPCR analysis. We found differential expression in *sxc*$^{-/-}$ pupae as compared to wild-type pupae of PRE-containing genes as well as most of the other genes, including the heterochromatic genes, that were found at unique non-PRE O-GlcNAc marked loci (Figure 2.20, d and Figure 2.22). Expression of *betaTub56D* was unchanged. It is possible that loci at which we observed O-GlcNAc and Pho may in some cases represent previously unannotated PREs. However, the binding of O-GlcNAc to loci where no PcG proteins or H3K27me3 marks are present raises the intriguing possibility that many of the latter may be subject to regulation by OGT in a manner that is independent of PcG proteins including Ph.

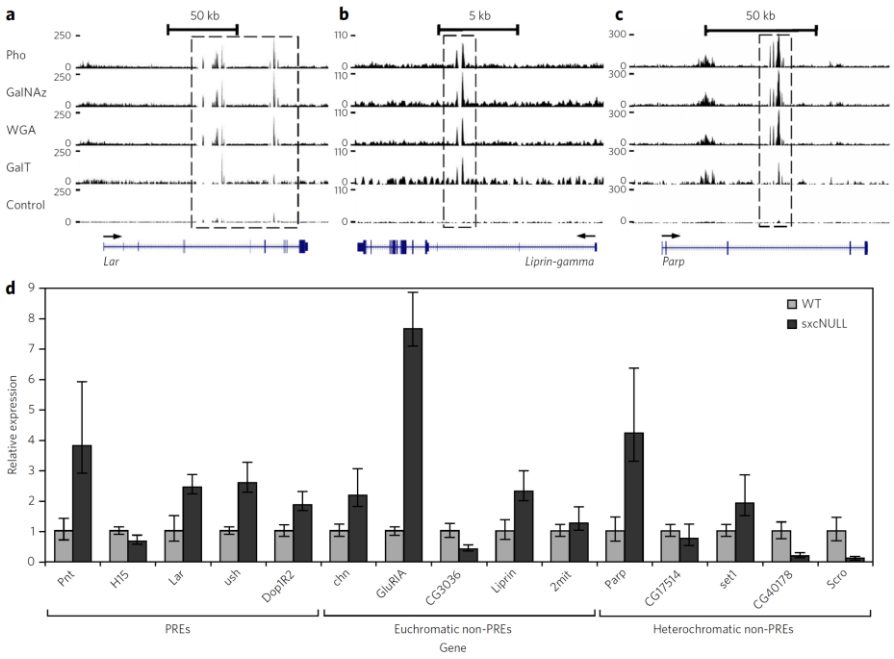


Figure 2.20. O-GlcNAcylated proteins are distributed to genomic loci in *Drosophila* that contain PREs as well as those that lack PREs and gene expression from these diverse loci is regulated by OGT.
ChIP-seq of normalized read density in S2 cells obtained using Pho antibody (Pho - top track), Ac_4GalNAz feeding (GalNAz), WGA pull down (WGA), GalT labeling (GalT), and non-enriched genomic DNA (Control - bottom track). Genes are labeled and depicted in blue with exons shown as darker boxes, introns as thin lines, and the transcriptional start site marked by an arrow. Sequencing data for (a) the known PRE *Lar*, (b) non-PRE *Liprin-gamma*, (c) and non-PRE heterochomatic gene *Parp*. Peaks are highlighted in dashed boxes. (d) Gene expression in *sxc*$^{-/-}$ and wild type white pre-pupae normalized against the house-keeping gene (*betaTub56D*) as determined using qPCR for the known PREs: *pnt, H15, Lar, ush, Dop1R2*, the non PREs: *chn, GluRIA*, *CG3036, Liprin-gamma (liprin), 2mit* as well as the non-PRE heterochomatic genes: *Parp, CG17514, set1, CG40178*, and *Scro* (n=3, data represent mean values ± s.d.). Data is representative of three independent biological replicates (Figure 2.22).

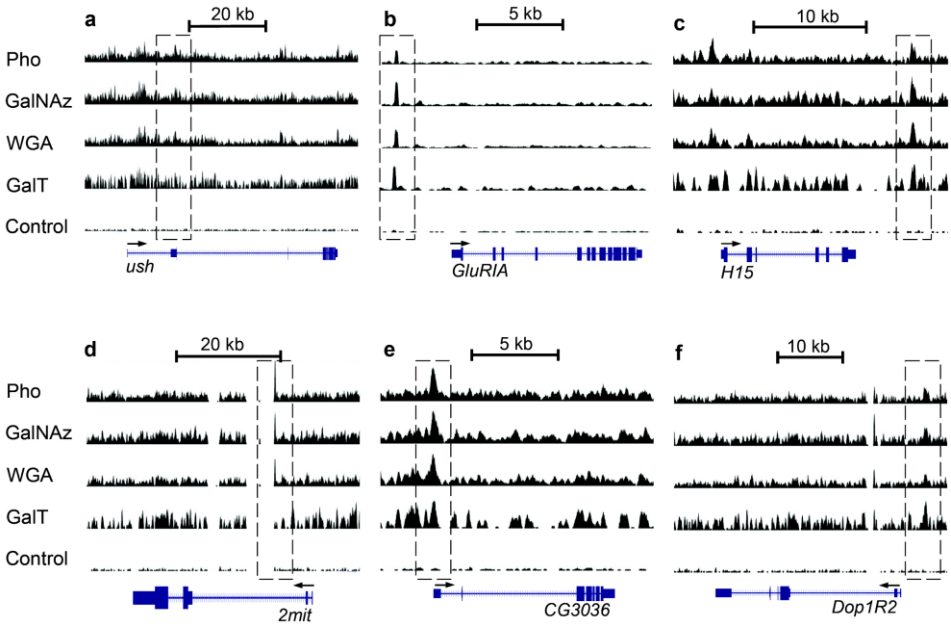


Figure 2.21. ChIP-seq tracks of several O-GlcNAcylated loci in S2 cells.
ChIP-seq of normalized read density of Pho (top track), Ac_4GalNAz, WGA, GalT labeling, and genomic DNA control (bottom track) in S2 cell. OGlcNAcylated loci *ush* (scale = 0-110) (a), *GluRIA* (scale = 0-150) (b), *H15* (scale = 0-50) (c), *2mit* (scale = 0-60) (d), *CG3036* (scale = 0-60) (e), and *Dop1R2* (scale = 0- 60) (f). These genes, among others, showed increased mRNA transcript levels in *sxc*$^{-/-}$ pupae (Figure 2.20, d). Peaks are highlighted with dashed boxes.

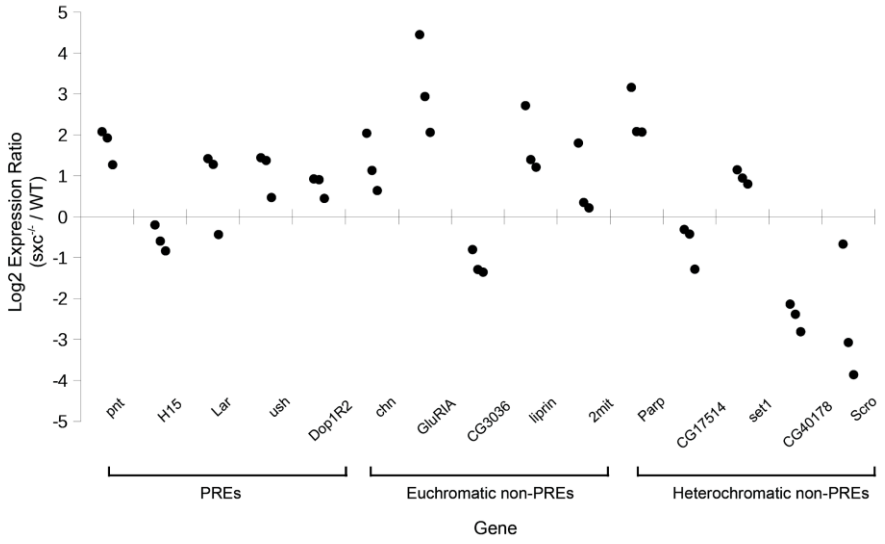


Figure 2.22. O-GlcNAc bound loci show differential gene expression upon loss of *sxc*.
Gene expression in *sxc*$^{-/-}$ and wild type white pre-pupae normalized against the housekeeping gene (betaTub56D) as determined using qPCR for the known PREs: *pnt, H15, Lar, ush, Dop1R2*, the non PREs: *chn, GluRIA, CG3036, Liprin-gamma* (*liprin*), *2mit* as well as the non-PRE heterochomatic genes: *Parp, CG17514, set1, CG40178,* and *Scro*. Data shows Log2 transformed relative quantity values of wild type/*sxc*$^{-/-}$ mRNA levels of three biological replicates, each shown as a separate dot.

2.3 Discussion and conclusions

Here we described a chemical method for genome-wide mapping of O-GlcNAcylated proteins in *Drosophila melanogaster* that eliminates the need for lectins and antibodies, which have recognized limitations[250,273]. Previous studies in which O-GlcNAc has been mapped to the genome have used the lectin wheat germ agglutinin (WGA) and antibodies that recognize O-GlcNAc in certain contexts[153,230,274]. Comparative analysis of this metabolic labeling method for mapping of O-GlcNAcylated proteins to the genome with the use of WGA or GalT chemoenzymatic labeling, shows that of these three methods the metabolic labeling strategy correlates most closely with predicted PREs. This chemical strategy may accordingly aid the genome-wide identification of PREs and facilitate greater understanding of the role of O-GlcNAc in repression of genes associated with PREs. Notably, however, we also observe O-GlcNAc at many loci that are not annotated as containing PREs.

The apparent limited distribution of O-GlcNAcylated proteins to specific loci across the genome that we observe using this metabolic feeding approach is somewhat surprising given reports of many chromatin associated proteins being O-GlcNAcylated[253], particularly within mammals[275]. One might reasonably expect these diverse O-GlcNAcylated proteins to be broadly distributed on the genome. We considered whether the ChIP-seq profile obtained using Ac_4GalNAz may therefore be biased so as to preferentially enrich loci bearing proteins that incorporate O-GlcNAc over the time frame of the feeding experiments. In this regard, however, it is notable that the limited distribution of O-GlcNAcylated proteins that we observe within flies resembles the distribution we see in S2 cells. This similarity is significant since progeny, having been raised on medium containing Ac_4GalNAz, would be expected to have newly biosynthesized proteins labeled with O-GlcNAz. Furthermore, though showing limited specificity, the GalT chemoenzymatic labeling method as well as the more specific lectin WGA method both detect native O-GlcNAc, yet they too revealed a similarly limited distribution of O-GlcNAc throughout the genome. These observations collectively suggest, somewhat surprisingly, that O-GlcNAcylated proteins are limited to specific loci including, among others, those possessing PREs.

Recent studies show that *Drosophila* OGT acts as a PcG protein through its ability to modify and stabilize Polyhomeotic (Ph) against aggregation[163]. Though the phenotype of *sxc*$^{-/-}$ embryos resembles that of embryos expressing non-glycosylated Ph[163], it is possible that this effect is epistatic to the effects of aberrant O-GlcNAcylation of other PcG proteins or the loss of non-catalytic functions of OGT. In agreement with this possibility are recent findings showing that *sxc* also acts in

Drosophila development outside of its ability to glycosylate Ph[225]. Consistent with this scenario is that we observe O-GlcNAc at many loci that are not annotated or predicted to contain PREs and do not bear PcG proteins or the H3K27me3 mark. Further, loss of OGT in *sxc*$^{-/-}$ flies leads to altered expression of genes at these non-PRE-containing O-GlcNAcylated loci, which underscores the potential for O-GlcNAc to act independently of Ph and other PcG proteins to regulate gene expression. Collectively, these observations further suggest OGT functions are not restricted to the regulation of genes known to be PcG targets; rather they support the notion that OGT acts through multiple regulatory mechanisms. Of particular interest is the potential role for OGT to regulate the expression of heterochromatic genes, since mechanisms of control of specific heterochromatic genes remains largely undefined. Accordingly, we anticipate that our metabolic ChIP–seq strategy will aid the understanding of potential novel roles of OGT in gene regulation.

In summary, this metabolic feeding approach using highly selective ligation probes to map O-GlcNAc will be a valuable strategy for delineating the roles of OGT and O-GlcNAc in regulating the structure and function of chromatin. This strategy does not rely on genetic engineering of cell lines, as required for chemical mapping of protein methylation within cells[201], and shows high enrichment of DNA at specific loci. Accordingly, we anticipate that this chemical approach will prove useful for researchers interested in the role of O-GlcNAc in the regulation of gene expression, not only in *Drosophila* but also in mammalian tissues. Future studies on the use of this technology for ChIP-seq will likely open the possibility of new experimental modes. Finally, we anticipate that this efficient antibody-free ChIP-seq approach will be amenable to the study of other post-translational modifications found on proteins associated with the genome.

2.4 Experimental methods

2.4.1 S2 cell culture and azido sugar labeling

Drosophila melanogaster Schneider S2 cells were cultured in Sf-900 II SFM medium (Invitrogen) supplemented 100 U/ml Penicillin and 100 mg/ml Streptomycin at 25°C. Cells were passaged at 1:4 ratio every two days to keep logarithmic growth. For labeling, media was aspirated and the cells were washed with PBS. DMSO stocks of Ac_4GlcNAz, Ac_4GalNAz or Ac_4ManNAz were added to achieve the final treatment conditions and incubated for 16-24 h, with vehicle-only controls always included. ~$3.3x10^6$ S2 cells were used as a unit of cells for each experiment. This amount of cells yielded more than enough DNA and protein for downstream analysis and

ensured a large biological sample size. S2 cells were purchased from ATCC, cell line: Schneider's Drosophila Line 2 [D. Mel. (2), SL2] (ATCC® CRL-1963™).

2.4.2 Preparation of nuclear extract and biotin-conjugation reaction

S2 cells were washed twice in ice cold PBS at the time of harvest, then resuspended in solution A (10 mM HEPES pH7.9, 10 mM KCl, 0.1 mM $MgCl_2$, 0.1 mM EDTA, 0.1 mM DTT, 0.5 mM PMSF, Roche protease inhibitor cocktail) and passed 7 times through a 25G syringe. Crude nuclei were pelleted by centrifugation (2,000 rpm, 10 min, 4°C). The crude nuclear pellet was washed in solution A four times and resuspended in solution B (10 mM HEPES pH 7.9, 400 mM NaCl, 1.5 mM $MgCl_2$, 0.1 mM EDTA, 0.1 mM DTT, 0.5 mM PMSF, 5% glycerol, Roche protease inhibitor cocktail) and incubated on ice for 30 min with occasional flicking. After centrifugation at 12,000 rpm, 20 min, 4°C the supernatant was collect as nuclear fraction. Protein was precipitated from the nuclear extract by chloroform/methanol precipitation. Precipitated proteins were pelleted by centrifugation at 14,000 rpm for 5-10 min at RT and washed twice with four volumes of methanol. The supernatant was removed without disturbing the pellet and the pellet was then air dried. The protein pellet was resuspended in PBS containing 1% SDS for a final concentration of 5-10 mg/mL. To specifically biotinylate O-GlcNAz modified proteins, a Staudinger capture reaction using biotinylated phosphine capture reagent (Biotin-*azo*-phosphine) was performed on nuclear protein. For a 100 µL solution consisting of 150-200 µg nuclear protein and 200 µM Biotin-*azo*-phosphine in 1% SDS/PBS was reacted for overnight at room temperature. Unreacted probe was removed by chloroform/methanol precipitation; 20 µg of each biotinylated sample was analyzed by Streptavidin (Strvn) blot using Odyssey (LI-COR Biosciences).

2.4.3 Streptavidin enrichment of azide-modified nuclear proteins and sodium dithionite ($Na_2S_2O_4$) elution

Before each enrichment of O-GlcNAz modified nuclear proteins, 500 µL solution (1% SDS/PBS) consisting of 800-1,000 µg nuclear protein was mixed with 50 µL Streptavidin-agarose slurry (Sigma), incubated for 1 hour at 4°C for preclearing. The supernatant was transferred to a new tube and reacted with 200 µM Biotin-*azo*-phosphine overnight at room temperature. Unreacted probe was removed by chloroform/methanol precipitation. Air-dried protein pellets were resuspended in 6 M urea/2 M thiourea/10 mM HEPES (pH 8.0) and enriched by Streptavidin-agarose slurry (Sigma) for a couple of hours at 4°C with gentle rocking. The beads were sequentially washed three times with 5-10 volumes of 6 M urea/2 M thiourea/10 mM

HEPES (pH 8.0), PBS, and 1% SDS/PBS. Centrifugation of the beads between washing steps was carried out (7,500 rpm, 2 min). Bound proteins were cleaved from the beads by treating with one bead slurry volume of elution buffer (100 mM $Na_2S_2O_4$ in 1% SDS/PBS) for 30 min three times. Eluants from separate tubes were collected and combined. Removal of the majority of $Na_2S_2O_4$ from the eluants was achieved by precipitating them with chloroform/methanol method; 10% of each eluted sample was analyzed by Immunoblot using Odyssey (LI-COR Biosciences).

2.4.4 Antibodies

The following antibody ratios were used for immunoblots: Ph (rabbit serum): 1:600, Pho (rabbit serum): 1:800, HCF (rabbit serum): 1:1000, actin: 1:5000. Actin antibody purchased from Santa Cruz Biotechnology: β-Actin Antibody (C4): sc-47778. Pho antibody was kindly provided by Dr. Judith Kassis and Ph antibody was provided by Dr. Judith Kassis, originally from Dr. Donna Arndt-Jovan[276]. HCF antibody was kindly provided by Dr. Hugh Brock, originally from Dr. Winship Herr[277].

2.4.5 BtOGA digests

For BtOGA treatment experiments, the O-GlcNAcylated proteins precipitated from S2 cells were treated with WT-*Bt*OGA (*Bacteroides thetaiotaomicron Bt*GH84 O-GlcNAcase) or with an inactive variant Mut-*Bt*OGA (final concentrations of 4 M) and incubated in PBS (pH=7.4) overnight at 37°C.

2.4.6 O-GlcNAz modified and vehicle-only chromatin preparation from S2 cells

For each preparation, ~$2x10^7$ Ac_4GalNAz-treated and vehicle-only control cells were fixed for 10 min at room temperature by gently mixing in 10 mL crosslinking solution (1% formaldehyde, 50 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA), crosslinking was quenched by adding 2M glycine to a final concentration of 240 mM. The cells were sequentially collected by centrifugation at 700 x g for 10 min and washed three times with 10 mL ice-cold PBS. These cell pellets were sonicated in parallel to obtain chromatin fragments of ~200 to 700 bp, each in 4 mL sonication buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.5 mM EGTA, Roche protease inhibitor cocktail) with a Sonic Dismembrator Model 500 (Fisher Scientific, 10 x 30 seconds on / 45 seconds off cycles, 50% power settings). After sonication, debris was removed by centrifugation at 4°C for 10 minutes at 13,000 rpm, and equal amount of 6M urea was added to the solution then incubated for 10 minutes at 4°C on a rotating wheel. The soluble chromatin was dialysed using a membrane with a molecular weight cut-off of 3.5 kDa, at 4°C against 2 L dialysis

buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 4% glycerol) overnight. Insoluble debris was removed by quick centrifugation (7,500 rpm, 2 min). The chromatin solution was precleared with 100 µL Streptavidin-agarose slurry (Sigma) or 100 µL protein A/G agarose beads (Calbiochem) for O-GlcNAz modified or vehicle-only chromatin separately and incubated for 1 hour at 4°C on a rotating wheel. The precleared chromatin was aliquoted after centrifugation at 4°C for 2 min at 7500 rpm, snap-frozen in liquid nitrogen and stored at -80°C for future use.

2.4.7 Galactosyltransferase labeling

To label *O*-GlcNAcylated proteins with Ac_4GalNAz, the Click-iT™*O*-GlcNAc Enzymatic Labeling System was used (Invitrogen). Briefly, Gal-$T1^{Y289L}$ was incubated with proteins in labeling buffer (containing 20 mM HEPES, pH 7.9; 50 mM NaCl; 2% NP-40; 5.5 mM $MnCl_2$; 25 µM UDP-GalNAz), according to manufacturer's recommendations. Reaction was performed at 4°C under gentle agitation for 24 h. All reagents were provided in the kit. Once labeling was achieved, proteins were chloroform/methanol precipitated. Note that the volume of each reagent was adjusted for higher proteins quantity, i.e., when 500 µg of proteins were labeled.

2.4.8 ChIP assays

DNA concentration of chromatin solution was determined using the NanoDrop (Thermo Scientific). ~120 µg DNA (corresponding to ~$3.3x10^6$ cells) was used for one immunoprecipitation (IP). DNA enrichment of Ac_4GalNAz fed S2 cells followed by click chemistry or Staudinger ligation was calculated based on a ratio of enriched DNA compared to controls lacking sugar. Details on precipitation methods are listed below. All beads were resuspended in 100 µL TE and incubated with 10 µg /mL RNase A at 37°C for 30 min, then adjusted to 0.5% SDS / 0.15 mg/mL proteinase K and incubated at 55°C for 4 hours to overnight. Cross-links were reversed by incubating overnight at 65°C. DNA was purified QIAquick PCR Purification Kit and stored at -20°C.

2.4.8.1 Ac_4GalNAz and GalT IP

IPs using the independent control and O-GlcNAz chromatin preparations was performed in parallel. The O-GlcNAz chromatin preparation was adjusted to 0.5% SDS-containing condition and reacted with 200 µM biotinylated phosphine probe (Phosphine-PEG3-Biotin, Thermo Scientific) overnight at room temperature on a rotating wheel. Unreacted probe was removed by two rounds of size exclusion chromatography using a Bio-Gel P-10 (Bio-Rad) gravity flow column or dialysed

using a membrane with a molecular weight cut-off of 3.5 kDa, at 4°C against 1 L dialysis buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 2% glycerol) for 4 hours to overnight. Biotinylated probe-chromatin complexes were captured by incubation with 100 µL Streptavidin-agarose slurry (Sigma), at 4°C for 3 h. Beads were washed for 5-10 min at 4°C with 1 mL of the following buffers: 6 washes with low salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8.1, 150 mM NaCl), followed by 3 washes with high salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8.1, 500 mM NaCl), then 2 washes with lithium wash buffer (0.25 M LiCl, 1% NP-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.0*;* then 2 washes with TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA).

2.4.8.2 WGA IP

120 µg of extract were incubated in a final volume of 1 mL IP buffer (15 mM Hepes pH 7.9 / 200 mM KCl / 1.5 mM $MgCl_2$ / 0.2 mM EDTA pH 8 / 0.25% NP-40 / 20% glycerol / 0.3 mM DTT / 1x “Complete” protease inhibitor cocktail / 1 mM PMSF) with 100 µl of a 50% slurry of washed succinylated WGA-agarose resin (Vector Labs) for 12 hours at 4°C. Note: The length of these washes can range from 10 min to overnight. An overnight wash can significantly reduce background signal. The length of washing was optimized by examination of the background present in the control. Succinylated WGA has high affinity for GlcNAc, and a reduced affinity for sialic acids. Beads were washed with IP buffer containing 0.5 mM DTT and 0.4% NP-40, followed by a 1 hour incubation with 1 M GlcNAc (Galab) on ice to elute resin-bound proteins.

2.4.8.3 Pho IP

120 µg of extract were incubated in a final volume of 1 mL IP buffer (15 mM Hepes pH 7.9 / 200 mM KCl / 1.5 mM $MgCl_2$ / 0.2 mM EDTA pH 8 / 0.25% NP-40 / 20% glycerol / 0.3 mM DTT / 1x “Complete” protease inhibitor cocktail / 1 mM PMSF) with 1:100 anti-Pho antibody (kindly provided by Judith Kassis[276] for 12 hours at 4°C. 100 µL protein A/G agarose beads (Calbiochem), previously blocked with 1 mg/mL BSA, was added to each chromatin at 4°C overnight. Washes were then performed as for the Streptavidin precipitates.

2.4.9 ChIP-PCR to determine binding at specific chromosomal locations

These semi-quantitative PCR analyses using PREs primers for screening were performed in 10 µL, using 3 ng DNA per reaction and 12 pmol of each primer. The PCR conditions were: 1 min 94°C initial denaturation, followed by 20 cycles of

94°C for 30 sec, 52°C for 35 sec, 72°C for 30 sec, with a final extension of 2 min at 72°C and storage at 16°C. The following PCR primer pairs were used[153]. Distances (in kilobases) of the middle nucleotides in the amplified regions are given relative to the gene's first transcription start site:

Abd-B (+2.1 kb): forward, 5'-CTGCTGGTACATTTGCACGG-3', reverse, 5'-TCTGTGTCTCTAATGGCTGCG-3';

Abd-B (+72.3 kb): forward, 5'-GGAATACCGCACTGTCGTAGG-3', reverse, 5'-GCAGCCATCATGGATGTGAA-3';

Dll (-1 kb): forward, 5'-CCTAGCCACAAAGCGACATT-3', reverse, 5'-CCCTGCTGAGAGCAGAAACT-3';

en (-0.3 kb): forward, 5'-GTTCACTCCCTCTGCGAGTAG-3', reverse, 5'-GAAAACGCAGATTGAAACGTC-3';

pnr (+3.9 kb): forward, 5'-GAGCAGGGGTGTTGAGACA-3', reverse, 5'-TCTTTCCTTCAGGGACTGTCA-3';

Scr (+0.2 kb): forward, 5'-GAAGTGCGCCACGTTCAAT-3', reverse, 5'-TCCTCTCTCTCGCACTCGTT-3';

tsh (+19.4 kb): forward, 5'-AAGGATTTCCACTTGCAACC-3', reverse, 5'-AGGTCCCAAAACGCAGATACT-3';

Ubx (-29.6 kb): forward, 5'-TAGTCTTATCTGTATCTCGCTCTTA-3', reverse, 5'-CAGAACCAAAGTGCCGATAACTC-3';

Ubx (-29.4 kb): forward, 5'-AAGGCGAAAGAGAGCACCAA-3', reverse, 5'-CGTTTTAAGTGCGACTGAG-3';

dpr12 (-3.2 kb, euchromatic control): forward, 5'-CCGAACATGAGAGATGGAAAA-3', reverse, 5'-AAAGTGCCGACAATGCAGTTA-3';

CG11665 (+12.5 kb, heterochromatic control): forward, 5'-CAGTTGATGGGATGAATTTGG-3', reverse, 5'-TGCCTGTGGTTCTATCCAAAC-3'.

Data was reproducible from at least three independent biological replicates.

2.4.10 Sequencing library preparation and Illumina sequencing

The libraries were prepared according to the manufacturer's instructions using the NEBNext kit (E6200). Briefly, DNA was fragmented by sonication to a maximum of 300 bp. Next the ends of the fragments were repaired with a combination of fill-in reactions and exonuclease activity to produce blunt ends that were then tailed with an A-base. Illumina-specific adaptors were ligated followed by removal of unligated adaptors using AMpure XP beads (Beckman Coulter). Finally a PCR with 12-15 cycles for both kits was performed to enrich final adaptor-ligated fragments. Quality and quantity were assessed on an Agilent Bioanalyzer Chip DNA

7500 or High Sensitivity. For all libraries, sequencing was performed on the MiSeq platform, using v2, 300-cycle reagent kits (Illumina). ChIP-seq reads were aligned against *D. melanogaster* reference dm3 using the Burrows-Wheeler Aligner[130] v.0.7.8 'bwasw' algorithm. BAM-format alignment files were generated with Samtools v.0.1.18. Duplicate read-pairs were removed with Samtools[278]. Homer v1.4[279], MACS v1.4.2[280], and MACS v2.1 peak calling software were used at different parameters to assess software stringency and data quality. Peak calling for downstream analysis was done with MACS v1.4.2[280] using the following parameters: (-g dm -B -S -m 4,50 -p 0.005). The parameters for the –m flag used enabled detection of peaks with a slightly wider range of enrichment, compared to control, than the default setting allows.

2.4.11 Bioinformatics analysis

Unix text manipulation commands and bedtools[281] were used to overlap RefSeq Genes with ChIP-seq peaks. MACS peaks were overlapped with known dm3 RefSeq genes and then merged to remove redundancy of isoforms. The non-redundant peaks were then compared in downstream analyses. GenometriCorr[282] analysis was performed in R version 3.2.1. The S2 ChIP-on-chip and EpiPredictor PRE-containing genes were used as the query and Ac_4GalNAz and Pho peaks as the reference. Differential expression was computed using DEseq2[269] in R version 3.2.1 using PCR duplicate filtered BAM files. BAM file reads from ChIP-seq experiments were compared to each other in 100 bp bins genome wide. For CEAS[283] analysis, version 0.9.9.7 was used. Enrichment of reads at peaks was calculated using Samtools[278] depth command. DAVID[284] tool was used to gather gene ontology information.

2.4.12 External data

Drosophila genome localization of Ph and O-GlcNAc by WGA in larval imaginal disks[153], Pho and dSfmbt in larval imaginal disks[270], H3K27me3 in pupae[271], annotated PREs in Kc cells[268] and in embryos[264], and predicted PREs in S2 cells by ChIP-on-chip[266] and by an *in silico* analysis from EpiPredictor[267].

2.4.13 *Drosophila* stocks and culture conditions

For a description of wild-type (Oregon R) and *sxc*/*Ogt*-null mutants used[163,229]. The *Ogt* null mutations were balanced over a *Cy*, *Tb* balancer to facilitate the selection of trans-heterozygous mutant (Tb^+) pupae. For this study, sxc^6/sxc^7 white null mutant pre-pupae (hereafter referred to as $sxc^{-/-}$ pupae) were generated via a

standard cross. All crosses were repeated at least twice. Crosses were grown at 25°C on cornmeal-molasses-yeast medium supplemented with the mold inhibitors tegosept and propionic acid. Ac_4GalNAz media was made by adding Ac_4GalNAz at 55°C to a final concentration of 100 μM. Parents mated on this food and the progeny grew up on it since birth. Approximately 25 white pre-pupae were harvested for ChIP-seq library prep which was accomplished in a similar process as Ac_4GalNAz fed S2 cells.

2.4.14 qPCR analysis

Total RNA from ~25 wild-type and *sxc*$^{-/-}$ pre-pupae was extracted with TRIzol Plus RNA Purification System (Invitrogen) or PureLink RNA Mini Kit (Ambion) following the manufacturer's instructions. Two micrograms of total RNA from each sample was subject to cDNA synthesis using ThermoScript RT-PCR System (Invitrogen). The Applied Biosystems StepOne was used for qPCR. 15 ng cDNA was subsequently used as a template for qPCR amplification (20 sec 95°C initial denaturation, followed by 40 cycles of 95°C for 3 sec, 55°C for 10 sec, 60°C for 20 sec, with primer sequences as follows:

CG3036: forward, 5'-TCCAACACGGAGAAAGGAGGT-3', reverse, 5'-ATGAAGCCGAGCATGGTCAG-3' (89 bp);

chn: forward, 5'-CCTGATGCAGAACGGTTTCG-3', reverse, 5'-CCAGACCAGATCCGTCCTGA-3' (92 bp);

Dop1R2: forward, 5'-GTTCTCCTTCGCCACGGTTT-3', reverse, 5'-GGCTGGTGATGAAGTAGTTGG-3' (98 bp);

GluRIA: forward, 5'-CTATCATTACCTGCT-3', reverse, 5'-AGTCGACAATCCG-3' (101 bp);

H15: forward, 5'-CTCGTGCAACTGCGACGAT-3', reverse, 5'-TTCATGGAACTTATCCCACAGC-3' (76 bp);

Lar: forward, 5'-GGAGGGTGTGGTTGGATCAG-3', reverse, 5'-AGCCGTCACAGTAATTGGCA-3' (50 bp);

Liprin-gamma: forward, 5'-AAATCTCAGCCCGAATCCATCG-3', reverse, 5'-GTGGAACCTTGACTCGTCGTT-3' (124 bp);

ush: forward, CGCCCAAGTACCCCAAAGT-3', reverse, AGCCAGGAGACGAGAGTCC-3' (104 bp);

2mit: forward, 5'-GACGTAGGACGGCTCAGCTA-3', reverse, 5'-TCTGTTGGTATCGCAATTCCC-3' (94 bp);

pnt: forward, 5'-CAGCGTATATGAGCA-3', reverse, 5'-GTCCAGCAGCAATTC-3' (131 bp);

betaTub56D: forward, 5'-AGTGCTCGATGTTGT-3', reverse, 5'-GGAAATCAGCAGGGT-3' (115 bp).
scro: forward, 5'-CGTACGCTATGGGAACCTTA-3', reverse, 5'-GACCAGCCAGAGATAGGTTA-3' (81 bp)
CG17514: forward, 5'-GCATTCTCCACTATCGCTTTC-3', reverse, 5'-ATCTTCTCGGTCACCAAGTC-3' (93 bp)
CG4017: forward, 5'-GGCGGTTAGAAGGTATCGTA-3', reverse, 5'-GAGGATGGAGTGACAACAGA-3' (80 bp)
PARP: forward, 5'-CTACTTCAGGTTTCGCGATG-3', reverse, 5'-CCGAATGTCAGTAAATCGGC-3' (141 bp)
Set1: forward, 5'-CGAAGCTCGCTCAAACCAGA-3', reverse, 5'-CAGCTATGGGCTCCATTGC-3' (161 bp)

Data are displayed as relative quantification against wild type white pre-pupae and normalized against *betaTub56D*. Error bars show standard deviation from technical replicates in Figure 2.20, d. Data was reproducible from independent three biological replicates prepared at different times using either RNA purification method. The independent biological replicates are shown in Figure 2.22.

Chapter 3: **Software for Time Dependent ChIP-sequencing Analysis (TDCA)**

3.1 Background

In recent years ChIP-seq has become a hallmark strategy to define genomic loci that are bound by particular proteins[121–124]. Genome organization and regulation of gene expression are dynamic processes and enable adaptation to changes in cellular signaling, physiology, and environmental cues. Therefore, there has been increasing interest in understanding the time-dependent changes in binding of proteins to the genome. Such studies depend on quantifying the number of sequencing reads at a given locus as a function of time in a series of parallel experiments. Using such data, changes in the number of sequencing reads at specific loci can be compared to changes at other loci, allowing one to evaluate changes in the abundance of proteins associated with specific genomic loci. Accordingly, such analyses are of increasing interest because uncovering genomic loci that are particularly responsive or impervious to a diverse range of stimuli will enable improved understanding of the mechanistic basis behind dynamic changes within the genome that enable adaptive responses.

Several reports have described time course (TC) ChIP-seq studies performed using a variety of techniques. The current scope of TC experiments have involved metabolic feeding of unnatural amino acids[285], induction of engineered genes bearing epitope tags[286–293], stimulus with known effectors of protein-DNA binding[294,295], induction of DNA cleavage by activation of proteins fused to nucleases[296,297], and examining the repair of DNA damage[298]. The development of novel tools to enable TC ChIP-seq analysis of new targets is an area of growing interest and such methods will facilitate a host of studies that should uncover new mechanisms contributing to the activation and repression of genes.

Although new TC ChIP-seq experimental strategies continue to be developed[215], the strategies for analysis of TC data vary widely. Indeed, there is no standard method for analysis within the field and this stems in part from the lack of software dedicated to such analyses. To our knowledge, there are three publications that offer analysis scripts for TC ChIP-seq data processing, mostly with limited functionality, documentation, applicability and none of these offer modelling options[289,294,296]. Manual analysis strategies are more common. Researchers have

estimated rates of turnover at genomic loci by manually fitting sequencing depth data at each locus over time to an inverse of a negative exponential function[285]. Strategies to calculate sequencing depth at loci in TC ChIP-seq experiments over time using a multi-linear regression has also been explored[290,293]. Other TC ChIP-seq analysis strategies instead focused simply on trends in the depth of sequencing reads over time[296,298]. Strategies involving data fitting are appealing because they enable researchers to reduce large amounts of complex data to a limited set of theoretically important values. Furthermore, using data fitting methods ensures that data at all loci are fit in a consistent manner, increasing the consistency of analyses and avoiding experimenter bias. However, complicating issues can arise when data cannot be fit by the proposed functions or if the model is overly simple. These problems can lead to loss of important information and missing insights that could otherwise be gleaned. Given the decreasing costs of sequencing, coupled with the high value of TC data for understanding physiological responses manifesting within the genome, TC studies are an area of growing interest. Accordingly, simple automated methods that facilitate analysis of such data will facilitate the adoption of TC methods by researchers new to the TC field as well as by non-specialists considering implementing ChIP-seq studies in their own research programs.

Here we describe the development and validation of software that greatly facilitates analysis of a wide range of TC data in a robust automated manner. We call this software the Time-Dependent ChIP-sequencing Analyser (TDCA). TDCA analyzes the sequencing read depth at a series of time points and uses this data to calculate protein binding half-lives at genomic loci by modelling TC sequencing depth to sigmoidal curves. We provide a comprehensive manual containing full algorithm details as well as installation procedures with our software, which is publicly available at: www.github.com/TimeDependentChipSeqAnalyser/TDCA. The following manuscript focuses on describing the accuracy, versatility, and utility of TDCA. We demonstrate the accuracy and versatility of TDCA by testing simulated data sets, as well as by replicating key findings and providing new insights from previously published data sets that were obtained using diverse methods. These data sets include: 1) TC ChIP-seq of doxycycline inducible HA-tagged histone 3.3 (H3.3) variant in MEF cells[290], 2) Chromatin endogenous cleavage followed by sequencing (Chec-seq) of Abf1 in yeast[296], and 3) eXcision repair sequencing (XR-seq) on (6-4)pyrimidine-pyrimidone photoproducts ([6-4]PP) in a normal fibroblast cell line (NHF1) and a DNA damage prone cell line (CS-B) in humans[298]. Data analysis by TDCA yields intuitive parameters that describe behavior at genomic loci and offers

customizable analysis with publication-ready graphical outputs, thus making TDCA of particular value for researchers.

3.2 Results

3.2.1 Strategy

Given that the amount of any specific protein bound to any given genomic locus must have an upper limit to its occupancy, we felt that using an inverse of a negative exponential function for data modelling should accurately reflect the eventual saturation or steady-state occupancy that should occur at loci over time. We also reasoned that protein binding to genomic regions should reach a lower limit defined by either complete vacancy or, in some cases, a low basal level. Finally, we reasoned that many methods applied to TC ChIP-seq, including for example the induction of tagged proteins, will involve a delay in responses that are not accounted for by a simple inverse negative exponential function. To account for this induction period, while incorporating the upper and lower limits of protein binding to the genome, we opted to fit data to sigmoidal curves. Fitting to a sigmoidal curve readily enables the definition of parameters that also define the speed at which occupancy of a given protein changes at any genomic locus. Finally we also considered that such sigmoidal curves may be asymmetric since, for example, induction of protein expression and saturation binding of a locus will vary. To account for such scenarios we introduced a parameter that can account for such behavior. We therefore considered that such sigmoidal fits should yield basic parameters that define the properties of binding of a given protein of interest at any genomic locus. These biologically relevant parametric outputs are reported to users as raw data. This approach accordingly enables users to reduce complex sequencing experiments to a few key features, clarifying research questions and enabling focused data analysis.

To model TC sequencing data, we implement the R package *drc*[299]. Important characteristics of this package include consistent parameter output that are reliably extracted for downstream analysis, a deterministic modelling approach that enables consistent results between analyses, and a hierarchical model structure that enables TDCA to confine parameters based on the data at a particular locus.

3.2.2 Implementation and core algorithm

TDCA models[299] normalized sequencing depth[278] to five parameter (5P) sigmoidal curves, at user specified loci, across multiple ChIP-seq TC experiments. TDCA accepts TC sequencing data in BAM file format and loci coordinates in standard BED file format. Raw sequencing data can be aligned to a reference

genome using a variety of published software[130,300] and converted to BAM files using Samtools[278]. Loci at which precipitated proteins bind DNA at significant levels, or ChIP-seq "peaks", can be defined using published software[279,280] or through custom analysis strategies. The equation and description of parameters for a 5P sigmoid are shown in equation 1 (Eqn. 1). We have also created an option within TDCA to model data to a four parameter sigmoid curve where the asymmetry factor is fixed to 1, should the user wish to exclude any asymmetry correction.

$$f(x) = d + \frac{a-d}{(1+e^{b(x-c)})^f} \qquad \text{Eqn. 1}$$

Where,
a = Lower asymptote (baseline protein binding)
b = Incorporation rate index (IRI, a measure of the slope at the inflection point)
c = Inflection point (time at which the curve reaches the incorporation rate Index)
d = Upper asymptote (maximal protein binding)
f = Asymmetry factor (recruitment influenced either favorably or adversely by previous binding)

During fitting, each locus in a TC ChIP-seq experiment is defined as one of six characteristic TDCA categories of change in sequencing depth as a function of time. These six categories of behavior are defined as follows:

1) Rises: Sequencing depth increases over time and data are modeled to a single 5P sigmoid having a negative incorporation rate index.
2) Falls: Sequencing depth decreases over time and data are modeled to a single 5P sigmoid with a positive incorporation rate index.
3) Hills: Sequencing depth increases and then decreases over time and data are modeled to two 5P sigmoids - a rise then a fall.
4) Valleys: Sequencing depth decreases and then increases over time and data are modeled to two 5P sigmoids - a fall then a rise.
5) Undefined: Loci that do not display the behavior of the previous categories but are nevertheless modeled as either a single rise or fall.
6) Eliminated: Loci that are predicted to behave as a certain category but do not.

We have enabled TDCA to normalize sequencing depth data before modelling. This normalization can be done in two ways. Firstly, the depth values at each locus are normalized by the maximum sequencing depth at non-peak loci for all time points collected in a TC series. Using non-peak loci enables capturing levels of

true background sequencing. Additionally, TDCA can accommodate use of an input standard for normalization of data sets obtained at each time point. 'Input' refers to sequencing data for a control experiment wherein the protein-DNA complexes are not immunoprecipitated by a specific antibody and the sequencing results therefore provide a baseline sequencing depth distribution. If input control data is provided, the input is normalized in the same manner except the sequencing depth across the entire genome is used since there are no expected peaks. Sequencing depth at each time within the input data is then subtracted from experiment data to a lower limit of zero. TDCA has the capacity to handle any number of replicate data sets as well as any amount of input data. Notably, in order to accommodate novel normalization strategies that are emerging[301], we also provide users with the option to normalize data to a defined set of values (see manual for details).

To model data, we have designed TDCA to use a prediction algorithm that is based on the times at which the normalized absolute minimum and maximum sequencing depth values are observed at each locus. TDCA checks if there are either trailing data points (occurring later in time) and leading data points (occurring earlier in time) for the time points containing the absolute minimum and maximum sequencing depth to identify lower and upper asymptote boundaries and to determine if the behavior at a locus is a candidate for modelling using a double sigmoid as seen in "hills" or "valleys" or whether the behavior at the locus is described by a "rise" or "fall" and modeled by a single sigmoid (Figure 3.1, a). We enabled TDCA to use a user-defined "plateau range threshold" and "leading/trailing points threshold", which control the tolerated variation in sequencing depth that can be used to define a lower or upper asymptote boundary. Briefly, the plateau threshold allows users to define the tolerated differences in sequencing depth that is used to determine if the leading and trailing data points are within range to be considered asymptotes (i.e. if the differences are simply fluctuations of data points which have reached a plateau), or if the points are in fact changing meaningfully over time. If the latter is the case, then these data points are defined as genuine leading or trailing time point that permits defining an upper or lower asymptote boundary (for each side of the valley or hill) and corresponding assignment of behavior at a locus to either a hill or valley. The user defined leading/trailing points threshold allows users to define how many genuine leading or trailing data points (as determined by the plateau threshold) are necessary to shift the modelling of a loci from a single to a double sigmoid (Figure 3.1, b). The ability of TDCA to model a single locus to a range of specific categories based on a user-adjustable prediction algorithm allows one to gain important insights from available data. Furthermore, the categories we

have defined are biologically relevant, as shown through description provided below for the TDCA automated analyses of several published data sets.

After the categorization of each locus is completed, TDCA models the data at each locus and the time points used are partitioned according to the category of behavior predicted. If the result does not match the prediction, the locus is eliminated. This procedure provides a two-fold verification of locus behavior that effectively eliminates loci that are false positives. A visual representation of our algorithm is shown (Figure 3.1, a) and the computer system requirements for operation of TDCA as well as a visual of the TDCA modelling process (Figure 3.1, b). We have also optimized TDCA to operate using parallel processor libraries (Figure 3.2).

TDCA provides the results of the modelling as an output file. We have created TDCA to offer various graphical outputs[302], predominantly using the turnover time index (TTI), which is the inflection point obtained from the modeled data adjusted by the asymmetry factor. The TTI is indicative of the binding half-life of a protein at a particular locus and, for this reason, we find it to be a biologically interesting variable on which to focus attention.

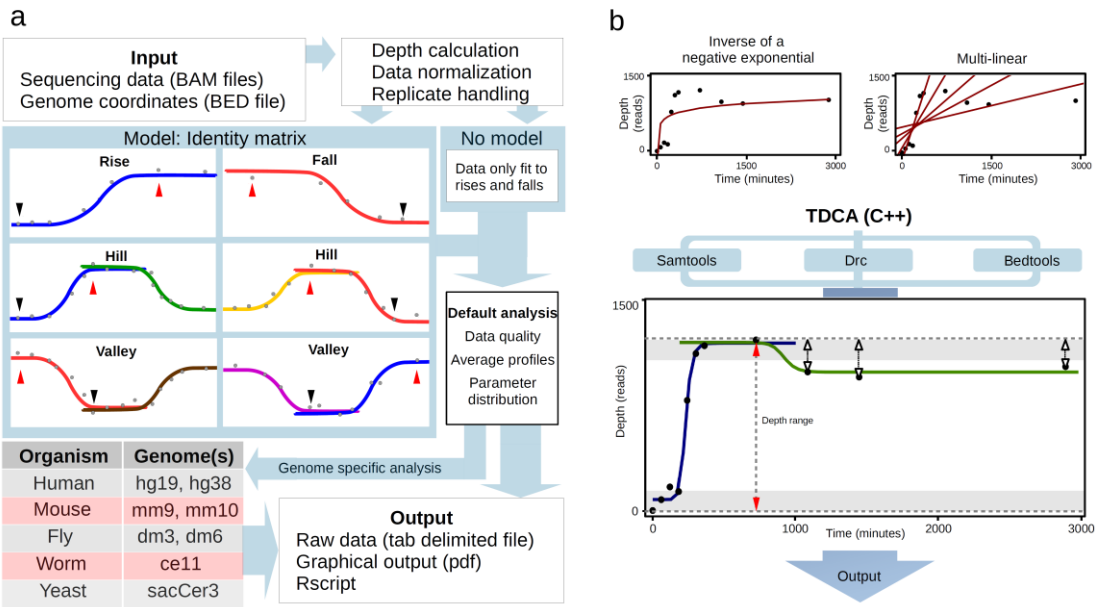


Figure 3.1. **TDCA analysis workflow and requirements.**
(a) Simplified work flow. Required input data are genomic coordinates in BED format and folders containing BAM TC sequence files. TDCA normalizes data based on total sequencing depth of each time point and also handles input files and replicates using additional normalization procedures. Loci can be modeled as the following categories of signal change: rise, fall, hill, or valley. An identity matrix that predicts loci category is based on the time at which absolute minimum sequencing depth (black arrows ▼) and absolute maximum sequencing depth (red arrows ▲) occurs as set by user defined thresholds. Each sigmoid colour indicates a rise or fall with different

combinations of absolute maximum and absolute minimum depth positions in time with genuine leading and trailing points. Alternatively, users can model all their data to a single sigmoidal curve. The resulting parameters from data fitting are then reported to the user along with raw depth calculations. Graphical output is provided to the user which can be enriched by specifying genome and genes. R scripts are provided in case users would like to change the look of default figures. (b) Plots show sequencing depth (y-axis) over time (x-axis) at locus with coordinates of chromosome 1:5012338-5013264 obtained from a H3.3 ChIP-seq experiment[290] using previously applied modelling strategies of inverse negative exponential (upper left) and multi-linear (upper right), and the sigmoidal fitting used by TDCA (lower). TDCA requires on terminal access to Samtools[278] for depth calculation of BAM files, bedtools[281] for BED file manipulations, and R with the drc[299] package for curve fitting. In the example shown here, parameters that govern data modelling by TDCA can be fine-tuned to result in either a single or double sigmoid. The lower and upper horizontal dashed lines represent absolute minimum depth and absolute maximum depth values, respectively. The overall sequencing depth range at a locus is shown as a vertical dashed line with red arrows. In this case, the three data points marked with white arrows exceed the plateau range threshold (gray boxes) and are defined as genuine absolute maximum trailing data points. This results in double sigmoid modelling as shown here. Parameters for both sigmoids are reported to users. The plateau range threshold and leading/trailing threshold could be adjusted such that the locus is modeled to a single sigmoid.

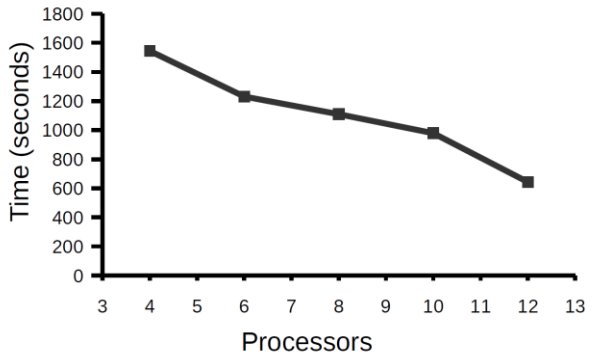


Figure 3.2. **TDCA is optimized to run on parallel processors.**
Processing times required for TC analysis of H3.3 bound loci on chromosome 10 using eleven time points as a function of variable numbers of processors used for computation. TDCA utilizes openmp to parallelize various algorithms in the program which requires an appropriate C++ compiler.

3.2.3 Analysis of simulated ChIP-seq time course data

To test the accuracy of TDCA, we generated simulated TC ChIP-seq data describing both rises and falls (see methods for details). Briefly, we varied different parameters for 1000 loci located on three chromosomes of the *Drosophila* genome. On chromosome 2L we assigned loci to vary in the time of the inflection point, defined as the turnover time index (TTI) and the magnitude of the slope at the TTI, defined as the incorporation rate index (IRI). On chromosome 2R we varied the length of the peaks. On chromosome 3R, we varied the position of the upper asymptote, which defines the depth of sequencing at loci. Calculated values for each

of the 3000 loci were converted into sequencing depth values for 11 different time points[281], and different random noise was added to each time point using standard methods[303]. We provide raw tracks of the simulated data[304] (Figure 3.3, a-c), and summaries of the simulated data (Figure 3.4, a-d). Our simulated data generation method allowed us to generate a constant level of noise which we believed would reflect random noise observed within real experiments (Figure 3.5 and Figure 3.6). We analyzed the simulated data using TDCA and focused on how well it could model the position of the inflection point (TTI), since this is a biologically interesting parameter akin to the half maximal change in the binding of a protein to a particular locus. To perform this study we evaluated the percent difference of the true inflection point based on the simulated calculations with the TTI calculated by TDCA using the TC data augmented with noise.

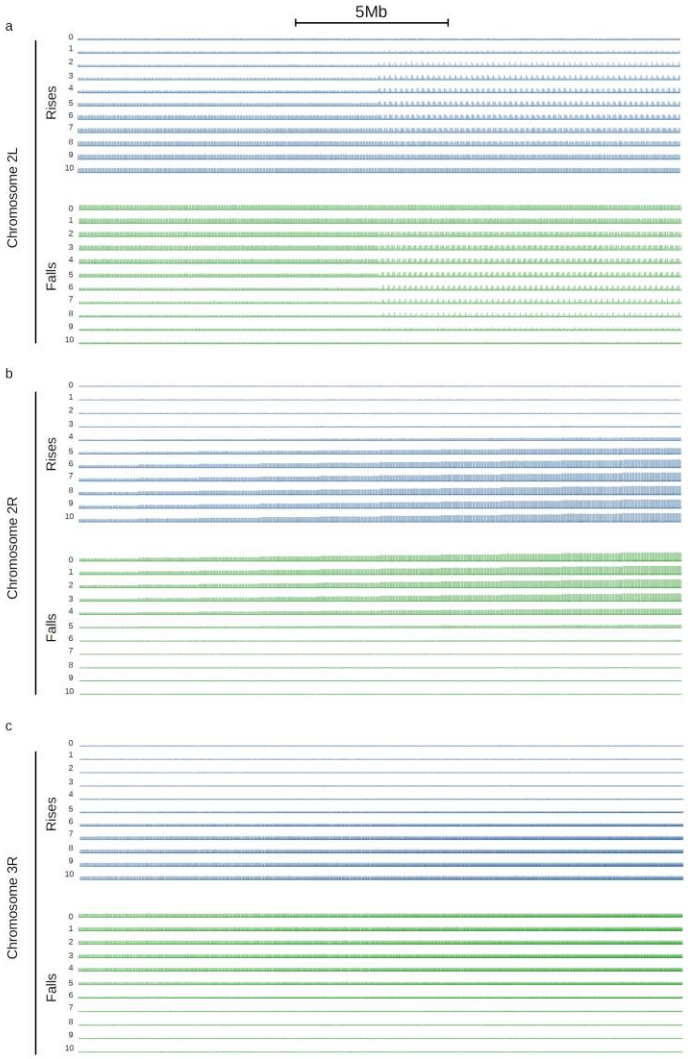


Figure 3.3. UCSC snapshots of simulated data.
Read density of rises (blue) and falls (green) of simulated data for chromosomes 2L (a), 2R (b), and 3R (c). Time of each ChIP-seq experiment

is written to the left of each track (relative units). A total of 11 time points were generate: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Scale bar are written above.

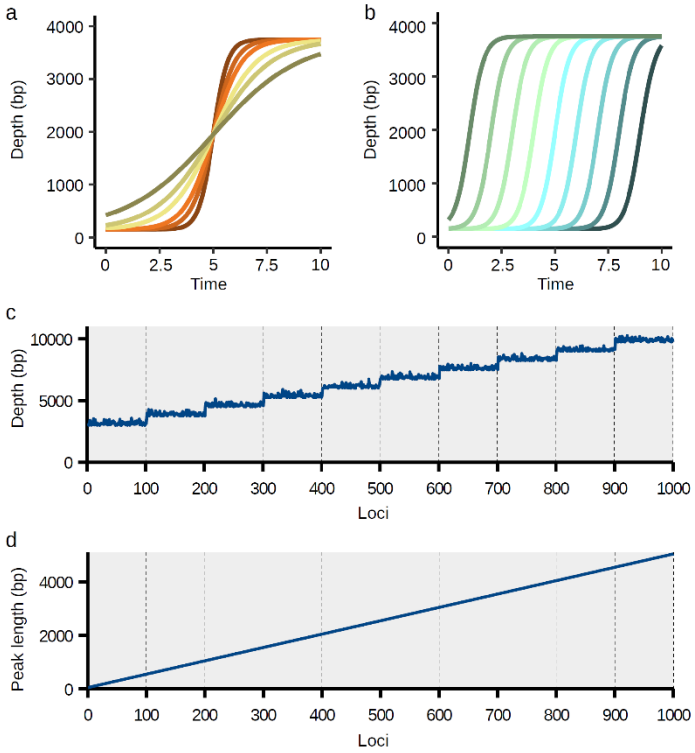


Figure 3.4. **Summary of simulated rise data.**
(a) Chromosome 2L contained 1000 loci with the first 500 loci containing variable incorporation rate index, resulting in data that behave as a sigmoid with steep and mellow slopes at the inflection point. Incorporation rate indices were set to: -0.5, -0.75, -1.0, -1.5, -2.0, and -3.0. (b) The second 500 loci of chromosome 2L contained loci with variable inflection points. Inflection points were set to: 1, 2, 3, 4, 5, 6, 7, 8, and 9 (relative time units). (c) Chromosome 2R contained 1000 loci of variable upper asymptote, shown here across loci. (d) Chromosome 3R contained 1000 loci of variable peak length, shown here across loci. Simulated fall data behaves similarly.

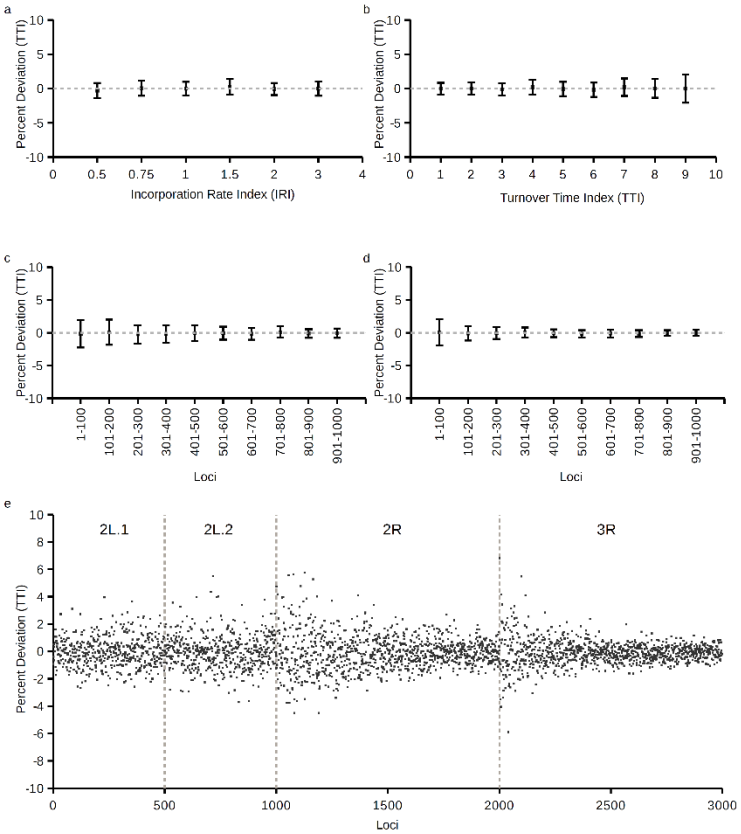


Figure 3.5. Simulated rise data noise analysis.
Noise was measured as a ratio of the sum of depth across each time point for simulated data and the sum of expected depth values across each time point, as a percent of expected depth values. Average and standard deviation is shown for chromosome 2L.1 (a), chromosome 2L.2 (b), chromosome 2R (c), and chromosome 3R (d). Noise for each locus is shown in (e).

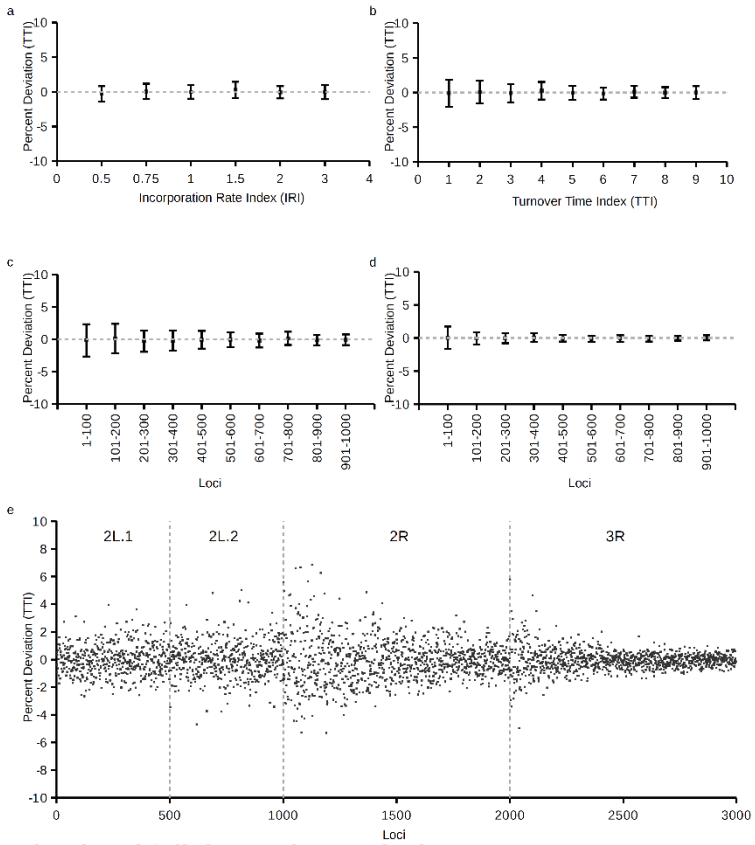


Figure 3.6. Simulated fall data noise analysis.
Same as in Figure 3.5 except for simulated fall data.

Analysis of the 3000 loci with simulated rise and fall data revealed that the TTI modeled by TDCA accurately predicts the true inflection point of the large majority of data (Figure 3.7). TDCA shows increased percent deviation from the true inflection point when data behaves more linearly, with a low absolute incorporation rate index (Figure 3.7, a-b), or when inflection points occur very near the first or last time points for which data is obtained (Figure 3.7, c-d). This behavior is summarized for simulated data describing rises on the first part of chromosome 2L (2L.1), where the incorporation rate index systematically changes across loci (Figure 3.8, a) and on the second part of chromosome 2L (2L.2), where inflection points systematically change across loci (Figure 3.8, b). Interestingly, we also observed more accurate TTI predictions of chromosome 2R loci with higher relative saturation (Figure 3.7, e-f). We reasoned that this behavior arises from the added noise contributing less significantly to data with overall greater sequencing depth, since greater sequencing depth would improve the signal to noise ratio. Therefore, both noise and sequencing depth are important factors to consider in TDCA modelling accuracy. Finally, we found that peak length had no noticeable effect on accuracy of modelling (Figure 3.7, g-h). Based on these analyses, we note that there are important factors to consider in TDCA modelling accuracy, and indeed analysis of TC ChIP-seq data in general,

including the extent of noise, sequencing depth, and the time points collected in the context of expected changes in protein binding to the genome. Regardless, deviation of fitted models to the simulated data sets revealed small (±10%) differences and we therefore consider the overall modelling accuracy of TDCA to be satisfactory.

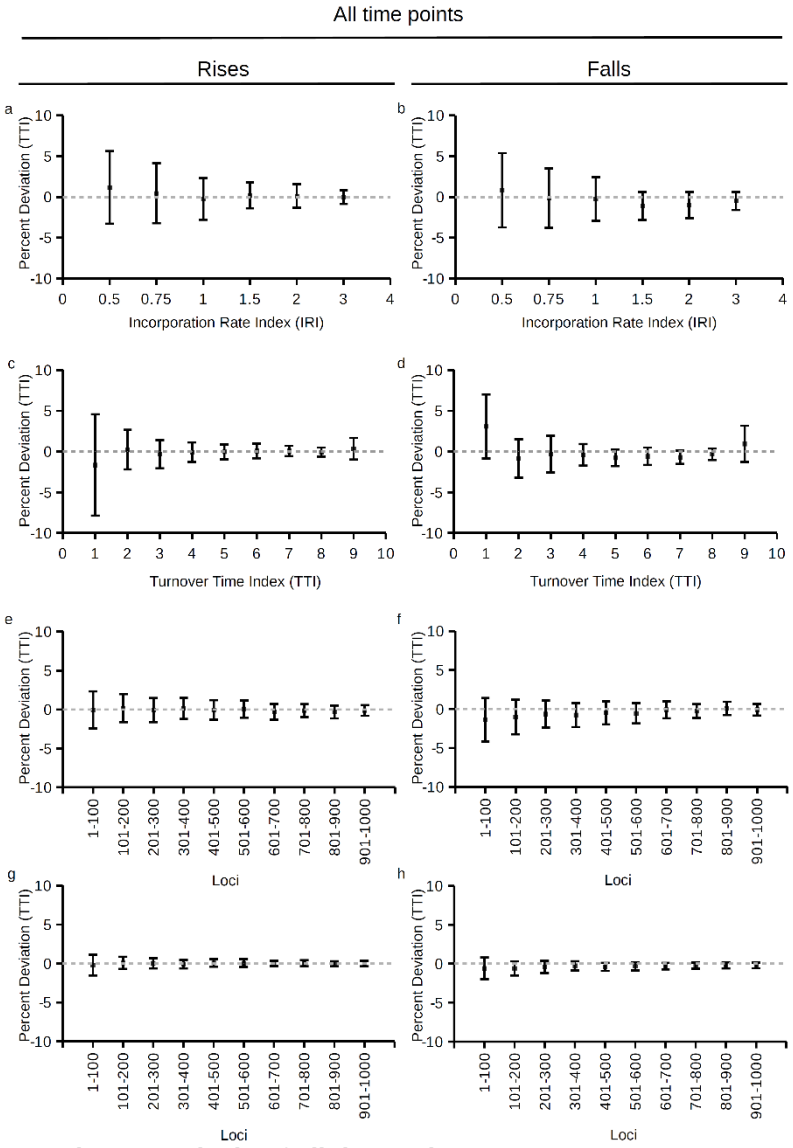


Figure 3.7. **Variance analysis of all time points.**
Percent deviation of TDCA modeled TTI to true TTI for rises (left) and falls (right). Data separated for chromosome 2L.1 (a-b), chromosome 2L.2 (c-d), chromosome 2R (e-f), and chromosome 3R (g-h).

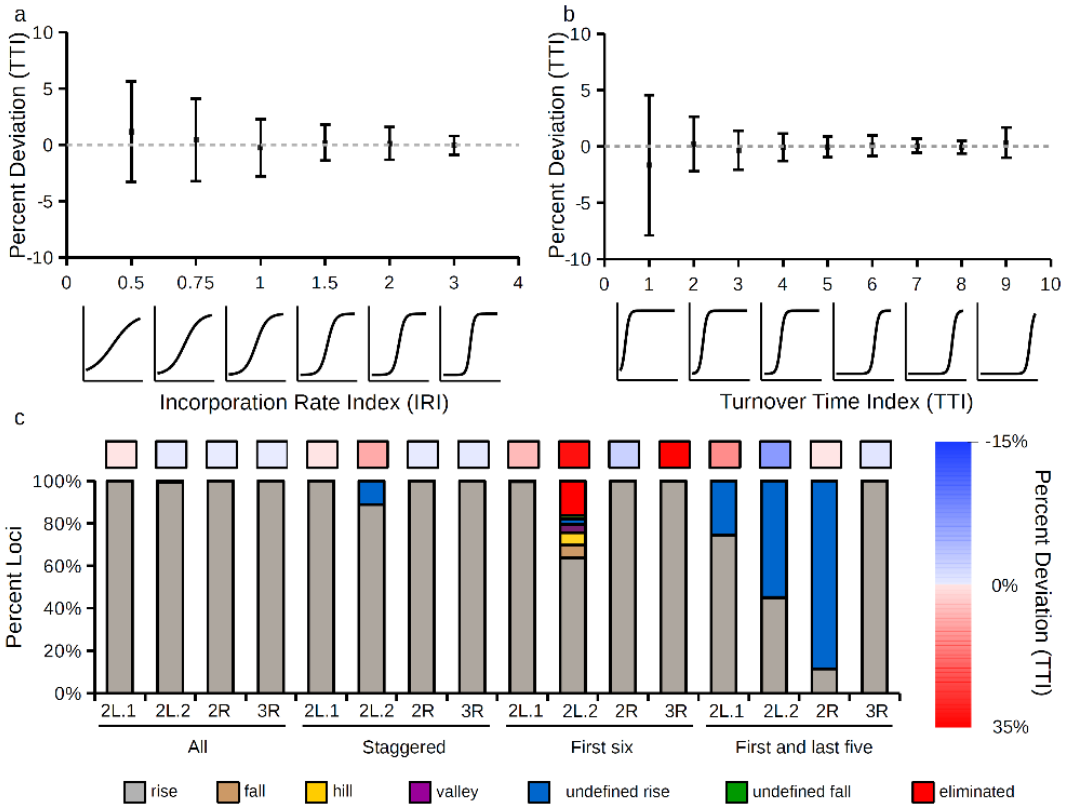


Figure 3.8. **Summary of simulated rise data analysis.**
(a) Percent deviation of TDCA modeled TTI to true TTI of simulated rises on chromosome 2L.1 using all time points binned by the absolute IRI. Representations of true data for different IRI values is shown underneath the deviation plots. (b) Percent deviation of TDCA modeled TTI to true TTI of simulated rises on chromosome 2L.2 using all time points binned by the true TTI value. Representations of true data for different TTI values is shown underneath the deviation plots. (c) Identification of loci categories in simulated rise data using different combinations of time points (all points, staggered points, first six points, and first and last five points). Upper boxes indicate average percent deviation of TDCA modeled TTI to true TTI with a scale shown to the right.

Given the value of having adequate time points to flank the TTI as noted above, we next evaluated how accurately TDCA would model our simulated data sets when only select time points were used. This analysis should provide useful guidance as to how many and at which times one should collect experimental data to realize reliable modelling of data by TDCA. We tested evenly staggered time points (0, 2, 4, 6, 8, and 10), the first six time points (0, 1, 2, 3, 4, and 5), and the first single and last five time points (0, 6, 7, 8, 9, and 10). These tests stem from practical situations that may arise at specific loci, where a researcher may have collected fewer time points (staggered), may have unknowingly ended collection prematurely (first six), or may have missed a block of time points or preferred to collected later data sets (first and last five).

Using these more sparse simulated data sets, we analyzed the percent deviation of the true simulated inflection point to the TTI modeled by TDCA at each locus (Figures 3.9-3.11). We found that the percent deviation was most significant at loci that contained true inflection points that were beyond the last available time point or within gaps of available time points. For example, using staggered time points we noticed that loci on chromosome 2L.2 with inflection points at time point 1 increased in percent deviation (Figure 3.9, c-d). When we modeled data using the first six time points, there was an expected and clear loss in accuracy for loci at chromosome 2L.2 having a TTI at a time greater than time point 5, which was the last time point included in this truncated analysis (Figure 3.10, c-d). Similarly, we noticed during analyses of the data sets containing data for the first time point along with the data for the last five time points, a notable loss in accurate modelling of the TTI at loci having inflection points that occurred within the gap of time points (1-5) (Figure 3.11, c-d). Interestingly, when analyzing the truncated data set containing only the first six time points, there was a larger deviation in accurate modelling of TTI for loci on chromosomes 2R and 3R, with inflection points of 4.5 and 5.5, respectively, in simulated rise data compared to simulated fall data (Figure 3.10, e-h). We reasoned that this effect stemmed from difficulty TDCA had in pinpointing the upper asymptote of rises, whereas those of falls could more easily be determined due to the constraint of requiring placement of the lower asymptote at a non-negative value.

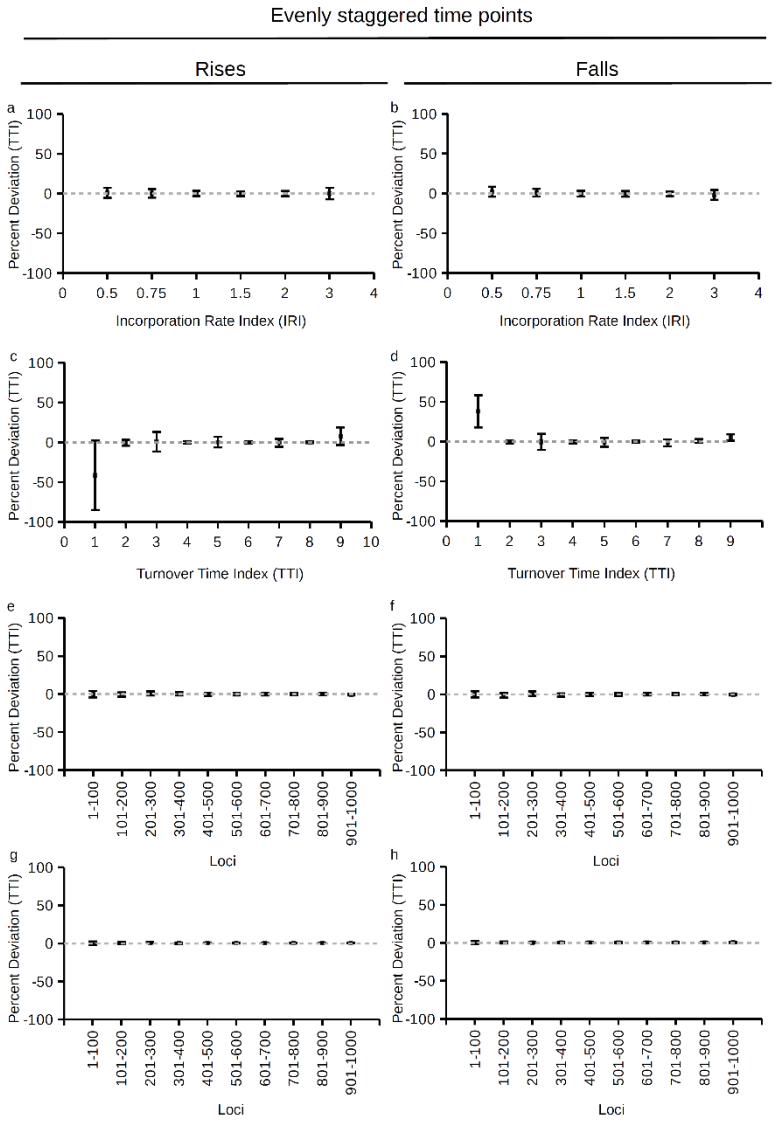


Figure 3.9. **Variance analysis of evenly staggered time points.**
As in Figure 3.7 except for evenly staggered time points.

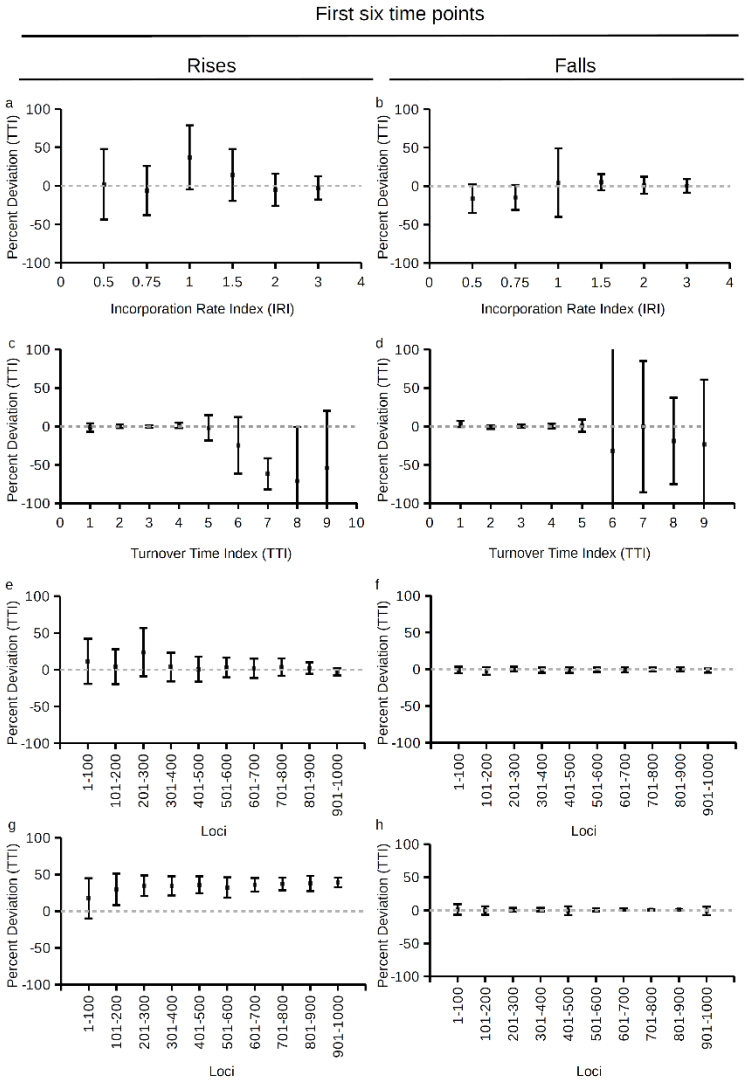


Figure 3.10. Variance analysis of first six time points.
As in Figure 3.7 except for first six time points. Variance for data point with inflection points 8 and 9 in plot (c) are 8: -71.3±70.6 and -54.1±74.3, respectively. Variance for data point with inflection points 6 and 9 in plot (d) are -32.1±333.8 and -23.4±84.3, respectively.

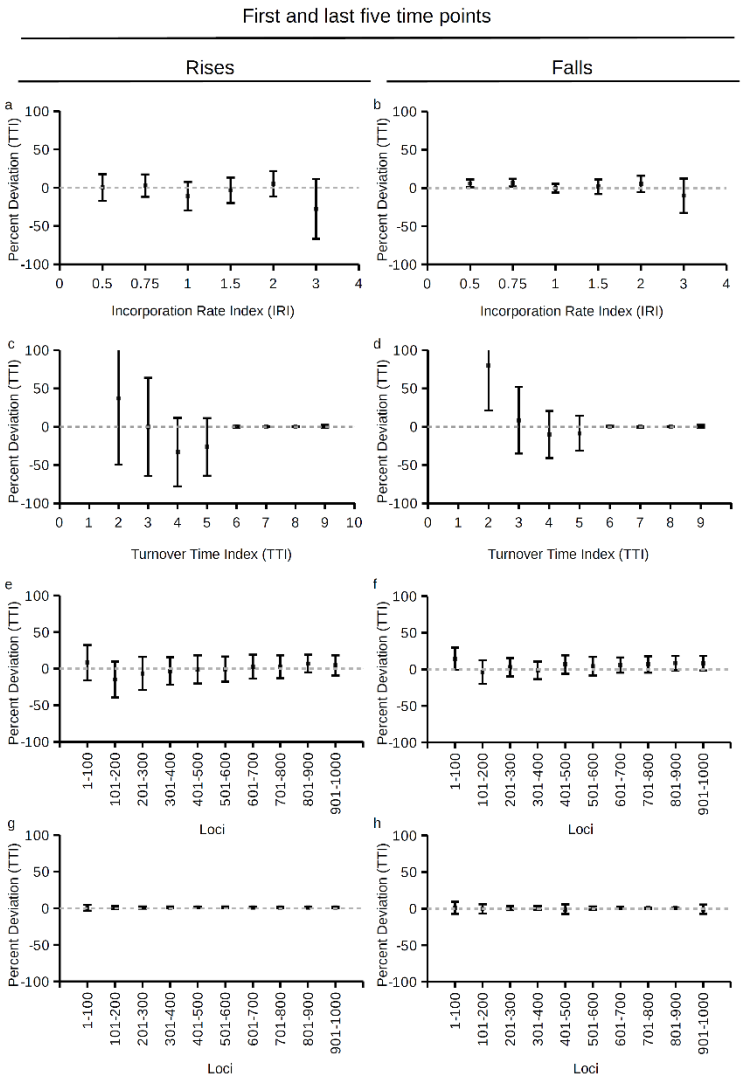


Figure 3.11. Variance analysis of first and last five time points.
As in Figure 3.7 except for first and last five time points. Variance for data point with inflection points 1 and 2 in plot (c) are 159.1±168.4 and 37.2±86.7, respectively. Variance for data point with inflection points 1 and 2 in plot (d) are 264.2±122.0 and 80.0±59.0, respectively.

Given that current recommendations regarding TC sequencing experiments calls for late time points to satisfy saturation of captured loci[305], modelling late TTI values should not be a major problem for researchers so long as this recommendation is followed. In order to circumvent issues in modelling early TTI values, we recommend limited preliminary studies that enable selection of suitable time points chosen to flank the TTI and then perform deeper sequencing studies for TC ChIP-seq experiments and modelling.

In our simulated TC experiments we also describe the accuracy of predictions returned by TDCA with regard to locus categorization for each simulated data set (Figure 3.12). Fundamentally, these results reflect the accuracy of the prediction of inflection points. As shown (Figure 3.8, c), the locus category prediction for simulated

rises is most sporadic at chromosome 2L.2 when using only the first six time points. TDCA has difficulty predicting locus categories when using only the first single time point along with the last five time points. This situation leads to a large occurrence of loci assigned as being undefined, however, the correct category of signal change is predicted (rises and falls are categorized as undefined rises and falls, respectively). Overall, loci that are correctly predicted by TDCA as being in their true category are more likely to be accurately modeled, indicating an important aspect of category predictions that should help guide favorable experimental TC ChIP-seq study design.

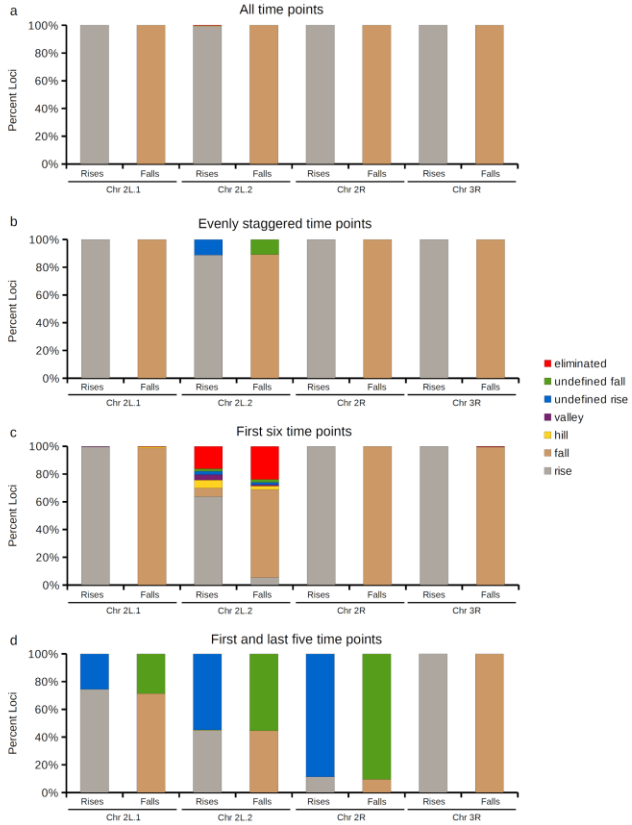


Figure 3.12. Loci type identification in simulated data.
Stacked bar charts of rises and falls separated by chromosomes. Y-axis indicates percent loci. Proportions of loci type are shown for all (a), staggered (b), first six (c), and first and last five (d) time points.

3.2.4 Analysis of inducible HA-tagged histone H3.3 variant in MEF cells

To showcase key features of our program we analyzed a robust TC ChIP-seq experiment performed using an engineered MEF cell line that produces HA-tagged H3.3 variant in the presence of doxycycline in a time dependent manner[290]. This data set contains two independent replicates at each of eleven time points, as well as an input control. We analyzed the replicates separately and found that the log2 TTI ratio of replicates across loci predominantly centered around zero (Figure 3.13, a) with

73.4% of loci within ±20% and 94.4% of loci within ±50% of the reported TTI value (Figure 3.14). This analysis supports good reproducibility of the replicate experiments.

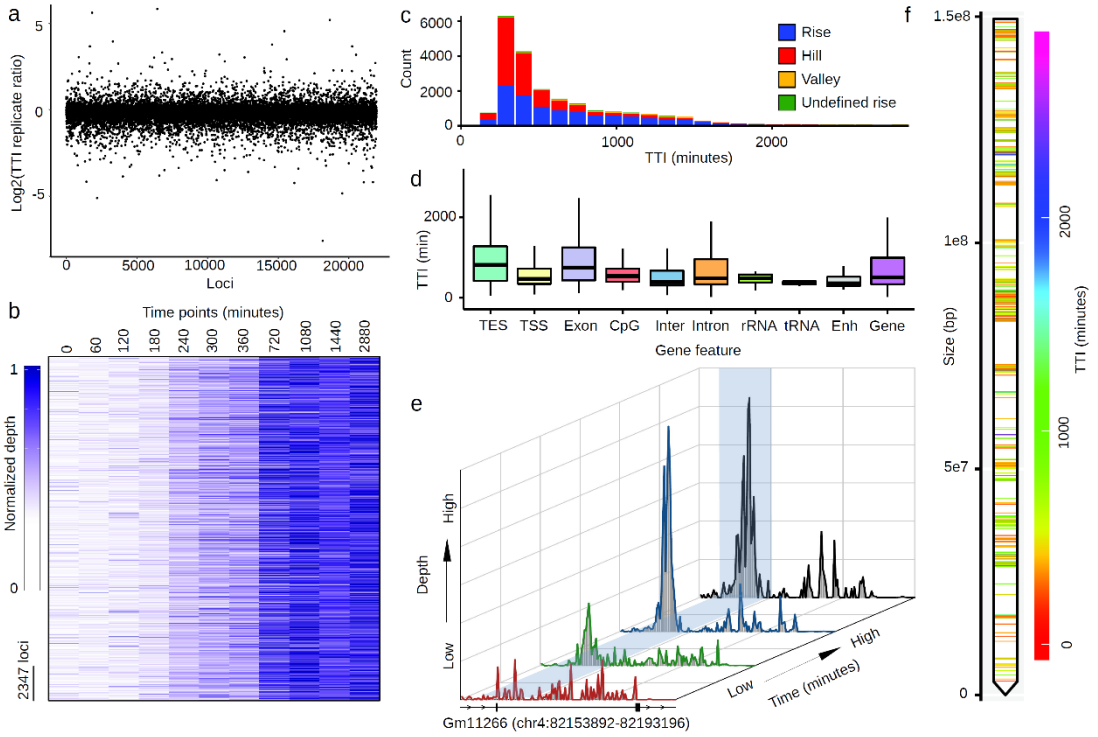


Figure 3.13. HA tagged H3.3 doxycycline inducible TC ChIP-seq analysis in MEF cells.

TDCA analysis of data from reported HA-tagged H3.3 doxycycline inducible TC ChIP-seq experiments performed in MEF cells. (a) Log2 ratio of TTI values from replicates 1 and 2 across each locus. (b) Depth heat map across time points for 23475 loci. Data for each locus is normalized from 0 (absolute minimum depth) to 1 (absolute maximum depth) so that loci can be compared with each other by visual inspection. (c) Distribution of loci that display signal increase are grouped within the defined modelling categories. TTI is shown on the x-axis and locus count on the y-axis. (d) Distribution of TTI values for loci that display increased signal at specific genome features. Lower lines, lower part of box, midline, upper part of box, and upper line are 1st quartile, 2nd quartile, median, 3rd quartile and 4th quartile respectively. The following genomic features are displayed: 3'UTR to 1000 bp downstream (TES), 5'UTR to 1000 bp upstream (TSS), coding exons (Exon), CpG islands (CpG), intergenic regions (Inter), introns (Intron), rRNA genes (rRNA), tRNA genes (tRNA), enhancers (Enh), and whole genes (Gene). (e) 3D plot of sequencing depth for the gene Gm1266 (chr4:82153892-82193196). Black boxes indicate exons, dark lines indicate introns, and lines with arrows indicate 1000 bp upstream and downstream regions. Highlighted region shows the position of two loci with TTI values of 338.9 and 322.3 minutes. (f) Ideogram heat map of chromosome 6. Bands indicate the positions of H3.3 bound loci and the colour scale indicates the TTI value.

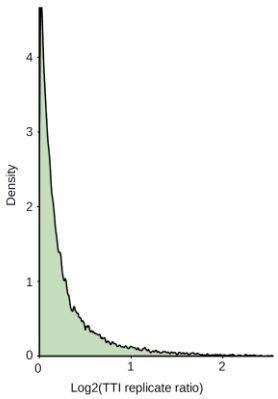


Figure 3.14. Replicate analysis of H3.3 TC data.
Density plot of log2 ratio of replicate 1 TTI/ replicate 2 TTI. 73.4% of loci are within ±20% and 94.4% of loci are within ±50% of the reported TTI value.

We next proceeded to analyze H3.3 loci using both replicates, along with the input control. Included in the default graphical output of TDCA is a genome wide heat map of normalized sequencing depth across time points (Figure 3.13, b). This is a useful chart to visualize the overall quality of data. We observed a general trend of increasing sequencing depth over time (Figure 3.13, b), which is expected as doxycycline treatment leads to a gradual increase of the tagged H3.3 and its recruitment to the genome. Other default graphs generated by TDCA includes a pie chart showing the percentage of loci that are assigned into one of the six TDCA categories of behavior and a bar chart showing the percent incidence of absolute minimum and absolute maximum sequencing depth values over all collected time points (Figure 3.15). We found that the H3.3 TC data contained 49.7% rises and 41.2% hills, accounting for 90.9% of loci. Importantly, the occurrence of decreasing signal after a maximum (defined as a being a hill) was also observed in the original analysis of the data[290], supporting the accuracy of the automated analysis and locus categorization performed by TDCA. We also observed an increased occurrence of absolute minimum depth near the early time points and an increased occurrence of absolute maximum depth at late time points. Overall, the quality charts support the expectation of increased signal over time.

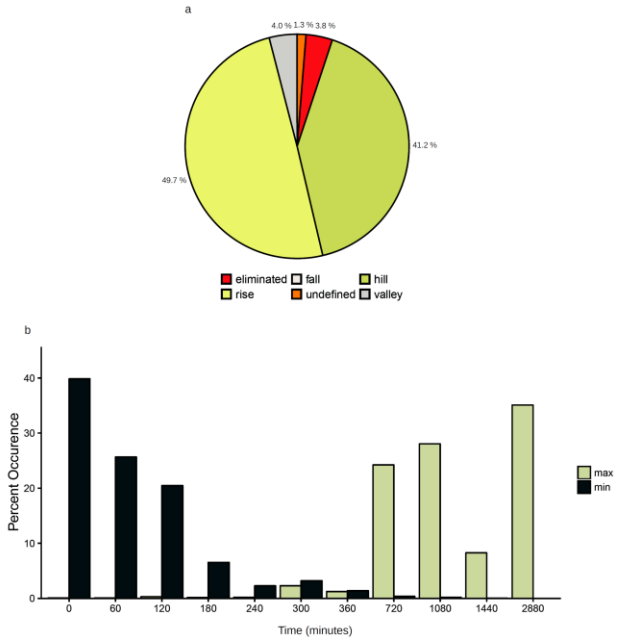


Figure 3.15. Quality analysis of H3.3 TC data.
(a) Pie chart of loci separated by different model types. (b) Bar chart showing percent occurrence of absolute minimum (min) and absolute maximum (max) depth values of all loci.

TDCA offers many default graphs to facilitate data analysis and interpretation. Of particular use is a count of loci that fall within binned TTI regions, which can be separated by the category assigned for a given locus (Figure 3.13, c). During analysis of this H3.3 data set, we observed a right tailed skewed distribution of TTI values centered around 300 minutes. From this observation, we noticed that the distribution of the TTI of the incline of the hills were faster than those of rises. This is an interesting and previously unobserved property of these data that may have functional significance that merits closer study. TDCA also automatically displays average profiles for each category of locus and we illustrate this output showing the relevant categories, hills and rises, for this H3.3 data set (Figure 3.16).

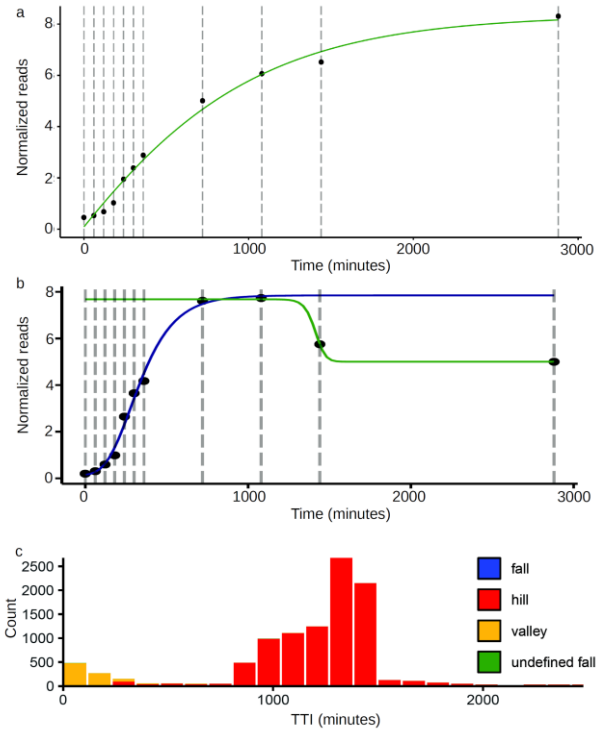


Figure 3.16. Behaviour and genomic distribution of H3.3 TC data.
(a) Average profile of loci that model as rises. Time in minutes is shown on the x-axis. Depth at each time point are shown as black dots. Normalized depth is a ratio of absolute depth averaged across all loci that behave as rises. (b) Average profile of loci that model as hills. The curve that models the incline of hills is shown in blue and the curve that models the decline in green. Time in minutes is shown on the x-axis. Depth at each time point are shown as black dots. Normalized depth is a ratio of absolute depth averaged across all loci that behave as hills. (c) Distribution of loci that display signal decrease grouped as different types. TTI is shown on the x-axis and loci count on the y-axis.

We expanded the customizable built in mouse gene feature library within TDCA to include analysis of loci comprising genes that encode tRNA and rRNA[306], as well loci encompassing enhancers[307] (see manual). These gene features were previously analyzed and found to exhibit unusually fast turnover of H3.3. Here, using TDCA, we rapidly replicated these results in a single automated step and include the distribution of TTI at other default gene features included in TDCA at loci that show an increase in signal change (Figure 3.13, d).

TDCA also provides the useful option of graphing, in a compressed 3D format, the normalized read depth at specific loci. Figure 3.13 (e) shows the 3D profile of the gene Gm11266 (chr4:82153892-82193196), which contains two loci bound by H3.3, which according to the raw data output, have TTI values of 338.9 and 322.3 minutes. As shown, saturation is observed at the last two compressed depth values. Conversely, the 3D profile of the gene Sgk1 (Figure 3.17), which also contains two loci bound by H3.3, does not appear to become saturated with tagged

H3.3. Consultation with the raw data supports this conclusion, revealing TTI values of 1868.4 and 1732.5 minutes for Sgk1. Overall, these 3D profiles are visually informative and provide users with a quick and intuitive way to examine the behavior of genes of particular interest.

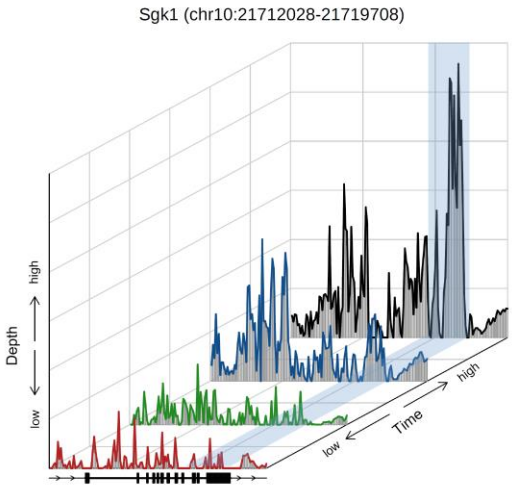


Figure 3.17. 3D plot of sequencing depth for Sgk1.
Black boxes show exons, dark lines introns, and lines with arrows are 1000bp upstream and downstream regions. Highlighted region shows the position of 2 loci with TTI values of 1868.1 and 1732.5 minutes.

Lastly, TDCA provides the distribution of loci to which H3.3 is bound along chromosomes, along with their TTI as an additional dimension shown colourimetrically as illustrated here for chromosome 6 (Figure 3.13, f) and genome wide (Figure 18, a). This ideogram heat map allows users to quickly scan the genome-wide distribution of their loci while simultaneously considering TTI values to decide if clustering analyses, such as the discovery of hotspots describing clusters of fast (low TTI) or slow (high TTI) loci exist within the data set. We binned the mouse genome into 200,000 bp bins and overlapped H3.3 loci at each bin. We found 30 bins that contained 30 or more H3.3 loci, which we defined as being clusters. We then plotted the average TTI and corresponding standard deviation within each of these clusters (Figure 3.18, b). Not surprisingly, since H3.3 shows a relatively bland TTI distribution, we find no drastic differences in TTI averages at clusters after considering the standard deviation. However, some clusters contain much smaller standard deviations than others, which suggests that some clusters are more tightly co-regulated in terms of H3.3 binding or turnover.

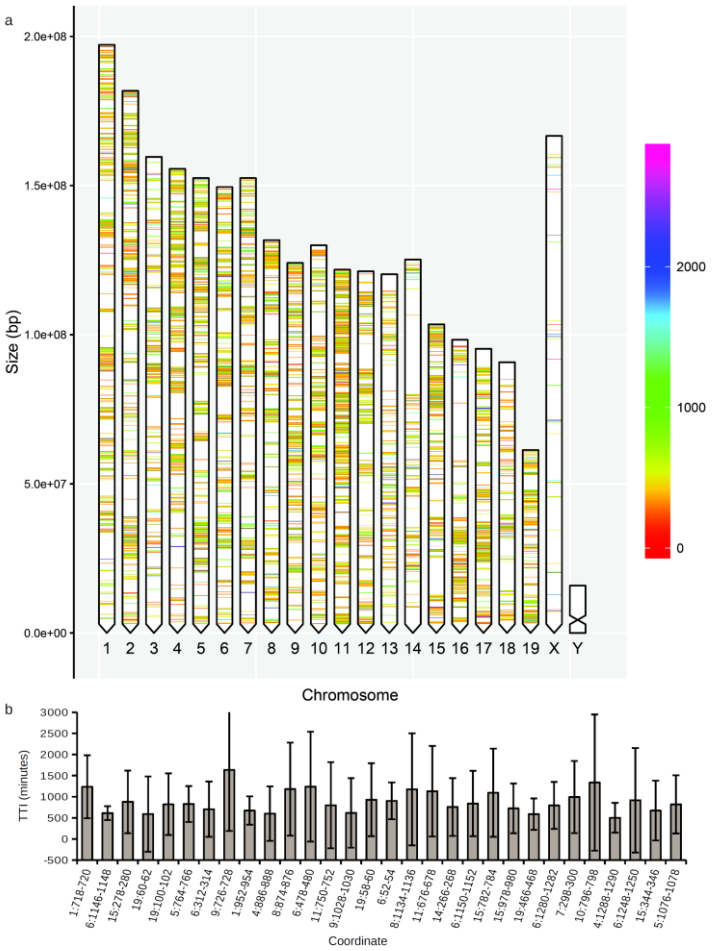


Figure 3.18. Behaviour and genomic distribution of H3.3 TC data.
(a) Ideogram heatmap. Bands indicate where H3.3 is bound and the colour scale indicates the TTI value. (b) Average and standard deviation of TTI (minutes) of 200,000 bp clusters that contain 30 or more loci (clusters). Coordinates are written as chromosome: start (base pairs 10^5) - end (base pairs 10^5).

3.2.5 Analysis of Abf1 time course Chec-seq in yeast

Recently, an interesting ChIP-seq-like technique called Chec-seq, escaping the general requirement of using antibodies for IP and for DNA fragmentation, has been described. This strategy relies on genetically engineered proteins of choice fused to calcium dependent endonucleases. Researchers can study the kinetics of binding of these fusion proteins along the genome by treating cells with calcium at various time points and for varying times. Although not a ChIP-seq experiment *per se*, the resulting data is completely amenable for analysis by TDCA.

We decided to test the performance of TDCA on a published Chec-seq experiment in which an Abf1 fusion protein was used in yeast[296]. This data set contains progressively longer treatments with calcium. This experiment should theoretically result in gradually increasing levels of DNA fragments that in time reach some upper limit, which would result in the TCDA loci categorization of rises to

predominate. However, the authors did note that for some loci, there was an increase in signal over time and then a disappearance, theoretically resulting in the TCDA loci category of hills. Because TDCA can model loci in the same data set as different categories a clear advantage can be gained using this software for automated analysis. We analyzed the Abf1 Chec-seq data set and found that 11715/12351 loci (94.9%) identified as rises or hills which contained positive TTI values on the signal increase modeled sigmoid. This encouraged us to proceed to reproduce key findings in the published data set to prove the accuracy of TDCA, as well as to highlight novel insights gained only through TDCA usage.

Previously, the Abf1 data set was categorized into two major clusters by k-means clustering and these categories were defined as being fast and slow. This categorization was based on whether the time point at which the absolute maximum depth after normalization occurred either early (fast category) or late (slow category). Focus was then directed on analyzing DNA sequence motifs and their abundance at both fast and slow loci. The authors found that fast and slow loci showed a tendency to contain high and low scoring motifs, respectively. Notably, TDCA uncovered a more complex distribution of the kinetic binding patterns of Abf1, as shown in the distribution of TTI values (Figure 3.19, a). When we used k-means clustering[308] to bin the TTI values obtained using TDCA into fast and slow categories we replicated the key observation that there is an increase in the motif scores of fast loci compared to slow. This effect, however, was more modest, and not as great as previously reported based on the time of absolute maximum sequencing depth (Figure 3.19, b). Notably, we also found that the previously clustered fast and slow loci do show an overall lower and higher TTI distribution, respectively (Figure 3.19, c). TDCA is therefore in general agreement with this previous analysis strategy and the reported Abf1 data set.

We next took the clustering based on the time point at which the absolute maximum depth after normalization occurred to its greatest limit by creating the smallest possible clusters. These smallest clusters are simply each time point used. We observed a general trend of increasing motif averages as the bins neared zero (Figure 3.19, d). Binning loci based on the TDCA obtained TTI value corresponding to the time points of calcium treatment did not show as great a trend for average motif scores as previously described (Figure 3.19, e). We reasoned that this apparent difference was due to a large proportion of loci containing TTI values occurring within 1 minute (Figure 3.19, a). We therefore ordered loci based on fastest to slowest TTI values and created bins containing 1000 loci. The average motif scores at these ordered bins re-captured similar average motif scores of clustered

data based on the time point at which the absolute maximum depth occurred (Figure 3.19, f). Strikingly, when we decreased the bin size to 500 loci (Figure 3.20, a), we observed an even greater average motif score at the fastest TTI bin, with local minima and maxima bin clusters. This resolution could not be obtained using the previously published strategy. We show that there are progressively dramatic leaps in the average motif scores as we observe the top 200, 100, 50, and 25 TTI loci. This marked increase in the motif score that stems from narrowing the bin size of the loci having the greatest TTI values highlights the importance of increasing resolution and speaks to the utility and accuracy of the TTI value in analyzing data sets.

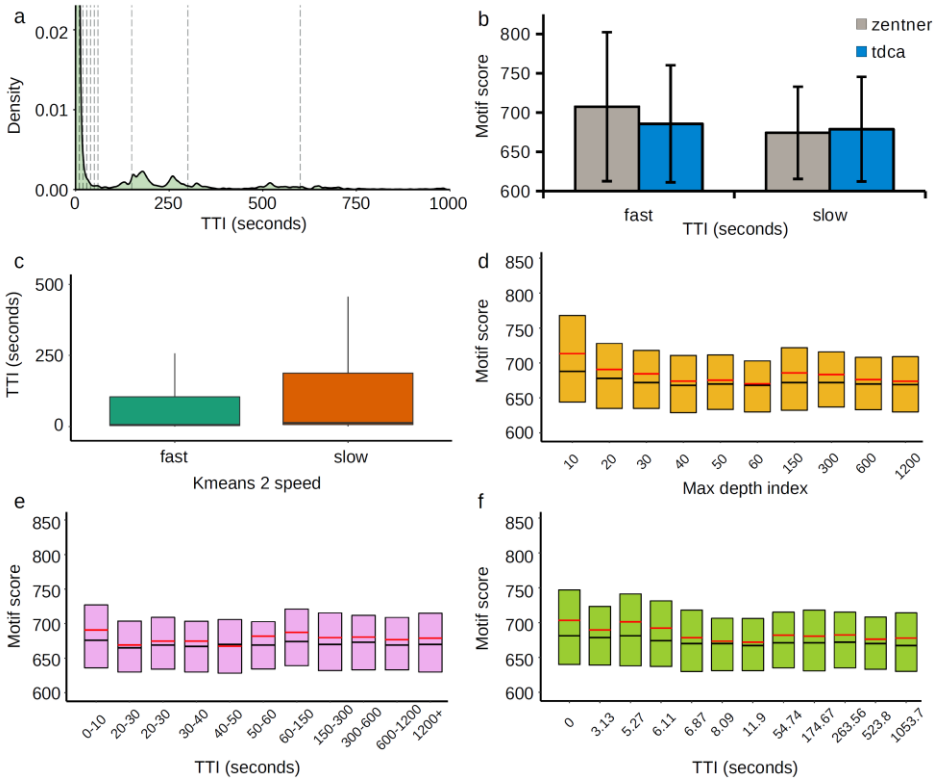


Figure 3.19. Chec-seq analysis of Abf1 hills and rises in yeast.
(a) Distribution of TTI values (seconds) of loci. (b) Average motif scores of loci that were originally categorized as fast and slow by k-means = 2 clustering (zentner)[308] and loci clustered by k-means = 2 by TTI (tdca). (c) Distribution of TTI values in the originally categorized fast and slow loci. (d) Distribution of motif scores of loci categorized by time point (seconds - x-axis) at which the absolute maximum depth occurs. Black midline indicates median and red midline indicates average. Quartiles 2 and 3 are lower and upper fractions of the box divided by the median. (e) Distribution of motif scores of loci binned by the time point used in the TC experiment. Black midline indicates median and red midline indicates average. Quartiles 2 and 3 are lower and upper fractions of the box divided by the median. (f) Distribution of motif score of loci ordered from fastest to slowest TTI and binned into groups of 1000. TTI in seconds is shown on the x-axis. Black midline indicates median and red midline indicates average. Quartiles 2 and 3 are lower and upper fractions of the box divided by the median.

Lastly, we ordered all loci based on their TTI from fastest to slowest and created bins of 1000 loci for which we then produced motifs (Figure 3.21). We were able to reproduce specific motifs[309] at loci having early TTI values (Figure 3.20, c), which eventually reduced to poly-A repeats, as noted in the initial report[296]. Because of our increased resolution, we also captured additional motifs that were not previously observed (Figure 20, d-e). Interested researchers would easily be able to pursue this type of discovery using the high level of automation and customizability offered by TDCA.

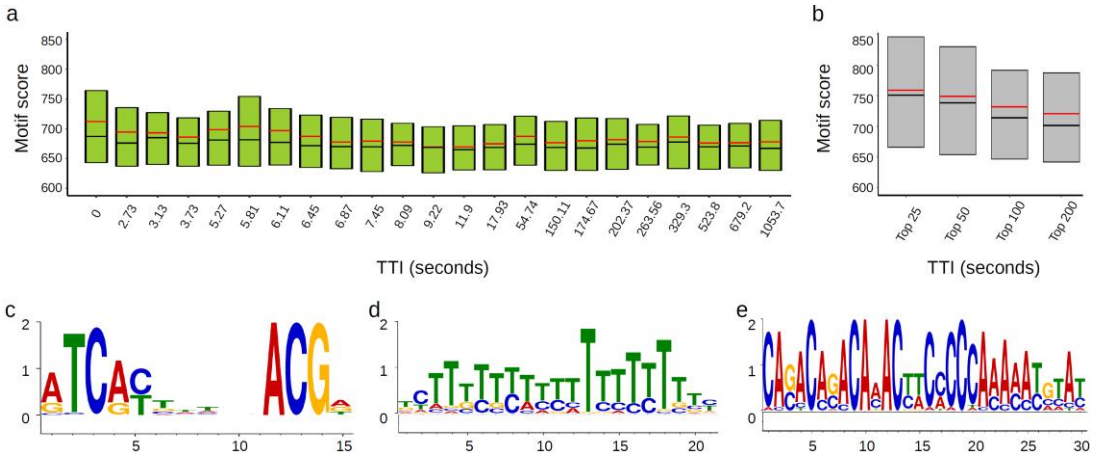


Figure 3.20. Summary of Chec-seq analysis of Abf1 hills and rises in yeast.
(a) Distribution of motif score of Abf1 Chec-seq loci that are ordered from fastest to slowest based on their TTI values and ranked into bins containing groups of 500 loci. The lower boundary for each bin is defined by the lowest TTI value in seconds and is shown on the x-axis. Black midline indicates median and red midline indicates average. Quartiles 2 and 3 are lower and upper fractions of the box divided by the median. (b) Average motif score of Abf1 Chec-seq loci ordered by TTI of the top 25, 50, 100, and 200 fastest TTI loci. Top scoring motifs (most significant) of the first (c), second (d), and third (e) bins of 1000 Abf1 Chec-seq loci ordered by their TTI values.

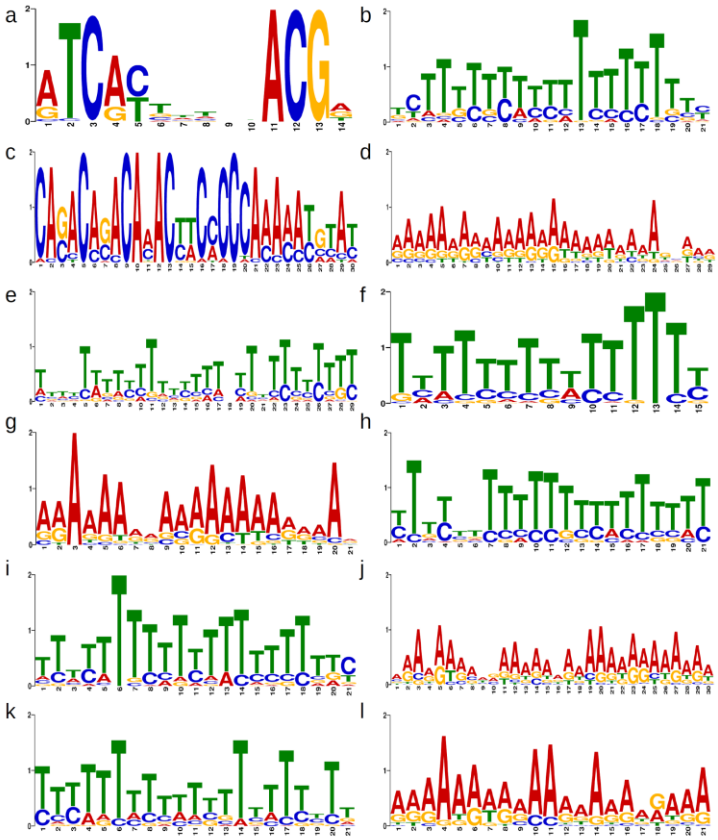


Figure 3.21. Abf1 Chec-seq motifs.
Top scoring motifs (most significant) of loci ordered from fastest to slowest TTI and binned into groups of 1000 (a-l).

3.2.6 Analysis of time course XR-seq on [6-4]PP in NHF1 and CS-B human cells

In humans, UV damaged DNA is removed through the action of the nucleotide excision repair pathway[310]. By monitoring DNA repair following UV treatment in a TC XR-seq experiment it has been shown that the time at which excision occurs after UV exposure varies depending on the locus and that excised fragments, which can be identified and quantified by sequencing, degrade over time[298]. This observation suggests that resulting TC sequencing data analyzed by TDCA should categorize predominantly as both rises and hills, depending on the rate of degradation of excised DNA fragments. We used MACS[280] to determine loci containing excised [6-4]PP, using the longest time point (240 minutes) and the shortest time point (5 minutes) as the signal and baseline, respectively. We viewed this process as leading to the identification of loci that release excision products at a relatively late time. Accordingly, we found that 96.2% (7565/7860) of NHF1 and 97.2% (5121/5268) CS-B loci are identified as rises.

To showcase the plateau range threshold option of TDCA we described previously, we performed an analysis of [6-4]PP loci using a range of plateau range

thresholds. As expected, we found there to be a modest but consistent increase in the number of loci that were categorized as rises as the plateau range threshold became looser (Figure 3.22, a), for both NHF1 and CS-B cell lines. We also used TDCA for analysis with input files containing sets of loci that had been called by MACS using different p-value thresholds[280]. While holding the plateau range threshold at a constant value and specifying more stringent MACS p-values, there was a general increase in the percent of loci that identified as rises (Figure 3.22, b). This is meaningful since the loci called at lower p-values should be more accurate.

In order to show that peaks called using the longest time point (240 minutes) and the shortest time point (5 minutes) as the signal and baseline, respectively, specifically result in rises, we analyzed three different randomly permuted[281] coordinates of [6-4]PP loci in NHF1 and CS-B cells while keeping the depth normalization constant (see methods). We found that the identity of these random loci were not specifically enriched in rises (Figure 3.22, c). This type of analysis is important to help demonstrate the specificity of behavior at loci having user defined coordinates.

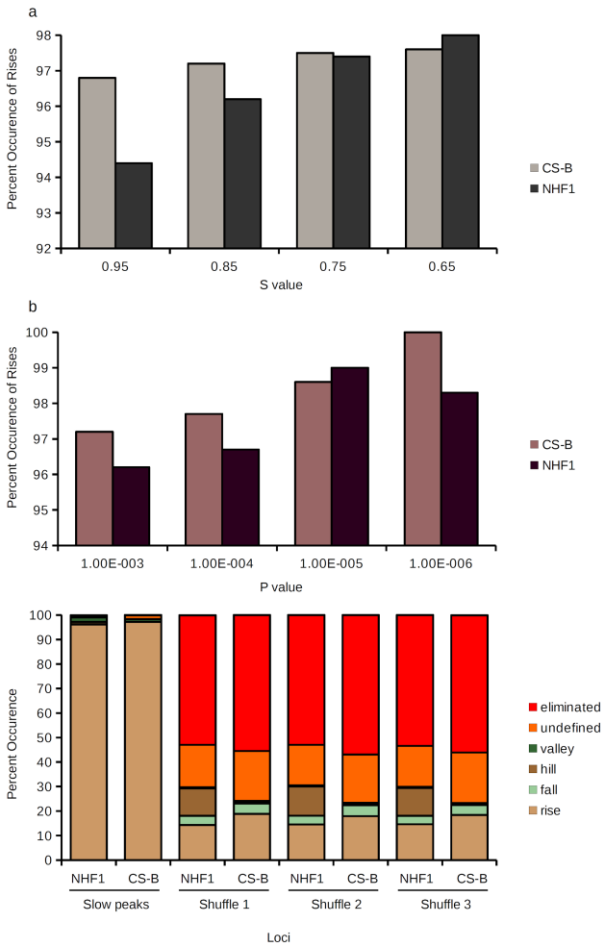


Figure 3.22. Analysis of slow [6-4]PP XR-seq loci.
(a) Percent occurrence of rises as the TDCA plateau range threshold (s-value) is altered. (b) Percent occurrence of rises as the MACS peak calling p-value is altered. (c) Category of loci modeled using constant normalization values (see -dm flag in manual) in slow peaks and loci that were generated by shuffling the positions of the slow peaks in three random permutations (shuffle 1 to 3).

To demonstrate the behavior of all [6-4]PP loci we analyzed the top 1% 500 bp bins in NHF1 and CS-B that showed the greatest change in sequencing depth over time for chromosomes 21 and 22, as done previously[298]. We found that 48.8% NHF1 and 50.6% CS-B loci identified as rises, suggesting that excision products from DNA resulting from excision at some loci may persist for longer than others, which are presumably degraded more quickly by nucleases. Notably, we also found that 24.7% NHF1 and 28.0% CS-B loci identified as falls, although this is likely an artifact since the first time point at which sequencing was performed was only 5 minutes after UV exposure. Alternatively, however, falls may represent exceptionally quickly excised loci and may be of functional significance.

To showcase TDCA, we decided to focus our analysis on the hills present in the NHF1 cell line. 13.6% (271/1990) of loci from this data set were categorized by

TDCA as hills, which is a much greater fraction than that found in CS-B cells (41/1990 = 2.5%). Plotting the difference in the TTI values for the declines and inclines of each hill in NHF1 cells revealed an average difference and standard deviation of 83.2±19.9 minutes (Figure 3.23, a). We find this is a reasonably tight time range and we hypothesize that the clearance of excision products at loci that identified as hills occurred within a certain limited time frame.

Next, we wanted to determine if the TTI values of the hill inclines and hill declines were correlated in some manner. At each locus, the inclines and declines seemed to cluster by visual inspection, which was corroborated by k means clustering[308] (Figure 3.23, b). We plotted a hill that had a relatively fast incline (TTI_{rise}) and decline (TTI_{fall}) as defined by the TTI values for each fitted sigmoid (Figure 3.23, c). We also plotted a hill with a relatively slow incline (TTI_{rise}) and decline (TTI_{fall}) as defined by the TTI values for each fitted sigmoid (Figure 3.23, d). These loci share similar clearance rates of excised product, as defined by the difference of fall and rise TTI of hills (TTI_{fall}-TTI_{rise}) yet excision appears to start at different times. This observation could potentially direct researchers to identify the molecular basis for why loci within this data set cluster in this way and why some loci seem to show delayed excision. Notably, TDCA greatly enabled these analyses in an automated and intuitive manner and we anticipate TDCA can similarly be applied to facilitate analysis of a wide variety of experimental TC sequencing data studies

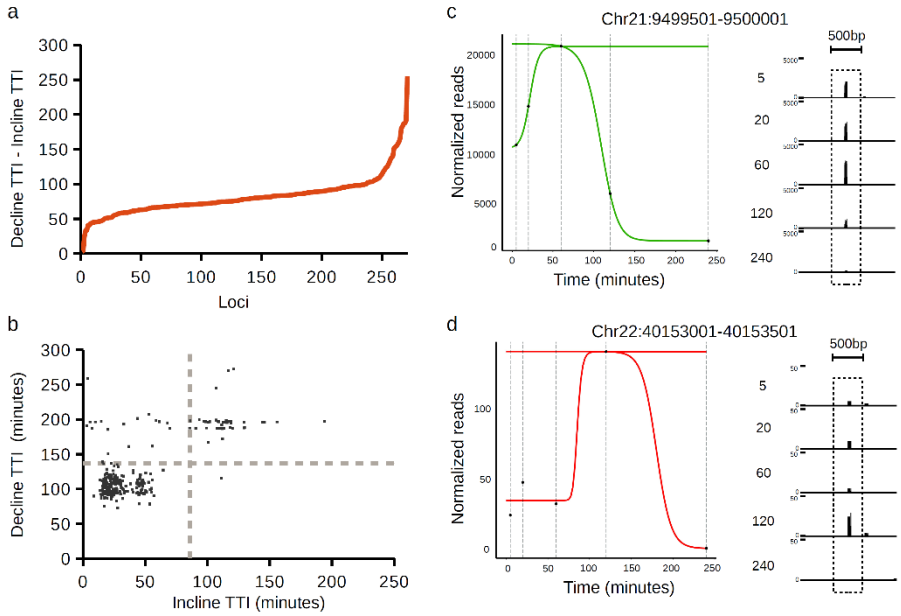


Figure 3.23. TC XR-seq analysis in NHF1 cells.
(a) 271 hills from chromosomes 21 and 22 ordered by the difference of their TTI_{fall} and TTI_{rise}. This difference (TTI_{fall} - TTI_{rise}) reflects the clearance rate of DNA excised from these loci. (b) Scatter plot of the signal increase (incline) TTI (TTI_{rise}) and signal decrease (decline) TTI (TTI_{fall}) of 271 hills from

chromosomes 21 and 22. Two clusters are apparent from this analysis that indicating either slow (top right quadrant) or fast (bottom left quadrant) excision of DNA at loci. (c) TDCA modelling of the locus at chromosome 21:9499501-9500001. Individual depth values are shown as black dots and can be visually seen in the sequencing tracks of normalized read density shown to the right. (d) As in (c) except for the locus at chromosome 22:40153001-40153501.

3.3 Discussion and conclusions

We describe a novel algorithm that we have developed called TDCA, which models changes in sequencing depth of individual loci within time course (TC) ChIP-seq, or conceptually related experiments, as a function of time. The behaviors of these changes are categorized as rises, falls, hills, and valleys. Such sigmoidal modelling of TC ChIP-seq data has, to our knowledge, not been performed. We believe such modelling of TC ChIP-seq data has a reasonable basis in underlying biological principles and that the outputs obtained from TDCA provide intuitively relevant parameters. Our analysis of three published and publicly available data sets support this view, illustrating that rises and hills are biologically meaningful ways to describe behavior of TC ChIP-seq data. Although the data sets explored here differ widely in terms of the experimental means used to obtain them, the outputs provided by TDCA proved in all cases to be readily applicable. Indeed, our automated analyses of these data sets reveal the speed by which TDCA can be applied to obtain insights that are of potential biological importance. Using TDCA, we rapidly recapitulated key findings from these data sets as well as being able to detect previously unobserved behaviors of potential biological significance. Notably, other published data sets also support falls as a biologically meaningful behavior[293] and hills have been demonstrated in other types of experiments[311]. We anticipate that valleys may be observed in the future as the scope of such TC sequencing studies grow and new experimental designs are applied.

TDCA offers many customizable options, such as the ability to tune modelling parameters, include genome specific analyses, and specify normalization constants. These properties confer on TDCA considerable versatility and should permit its use in analysis of nearly any type of TC ChIP-seq study or conceptually related TC sequencing experiments. TDCA is also amenable for analysis of developmental ChIP-seq studies. Considering the available well-documented protocols[312] and the apparent global changes found in developmental ChIP-seq studies[271], this should be a straightforward application for TDCA, and provide useful insights into such experiments. Furthermore, we recognize that investigators could readily use TDCA to model dose-response ChIP-seq experiments. Treating cells with inhibitors of

chromatin associated proteins for different time periods or in a dose-dependent manner followed by ChIP-seq of the inhibited proteins or their chromatin marks could, for example, distinguish between inhibitor sensitive and insensitive loci. Furthermore, novel advances in mapping genome associated small molecules[313] applied in dose-response ChIP-seq experiments could also complement or provide new insights into dose-responsive behaviour of genomic loci. TDCA could be applied to both time- and dose-responsive ChIP-seq strategies with only minor adjustments to the labeling of the X-axis.

TDCA does have some limitations with regard to the type of behavior expected in a time-course experiment. Any kind of behavior that does not fall into a typical fall, rise, hill or valley model category may not be accurately fitted. For instance, in a TC experiment where multiple hills or valleys occur within TC data for one locus, TDCA would not be able to properly model such oscillating data. Another scenario that would present problems would be a rise that reaches a plateau that is followed by another rise; this step-like behavior would be modeled as a single rise. The analogous step fall-plateau-fall behavior would also be poorly modeled. Though these scenarios are currently unknown and seem improbably, users will ultimately need to consider if their experimental design is suitable to analysis by TDCA.

To guide users in experimental design and to highlight the strength and limitations of TDCA, we have also analyzed simulated data. These analyses provide guidance for the number of time resolved data sets needed for successful analysis. They also help define which distributions of time resolved data points are beneficial for proper and reliable modelling. Accordingly, the results of these efforts, along with previously recommended protocols[305] should be considered when designing TC ChIP-seq experiments or other conceptually related TC studies.

The default analysis produced by TDCA and custom analysis using the TTI output described here are intended to stimulate activity in the field of TC ChIP-seq by providing computation support for analysis of resulting TC data. This work is also intended to provide conceptual impetus to more deeply consider analysis strategies and enable rapid exploration of various analysis parameters so as to enable the community to glean as many insights as are available within the growing stream of data being generated within this important and rapidly developing field.

To stimulate further research in the area of time course (TC) ChIP-seq experiments, we have developed the first robust automated tool for analysis of such data. TDCA accepts sequence alignment data in BAM file format and loci in BED format. The graphical and raw data output provided by TDCA provides users with biologically relevant data and will facilitate research as well as inspire future effort in

this growing area of study. While we have described the use of TDCA in the context of ChIP-seq experiments to monitor protein binding to the genome, the term protein is used throughout for simplicity and analysis by TDCA is applicable to analysis of any molecular feature associated with the genome that can be detected in a selective manner. Moreover, we show that TDCA has the potential to be applied to many TC sequencing experiments. Accordingly, as strategies applicable to TC sequencing studies develop, existing strategies improve, and costs of sequencing continue to decline, we expect TDCA will prove broadly useful for a wide range of new experiments as well as providing a benchmark system to help guide optimization of data collection.

3.4 Experimental methods

3.4.1 TDCA Design and dependencies

Samtools[278] is required for depth calculation of BAM files. The Samtools depth command is used for this and is called to the terminal within TDCA. The bedtools intersect[281] command is used for the genome specific analysis. User defined peak coordinates are intersected with genome feature BED files and the TTI values are reported as a boxplot. TDCA uses the R package dose-response curve (drc)[299] for data modelling to a sigmoidal curve.

The generation of graphs requires the following R packages: ggplot2[302], scales, and grid. In addition, plot3D and rgl are required for the construction of the 3D scatterplot of user specified genes when the -3d flag is called.

Much of TDCA is parallelized including commands called by TDCA such as bedtools, Samtools and drc. We used openmp for this which requires an appropriate compiler (see www.openmp.org).

3.4.2 Simulated data generation

We assigned 1000 loci to three chromosomes in the *Drosophila* genome, 2L, 2R, and 3R. The inflection points of loci on the first half of chromosome 2L (2L.1) were fixed at 5 with variable incorporation rate indices (Hill coefficients) of: -0.5, -0.75, -1.0, -1.5, -2.0, and -3.0 for rises and the corresponding absolute values for falls. The inflection points of loci on the second half of chromosome 2L (2L.2) were set to: 1, 2, 3, 4, 5, 6, 7, 8, and 9, with incorporation rate indices set to -3 for rises and the corresponding absolute value for falls. The inflection point of loci on chromosome 2R and 3R were held constant at 4.5 and 5.5, respectively, with incorporation rate indices set to -1.5 for rises and the corresponding absolute value for falls.

With these calculated values in mind we created BAM files that satisfied the required depth for 11 time points (0-10, relative units). To do this, we iteratively concatenated coordinates of loci to each other, for each time point, and converted to a BAM file (bedtools bedToBam[281]). Each concatenation increased depth by the length of the bed file coordinate; therefore, the concatenation never truly reached the exact value of required depth. This was intentional and permitted an intrinsic aspect of noise. However, the intrinsic noise was found to be negligible, so different simulated background noise was merged with each time point using simulated 1X coverage of the entire *Drosophila* genome using ART[303].

We ran TDCA on simulated rise and fall data using all time points (0-10), staggered time points (0, 2, 4, 6, 8, and 10), the first six time points (0-5), and the first and last five time points (0 and 6-10). This was done using the command: tdca -bed 3000-loci.bed -bam <folder_name> -L4. Using the -L4 flag (modelling to a four parameter sigmoid) enabled direct comparison of modeled TTI values to calculate inflection points since the asymmetry factor is equaled to one in a four parameter sigmoid. We converted BAM files to bdg files (bedtools genomecov[281]), added headers, and used UCSC to visualize tracks[304].

3.4.3 Analysis of External Data

The general procedure followed to obtain processable external data was to convert SRA files into fastq format using the fastq-dump command from the SRA toolkit[314]. The files were then aligned to the appropriate reference genome using the bwa mem command from the Burrows-Wheeler aligner[130]. The resulting sequence alignment map files were then converted to binary, sorted, cleared of duplicate reads, and indexed using the Samtools view -bS, sort, rmdup, and index commands, respectively[278]. Sources of data retrieval and additional specific processing instructions are listed below. Default TDCA parameters were used (tdca -bed loci.bed -bam <folder_name>) unless otherwise specified.

3.4.3.1 H3.3

H3.3 replicate 1 TC data was obtained from GEO accession numbers GSM1246648-GSM1246659. H3.3 replicate 2 TC data was obtained from GEO accession numbers GSM1246660-GSM1246670. Input TC data was obtained from GEO accession numbers GSM1246671-GSM1246682. Data was aligned to the mm9 genome. Time points at 72 hours for rep1 and input were used for peak calling only, not in tdca analysis. Peaks were called using MACS2 callpeak q-value of 1e-7[280].

For the expansion of the TDCA genome feature library, tRNA and rRNA gene coordinates were curated from UCSC[306] and enhancer loci from VISTA Enhancer[307].

3.4.3.2 Abf1 Chec-seq

Free MNase Chec-seq (input) data was obtained from GEO accession numbers GSM1647289-GSM1647299. Abf1 Chec-seq data was obtained from GEO accession numbers GSM1647300-GSM1647312. Data was aligned to the sacCer3 genome. Peaks were not called for the Chec-seq data. Instead, loci were obtained from supplementary dataset 1 from the original manuscript[296]. TDCA was run using -t 0 because hills emerged as early as the second time point. K-means clustering was performed using mlpack_kmeans[308]. We used motif scores from the original manuscript so that our results would be highly comparable. For motif discovery, we used MEME-ChIP[309].

3.4.3.3 eXcision Repair-sequencing (XR-seq) on (6-4)pyrimidine-pyrimidone photoproducts [(6-4)PPs]

Replicates 1 and 2 in NHF1 cells (6-4)PP XR-seq were obtained from GEO accession numbers GSM1985857-GSM1985866. Replicates 1 and 2 in CS-B cells (6-4)PP XR-seq were obtained from GSM1985867-GSM1985874. Data was aligned to the hg19 genome.

Slow loci were called with MACS2 using cells that had 240 minutes to repair after UV treatment as the experiment file and cells that had 5 minutes to repair after UV treatment as the control file. Peaks were called for replicates separately, then concatenated and merged to remove redundancy. For the depth threshold analysis and the shuffled loci analysis, peaks called with a p-value threshold of 1e-3 were used. For the p-value analysis, peaks were called using p-values of, 1e-3, 1e-4, 1e-4, and 1e-5. TDCA analysis with shuffled loci was performed using the -dm flag and a file containing the same normalization depth values as non-shuffled loci so that the shuffled and non-shuffled loci could be directly compared.

For loci that displayed variable signal across time, we binned the hg19 genome into 500 bp bins and analysed the top 1% loci at chromosomes 21 and 22 that had the largest range in sequencing depth at each bin after normalizing for variability in total sequencing depth.

Chapter 4: **Time resolved ChIP-seq of O-GlcNAcylated proteins in live flies**

4.1 Background

Chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) is used to define the genome wide distribution of chromatin associated proteins. An area of emerging interest is to study time resolved changes in the distribution of chromatin associated proteins using ChIP-seq experiments. Time course (TC) ChIP-seq enables the differentiation of protein binding dynamics at loci on a genome, allowing researchers to monitor the binding behaviour of a protein. Currently, there have been few time course ChIP-seq experiments published due to technical limitations. However, recent advances[215] provide promising direction for the improvement and expansion of TC ChIP-seq experiments. General TC ChIP-seq strategies use either feeding with bioorthogonal amino acids[285], induction of various tagged engineered genes[286–293], stimulus with known effectors of protein-DNA binding[294,295], induction of DNA cleavage by activation of nuclease fused to proteins of interest[296,297], or repair of DNA damage[298]. Of the handful of TC ChIP-seq experiments that have been performed, all have used cells from culture. New strategies that could potentially be applied to TC ChIP-seq are a topic of high current interest. The benefit of using bioorthogonal chemicals for TC ChIP-seq include antibody free sample preparation, the ability to avoid using genetically engineered cells or organisms, and the potential for very high selectivites.

Recently, novel methodology developed using metabolic feeding with Ac_4GalNAz, a substrate that is metabolized to GlcNAz and incorporated into endogenous O-GlcNAcylated residues[210,215], opened theoretical possibilities for TC ChIP-seq experiments of this interesting post-translational modification. Furthermore, chemical mapping of O-GlcNAcylated proteins in the *Drosophila* genome has been shown to work in live flies, allowing a never before explored experimental mode for TC ChIP-seq. We envisioned that TC ChIP-seq using Ac_4GalNAz would provide a rich understanding of how O-GlcNAc modified proteins bind to genomic loci. Therefore, we set out to design experiments investigating TC ChIP-seq in Ac_4GalNAz fed *Drosophila*.

The first high-throughput sequencing experiment investigating protein-DNA binding dynamics using chemical probes used ChIP on chip[285]. These authors fed

cells with azidohomoalanine (AHA), a methionine derivative that is incorporated into proteins in place of methionine in a manner that depends on the amount of time cells are treated with the amino acid analogue. The researchers purified AHA incorporated proteins from cells that were treated with AHA for differing amounts of time, and then purified nucleosomes using a salt gradient. The nucleosome bound DNA was then sequenced, producing time resolved nucleosome ChIP-chip data.

Similar to AHA feeding, flies can be fed with Ac_4GalNAz for different amounts of time. This would allow for time dependent ChIP-seq analysis of O-GlcNAcylated proteins. However, one would have a difficult time distinguishing between loci at which proteins are simply degraded from loci that contain proteins that are de-glycosylated. Notably, O-GlcNAc is removed from residues by just a single enzyme, OGA. Accordingly, time-dependent Ac_4GalNAz feeding experiments in wild-type (WT) and OGA knockout (ogaKO) flies would effectively enable one to differentiate O-GlcNAc removal by OGA and loss of O-GlcNAc proteins by protein degradation.

Additionally, one can envision performing such an experiment in two ways. One can treat flies with the sugar analogue and then transfer them to media with no sugar or one may feed flies with Ac_4GalNAz for varying amounts of time. Using the first approach to metabolically feed *Drosophila* with Ac_4GalNAz, we have generated the first ever TC ChIP-seq experiment in a live organism. Furthermore, this method provides the first TC ChIP-seq data of a post translational modification (PTM) and could be used as a template to guide TC experiments focusing on other PTMs using different chemical probes. Here we show that some loci are bound by O-GlcNAc modified proteins that were affected by ogaKO, whereas others were unaffected. We discovered developmentally specific genes in larvae that are bound by O-GlcNAcylated proteins and provide an *in silico* analysis of putative substrates and/or protein binding partners. Overall, we provide a never before explored area of epigenetics with respect to O-GlcNAc protein DNA interactions. We expect a wide interest in this TC ChIP-seq methodology, as future experiments could be modified to vary nutrient dosage or enable studies in a range of genetic knockout models. Such research could potentially provide insight into nutrient regulated gene expression and involvement of O-GlcNAc in diseases such as diabetes.

4.2 Results

4.2.1 Ac_4GalNAz time course proof of concept

In order to demonstrate that a TC ChIP-seq experiment using Ac_4GalNAz is possible, we aimed to feed live flies with Ac_4GalNAz and show time dependent changes in O-GlcNAc levels by immunoblot. We envisioned that the best way to do

this would be to feed *Drosophila* from the time they are hatched. This theoretically enables O-GlcNAc proteins to be highly labeled with Ac_4GalNAz. The time course could then be obtained by transferring Ac_4GalNAz labeled flies to Ac_4GalNAz free media, allowing us to monitor decreases in O-GlcNAcylation over time. Early time points would then start with the highest relative signal which would decrease over time to a lower limit of 0. This confined the extent of O-GlcNAc levels between an upper initial limit and zero, which we thought would be of great benefit, especially in considering the modelling accuracy of TC ChIP-seq data. The second crucial consideration was to choose an optimal developmental period in *Drosophila* such that flies would continuously feed on media, presumably allowing continual renewal of labeled O-GlcNAc (O-GlcNAz) on proteins by incorporation of fresh nutrients. Considering this, we speculated that the larval stage, characterized by heavy feeding and growth, would be optimal for this purpose.

We designed a preliminary Ac_4GalNAz feeding timeline where parents were allowed a 2 hour window to lay eggs on Ac_4GalNAz containing media (Figure 4.1). This timing enabled a small difference in age between progeny, yet permitted the collection of enough *Drosophila* for sequencing and immunoblot analysis. After the 2 hour egg laying window, parents were removed and eggs were incubated on the Ac_4GalNAz media for 36 hours. Considering that the time in embryo is approximately 24 hours at 25°C, this gave larvae approximately 12 hours to feed on Ac_4GalNAz. We then transferred larvae at the 36 hour time point post-egg lay to Ac_4GalNAz free media, or alternatively harvested larvae at this time. We denoted this time as time point 0 hour, as in 0 hours of transfer to Ac_4GalNAz free media (Figure 4.1). Larvae fed on Ac_4GalNAz free media were subsequently collected at 4, 8, 12, and 36 hours post transfer. Therefore, from egg lay to the first collection time the larvae were 36-38 hours old and from egg lay to the final collection time the larvae were 72-74 hours old. Consistent with this timeline, the larvae at the final collection time were noticeably larger than those of the first collection. We considered this window optimal for saturating O-GlcNAc loci yet allowing enough time for removal of O-GlcNAz. We also included a control where flies were not fed with Ac_4GalNAz (no feed - Figure 4.1). We prepared wild type Oregon R (WT) and OGA knockout (ogaKO)[315] *Drosophila* in accordance with this Ac_4GalNAz feeding regimen.

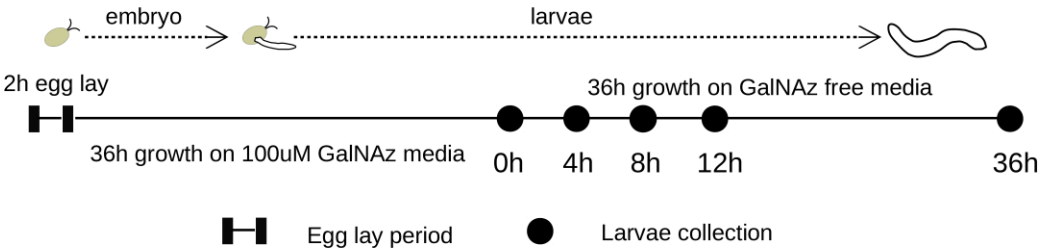


Figure 4.1. **Timeline of Ac_4GalNAz feeding.**
Oregon R wild type (WT) or OGA knockout (ogaKO) *Drosophila* adults were given a two hour window to lay eggs. This pre-laying window permitted better synchronization of progeny that were collected. The second two hour egg laying period is shown in the legend. Eggs were laid on 100 uM Ac_4GalNAz media for the time course experiments, and on non-Ac_4GalNAz media for the no feed control. After 36 hours of growth, larvae were collected (0h) or transferred onto non-Ac_4GalNAz media. The no feed control was collected at this time as well. Subsequent growth on non-Ac_4GalNAz media and collections at 4, 8, 12, and 36 hours post-transfer ended the collections.

Next, we performed an immunoblot using lysates from WT and ogaKO Ac_4GalNAz time course fed *Drosophila* as we have done previously[215] (Figure 4.2, a). As expected, we observed a decrease in immunoreactivity over time. Moreover, we see that ogaKO flies retain higher levels of O-GlcNAc immunoreactivity at late time points. We also verified that ogaKO flies contained less OGA mRNA using RT-PCR (Figure 4.2, b). These results encouraged us to perform TC ChIP-seq in our Ac_4GalNAz fed *Drosophila*.

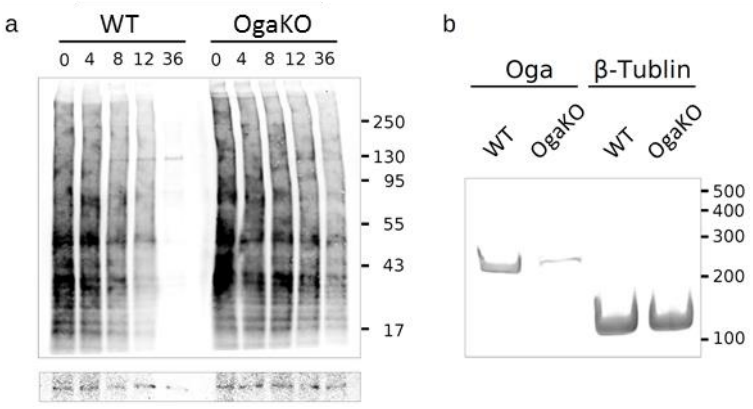


Figure 4.2. **Western blot of time course Ac_4GalNAz fed larvae and confirmation of ogaKO flies.**
(a) Larvae proteins were extracted and conjugated to biotin. Upper panel shows blot with streptavidin, indicative of Ac_4GalNAz protein quantity (ladder in kDa). Lower panel shows actin loading control (band is between 55-43 kDa). (b) RT-PCR of OGA and tubulin in WT and ogaKO flies. Ladder sizes in bp.

4.2.2 ChIP-seq data pre-processing and selection of loci

For each time point in WT and ogaKO, we prepared DNA libraries for Illumina sequencing, sequenced DNA using an Illumina Miseq, and aligned[130,300] the resulting DNA to the *Drosophila* dm3 genome. Figure 4.3, a-c shows the total number of sequenced reads, the percentage of aligned reads and total aligned reads for both WT and ogaKO flies. Although there are noticeable differences between these parameters, the ratio of aligned reads relative to the time point with the absolute maximum aligned reads for both WT and ogaKO TC experiments showed similar decreases in numbers of reads over time (Figure 4.3, d). We noticed that the time point at 8 hours did not follow the relative change in sequencing depth over time for WT or ogaKO. In fact, when we tested the Spearman correlation of WT and ogaKO normalized depth values we found a poor correlation (Spearman correlation of 0.3 with a p-value of 0.68). Removing the 8 hour time point from each set resulted in the Spearman correlation increasing to 0.8 (p-value = 0.33). A Spearman correlation of 1 indicates 100% positive correlation, while -1 indicates complete negative correlation. A Spearman correlation of 0 indicates neither a positive nor negative correlation. Because we wanted fairly similar quality sequencing data (positive Spearman correlation) as input for our subsequent analyses of WT and ogaKO samples, so as to limit analysis bias, we decided to exclude the 8 hour time point from analysis.

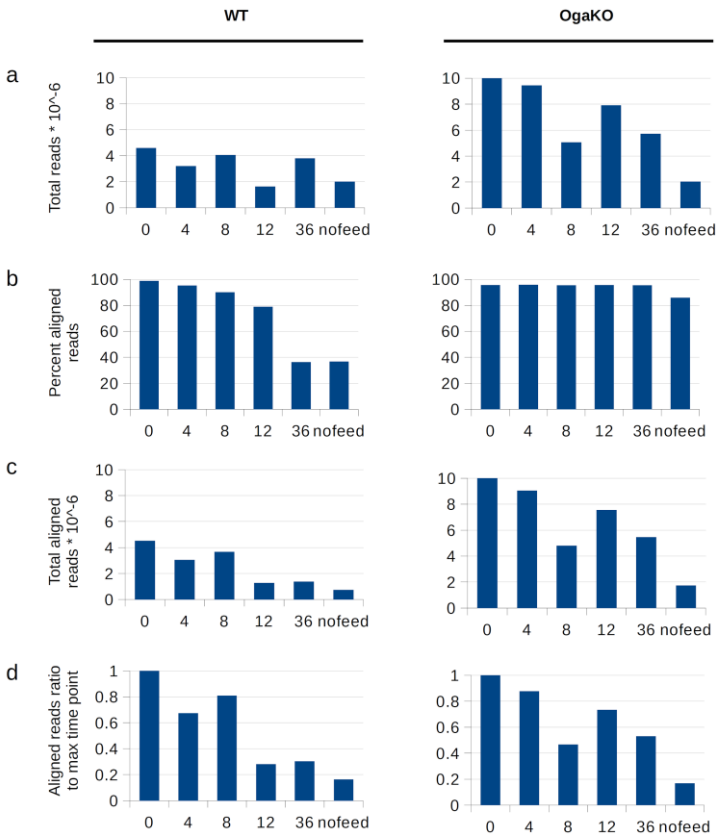


Figure 4.3. **$Ac_4GalNAz$ time course sequencing quality.**
(a) Total reads for WT and ogaKO $Ac_4GalNAz$ TC experiments. (b) Percent aligned reads to the dm3 genome for WT and ogaKO $Ac_4GalNAz$ TC experiments. (c) Total aligned reads to the dm3 genome for WT and ogaKO $Ac_4GalNAz$ TC experiments. (d) Total aligned reads to the dm3 genome for WT and ogaKO $Ac_4GalNAz$ TC experiments represented as a ratio to the time point with the absolute maximum value. The x-axis indicates time at which *Drosophila* samples were collected.

The next important factor to consider was how to characterize loci that are bound by O-GlcNAz modified proteins. We explored different strategies to identify these loci. First, MACS software[280] was used to identify loci at which many sequencing reads are observed, which are termed “peaks”. To do this we used the 0 hour time point as the treatment file and 0 hour no feed as the control. This process was repeated for WT and ogaKO. Just over 1000 peaks were detected for both WT and ogaKO (Figure 4.4, a) of which only a fraction (179) overlapped. It was important to only note the identity of the overlapping peaks of WT and ogaKO so that downstream processing would allow a pairwise comparison of loci. This procedure was repeated using additional peak calling strategies. Although we used a relatively loose p-value threshold, a limited number of peaks were identified.

An alternative strategy whereby the genome was divided into 500 bp bins was used. In this approach, the bins having the greatest ratios of sequencing depth at time 0 hour compared to no feed controls was explored. We found in the top 5%

and 10% bins similarly low overlap between loci identified in WT and ogaKO using this strategy (Figure 4.4, b-c). Finally, we used an enrichment strategy to call peaks, as was done previously[296]. This strategy was done by iterating through each base pair using the 0 hour time point and the no feed control. Peaks were called if the base had 5 times the average sequencing depth in the experiment as compared to the control. Peaks identified in this way that were within 30 bp were then merged together[281]. In order to eliminate false positives, peaks in the no feed control were eliminated from peaks called using the 0 time point. Overlap of WT and ogaKO peaks resulted in 496 conserved loci (Figure 4.4, d).

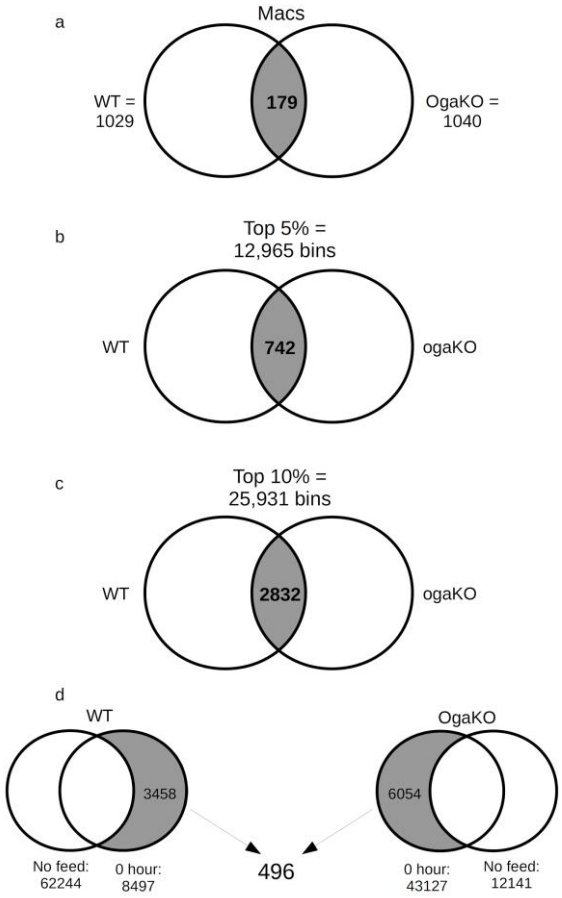


Figure 4.4. Exploration of peak calling strategies for $Ac_4GalNAz$ time course experiment.

(a) MACS[280] software was used to call peaks in WT 0 hour against WT no feed and ogaKO 0 hour against ogaKO no feed (p=0.05). 179 peaks were conserved between the two groups. (b) The *Drosophila* genome was divided into 500 bp bins and at each bin, the normalized read depth of time 0 hour and no feed was calculated. This was done for both WT and ogaKO. The bins were then sorted from most to least enriched (relative to time 0) and the top 5% were then overlapped in WT and ogaKO resulting in 742 loci. (c) As in (b) except the top 10% bins were used. (d) The average depth was calculated per base for time 0 hour and no feed in WT and ogaKO, and bases that exceeded 5X the average were kept as peaks. Peaks were then merged if they were within 30 base pairs of each other. For WT and ogaKO,

no feed peaks were subtracted from time 0 hour peaks and the filtered 0 hour peaks were overlapped to obtain 496 loci.

Visual inspection of a large majority of peaks called from each peak calling strategy revealed that loci from the last strategy, the enrichment approach, showed the greatest ratio of signal to noise. Furthermore, this peak calling strategy was previously published[296], and resulted in a number of loci (496) that would not drastically reduce the power of genome wide analyses. For these reasons, we carried out the remaining analysis using this 5X enriched peak strategy.

4.2.3 TDCA analysis of WT and ogaKO Ac_4GalNAz time course

Using TDCA we next analysed Ac_4GalNAz TC ChIP-seq data using WT and ogaKO flies using 0, 4, 12, and 36 hour time points at the 5X enriched loci. Quality analyses revealed that the majority of loci in both WT and ogaKO showed a decrease in signal over time, that is, they behaved as falls (Figure 4.5, a-b). This was supported by the preponderance of absolute maximum sequencing depth being obtained at time point 0 for both WT and ogaKO (Figure 4.5, c-d). Furthermore, we observed that the absolute minimum sequencing depth occurred, as expected, mostly at 12 and 36 hour time points. For WT, minimum absolute depth was observed at 12 hours (39.5%) and 36 hours (60.3%), whereas for ogaKO, the percent occurrence of absolute minimum sequencing depth for these time points was 13.1% and 82.3%, respectively (Figure 4.5, c-d). These data indicate that WT reached the lower asymptote representing complete removal of Ac_4GalNAz at more loci than ogaKO, which was expected.

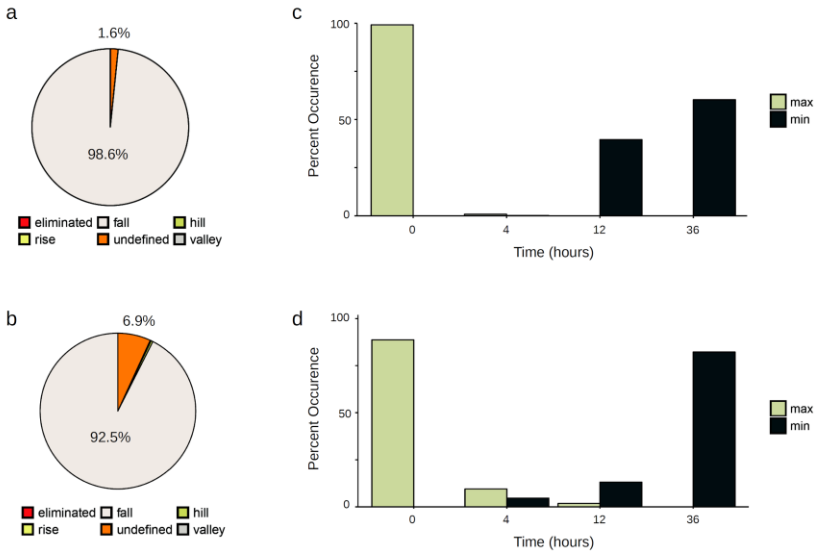


Figure 4.5. **TDCA quality analysis of $Ac_4GalNAz$ time course data.**
(a) Pie chart indicating the behaviour type that WT *Drosophila* are categorized as. (b) As in (a) except for ogaKO. (c) Bar chart showing the average occurrence of the absolute maximum sequencing depth and absolute minimum sequencing depth across loci for WT *Drosophila*. (d) As in (c) except for ogaKO

TDCA provides a heatmap where sequencing depth at each locus is normalized from 0 to 1 across time points so that comparisons can be rapidly made by visual inspection across loci. We observed faster depletion of signal for WT compared to ogaKO (Figure 4.6, a-b), which corroborated our immunoblot data (Figure 4.2, a). Average fall profiles in WT and ogaKO samples also supported overall more persistent signal from ogaKO (Figure 4.7, a-b).

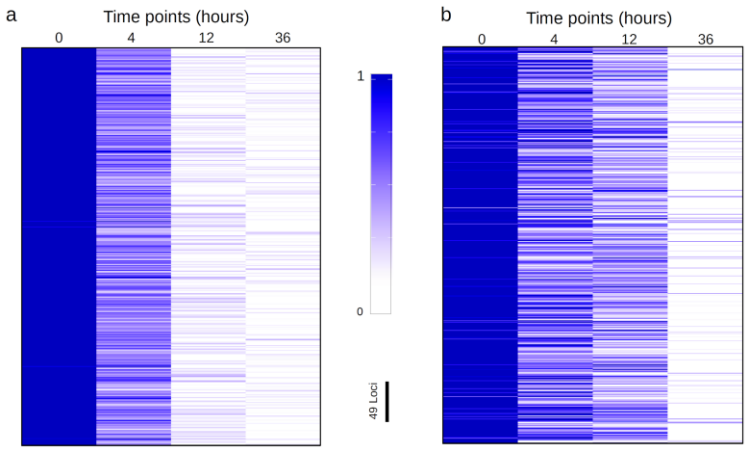


Figure 4.6. **Normalized sequence depth heatmap.**
Normalized sequence depth, represented as a ratio to the time point with the absolute maximum sequence depth, for each loci, shown in rows for (a) WT and (b) ogaKO flies.

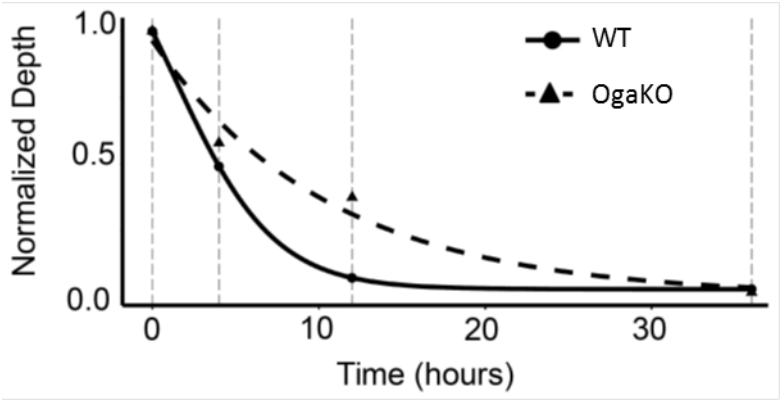


Figure 4.7. Average fall profiles.
Average fall profiles of WT and ogaKO represented as an average fold change to the time point at which the absolute minimum sequence depth occurs. Lines show how the data is modelled and the circles and triangles indicate average sequence depth fold change.

At a given locus, TDCA models sequencing depth at each time point to a 5 parameter sigmoidal curve. The inflection point, adjusted by the asymmetry factor, provides a measure of time at which half maximal protein-DNA binding occurs. This parameter is denoted as the turnover time index (TTI) and was used for the remaining analyses. A density plot of TTI values in WT and ogaKO flies was created to show the relative change in distribution (Figure 4.8). We observed a sharp peak for WT around 4 hours. On the other hand, ogaKO had a right tailed skewed distribution, indicating the persistence of longer lasting O-GlcNAz modified proteins on the genome.

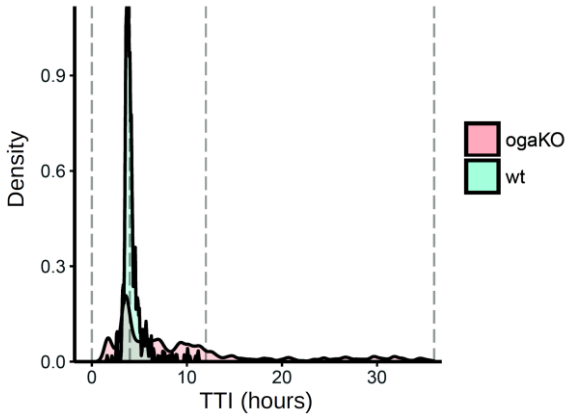


Figure 4.8. Density plot of TTI values from loci that behave as falls in WT and ogaKO.
TTI distribution of WT *Drosophila* larvae is indicated in blue and ogaKO *Drosophila* in red.

We provide three snapshots of loci on different chromosomes (Figure 4.9-4.11) with each time point for WT and ogaKO being shown. The highlighted regions

are 5X enriched peaks, with TTI values recorded at top in hours. We also provide TDCA modelling results of the highlighted regions. Here, it is apparent that some loci contain similar TTI values in both WT and ogaKO flies, such as loci chr2R:6964279-6964331 and chr3L:24056572-24056669 (Figure 4.9, c and Figure 4.11, b), whereas other loci contain vastly different TTI values, such as loci chr2R:6956302-6956326 and chr2R:7092645-7092675 (Figure 4.9, b and d).

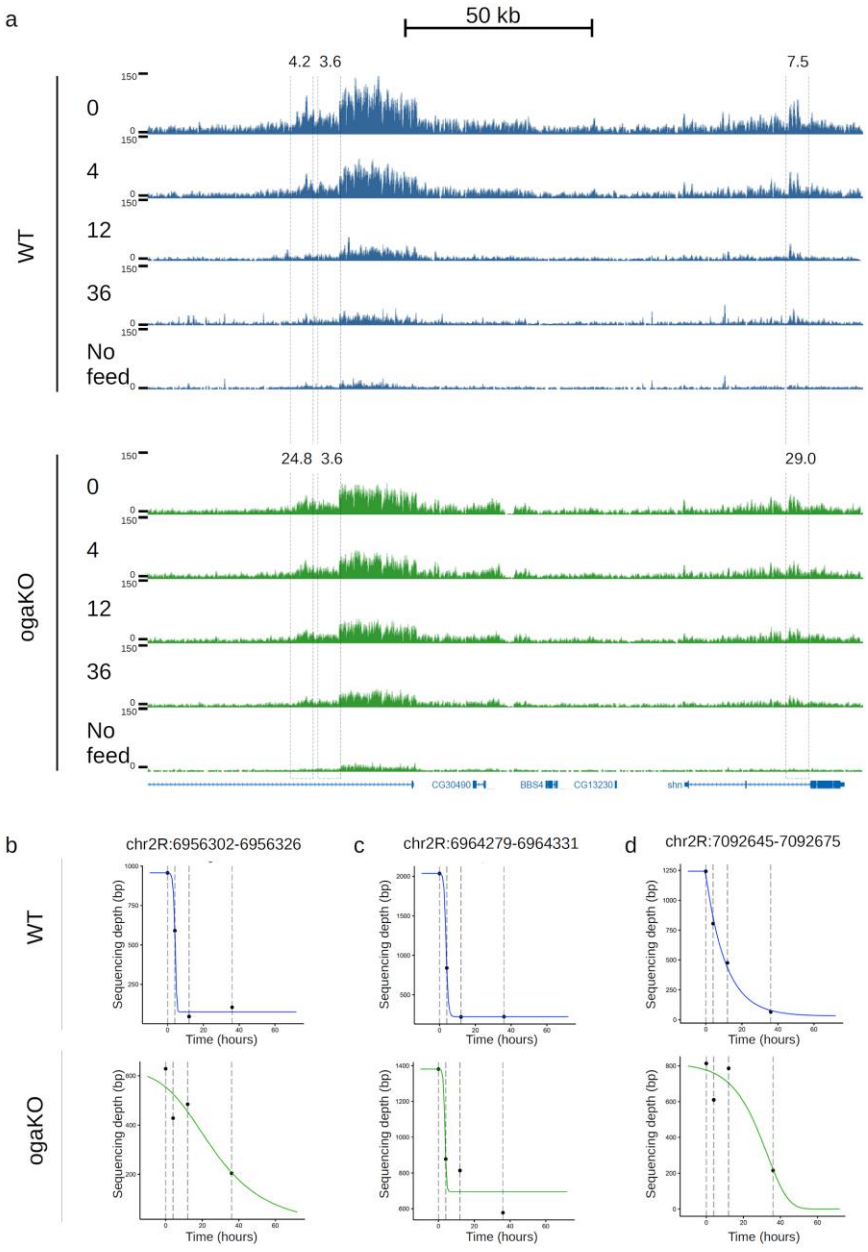


Figure 4.9. **Tracks and data modelling of loci on chromosome 2L.**
(a) ChIP-seq tracks of read density of WT (upper panel - blue) and ogaKO (lower panel - green) Ac_4GalNAz time course. Time points are indicated to the left as well as sequence depth intensity. Peaks are highlighted by dashed lines with TTI values. Genes are shown at the bottom of the tracks and a scale is shown on top. The three highlighted peaks are modelled in (b), (c),

and (d). Each time point contains a sequencing depth value which is modelled to a 5 parameter logistic curve. WT data is shown on top panels and ogaKO is shown at the bottom.

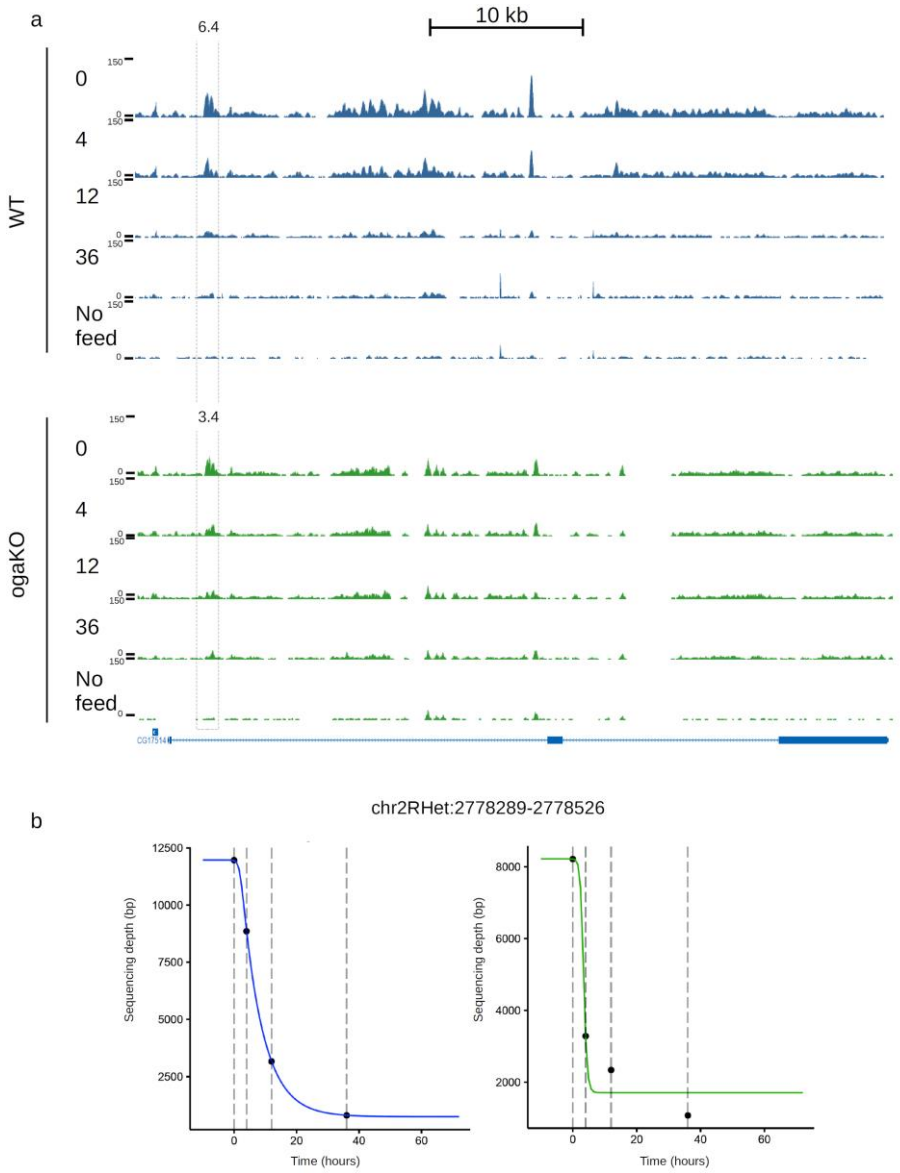


Figure 4.10. Tracks and data modelling of loci on chromosome 2RHet. Data is shown as in Figure 4.9 (a-b).

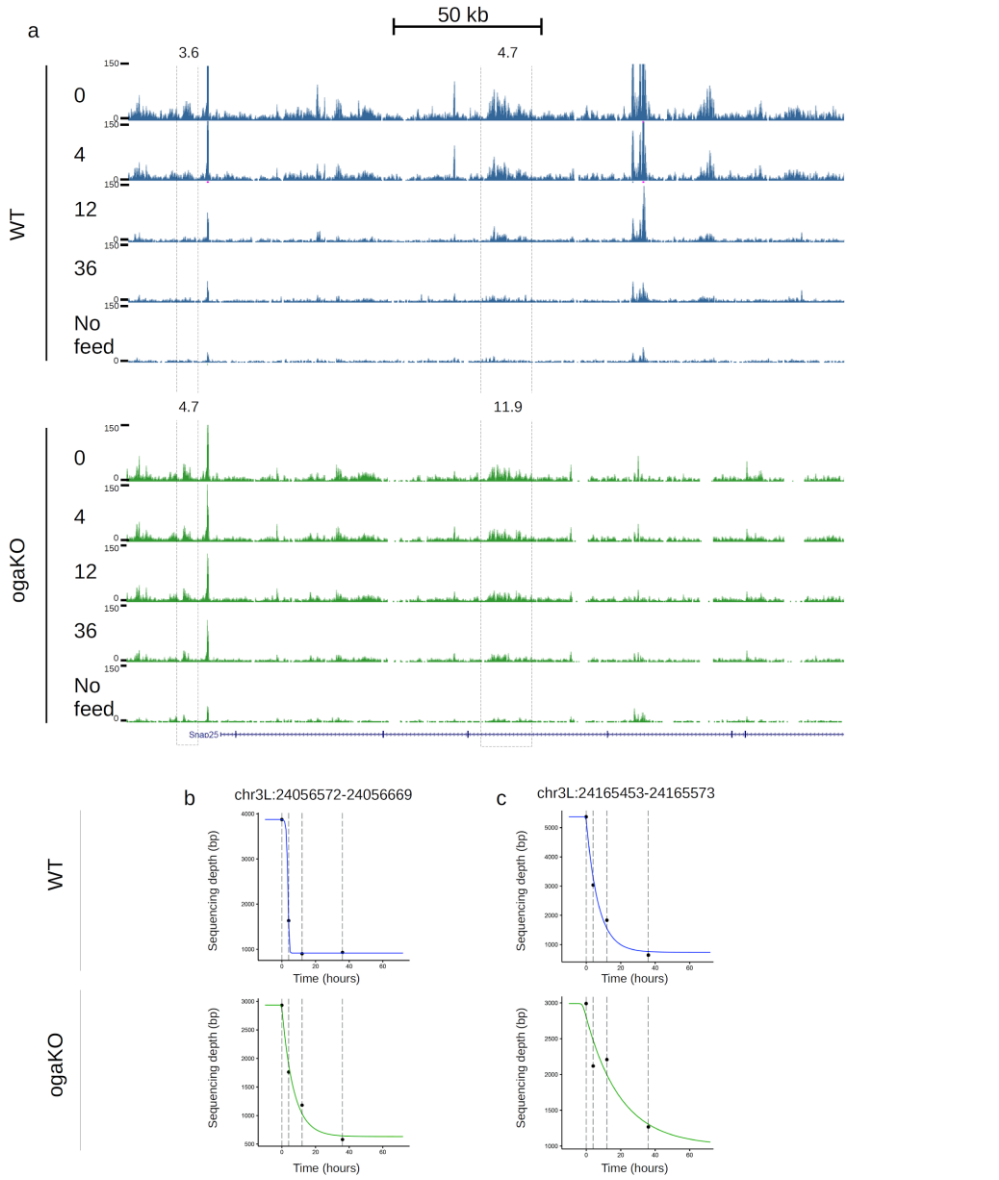


Figure 4.11. Tracks and data modelling of loci on chromosome 3L.
Data is shown as in Figure 4.9 (a-b).

Next, we explored the distribution of TTI values in WT and ogaKO at gene features (Figure 4.12). Interestingly, there was a homogeneous TTI average and standard deviation of approximately 4±2 hours across the 13 gene features we tested in WT. OgaKO showed an overall increase in TTI average and standard deviation, indicating that differences in WT and ogaKO TTI values likely do not stem from association with the analysed gene features.

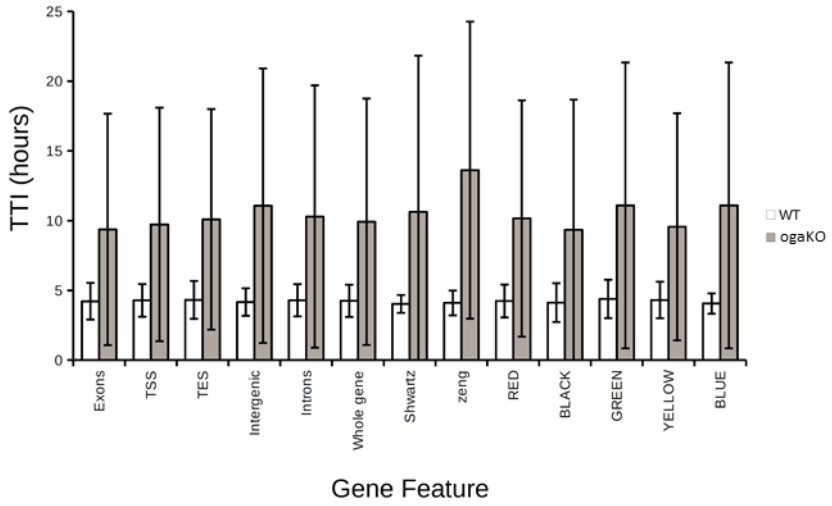


Figure 4.12. Gene feature TTI analysis.
Average and standard deviation of TTI values in WT (white) and ogaKO (grey) flies. Gene features include: coding exons (Exons), transcriptional start site - defined as the 3' UTR and 1000 bp upstream (TSS), transcriptional end site - defined as the 5' UTR and 1000 bp downstream (TES), intergenic regions - defined as DNA that does not overlap gene bodies and 1000 bp upstream and downstream of gene bodies (Intergenic), Introns (introns), Whole gene (whole gene), PRE data set defined by Shwartz[266], PRE data set defined by Zeng[267], and various heterochromatic regions defined by Fillion[49] (RED, BLACK, GREEN, YELLOW, BLUE).

A cluster analysis was then performed in order to determine if large domains in the *Drosophila* genome were differentially bound by O-GlcNAz modified proteins in WT and ogaKO flies. Figure 4.13, a-b shows a side by side ideogram heatmap of TTI values in WT and ogaKO flies. Careful inspection reveals that some loci do indeed show differences in TTI in these ideograms. Next, we divided the *Drosophila* dm3 genome into 200,000 bp bins and overlapped the bins with the 5X enriched peaks. Bins containing 5 or more peaks were considered as being clusters. The average and standard deviation TTI from both WT and ogaKO was then plotted at these regions (Figure 4.14, a). We found a similar trend to the gene feature analysis, where average WT TTI values centered around 4 hours and ogaKO had a general increase in both average and standard deviation TTI. Notably, however, in some clusters, ogaKO retained a fairly similar average TTI and standard deviation relative to WT flies. This observation indicated that there are clusters of loci that are not affected by loss of OGA in the ogaKO flies. We did not observe any clusters in ogaKO with both a higher average TTI compared to WT and having a tight standard deviation such that the limits of standard deviation were outside that of WT. This data indicates that clusters that do contain peaks with long TTI values in ogaKO, relative to WT, are relatively close in proximity (within the same cluster boundary) to peaks that have

similar TTI values in both WT and ogaKO. This observation was also reproduced using smaller (50,000 bp) bins (Figure 4.14, b).

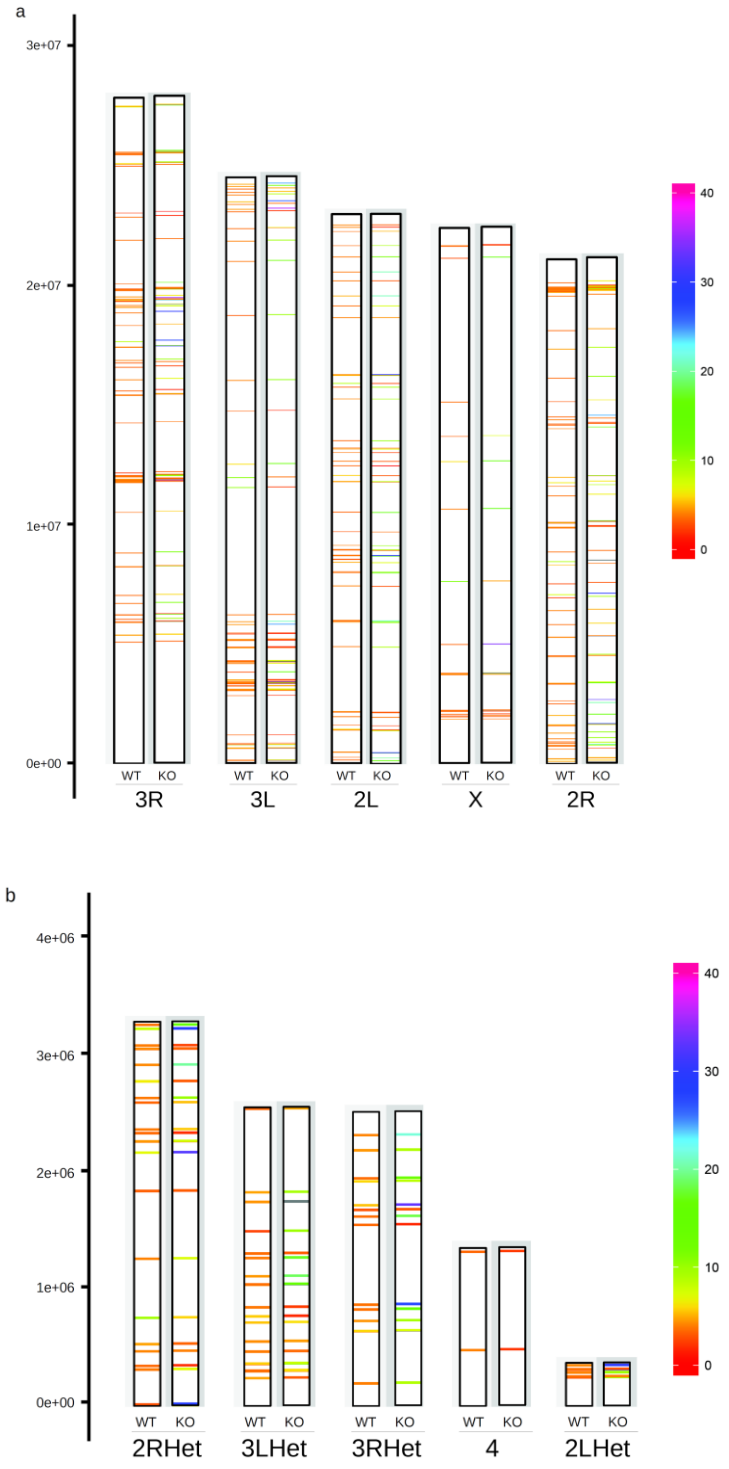


Figure 4.13. Heatmap ideograms of WT and ogaKO TTI values.
(a) Chromosomes 3L, 3R, 2L, X, and 2R. (b) Chromosomes 2RHet, 3LHet, 3RHet, 4, and 2LHet. TTI scale and chromosome base pair scale are shown to the right and left, respectively.

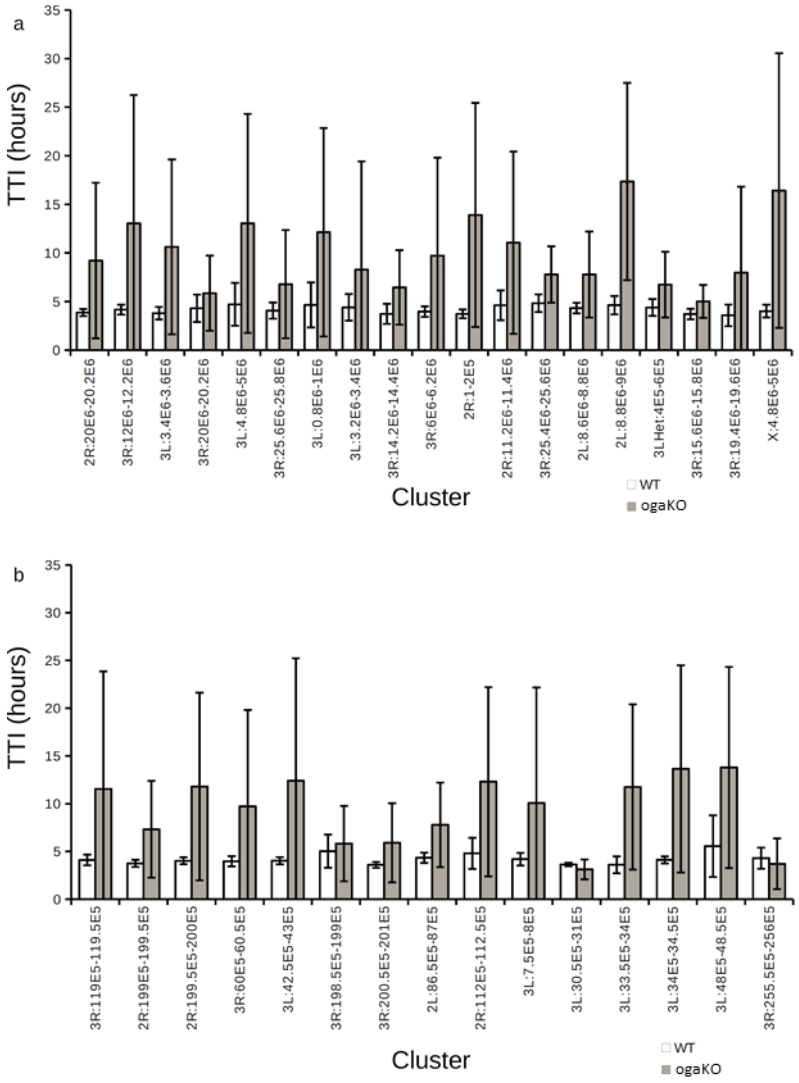


Figure 4.14. TTI cluster analysis in WT and ogaKO flies.
The dm3 genome was divided into 200,000 bp bins (a) or 50,000 bp bins (b) and each bin was overlapped with 5X enriched loci. Those bins that contained >= 5 loci were considered clusters and are shown here with the average and standard deviation of TTI values for WT (white) and ogaKO (grey) flies are shown at each of the bins. E indicates base 10. Chromosomes are indicated before the colon.

4.2.4 O-GlcNAc protein-DNA binding kinetics in WT and ogaKO *Drosophila* diverge at certain loci

In an attempt to provide a more intuitive way of understanding the behaviour of loci that bear O-GlcNAc modified proteins that are affected by ogaKO and those that are not, we performed a pairwise analysis of loci that behaved as falls in samples of both WT and ogaKO flies. Of the 496 5X enriched loci, 455 (91.7%) behaved as falls in both WT and ogaKO. We then took the difference of ogaKO TTI values and WT TTI values, which we call the TTI difference (ΔTTI), at each of these loci. After sorting TTI differences from least to greatest and plotting the ΔTTI for each

locus (Figure 4.15, a), it was apparent that approximately half of the loci centered close to 0 and half did not. We observed a reasonably large occurrence of loci with a positive TTI difference and only a small fraction of loci with negative TTI differences. A positive TTI difference is indicative of ogaKO having a longer TTI value than WT, while a negative TTI difference is indicative of WT having a longer TTI value compared to ogaKO. By this logic, a TTI difference of zero indicates no change in TTI when comparing WT and ogaKO samples. Therefore, we were encouraged to see approximately half of the peaks containing a positive TTI difference, with greater overall absolute values as compared to the few loci having negative ΔTTI values. This observation supported all previous data indicating loci in ogaKO flies have longer TTI values as compared to WT flies. We decided to pursue this measure of TTI difference in the subsequent analyses.

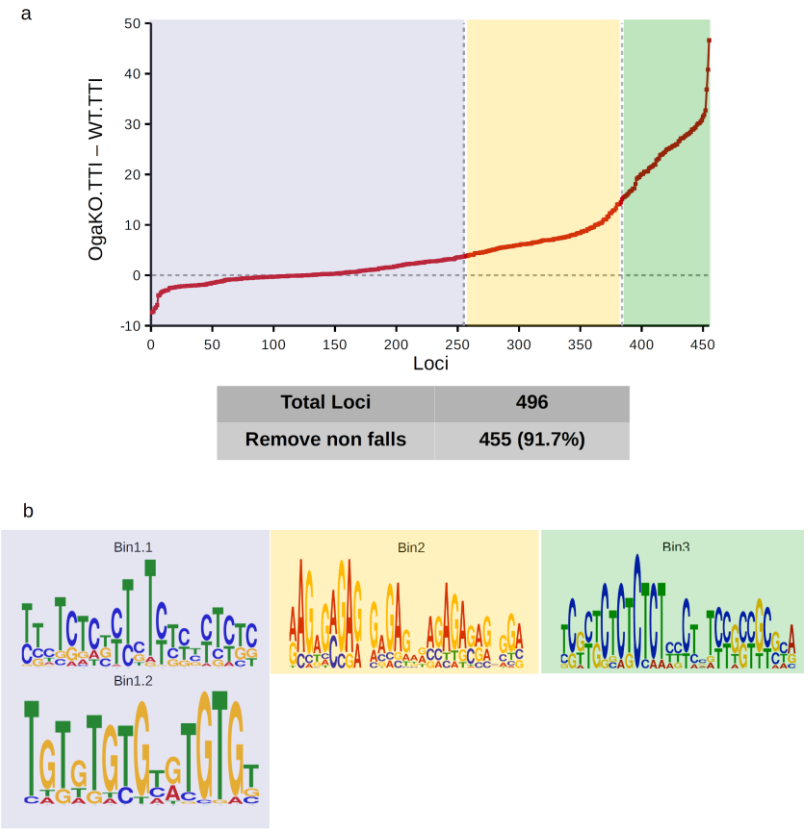


Figure 4.15. Motif analysis in Ac_4GalNAz time course data.
(a) Difference of ogaKO and WT TTI values sorted across loci from least to greatest. 91.7% of loci behaved as falls in both WT and ogaKO data sets. Horizontal dashed line indicates the position of 0. Loci k-means clustered into 3 categories, separated by vertical dashed lines. The blue region was characterized as loci whose O-GlcNAc bound proteins are unaffected by ogaKO. The yellow region was characterized by loci whose O-GlcNAc bound proteins were moderately affected by ogaKO. The green region was characterized by loci whose O-GlcNAc bound proteins were highly affected by ogaKO. (b) Motifs identified in each region.

First, we k-means clustered[308] the TTI differences into 3 clusters. This resulted in loci with TTI differences less than 3.73 hours to be considered as being in the same category. The two other clusters were loci with TTI differences between 3.73 hours and 15.06 hours and loci with TTI differences greater than 15.06 hours. This is shown graphically in Figure 4.15, a. We designated loci in the first cluster (TTI differences less than 3.73 hours - highlighted in blue in Figure 4.15, a) as those that contain O-GlcNAc proteins that are not affected by loss of OGA. Loci in the second cluster (TTI differences between 3.73 hours and 15.06 hours - highlighted in yellow in Figure 4.15, a) were considered to be moderately affected by loss of OGA. Finally in cluster 3 (TTI differences greater than 15.06 hours - highlighted in green in Figure 4.15, a) were considered to be highly affected by loss of OGA. A motif analysis[309] of 100 bp sequences around the center of each peak[281] within each cluster revealed two motifs in cluster one, and one motif in clusters two and three, with p-values less than 0.05 (Figure 4.15, b). We proceeded to use tomtom[316] to predict potential proteins that bind to these motifs in *Drosophila*. Our analysis revealed 17 potential proteins that putatively bind to the four collective motifs, which is shown as a colour map in Figure 4.16.

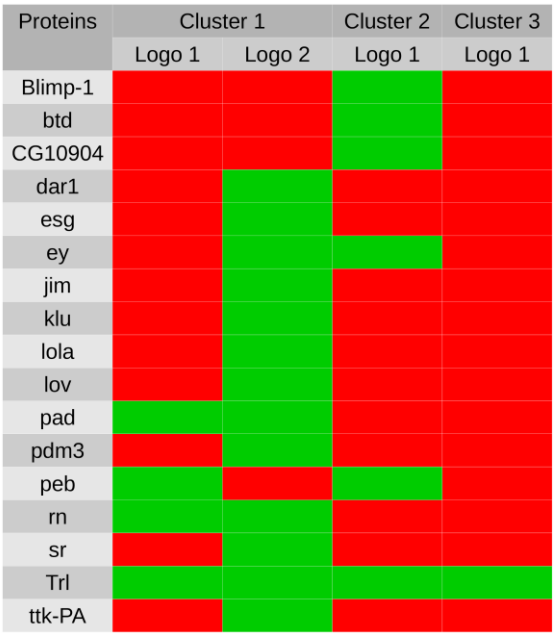


Figure 4.16. Proteins predicted to bind to O-GlcNAc motifs.
DNA-binding proteins listed to the right. A green filled quadrilateral represents a protein that is predicted to bind to a motif.

Next, we found 252 unique genes that intersect with the set of 5X enriched peaks. We considered these as genes that are putatively regulated by O-GlcNAcylated proteins in 36-72 hour old larvae. Notably, there were a number of

PRE containing genes previously known to be bound by O-GlcNAc modified proteins[153], providing additional support for the accuracy of using the 5X enriched loci strategy. We also performed a gene ontology analysis[284] of genes within each TTI difference cluster. We found that there were vastly different ontology terms within certain clusters (Table 4.1). We speculate that loss of OGA may therefore have the greatest effect on loci that bear genes with specific functions.

Table 4.1. Gene ontology analysis of genes in larvae that are bound by O-GlcNAc and are differentially affected by loss of OGA.
Only gene ontology terms with highly significant association (p-value <= 0.01) are reported. Cluster 1 contains genes that intersect loci containing O-GlcNAc modified proteins that are not affected by loss of OGA. Cluster 2 contains genes that intersect loci containing O-GlcNAc modified proteins that are moderately affected by loss of OGA. Cluster 3 contains genes that intersect loci containing O-GlcNAc modified proteins that are highly affected by loss of OGA. Percent occurrence in a single column can sum to be greater than 100 because some proteins fall within multiple ontology terms.

GoTerm	Percent Occurrence		
	Cluster 1	Cluster 2	Cluster 3
ATP binding	-	-	21.3
axon guidance	7.7	6.2	12.8
border follicle cell migration	-	6.2	8.5
chaeta development	-	5.2	-
chaeta morphogenesis	-	4.1	-
dendrite morphogenesis	7.1	6.2	-
determination of adult lifespan	4.7	-	10.6
dorsal appendage formation	3.0	-	-
dorsal closure	-	-	10.6
gonad development	3.0	-	-
imaginal disc-derived wing morphogenesis	5.9	11.3	17.0
larval somatic muscle development	3.0	-	-
metal ion binding	13.0	-	-
metamorphosis	-	4.1	-
muscle fiber development	1.8	-	-

GoTerm	Percent Occurrence		
	Cluster 1	Cluster 2	Cluster 3
muscle organ development	3.6	-	-
negative regulation of transcription from RNA polymerase II promoter	4.1	-	10.6
negative regulation of transcription, DNA-templated	5.3	6.2	10.6
nucleoplasm	-	6.2	-
nucleus	23.7	-	29.8
nurse cell apoptotic process	-	3.1	-
ovarian follicle cell development	5.3	5.2	-
peripheral nervous system development	-	5.2	10.6
phagocytosis	-	7.2	-
plasma membrane	11.2	-	-
positive regulation of cell proliferation	-	-	6.4
positive regulation of organ growth	-	-	6.4
positive regulation of transcription from RNA polymerase II promoter	-	7.2	-
positive regulation of transcription, DNA-templated	-	6.2	-
protein binding	-	17.5	-
protein homodimerization activity	5.3	-	-
R3/R4 cell fate commitment	-	3.1	-
R7 cell development	2.4	-	-
regulation of glucose metabolic process	-	7.2	-
regulation of transcription from RNA polymerase II promoter	-	9.3	-
regulation of transcription, DNA-templated	7.1	-	-
repressing transcription factor binding	-	3.1	-

GoTerm	Percent Occurrence		
	Cluster 1	Cluster 2	Cluster 3
RNA polymerase II transcription factor activity, ligand-activated sequence-specific DNA binding	-	3.1	-
sequence-specific DNA binding	7.1	10.3	14.9
signal transduction	-	-	10.6
tracheal outgrowth, open tracheal system	-	4.1	-
transcription factor activity, sequence-specific DNA binding	7.7	12.4	-
transcription factor binding	-	6.2	-
transcription, DNA-templated	7.1	9.3	-
transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	-	5.2	-
wing disc dorsal/ventral pattern formation	-	4.1	-
zinc ion binding	-	12.4	-

4.2.5 O-GlcNAc bound genes in different *Drosophila* stages

Lastly, we performed an analysis on genes found to contain O-GlcNAc modified proteins in S2 cells[215], pupae[215], and larvae (this study) using Ac_4GalNAz metabolic feeding. We report a Venn diagram of genes found to contain O-GlcNAc bound proteins in these two developmental stages and cell type (Figure 4.17, a). In larvae and pupae, we further probed expression levels of O-GlcNAc containing genes in relation to the expression levels of all genes at those particular stages[317]. In pupae and larvae, we found that genes that are expressed at a moderately-high level are significantly enriched. In pupae only, we found that genes that are expressed at a very low and extremely low level are significantly reduced, as determined by the Chi-squared test (Figure 4.17, b).

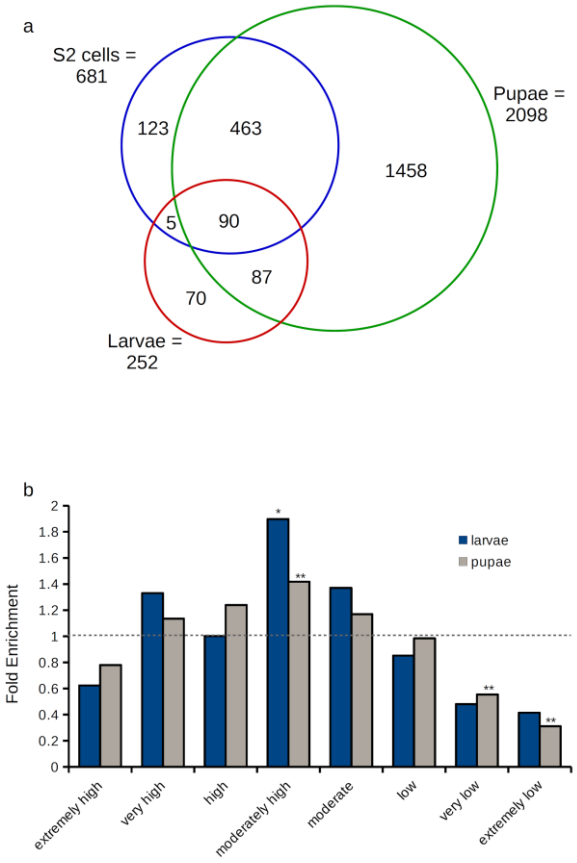


Figure 4.17. Analysis of genes found to contain O-GlcNAc bound proteins in S2 cells, pupae, and larvae.
(a) Venn diagram of FlyBase genes that contain O-GlcNAc bound proteins based on Ac_4GalNAz metabolic feeding in S2 cells, larvae, and pupae. (b) Enrichment of genes expressed at certain levels. Bars show comparison of genes containing O-GlcNAc bound proteins (observed) to all genes (expected) in larvae and pupae. * = p-value 0.01, ** = p-value 0.001 (Chi squared test).

4.3 Discussion and conclusions

Considering the Ac_4GalNAz feeding timeline, we realized that 24 hours may not have been long enough for some long lasting O-GlcNAc modified proteins that may have persisted and remained bound to the genome from embryo to lose O-GlcNAc or turnover. However, this was a limitation that we were willing to accept given the practical limitations in the developmental time frames of flies. Furthermore, this issue could be fully explored and potentially resolved through the use of endogenous O-GlcNAc capture strategies such as WGA in combination with genetically engineered flies in which OGT can be controlled. Also, given that the localization of chromatin bound proteins, and therefore O-GlcNAc bound proteins on chromatin, should be reasonably dynamic during development, each time point of larval collection presumably contains some number of unique loci. However, the peak calling strategies used here discern only those loci that are most different

between time point 0 hour and no feed. Therefore, this consideration is eliminated. Calling peaks individually at each time point with respect to the no feed controls would allow discovery of short lasting, developmental specific, O-GlcNAc loci, which could be of interest in future studies.

The exclusion of a time point at 8 hours was mathematically supported, however, this lessened the time points used and therefore lowered the overall power of modelling accuracy. Data for this sequencing experiment was intended to show that Ac_4GalNAz could be used as a tool for TC ChIP-seq and that differences could be seen in WT and ogaKO *Drosophila*. To this end, although the exclusion of the data point at time 8 hours was somewhat arbitrary, the current analysis is satisfactory. We are currently collecting new Ac_4GalNAz time course fed flies with additional time points, and with biological replicates. This should increase the confidence in data modelling and deliver more robust data. We also noticed increased human DNA contamination over time, which likely stems from an increased ratio of human DNA being observed over time as less *Drosophila* DNA is recovered. ChIP-seq spike-in normalization strategies should improve this artifact.

It is interesting that using three very different peak calling strategies consistently resulted in a low overall overlap between loci called from WT and ogaKO. Upon visual inspection of the tracks for the peaks revealed that this was likely due to sequencing biases stemming from using different sequencing kits for WT (V2) and ogaKO (V3). Considering that sequencing kits are known to have certain biases, it could be considered an added measure of accuracy that different kits were used, since this should help reduce false positives

It was surprising to see that the 13 gene features that were analysed for TTI showed homogenous distribution with relatively tight standard deviations in WT. ogaKO also did not show specific increases in TTI but rather a global increase in both average and standard deviation. Furthermore, when analysing the TTI average in 200,000 bp and 50,000 bp bins, we observed some clusters that maintained similar TTI averages in WT and ogaKO but no clusters in ogaKO that contained averages and standard deviations outside the range of WT. This indicated that loci that were affected by loss of OGA were in relatively close proximity to loci that were not affected by loss of OGA. The most dramatic differences were seen at genes binned by ontology. We found that genes involved in ATP binding, and dorsal closure, among others, to have high enrichment at 21.3% and 10.6%, respectively, in the cluster calculated to be most affected by loss of OGA, while these terms were entirely absent in other clusters (Table 4.1). In cluster 2, which is defined by loci that are moderately affected by loss of OGA, we seen an enrichment of genes involved in

protein binding (17.5%), which is an ontology category absent from the other clusters. This data may indicate that OGA is recruited specifically to genes that are involved in specific functions. Future studies could pursue this line of research and may potentially lead to interesting findings regarding regulation of gene expression by O-GlcNAc.

Our motif analysis identified 17 putative binding proteins at O-GlcNAc containing loci (Figure 4.16). Using a mass spectrometry technique to identify O-GlcNAc bound proteins on chromatin[318], we intend to verify if these predicted proteins are substrates of OGT in larvae. However, given that the ChIP-seq sample preparation requires cross-linking of proteins and their complex partners to DNA, it may be that the predicted proteins are recruiters of O-GlcNAc modified proteins to specific loci.

In our analysis of expression levels of O-GlcNAc bound genes in pupae and larvae, we found that moderately high expressed genes were enriched compared to the expected ratios. Further, pupal genes at which O-GlcNAc modified proteins are bound were depleted in very low and extremely low expressed genes and larval genes followed this trend but did not show significance in the Chi-squared test due to lower numbers. This observation is counterintuitive to current thinking that Ph is the predominant O-GlcNAc modified protein involved in gene expression in *Drosophila*, since Ph is expected to localize to genes that are repressed. This observation further supports our argument that O-GlcNAc proteins besides Ph are important in *Drosophila* development.

In conclusion, we provide the first ever TC ChIP-seq data set in a developmental specific stage of a live organism. Furthermore, this is the first ever TC ChIP-seq experiment on a PTM. Strikingly, there was a homogenous distribution of TTI values across gene features and domains containing clusters of peaks. We found that O-GlcNAc proteins binding at genes with specific function may be regulated by OGA. We therefore speculate that OGA may be specifically targeted to certain O-GlcNAc modified proteins bound to chromatin containing genes of particular ontology.

4.4 Experimental methods

4.4.1 Time course Ac_4GalNAz feeding

As mentioned, WT Oregon R and ogaKO[315] larvae were transferred and/or collected after 36 hours growth on Ac_4GalNAz or non-Ac_4GalNAz containing corn meal food. The no feed control was grown on non-Ac_4GalNAz containing food for 36 hours and flash frozen. The time course Ac_4GalNAz feeding experiment contained

Ac_4GalNAz fed larvae 36 hour since their initial lay. Ac_4GalNAz fed larvae fed on Ac_4GalNAz food for 36 hours and were then transferred to non-Ac_4GalNAz containing corn meal for 0, 4, 8, 12, and 36 hours. All larvae were flash frozen. Parents were allowed to mate for 2 hours on food. A 2 hour pre-lay was performed to increase the synchronization of embryos.

4.4.2 *In vivo* formaldehyde crosslinking of *Drosophila* larvae and O-GlcNAz modified chromatin purification

The larvae (Oregon R and Oga-null) were collected and rinsed with PBST (PBS + 0.1% Triton X-100) three times to remove adhering corn food. One volume of hexanes (a mixture of isomers) was equilibrated for 30 min against 0.175 vol of 37% formaldehyde and 0.130 vol of 10X PBS (1.37 M NaCl, 27 mM KCl, 43 mM Na_2HPO_4 and 14 mM KH_2PO_4), pH 7.5, to make the mixture 5% formaldehyde and 1X PBS. Only the upper organic phase was used for crosslinking. Larvae were fixed in 10 ml of buffered 5% formaldehyde/hexanes per gram of larvae by vigorous shaking for 5 min at room temperature. The larvae were allowed to settle and the formaldehyde/hexanes solution was removed by centrifugation (3,000 x g for 5 min at RT). The larvae were washed twice in a solution containing 1X PBS and 0.5% Triton X- 100 (0.5% PBST), using the five volumes as used for fixing. They were then stored at -80°C. Chromatin was purified by quickly thawing frozen larvae and resuspending them in buffer A (0.3 M sucrose, 15 mM NaCl, 5 mM $MgCl_2$, 15 mM Tris pH 7.5, 60 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 0.5 mM dithiothreitol (DTT) and 1 mM phenylmethylsulfonyl fluoride (PMSF)). They were homogenized with 20-25 strokes of a hand-held type B Dounce homogenizer. Triton X-100 was added to 0.3%, and the homogenate was centrifuged (2,000 x g for 15 min at 4°C). The pellet was resuspended in buffer B (100 mM NaCl, 10 mM Tris pH 7.9, 1 mM EDTA, 0.1% v/v NP-40 and 1 mM PMSF) and homogenized with ~5 strokes of a hand-held type B Dounce homogenizer. The homogenate was then sheared by 20 cycles (20 × 20 s on and 50 s off cycles, 40% power settings) through a sonicator (Sonic Dismembrator Model 500, Fisher Scientific) to an average size of ~200 to 700 bp chromatin fragments. After sonication, debris was removed by centrifugation at 4°C for 10 min at 13,000 r.p.m., and an equal amount of 6 M urea was added to the solution and then incubated for 10 min at 4°C on a rotating wheel. The soluble chromatin was dialyzed using a membrane with a molecular weight cut-off of 3.5 kDa, at 4°C against 2 L dialysis buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 4% glycerol) overnight. Insoluble debris was removed by quick centrifugation (7,500 r.p.m., 2 min). The chromatin solution was pre-cleared with 100 µl Streptavidin-agarose slurry

(Sigma) beads for O-GlcNAz modified or vehicle-only chromatin separately and incubated for 1 hour at 4°C on a rotating wheel. The pre-cleared chromatin was aliquoted after centrifugation at 4°C for 2 min at 7,500 r.p.m., snap-frozen in liquid nitrogen and stored at −80°C for future use. For click chemistry, 20-50 µg chromatin DNA was incubated with iodoacetamide (IAA) to a final concentration of 15 mM, agitate mildly for 30 min at RT then added DMSO solution of dibenzylcyclooctyne-S-S-PEG3-Biotin (DBCO-S-S-PEG3-Biotin, Jena Bioscience GmbH) to a final concentration of 40 µM. The samples were protected from light and agitated mildly for 1 hour at RT. The unreacted probe was filtered off by the Amicon Ultra-0.5 mL centrifugal filter (15K cut-off) (EMD Millipore). Biotinylated-probe-chromatin complexes were captured by incubation with 50 µl Streptavidin magnetic beads (NEB) at 4°C for 2 hours. Beads were washed with 0.5 ml of the following buffers: three washes with low-salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 150 mM NaCl), followed by three washes with high-salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 500 mM NaCl), then three washes with lithium wash buffer (0.25 M LiCl, 1% NP-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.0); then two washes with TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA). Beads were resuspended in 100 µl TE containing 0.5% SDS and incubated with 0.15 mg/ml proteinase K at 55°C for 3-4 hours. DNA-protein complex were decrosslinked by incubating overnight at 65°C (not more than 15 hours). ChIP-DNA was purified with QIAquick PCR Purification Kit and stored at -20°C.

4.4.3 RT-PCR

Semi-quantitative PCR analyses using a primer set specific for the last exon in OGA mRNA was performed in 10 µL, using 3 ng DNA per reaction and 12 pmol of each primer. The PCR conditions were: 1 min 94°C initial denaturation, followed by 25 cycles of 94°C for 30 sec, 52°C for 35 sec, 72°C for 30 sec, with a final extension of 2 min at 72°C and storage at 16°C. The following PCR primer pairs were used:
5’-TACGACGAGAGTAACCGTATCA-3’
5’-ATCAACACGGCAGGGAAG-3’

4.4.4 Library preparation and sequencing

The libraries were prepared according to the manufacturer’s instructions using the NEBNext kit (E6200). Briefly, DNA was fragmented by sonication to a maximum of 300 bp. Next, the ends of the fragments were repaired with a combination of fill-in reactions and exonuclease activity to produce blunt ends that

were then tailed with an A-base. Illumina-specific adaptors were ligated followed by removal of unligated adaptors using AMpure XP beads (Beckman Coulter). Finally a PCR with 12-15 cycles for both kits was performed to enrich final adaptor-ligated fragments. Quality and quantity were assessed on High Sensitivity dsDNA Agilent Bioanalyzer Chips and Quibit. WT libraries were sequenced on the MiSeq platform, using v2, 300-cycle reagent kits (Illumina). OgaKO libraries were sequenced on the MiSeq platform, using v3, 150-cycle reagent kits (Illumina).

4.4.5 Sequence alignment and pre-processing

ChIP-seq reads were aligned to the *D. melanogaster* reference dm3 genome using the Burrows-Wheeler Aligner[130] or SNAP[300]. BAM-format alignment files were generated (along with proper index files) and duplicate read-pairs were removed using Samtools[278].

4.4.6 Peak calling

MACS 1.4.2[280] (p-value = 0.05) was used to call peaks with 0 hour larvae as treatment file and no feed larvae as control file. Top percent enriched peaks were calculated using a custom made C++ script that called out to Samtools for depth calculations. The top 5 and 10 percent of 500 bp bins with highest scoring normalized sequencing depth ratios (0 hour / no feed) were considered peaks. For 5X enriched peaks, a custom made C++ script calculated the average sequencing depth of WT and ogaKO larvae at 0 hour and no feed. Peaks that were 5 times greater than the average for 0 hour and no feed were considered peaks. No feed peaks in WT and ogaKO were removed from 0 hour peaks for each genotype. The resulting filtered peaks were overlapped in WT and ogaKO to give the final number of peaks.

4.4.7 Time course ChIP-seq analysis

TDCA (www.github.com/TimeDependentChipSeqAnalyser/TDCA) was used to analyse data with the -nonorm flag. UCSC browser[304] was used to visualize tracks. MEME-ChIP[309] and tomtom[316] were used to generate motifs and predict motif binding proteins, respectively.

4.4.8 Flybase stage specific gene expression

Stage specific *Drosophila* genes were curated from FlyBase[317]. The expression of O-GlcNAz bound genes in larvae and pupae were then compared to all

genes at those stages (observed / expected). Chi-squared tests were performed accordingly.

Chapter 5: **Future directions**

Overall, this thesis highlights pivotal epigenetic roles of OGT and O-GlcNAc in *Drosophila* and describes the use of novel methodologies to study such phenomena. The chemical ChIP-seq strategy provided should be broadly applicable as novel and currently available probes are discovered/characterized. Furthermore, we anticipate that time course ChIP-seq using bioorthogonal probes will provide new insights into the binding kinetics of target proteins and DNA. Moreover, our program to process and analyze ChIP-seq data should prove useful for a broad range of time course studies. Thus, the studies should have a major impact on a broad range of scientific fields. A brief discussion of future directions follows, including advice for experimental design, coupled with some interesting observations.

5.1 O-GlcNAc regulation of proteins at a substrate and epigenetic level

Previous O-GlcNAc proteome studies using Ac_4GalNAz metabolic labeling in S2 cells revealed 501 unique proteins (see Tom Clark's MSc thesis, SFU 2014). Interestingly, there is some overlap between proteins that are apparently O-GlcNAc substrates and whose genomic loci contain O-GlcNAc bound proteins based on Ac_4GalNAz metabolic labeling followed by ChIP-seq. Figure 5.1 shows the overlap of O-GlcNAc substrates in S2 cells and genes with O-GlcNAc bound proteins in S2 cells, larvae, and pupae.

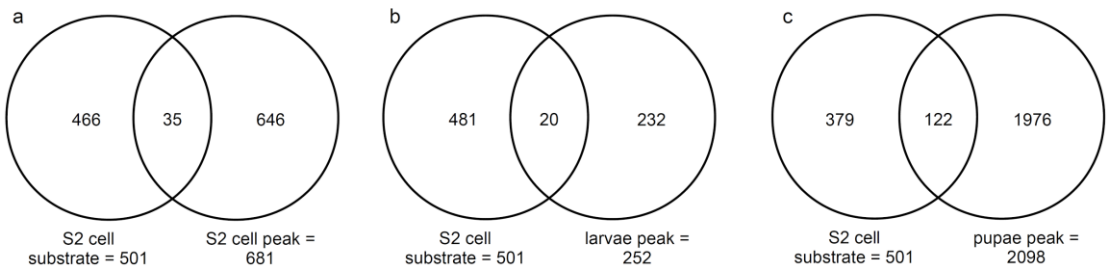


Figure 5.1. Overlap of O-GlcNAc substrates in S2 cells and genes with O-GlcNAc bound proteins in S2 cells, larvae, and pupae.
(a) Overlap of O-GlcNAc substrates in S2 cells (left) and O-GlcNAc ChIP-seq peaks in S2 cells (right). (b) Overlap of O-GlcNAc substrates in S2 cells (left) and O-GlcNAc ChIP-seq peaks in larvae (right). (c) Overlap of O-GlcNAc substrates in S2 cells (left) and O-GlcNAc ChIP-seq peaks in pupae (right).

This interesting observation suggests that a fraction of proteins are regulated at both an epigenetic and PTM level by OGT. Clearly this is of interest for future studies.

5.2 Detailed investigations of genes containing O-GlcNAc bound proteins

Since our results show that many loci contain O-GlcNAc bound proteins in S2 cells, larvae, and pupae (see Figure 4.17 a), it would be worthwhile to further investigate the regulation of these genes in detail. For example, many carbohydrate metabolic pathway genes were found to contain O-GlcNAc bound proteins. Enzymes involved in the HBP were probed for differential expression in wild type and *sxc*$^{-/-}$ pupae as well as in 4HT inducible OGT knockout MEF cells[166] (Figure 5.2). There was a pronounced mis-expression of genes in fly and mis-expression was also noted for some of the mouse homologues.

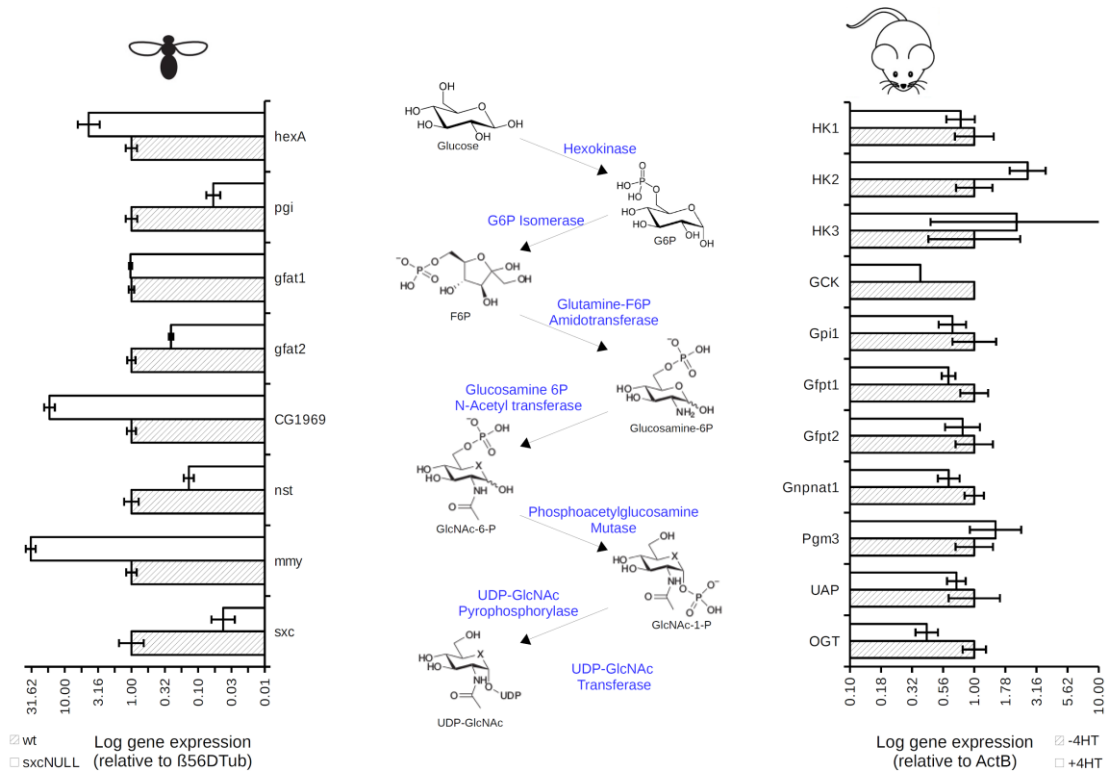


Figure 5.2. qPCR of HBP genes in *Drosophila* pupae and mouse cells.
Left: qPCR of HBP genes in wild type and *sxc*$^{-/-}$ pupae. Middle: HBP pathway. Right: qPCR of HBP genes in 4HT inducible OGT knockout MEF cells[166]. +4HT indicates 4HT treatment and -4HT is no 4HT treatment.

Especially striking is the global mis-expression of *Drosophila* genes encoding Glycolytic enzymes observed for *sxc*$^{-/-}$ versus wildtype pupae (Figure 5.3). Interestingly, many enzymes involved in the HBP, glycolysis, citric acid cycle, and

pentose phosphate pathway are O-GlcNAc substrates in S2 cells (Tom Clark's MSc thesis, SFU 2014). Previous studies have shown that glucokinase is modified and regulated by O-GlcNAc in mouse[319]. Our data raise the possibility of global regulation of carbohydrate metabolism at the PTM and epigenetic levels by OGT. Clearly this question warrants further study.

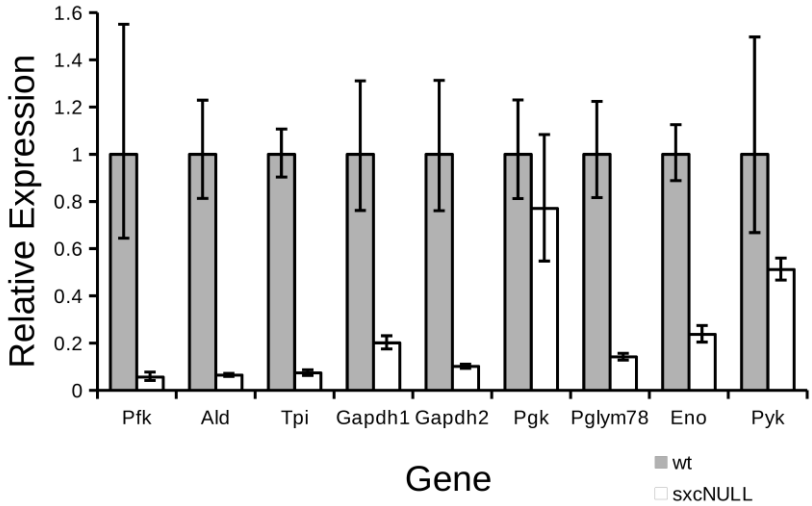


Figure 5.3. **qPCR of glycolysis genes in *Drosophila* pupae.**
qPCR of glycolysis genes in wild type and *sxc*$^{-/-}$ pupae.

5.3 Investigating the proteomes of O-GlcNAc loci

A key question pertaining to the results of this work is the identity of proteins bound at O-GlcNAc peaks. The identification and characterization of relevant O-GlcNAc-modified transcription factors could lead to a more complete understanding of underlying molecular mechanisms of O-GlcNAc epigenetic regulation. In this regard, some genomic loci present intriguing binding patterns; for example, the 100kb O-GlcNAc peak surrounding the genes CG1969 and Dop1R2 in S2 cells, larvae, and pupae (Figure 5.4). Investigating the proteins bound at such loci would be of significant interest. For these reasons, it is worthwhile to consider some possible strategies to delineate the protein identity at O-GlcNAc loci. Some possible methods tackling this topic are addressed below.

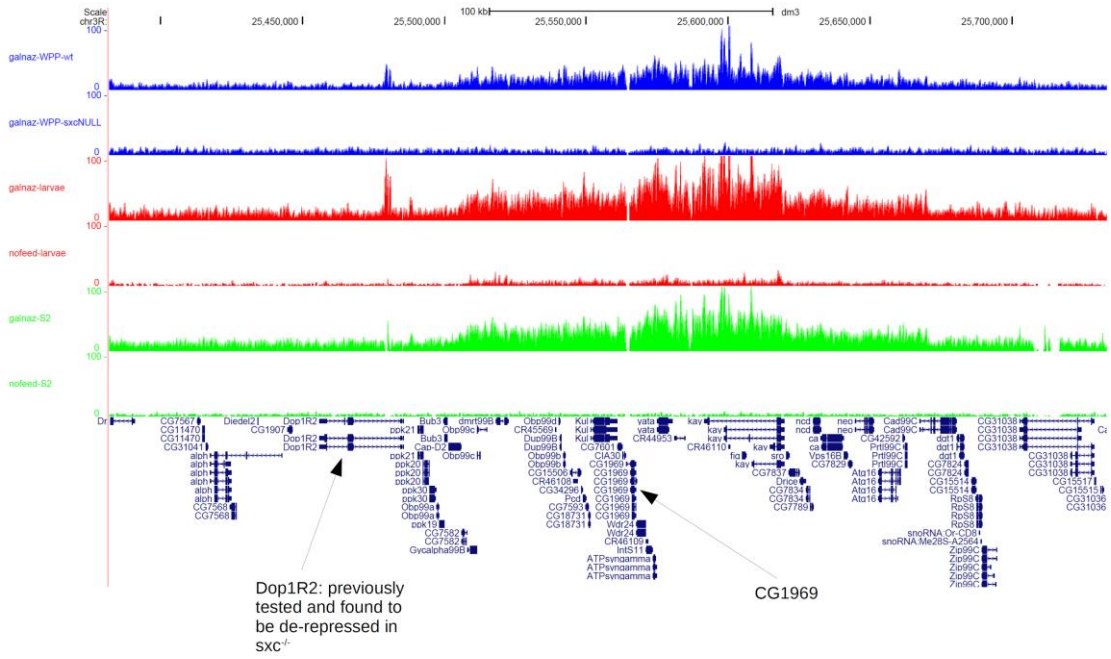


Figure 5.4. **100 kb O-GlcNAc locus in S2 cells, larvae and pupae.** ChIP-seq in Ac_4GalNAz fed wild type *Drosophila* pupae (top blue), Ac_4GalNAz fed *sxc*$^{-/-}$ *Drosophila* pupae (bottom blue), Ac_4GalNAz fed *Drosophila* larvae (top red), non-Ac_4GalNAz fed *Drosophila* larvae (bottom red), Ac_4GalNAz fed S2 cells (top green), non-Ac_4GalNAz fed S2 cells (bottom green). Genes shown in blue at bottom with boxes as exons and arrowed lines as introns.

Previously, techniques that distinguish multi-modification states of histones at single nucleosomes used the method of linking nucleosomes to a slide and treating them with fluorescent antibodies for different histone modifications[80]. The x,y coordinates of fluorescence of each antibody could then be used to pinpoint multi-modification states of histones of single nucleosomes. The platform for this experiment was designed such that DNA sequencing by synthesis could be done directly on the slide, enabling the correlation of DNA sequences to types of histone modifications. In order to adapt this approach to the identification of O-GlcNAc-modified DNA bound proteins, one would require an extensive panel of known O-GlcNAc protein antibodies. Such a panel is currently unavailable.

As an alternative, one could attempt to fractionate/purify O-GlcNAc-bound proteins, using standard methods (e.g. size, charge etc). If proteins were crosslinked to DNA prior to fractionation, one could make a library of each fraction and sequence the DNA. Each fraction could then be subjected to mass spectrometry analysis in order to gain insight into the protein identities. However, since the ultimate objective would be to fractionate O-GlcNAc proteins to single protein resolution, this could require hundreds of fractions. Furthermore, it is likely that in some cases, complexes of proteins would be purified together from a given locus, making it difficult to identify single proteins with precision. Nevertheless, this strategy might provide some

important insight into the identity of O-GlcNAc-modified proteins whose binding to genomic sites has been revealed in the current work.

To simplify matters, a researcher could define the loci bound by a known O-GlcNAc protein in its modified and unmodified states - *i.e.* case studies. For example, Ph is known to bind to DNA with a similar profile to WGA ChIP-seq. Given that O-GlcNAc is cycled on and off proteins, it could be that Ph is bound to DNA in O-GlcNAc modified and unmodified states. Using a double immunoprecipitation strategy whereby O-GlcNAcylated proteins are purified and re-purified with an antibody for an OGT substrate of choice, such as Ph, one could de-convolute the loci at which an OGT substrate is bound in modified and unmodified states. This would enable resolving genomic loci that any known OGT substrate binds in O-GlcNAc modified and un-modified state. This double immunoprecipitation strategy is highly targeted and should only be undertaken if one is certain that a binary O-GlcNAc system exists (modified and unmodified states of an O-GlcNAc substrate both bind to DNA).

Alternatively, one could expand a chemical labeling, as well as TC ChIP-seq approach by using AHA combined with antibodies, to target a broad spectrum of proteins. AHA labels all newly translated proteins with a targetable methionine azide homologue. One could purify newly translated proteins and re-purify a specific protein with an antibody of choice from either cells or animals treated with AHA, thereby allowing TC ChIP-seq on a specific protein. To bypass double immunoprecipitations, one could design an antibody for a small molecule clicked onto a protein of choice.

Another interesting pursuit might be to investigate the proteome of individual loci containing O-GlcNAc bound proteins. A brief discussion of available techniques for such studies is highlighted below.

Targeted chromatin purification (TChP) requires genetic insertion of a sequence that will bind to an exogenously expressed protein near a locus of interest[320]. Although TChP requires genetically engineered cells, the strategy has reported protein identification at a single locus[320]. This is a similar strategy to the older chromatin affinity purification with mass spectrometry (ChAP-MS), where an engineered LexA binding site near a locus of choice is used as a bait for LexA which is affinity purified, enabling detection of loci specific proteomes[321].

Proteomics of isolated chromatin segments (PICh) uses locked nucleic acid probes, which are oligonucleotides with an altered backbone that have increased stability as hybridized strands[322]. PICh has succeeded in defining the proteome of telomeric binding proteins, which are about 100 fold greater in abundance than a

single gene locus. A strategy called hybridization capture of chromatin associated proteins for proteomics (HyCCAPP), claims that desthiobiotin oligomer probes, which hybridize single gene loci, are sufficient for the identification of proteins at target loci[323]. A modified version of PICh, where chromatin is digested with restriction enzymes rather than sheared through sonication, also claims to permit proteomics at a single locus[324].

In another study, meganucleases were used to cleave a specific fragment of DNA containing an engineered bait sequence, which was then probed for by an engineered protein bait interaction. This allowed the identification of proteins bound to targeted loci[325]. More recently, researchers have developed an approach called CRISPR-based chromatin affinity purification with mass spectrometry (CRISPR-ChAP-MS)[326]. CRISPR-ChAP-MS uses guide RNA to target a catalytically inactive tagged Cas9 to a specific locus. Cross-linking and purification with an antibody for the Cas9 tag enables identification of proteins associated at a single locus. These techniques could be experimented with an attempt to purify proteins at a single locus.

5.4 TDCA maintenance and expansion

With minor modifications, many bioinformatics methods can be readily adapted for other relevant research. For example, TDCA could be applied to other types of ChIP-seq experiments, such as dose-response ChIP-seq, with only minor changes in the existing code. Alternatively, TDCA could be optimized with respect to speed and/or fitted with a graphical user interface. One innovation that would maximize speed would be to replace the dependency drc, since this is the slowest step in the process. It is common for additional versions of software to be published in brief communications in reputable journals. Additionally, the scientific community would be more inclined to use reliable, well written, easy to use software.

In addition, time course ChIA-PET or HiC analysis software would be an interesting route for new developments as these are intrinsically difficult data to analyse and communicate with others. Analysis software for this long distance DNA-DNA interaction and DNA-protein interaction data would allow for novel insights in chromatin architecture.

5.5 Conclusion

The novel chemical biology approach to study the epigenetics of O-GlcNAc modified proteins highlighted here should contribute significantly to the repertoire of tools for studying this interesting protein modification. Furthermore, the approach developed here could be used as a guide for other protein modification, thereby helping to provide insight into a variety of important biological questions.

References

1. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561–3 (1970).
2. Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.* **122,** 565–581 (2008).
3. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids. *Nature* **171,** 737–738 (1953).
4. Watson, J. D. & Crick, F. H. C. Genetic Implications of the structure of Deoxyribonucleic Acid. *Nature* **171,** 964–967 (1953).
5. McCarty, M. & Avery, O. T. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Ii. Effect of Desoxyribonuclease on the Biological Activity of the Transforming Substance. *The Journal of experimental medicine* **83,** 89–96 (1946).
6. Hershey, A. D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* **36,** 39–56 (1952).
7. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).
8. Sanger, F., Nicklen, S. & Coulson, a R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–7 (1977).
9. Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J. & Hood, L. E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* **13,** 2399–412 (1985).
10. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321,** 674–679 (1986).
11. Canard, B. & Sarfati, R. S. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene* **148,** 1–6 (1994).
12. Saiki, R. K. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230,** 1350–4 (1985).
13. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
14. Guo, J. *et al.* Four-colour DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 9145–50 (2008).
15. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
16. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281,** 363, 365 (1998).
17. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 13770–3 (1996).
18. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–6 (2003).
19. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–8 (2009).
20. Towbin, H., Staehelin, T. & Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 4350–4 (1979).

21. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
22. Rogler, C. E. *et al.* RNA expression microarrays (REMs), a high-throughput method to measure differences in gene expression in diverse biological samples. *Nucleic Acids Res.* **32,** e120 (2004).
23. Morin, R. D. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45,** 81–94 (2008).
24. Letsou, A. & Bohmann, D. Small flies?Big discoveries: Nearly a century ofDrosophila genetics and development. *Dev. Dyn.* **232,** 526–528 (2005).
25. Rubin, G. M. A Brief History of Drosophila's Contributions to Genome Research. *Science (80-.).* **287,** 2216–2218 (2000).
26. Jeibmann, A. & Paulus, W. Drosophila melanogaster as a Model Organism of Brain Diseases. *Int. J. Mol. Sci.* **10,** 407–440 (2009).
27. Reiter, L. T. A Systematic Analysis of Human Disease-Associated Gene Sequences In Drosophila melanogaster. *Genome Res.* **11,** 1114–1125 (2001).
28. Venken, K. J. T. & Bellen, H. J. Emerging technologies for gene manipulation in Drosophila melanogaster. *Nat. Rev. Genet.* **6,** 167–178 (2005).
29. Venken, K. J. T. & Bellen, H. J. Chemical mutagens, transposons, and transgenes to interrogate gene function in Drosophila melanogaster. *Methods* **68,** 15–28 (2014).
30. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci.* **36,** 344–355 (1950).
31. Rubin, G. M. & Spradling, A. C. Genetic transformation of Drosophila with transposable element vectors. *Science* **218,** 348–53 (1982).
32. Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118,** 401–15 (1993).
33. Rong, Y. S. Gene Targeting by Homologous Recombination in Drosophila. *Science (80-.).* **288,** 2013–2018 (2000).
34. Gong, W. J. & Golic, K. G. Ends-out, or replacement, gene targeting in Drosophila. *Proc. Natl. Acad. Sci.* **100,** 2556–2561 (2003).
35. Turan, S. *et al.* Recombinase-Mediated Cassette Exchange (RMCE): Traditional Concepts and Current Challenges. *J. Mol. Biol.* **407,** 193–221 (2011).
36. Turan, S., Zehe, C., Kuehle, J., Qiao, J. & Bode, J. Recombinase-mediated cassette exchange (RMCE) — A rapidly-expanding toolbox for targeted genomic modifications. *Gene* **515,** 1–27 (2013).
37. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391,** 806–11 (1998).
38. Mohr, S. E. RNAi screening in Drosophila cells and in vivo. *Methods* **68,** 82–88 (2014).
39. Boutros, M. *et al.* Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science* **303,** 832–5 (2004).
40. Adams, M. D. The Genome Sequence of Drosophila melanogaster. *Science (80-.).* **287,** 2185–2195 (2000).
41. Smith, J. Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. *Nucleic Acids Res.* **28,** 3361–3369 (2000).
42. Beumer, K. J. & Carroll, D. Targeted genome engineering techniques in Drosophila. *Methods* **68,** 29–37 (2014).
43. Boch, J. *et al.* Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science (80-.).* **326,** 1509–1512 (2009).
44. Boch, J. TALEs of genome targeting. *Nat. Biotechnol.* **29,** 135–136 (2011).
45. Horvath, P. & Barrangou, R. CRISPR/Cas, the Immune System of Bacteria

and Archaea. *Science (80-.).* **327,** 167–170 (2010).
46. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (80-.).* **337,** 816–821 (2012).
47. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
48. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–51 (2001).
49. Filion, G. J. *et al.* Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* **143,** 212–224 (2010).
50. Aguilar, C. A. & Craighead, H. G. Micro- and nanoscale devices for the investigation of epigenetics and chromatin dynamics. *Nat. Nanotechnol.* **8,** 709–18 (2013).
51. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat. Struct. Mol. Biol.* **20,** 1147–55 (2013).
52. Aranda, S., Mas, G. & Di Croce, L. Regulation of gene transcription by Polycomb proteins. *Sci. Adv.* **1,** e1500737 (2015).
53. De, S. & Kassis, J. A. Passing epigenetic silence to the next generation. *Science (80-.).* **356,** 28–29 (2017).
54. Ahuja, N., Sharma, A. R. & Baylin, S. B. Epigenetic Therapeutics: A New Weapon in the War Against Cancer. *Annu. Rev. Med.* **67,** 73–89 (2016).
55. Kelly, T. K., De Carvalho, D. D. & Jones, P. A. Epigenetic modifications as therapeutic targets. *Nat. Biotechnol.* **28,** 1069–78 (2010).
56. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184,** 868–71 (1974).
57. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389,** 251–60 (1997).
58. Balhorn, R. The protamine family of sperm nuclear proteins. *Genome Biol.* **8,** 227 (2007).
59. Hatch, C. L. & Bonner, W. M. The human histone H2A.Z gene. Sequence and regulation. *J. Biol. Chem.* **265,** 15211–8 (1990).
60. Whitehouse, I. *et al.* Nucleosome mobilization catalysed by the yeast SWI/SNF complex. *Nature* **400,** 784–787 (1999).
61. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4,** e1000138 (2008).
62. Li, G., Levitus, M., Bustamante, C. & Widom, J. Rapid spontaneous accessibility of nucleosomal DNA. *Nat. Struct. Mol. Biol.* **12,** 46–53 (2005).
63. Shao, Z. *et al.* Stabilization of chromatin structure by PRC1, a Polycomb complex. *Cell* **98,** 37–46 (1999).
64. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298,** 1039–43 (2002).
65. Czermin, B. *et al.* Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell* **111,** 185–96 (2002).
66. Francis, N. J., Kingston, R. E. & Woodcock, C. L. Chromatin compaction by a polycomb group protein complex. *Science* **306,** 1574–7 (2004).
67. Wang, H. *et al.* Role of histone H2A ubiquitination in Polycomb silencing. *Nature* **431,** 873–8 (2004).
68. Cao, R., Tsukada, Y. & Zhang, Y. Role of Bmi-1 and Ring1A in H2A Ubiquitylation and Hox Gene Silencing. *Mol. Cell* **20,** 845–854 (2005).
69. Lee, M. G. *et al.* Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. *Science* **318,** 447–50 (2007).
70. Scheuermann, J. C. *et al.* Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. *Nature* **465,** 243–247 (2010).

71. Abdel-Wahab, O. & Dey, A. The ASXL–BAP1 axis: new factors in myelopoiesis, cancer and epigenetics. *Leukemia* **27,** 10–15 (2013).
72. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nat. Biotechnol.* **28,** 1057–1068 (2010).
73. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469,** 343–349 (2011).
74. Dehennaut, V., Leprince, D. & Lefebvre, T. O-GlcNAcylation, an epigenetic mark. Focus on the histone code, TET family proteins, and polycomb group proteins. *Front. Endocrinol. (Lausanne).* **5,** 1–7 (2014).
75. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476,** 298–303 (2011).
76. Laprell, F., Finkl, K. & Müller, J. Propagation of Polycomb-repressed chromatin requires sequence-specific recruitment to DNA. *Science (80-.).* **356,** 85–88 (2017).
77. Coleman, R. T. & Struhl, G. Causal role for inheritance of H3K27me3 in maintaining the OFF state of a Drosophila HOX gene. *Science (80-.).* **356,** eaai8236 (2017).
78. Wang, X. & Moazed, D. DNA sequence-dependent epigenetic inheritance of gene silencing and histone H3K9 methylation. *Science (80-.).* **356,** 88–91 (2017).
79. Ng, M. K. & Cheung, P. A brief histone in time: understanding the combinatorial functions of histone PTMs in the nucleosome context. *Biochem. Cell Biol.* **94,** 33–42 (2016).
80. Shema, E. *et al.* Single-molecule decoding of combinatorially modified nucleosomes. *Science* **352,** 717–21 (2016).
81. Maity, S. K., Jbara, M. & Brik, A. Chemical and semisynthesis of modified histones. *J. Pept. Sci.* **22,** 252–259 (2016).
82. Bracken, A. P. & Helin, K. Polycomb group proteins: navigators of lineage pathways led astray in cancer. *Nat. Rev. Cancer* **9,** 773–84 (2009).
83. Brockdorff, N. Noncoding RNA and Polycomb recruitment. *Rna* **19,** 429–42 (2013).
84. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324,** 930–5 (2009).
85. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324,** 929–30 (2009).
86. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed. Engl.* **50,** 7008–12 (2011).
87. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333,** 1300–3 (2011).
88. He, Y.-F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (80-.).* **333,** 1303–7 (2011).
89. Blaschke, K. *et al.* Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500,** 222–6 (2013).
90. Hu, X. *et al.* Tet and TDG mediate DNA demethylation essential for mesenchymal-to-epithelial transition in somatic cell reprogramming. *Cell Stem Cell* **14,** 512–22 (2014).
91. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **6,** 1049–55 (2014).
92. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11,** 1–4 (2015).
93. Raiber, E.-A. *et al.* 5-Formylcytosine alters the structure of the DNA double helix. *Nat. Struct. Mol. Biol.* **22,** 44–9 (2015).
94. Song, C. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153,** 678–691 (2013).
95. Amouroux, R. *et al.* supplement De novo DNA methylation drives 5hmC

accumulation in mouse zygotes. *Nat. Cell Biol.* **18,** 1–5 (2016).
96. Zhang, G. *et al.* N6-methyladenine DNA modification in Drosophila. *Cell* **161,** 893–906 (2015).
97. Wu, H. & Zhang, Y. Charting oxidized methylcytosines at base resolution. *Nat. Struct. Mol. Biol.* **22,** 656–661 (2015).
98. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336,** 934–7 (2012).
99. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149,** 1368–1380 (2012).
100. Xu, Y. *et al.* Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. *Mol. Cell* **42,** 451–464 (2011).
101. Tan, L. *et al.* Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural progenitor cells by a new comparative hMeDIP-seq method. *Nucleic Acids Res.* **41,** 1–12 (2013).
102. Song, C. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29,** 68–72 (2011).
103. Tuesta, L. M. & Zhang, Y. Mechanisms of epigenetic memory and addiction. *EMBO J.* **33,** 1091–103 (2014).
104. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. **33,** 5868–5877 (2005).
105. Kono, T. *et al.* Birth of parthenogenetic mice that can develop to adulthood. *Nature* **428,** 860–4 (2004).
106. Liyanage, V. R. *et al.* DNA modifications: function and applications in normal and disease States. *Biol.* **3,** 670–723 (2014).
107. Feng, J. *et al.* Role of Tet1 and 5-hydroxymethylcytosine in cocaine action. *Nat. Neurosci.* **18,** 536–544 (2015).
108. Taqi, M. M. *et al.* Prodynorphin CpG-SNPs associated with alcohol dependence: Elevated methylation in the brain of human alcoholics. *Addict. Biol.* **16,** 499–509 (2011).
109. Liu, Y., Balaraman, Y., Wang, G., Nephew, K. P. & Zhou, F. C. Alcohol exposure alters DNA methylation profiles in mouse embryos at early neurulation. *Epigenetics* **4,** 500–511 (2009).
110. Maccani, J. Z., Koestler, D. C., Houseman, E. A., Marsit, C. J. & Kelsey, K. T. Placental DNA methylation alterations associated with maternal tobacco smoking at the RUNX3 gene are also associated with gestational age. *Epigenomics* **5,** 619–630 (2013).
111. Belot, M. P. *et al.* CpG Methylation Changes within the IL2RA Promoter in Type 1 Diabetes of Childhood Onset. *PLoS One* **8,** 1–7 (2013).
112. Dayeh, T. A. *et al.* Identification of CpG-SNPs associated with type 2 diabetes and differential DNA methylation in human pancreatic islets. *Diabetologia* **56,** 1036–1046 (2013).
113. Szulwach, K. E. *et al.* 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat Neurosci* **14,** 1607–1616 (2011).
114. Feng, J. *et al.* Dnmt1 and Dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons. *Nat. Neurosci.* **13,** 423–30 (2010).
115. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479,** 74–9 (2011).
116. Weng, Y. L., An, R., Shin, J., Song, H. & Ming, G. li. DNA Modifications and Neurological Disorders. *Neurotherapeutics* **10,** 556–567 (2013).
117. Fournier, A., Sasai, N., Nakao, M. & Defossez, P. A. The role of methyl-binding proteins in chromatin organization and epigenome maintenance. *Brief. Funct. Genomics* **11,** 251–264 (2012).
118. Blat, Y. & Kleckner, N. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric

region. *Cell* **98,** 249–259 (1999).
119. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science (80-.).* **290,** 2306–2309 (2000).
120. Iyer, V. R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409,** 533–538 (2001).
121. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316,** 1497–502 (2007).
122. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823–837 (2007).
123. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4,** 651–657 (2007).
124. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
125. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295,** 1306–11 (2002).
126. Lieberman-aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-.).* **326,** 289–293 (2009).
127. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462,** 58–64 (2009).
128. Ocampo-Hafalla, M., Muñoz, S., Samora, C. P. & Uhlmann, F. Evidence for cohesin sliding along budding yeast chromosomes. *Open Biol.* **6,** 150178 (2016).
129. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–10 (1990).
130. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
131. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
132. Karve, T. M. & Cheema, A. K. Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease. *J. Amino Acids* **2011,** 207691 (2011).
133. Wang, Y.-C., Peterson, S. E. & Loring, J. F. Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res.* **24,** 143–60 (2014).
134. Holt, G. D. & Hart, G. W. The subcellular distribution of terminal N-acetylglucosamine moieties. Localization of a novel protein-saccharide linkage, O-linked GlcNAc. *J. Biol. Chem.* **261,** 8049–57 (1986).
135. Hart, G. W., Housley, M. P. & Slawson, C. Cycling of O-linked β-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* **446,** 1017–1022 (2007).
136. Caramelo, J. J. & Parodi, A. J. Getting in and out from calnexin/calreticulin cycles. *J. Biol. Chem.* **283,** 10221–10225 (2008).
137. Freire-de-Lima, L. *et al.* Sialic acid: A sweet swing between mammalian host and Trypanosoma cruzi. *Front. Immunol.* **3,** 1–12 (2012).
138. Helle, F., Duverlie, G. & Dubuisson, J. The hepatitis C virus glycan shield and evasion of the humoral immune response. *Viruses* **3,** 1909–1932 (2011).
139. Anthony, R. M. *et al.* Recapitulation of IVIG anti-inflammatory activity with a recombinant IgG Fc. *Science* **320,** 373–6 (2008).
140. Kimura, T. & Finn, O. J. MUC1 immunotherapy is here to stay. *Expert Opin. Biol. Ther.* **13,** 35–49 (2013).
141. Jaeken, J. Congenital disorders of glycosylation. *Inborn Metab. Dis. Diagnosis Treat.* **1214,** 607–616 (2012).

142. Ghazarian, H., Idoni, B. & Oppenheimer, S. B. A glycobiology review: carbohydrates, lectins and implications in cancer therapeutics. *Acta Histochem.* **113,** 236–47 (2011).
143. Dalziel, M., Crispin, M., Scanlan, C. N., Zitzmann, N. & Dwek, R. A. Emerging Principles for the Therapeutic Exploitation of Glycosylation. *Science (80-.).* **343,** 1235681–1235681 (2014).
144. Maynard, J. C., Burlingame, A. L. & Medzihradszky, K. F. Cysteine S-linked N-acetylglucosamine (S-GlcNAcylation), A New Post-translational Modification in Mammals. *Mol. Cell. Proteomics* **15,** 3405–3411 (2016).
145. Torres, C. R. & Hart, G. W. Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. *J. Biol. Chem.* **259,** 3308–17 (1984).
146. Haltiwanger, R. S., Blomberg, M. A. & Hart, G. W. Glycosylation of nuclear and cytoplasmic proteins: Purification and characterization of a uridine diphospho-N-acetylglucosamine:polypeptide ??-N-acetylglucosaminyltransferase. *J. Biol. Chem.* **267,** 9005–9013 (1992).
147. Lubas, W. A., Frank, D. W., Krause, M. & Hanover, J. A. O-linked GlcNAc transferase is a conserved nucleocytoplasmic protein containing tetratricopeptide repeats. *J. Biol. Chem.* **272,** 9316–9324 (1997).
148. Lazarus, M. B., Nam, Y., Jiang, J., Sliz, P. & Walker, S. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* **469,** 564–567 (2011).
149. Shafi, R. *et al.* The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 5735–9 (2000).
150. O'Donnell, N., Zachara, N. E., Hart, G. W. & Marth, J. D. Protein Glycosylation Is a Requisite Modification in Somatic Cell Function and Embryo Viability Ogt - Dependent X-Chromosome-Linked Protein Glycosylation Is a Requisite Modification in Somatic Cell Function and Embryo Viability. *Mol Cell Biol* **24,** 1680–1690 (2004).
151. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43,** D512–D520 (2015).
152. Lazarus, M. B. *et al.* HCF-1 is cleaved in the active site of O-GlcNAc transferase. *Science* **342,** 1235–9 (2013).
153. Gambetta, M. C., Oktaba, K. & Muller, J. Essential Role of the Glycosyltransferase Sxc/Ogt in Polycomb Repression. *Science (80-.).* **325,** 93–96 (2009).
154. Vella, P. *et al.* Tet Proteins Connect the O-Linked N-acetylglucosamine Transferase Ogt to Chromatin in Embryonic Stem Cells. *Mol. Cell* **49,** 645–656 (2013).
155. Shi, F. T. *et al.* Ten-eleven translocation 1 (Tet1) is regulated by o-linked n-acetylglucosamine transferase (ogt) for target gene repression in mouse embryonic stem cells. *J. Biol. Chem.* **288,** 20776–20784 (2013).
156. Zhang, Q. *et al.* Differential regulation of the ten-eleven translocation (TET) family of dioxygenases by O-linked β-N-acetylglucosamine transferase (OGT). *J. Biol. Chem.* **289,** 5986–5996 (2014).
157. Kapuria, V. *et al.* Proteolysis of HCF-1 by Ser / Thr glycosylation-incompetent O -GlcNAc transferase : UDP-GlcNAc complexes. 960–972 (2016). doi:10.1101/gad.275925.115.
158. Vocadlo, D. J. O-GlcNAc processing enzymes: catalytic mechanisms, substrate specificity, and enzyme regulation. *Curr. Opin. Chem. Biol.* **16,** 488–97 (2012).
159. Love, D. C. Mitochondrial and nucleocytoplasmic targeting of O-linked GlcNAc transferase. *J. Cell Sci.* **116,** 647–654 (2002).
160. Coomer, M. & Essop, M. F. Differential hexosamine biosynthetic pathway

gene expression with type 2 diabetes. *Mol. Genet. Metab. Reports* **1,** 158–169 (2014).
161. Jacobsen, S. E. & Olszewski, N. E. Mutations at the SPINDLY locus of Arabidopsis alter gibberellin signal transduction. *Plant Cell* **5,** 887–896 (1993).
162. Zhu, Y. *et al.* O-GlcNAc occurs cotranslationally to stabilize nascent polypeptide chains. *Nat. Chem. Biol.* **11,** 319–25 (2015).
163. Gambetta, M. C. & Müller, J. O-GlcNAcylation Prevents Aggregation of the Polycomb Group Repressor Polyhomeotic. *Dev. Cell* **31,** 629–639 (2014).
164. Yuzwa, S. A. *et al.* Increasing O-GlcNAc slows neurodegeneration and stabilizes tau against aggregation. *Nat. Chem. Biol.* **8,** 393–9 (2012).
165. Gao, Y., Miyazaki, J. I. & Hart, G. W. The transcription factor PDX-1 is post-translationally modified by O-linked N-acetylglucosamine and this modification is correlated with its DNA binding activity and insulin secretion in min6 ??-cells. *Arch. Biochem. Biophys.* **415,** 155–163 (2003).
166. Kazemi, Z., Chang, H., Haserodt, S., McKen, C. & Zachara, N. E. O-Linked -N-acetylglucosamine (O-GlcNAc) Regulates Stress-induced Heat Shock Protein Expression in a GSK-3 -dependent Manner. *J. Biol. Chem.* **285,** 39096–39107 (2010).
167. Itkonen, H. M. *et al.* O-GlcNAc transferase integrates metabolic pathways to regulate the stability of c-MYC in human prostate cancer cells. *Cancer Res.* **73,** 5277–5287 (2013).
168. de Queiroz, R. M., Carvalho, Ã. & Dias, W. B. O-GlcNAcylation: The Sweet Side of the Cancer. *Front. Oncol.* **4,** 132 (2014).
169. Yi, W. *et al.* Phosphofructokinase 1 Glycosylation Regulates Cell Growth and Metabolism. *Science (80-.).* **337,** 975–980 (2012).
170. Ferrer, C. M. *et al.* O-GlcNAcylation regulates cancer metabolism and survival stress signaling via regulation of the HIF-1 pathway. *Mol. Cell* **54,** 820–31 (2014).
171. Champattanachai, V., Marchase, R. B. & Chatham, J. C. Glucosamine protects neonatal cardiomyocytes from ischemia-reperfusion injury via increased protein O-GlcNAc and increased mitochondrial Bcl-2. *Am. J. Physiol. Cell Physiol.* **294,** C1509–C1520 (2008).
172. Laczy, B., Marsh, S. A., Brocks, C. A., Wittmann, I. & Chatham, J. C. Inhibition of O -GlcNAcase in perfused rat hearts by NAG-thiazolines at the time of reperfusion is cardioprotective in an O -GlcNAc-dependent manner. *Am J Physiol Hear. Circ Physiol* **299,** H1715–H1727 (2010).
173. Jones, S. P. *et al.* Cardioprotection by N-acetylglucosamine linkage to cellular proteins. *Circulation* **117,** 1172–1182 (2008).
174. Fülöp, N., Zhang, Z., Marchase, R. B. & Chatham, J. C. Glucosamine cardioprotection in perfused rat hearts associated with increased O-linked N-acetylglucosamine protein modification and altered p38 activation. *Am. J. Physiol. Hear. Circ. Physiol.* **292,** H2227-36 (2007).
175. Yuzwa, S. A. *et al.* A potent mechanism-inspired O-GlcNAcase inhibitor that blocks phosphorylation of tau in vivo. *Nat. Chem. Biol.* **4,** 483–90 (2008).
176. Wilson, I. B., Gavel, Y. & von Heijne, G. Amino acid distributions around O-linked glycosylation sites. *Biochem. J.* **275 (Pt 2,** 529–534 (1991).
177. Gupta, R. & Brunak, S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput.* **322,** 310–22 (2002).
178. Chen, S.-A., Lee, T.-Y. & Ou, Y.-Y. Incorporating significant amino acid pairs to identify O-linked glycosylation sites on transmembrane proteins and non-transmembrane proteins. *BMC Bioinformatics* **11,** 536 (2010).
179. Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D. & Honavar, V. Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics* **8,** 438 (2007).

180. Hamby, S. E. & Hirst, J. D. Prediction of glycosylation sites using random forests. *BMC Bioinformatics* **9,** 500 (2008).
181. Li, F. *et al.* GlycoMine: A machine learning-based approach for predicting N-, C-and O-linked glycosylation in the human proteome. *Bioinformatics* **31,** 1411–1419 (2015).
182. Wang, J., Torii, M., Liu, H., Hart, G. W. & Hu, Z.-Z. dbOGAP - an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinformatics* **12,** 91 (2011).
183. Jia, C.-Z., Liu, T. & Wang, Z.-P. *O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. Molecular bioSystems* **9,** 2909–13 (2013).
184. Kao, H.-J. *et al.* A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs. *BMC Bioinformatics* **16,** S10 (2015).
185. Wu, H.-Y. *et al.* Characterization and identification of protein O-GlcNAcylation sites with substrate specificity. *BMC Bioinformatics* **15,** S1 (2014).
186. Lee, J.-S. & Zhang, Z. O-linked N-acetylglucosamine transferase (OGT) interacts with the histone chaperone HIRA complex and regulates nucleosome assembly and cellular senescence. *Proc. Natl. Acad. Sci. U. S. A.* 1600509113- (2016). doi:10.1073/pnas.1600509113
187. Chen, Q., Chen, Y., Bian, C., Fujiki, R. & Yu, X. TET2 promotes histone O-GlcNAcylation during gene transcription. *Nature* **493,** 561–4 (2013).
188. Carbone, M. *et al.* BAP1 and cancer. *Nat. Rev. Cancer* **13,** 153–9 (2013).
189. Su, M.-G. & Lee, T.-Y. Incorporating substrate sequence motifs and spatial amino acid composition to identify kinase-specific phosphorylation sites on protein three-dimensional structures. *BMC Bioinformatics* **14 Suppl 1,** S2 (2013).
190. Snow, C. M., Senior, A. & Gerace, L. Monoclonal antibodies identify a group of nuclear pore complex glycoproteins. *J. Cell Biol.* **104,** 1143–56 (1987).
191. Comer, F. I., Vosseller, K., Wells, L., Accavitti, M. A. & Hart, G. W. Characterization of a Mouse Monoclonal Antibody Specific for O-Linked N-Acetylglucosamine. *Anal. Biochem.* **293,** 169–177 (2001).
192. Isono, T. O-glcnac-specific antibody CTD110.6 cross-reacts with N-GlcNAc2-modified proteins induced under glucose deprivation. *PLoS One* **6,** (2011).
193. Teo, C. F. *et al.* Glycopeptide-specific monoclonal antibodies suggest new roles for O-GlcNAc. *Nat. Chem. Biol.* **6,** 338–343 (2010).
194. Burger, M. M. & Goldberg, a R. Identification of a tumor-specific determinant on neoplastic cell surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **57,** 359–366 (1967).
195. Monsigny, M., Roche, A. C., Sene, C., Maget-Dana, R. & Delmotte, F. Sugar-lectin interactions: how does wheat-germ agglutinin bind sialoglycoconjugates? *Eur. J. Biochem.* **104,** 147–53 (1980).
196. Ma, J. & Hart, G. W. O-GlcNAc profiling: from proteins to proteomes. *Clin. Proteomics* **11,** 8 (2014).
197. Leickt, L., Bergström, M., Zopf, D. & Ohlson, S. Bioaffinity chromatography in the 10 mM range of Kd. *Anal.Biochem.* **253,** 135–136 (1997).
198. Lienemann, M. *et al.* Characterization of the wheat germ agglutinin binding to self-assembled monolayers of neoglycoconjugates by AFM and SPR. *Glycobiology* **19,** 633–643 (2009).
199. Griffin, B. A., Adams, S. R. & Tsien, R. Y. Specific covalent labeling of recombinant protein molecules inside live cells. *Science* **281,** 269–72 (1998).
200. Kiick, K. L., Saxon, E., Tirrell, D. A. & Bertozzi, C. R. Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proc. Natl. Acad. Sci.* **99,** 19–24 (2002).
201. Wang, R. *et al.* Profiling genome-wide chromatin methylation with engineered posttranslation apparatus within living cells. *J. Am. Chem. Soc.* **135,** 1048–

1056 (2013).
202. Saxon, E. *et al.* Investigating cellular metabolism of synthetic azidosugars with the Staudinger ligation. *J. Am. Chem. Soc.* **124,** 14893–14902 (2002).
203. Kho, Y. *et al.* A tagging-via-substrate technology for detection and proteomics of farnesylated proteins. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 12479–84 (2004).
204. Sletten, E. M. & Bertozzi, C. R. Bioorthogonal Reactions. *Acc. Chem. Res.* **44,** 666–676 (2011).
205. Prescher, J. a & Bertozzi, C. R. Chemistry in living systems. *Nat. Chem. Biol.* **1,** 13–21 (2005).
206. Saxon, E. & Bertozzi, C. R. Cell Surface Engineering by a Modified Staudinger Reaction. *Science (80-.).* **287,** 2007–2010 (2000).
207. Laughlin, S. T. & Bertozzi, C. R. Metabolic labeling of glycans with azido sugars and subsequent glycan-profiling and visualization via Staudinger ligation. *Nat. Protoc.* **2,** 2930–2944 (2007).
208. McKay, C. S. & Finn, M. G. Click Chemistry in Complex Mixtures: Bioorthogonal Bioconjugation. *Chem. Biol.* **21,** 1075–1101 (2014).
209. Agard, N. J., Prescher, J. A. & Bertozzi, C. R. A strain-promoted [3 + 2] azide-alkyne cycloaddition for covalent modification of biomolecules in living systems. *J. Am. Chem. Soc.* **126,** 15046–15047 (2004).
210. Vocadlo, D. J., Hang, H. C., Kim, E., Hanover, J. A. & Bertozzi, C. R. A chemical approach for identifying O-GlcNAc-modified proteins in cells. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 9116–21 (2003).
211. Gloster, T. M. *et al.* Hijacking a biosynthetic pathway yields a glycosyltransferase inhibitor within cells. *Nat. Chem. Biol.* **7,** 174–181 (2011).
212. Gurcel, C. *et al.* Identification of new O-GlcNAc modified proteins using a click-chemistry-based tagging. *Anal. Bioanal. Chem.* **390,** 2089–2097 (2008).
213. Hang, H. C., Yu, C., Kato, D. L. & Bertozzi, C. R. A metabolic labeling approach toward proteomic analysis of mucin-type O-linked glycosylation. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 14846–14851 (2003).
214. Boyce, M. *et al.* Metabolic cross-talk allows labeling of O-linked beta-N-acetylglucosamine-modified proteins via the N-acetylgalactosamine salvage pathway. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 3141–6 (2011).
215. Liu, T.-W. *et al.* Genome-wide chemical mapping of O-GlcNAcylated proteins in Drosophila melanogaster. *Nat. Chem. Biol.* **13,** 1–26 (2016).
216. Laughlin, S. T. & Bertozzi, C. R. In vivo imaging of Caenorhabditis elegans glycans. *ACS Chem. Biol.* **4,** 1068–1072 (2009).
217. Li, J. *et al.* An OGA-resistant probe allows specific visualization and accurate identification of O-GlcNAc-modified proteins in cells. *ACS Chem. Biol.* **11,** 3002–3006 (2016).
218. Yu, S.-H. *et al.* Metabolic labeling enables selective photocrosslinking of O-GlcNAc-modified proteins to their binding partners. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 4834–4839 (2012).
219. Fang, B. & Miller, M. W. Use of galactosyltransferase to assess the biological function of O-linked N-acetyl-d-glucosamine: a potential role for O-GlcNAc during cell division. *Exp. Cell Res.* **263,** 243–253 (2001).
220. Khidekel, N. *et al.* A Chemoenzymatic Approach toward the Rapid and Sensitive Detection of O -GlcNAc Posttranslational Modifications. *J. Am. Chem. Soc.* **125,** 16162–16163 (2003).
221. Clark, P. M. *et al.* Direct In-Gel Fluorescence Detection and Cellular Imaging of *O* -GlcNAc-Modified Proteins. *J. Am. Chem. Soc.* **130,** 11576–11577 (2008).
222. Wang, Z. *et al.* Enrichment and site mapping of O-linked N-acetylglucosamine by a combination of chemical/enzymatic tagging, photochemical cleavage, and electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics* **9,** 153–60 (2010).

223. Pathak, S. *et al.* The active site of O-GlcNAc transferase imposes constraints on substrate sequence. *Nat. Struct. Mol. Biol.* **22,** 744–750 (2015).
224. Shi, J., Sharif, S., Ruijtenbeek, R. & Pieters, R. J. Activity Based High-Throughput Screening for Novel O-GlcNAc Transferase Substrates Using a Dynamic Peptide Microarray. *PLoS One* **11,** e0151085 (2016).
225. Mariappa, D. *et al.* Dual functionality of O -GlcNAc transferase is required for Drosophila development. *Open Biol.* **5,** 150234 (2015).
226. Iyer, S. P. N., Akimoto, Y. & Hart, G. W. Identification and cloning of a novel family of coiled-coil domain proteins that interact with O-GlcNAc transferase. *J. Biol. Chem.* **278,** 5399–5409 (2003).
227. Deng, R. P. *et al.* Global identification of O-GlcNAc transferase (OGT) interactors by a human proteome microarray and the construction of an OGT interactome. *Proteomics* **14,** 1020–1030 (2014).
228. Ingham, P. W. A gene that regulates the bithorax complex differentially in larval and adult cells of Drosophila. *Cell* **37,** 815–23 (1984).
229. Sinclair, D. A. R. *et al.* Drosophila O-GlcNAc transferase (OGT) is encoded by the Polycomb group (PcG) gene, super sex combs (sxc). *Proc. Natl. Acad. Sci.* **106,** 13427–13432 (2009).
230. Gambetta, M. C. & Müller, J. A critical perspective of the diverse roles of O-GlcNAc transferase in chromatin. *Chromosoma* **124,** 429–442 (2015).
231. Özcan, S., Andrali, S. S. & Cantrell, J. E. L. Modulation of transcription factor function by O-GlcNAc modification. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1799,** 353–364 (2010).
232. Yang, W. H. *et al.* Modification of p53 with O-linked N-acetylglucosamine regulates p53 activity and stability. *Nat. Cell Biol.* **8,** 1074–83 (2006).
233. Han, I. & Kudlow, J. E. Reduced O glycosylation of Sp1 is associated with increased proteasome susceptibility. *Mol. Cell. Biol.* **17,** 2550–2558 (1997).
234. Housley, M. P. *et al.* O-GlcNAc regulates FoxO activation in response to glucose. *J. Biol. Chem.* **283,** 16283–16292 (2008).
235. Kuo, M., Zilberfarb, V., Gangneux, N., Christeff, N. & Issad, T. O-glycosylation of FoxO1 increases its transcriptional activity towards the glucose 6-phosphatase gene. *FEBS Lett.* **582,** 829–834 (2008).
236. Kelly, W. G., Dahmus, M. E. & Hart, G. W. RNA polymerase II is a glycoprotein: Modification of the COOH-terminal domain by O-GlcNAc. *J. Biol. Chem.* **268,** 10416–10424 (1993).
237. Dey, A. *et al.* Loss of the tumor suppressor BAP1 causes myeloid transformation. *Science* **337,** 1541–6 (2012).
238. Campbell, R. B., Sinclair, D. a, Couling, M. & Brock, H. W. Genetic interactions and dosage effects of Polycomb group genes of Drosophila. *Mol. Gen. Genet.* **246,** 291–300 (1995).
239. Sakabe, K., Wang, Z. & Hart, G. W. Beta-N-acetylglucosamine (O-GlcNAc) is part of the histone code. *Pnas* **107,** 19915–19920 (2010).
240. Fujiki, R. *et al.* GlcNAcylation of histone H2B facilitates its monoubiquitination. *Nature* **480,** 557–60 (2011).
241. Gagnon, J. *et al.* Undetectable histone O-GlcNAcylation in mammalian cells. *Epigenetics* **10,** 677–691 (2015).
242. Sakabe, K. & Hart, G. W. O-GlcNAc transferase regulates mitotic chromatin dynamics. *J. Biol. Chem.* **285,** 34460–34468 (2010).
243. Capotosti, F. *et al.* O-GlcNAc transferase catalyzes site-specific proteolysis of HCF-1. *Cell* **144,** 376–388 (2011).
244. Deplus, R. *et al.* TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. *EMBO J.* **32,** 645–655 (2013).
245. Chu, C.-S. C.-S. *et al.* O-GlcNAcylation regulates EZH2 protein stability and function. *Proc. Natl. Acad. Sci.* **111,** 1355–1360 (2014).

246. Kassis, J. A. Unusual properties of regulatory DNA from the Drosophila engrailed gene: Three 'pairing-sensitive' sites within a 1.6-kb region. *Genetics* **136,** 1025–1038 (1994).
247. Schwartz, Y. B. & Pirrotta, V. Polycomb silencing mechanisms and the management of genomic programmes. *Nat. Rev. Genet.* **8,** 9–22 (2007).
248. Kassis, J. & Brown, J. Polycomb Group Response Elements in Drosophila and Vertebrates. *Adv. Genet.* 83–118 (2012). doi:10.1016/B978-0-12-407677-8.00003-8.Polycomb
249. Niessen, H. E. C., Demmers, J. a & Voncken, J. W. Talking to chromatin: post-translational modulation of polycomb group function. *Epigenetics Chromatin* **2,** 10 (2009).
250. Bradbury, A. & Plückthun, A. Reproducibility: Standardize antibodies used in research. *Nature* **518,** 27–29 (2015).
251. Peach, S. E., Rudomin, E. L., Udeshi, N. D., Carr, S. A. & Jaffe, J. D. Quantitative Assessment of Chromatin Immunoprecipitation Grade Antibodies Directed against Histone Modifications Reveals Patterns of Co-occurring Marks on Histone Protein Molecules. *Mol. Cell. Proteomics* **11,** 128–137 (2012).
252. Pindyurin, A. V, Pagie, L., Kozhevnikova, E. N., van Arensbergen, J. & van Steensel, B. Inducible DamID systems for genomic mapping of chromatin proteins in Drosophila. *Nucleic Acids Res.* **44,** 5646–5657 (2016).
253. Kelly, W. G. & Hart, G. W. Glycosylation of chromosomal proteins: localization of O-linked N-acetylglucosamine in Drosophila chromatin. *Cell* **57,** 243–51 (1989).
254. Zaro, B. W., Hang, H. C. & Pratt, M. R. Incorporation of Unnatural Sugars for the Identification of Glycoproteins. in *Methods in molecular biology (Clifton, N.J.)* (eds. Kohler, J. J. & Patrie, S. M.) **951,** 57–67 (Humana Press, 2013).
255. Sprung, R. *et al.* Tagging-via-Substrate Strategy for Probing O-GlcNAc Modified Proteins. *J. Proteome Res.* **4,** 950–957 (2005).
256. Verdoes, M. *et al.* Azido-BODIPY Acid Reveals Quantitative Staudinger–Bertozzi Ligation in Two-Step Activity-Based Proteasome Profiling. *ChemBioChem* **9,** 1735–1738 (2008).
257. Verhelst, S. H. L., Fonović, M. & Bogyo, M. A Mild Chemically Cleavable Linker System for Functional Proteomic Applications. *Angew. Chemie Int. Ed.* **46,** 1284–1286 (2007).
258. Chuh, K. N., Zaro, B. W., Piller, F., Piller, V. & Pratt, M. R. Changes in metabolic chemical reporter structure yield a selective probe of O -GlcNAc modification. *J. Am. Chem. Soc.* **136,** 12283–12295 (2014).
259. Lin, W., Gao, L. & Chen, X. Protein-Specific Imaging of O-GlcNAcylation in Single Cells. *ChemBioChem* **16,** 2571–2575 (2015).
260. Hiromura, M. *et al.* YY1 is regulated by O-linked N-acetylglucosaminylation (O-GlcNAcylation). *J. Biol. Chem.* **278,** 14046–14052 (2003).
261. Shen, D. L., Gloster, T. M., Yuzwa, S. A. & Vocadlo, D. J. Insights into O-linked N-acetylglucosamine (O-GlcNAc) processing and dynamics through kinetic analysis of O-GlcNAc transferase and O-GlcNAcase activity on protein substrates. *J. Biol. Chem.* **287,** 15395–15408 (2012).
262. Dennis, R. J. *et al.* Structure and mechanism of a bacterial β-glucosaminidase having O-GlcNAcase activity. *Nat. Struct. Mol. Biol.* **13,** 365–371 (2006).
263. Hédou, J., Bastide, B., Page, A., Michalski, J.-C. & Morelle, W. Mapping of O-linked β -N-acetylglucosamine modification sites in key contractile proteins of rat skeletal muscle. *Proteomics* **9,** 2139–2148 (2009).
264. Schuettengruber, B. *et al.* Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS Biol.* **7,** (2009).
265. Landt, S. & Marinov, G. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome …* 1813–1831 (2012).

doi:10.1101/gr.136184.111.
266. Schwartz, Y. B. *et al.* Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nat. Genet.* **38,** 700–705 (2006).
267. Zeng, J., Kirk, B. D., Gou, Y., Wang, Q. & Ma, J. Genome-wide polycomb target gene prediction in Drosophila melanogaster. *Nucleic Acids Res.* **40,** 5848–5863 (2012).
268. Tolhuis, B. *et al.* Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. *Nat. Genet.* **38,** 694–699 (2006).
269. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
270. Oktaba, K. *et al.* Dynamic Regulation by Polycomb Group Protein Complexes Controls Pattern Formation and the Cell Cycle in Drosophila. *Dev. Cell* **15,** 877–889 (2008).
271. Negre, N. *et al.* A cis-regulatory map of the Drosophila genome. *Nature* **471,** 527–531 (2011).
272. Eissenberg, J. C. & Elgin, S. C. R. HP1a: a structural chromosomal protein regulating transcription. *Trends Genet.* **30,** 103–10 (2014).
273. Cummings, R. D. & Etzler, M. E. Antibodies and Lectins in Glycan Analysis. in *Essentials of Glycobiology 2nd Edition* 633–648 (2009). doi:NBK1919 [bookaccession]
274. Akan, I., Love, D. C., Harwood, K. R., Bond, M. R. & Hanover, J. A. Drosophila O-GlcNAcase deletion globally perturbs chromatin O-GlcNAcylation. *J. Biol. Chem.* **291,** 9906–9919 (2016).
275. Bond, M. R. & Hanover, J. A. A little sugar goes a long way: The cell biology of O-GlcNAc. *J. Cell Biol.* **208,** 869–880 (2015).
276. Brown, J. L., Fritsch, C., Mueller, J. & Kassis, J. a. The Drosophila pho-like gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing. *Development* **130,** 285–294 (2003).
277. Rodriguez-Jato, S., Busturia, A. & Herr, W. Drosophila melanogaster dHCF interacts with both PcG and TrxG epigenetic regulators. *PLoS One* **6,** (2011).
278. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
279. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).
280. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
281. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
282. Favorov, A. *et al.* Exploring massive, genome scale datasets with the genometricorr package. *PLoS Comput. Biol.* **8,** (2012).
283. Shin, H., Liu, T., Manrai, A. K. & Liu, S. X. CEAS: Cis-regulatory element annotation system. *Bioinformatics* **25,** 2605–2606 (2009).
284. Huang, D. W., Lempicki, R. a & Sherman, B. T. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).
285. Deal, R. B., Henikoff, J. G. & Henikoff, S. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science* **328,** 1161–4 (2010).
286. Mito, Y., Henikoff, J. G. & Henikoff, S. Histone replacement marks the boundaries of cis-regulatory domains. *Science* **315,** 1408–11 (2007).
287. Mito, Y., Henikoff, J. G. & Henikoff, S. Genome-scale profiling of histone H3.3 replacement patterns. *Nat. Genet.* **37,** 1090–1097 (2005).
288. Dion, M. F. *et al.* Dynamics of replication-independent histone turnover in

budding yeast. *Science* **315,** 1405–8 (2007).
289. Lickwar, C. R., Mueller, F., Hanlon, S. E., McNally, J. G. & Lieb, J. D. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **484,** 251–5 (2012).
290. Kraushaar, D. C. *et al.* Genome-wide incorporation dynamics reveal distinct categories of turnover for the histone variant H3.3. *Genome Biol* **14,** R121 (2013).
291. Ha, M., Kraushaar, D. C. & Zhao, K. Genome-wide analysis of H3.3 dissociation reveals high nucleosome turnover at distal regulatory regions of embryonic stem cells. *Epigenetics Chromatin* **7,** 1–14 (2014).
292. Yildirim, O. *et al.* A System for Genome-Wide Histone Variant Dynamics In ES Cells Reveals Dynamic MacroH2A2 Replacement at Promoters. *PLoS Genet.* **10,** (2014).
293. Deaton, A. M. *et al.* Enhancer regions show high histone H3.3 turnover that changes during differentiation. *Elife* **5,** 1–24 (2016).
294. wa Maina, C. *et al.* Inference of RNA Polymerase II Transcription Dynamics from Chromatin Immunoprecipitation Time Course Data. *PLoS Comput. Biol.* **10,** 1–17 (2014).
295. Fiorito, E. *et al.* CTCF modulates Estrogen Receptor function through specific chromatin and nuclear matrix interactions. *Nucleic Acids Res.* **44,** 1–15 (2016).
296. Zentner, G. E., Kasinathan, S., Xin, B., Rohs, R. & Henikoff, S. ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nat. Commun.* **6,** 8733 (2015).
297. Grünberg, S., Henikoff, S., Hahn, S. & Zentner, G. E. Mediator binding to UASs is broadly uncoupled from transcription and cooperative with TFIID recruitment to promoters. *EMBO J.* **35,** 2435–2446 (2016).
298. Adar, S., Hu, J., Lieb, J. D. & Sancar, A. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* 201603388 (2016). doi:10.1073/pnas.1603388113
299. Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose-response analysis using R. *PLoS One* **10,** 1–13 (2015).
300. Zaharia, M., Bolosky, W. & Curtis, K. Faster and More Accurate Sequence Alignment with SNAP. *arXiv Prepr. arXiv …* 1–10 (2011).
301. Bonhoure, N. *et al.* Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Res.* **24,** 1157–68 (2014).
302. Wickham, H. *ggplot2*. *Elegant Graphics for Data Analysis* (2009). doi:10.1007/978-0-387-98141-3
303. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: A next-generation sequencing read simulator. *Bioinformatics* **28,** 593–594 (2012).
304. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
305. Lickwar, C. R., Mueller, F. & Lieb, J. D. Genome-wide measurement of protein-DNA binding dynamics using competition ChIP. *Nat. Protoc.* **8,** 1337–1353 (2013).
306. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** D493-6 (2004).
307. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** 88–92 (2007).
308. Curtin, R. R. *et al.* MLPACK: a scalable C++ machine learning library. *J. Mach. Learn. Res.* **14,** 801–805 (2013).
309. Machanick, P. & Bailey, T. L. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **27,** 1696–1697 (2011).
310. Marteijn, J. a, Lans, H., Vermeulen, W. & Hoeijmakers, J. H. J. Understanding

nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15,** 465–81 (2014).
311. Chen, J. *et al.* Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell* **156,** 1274–1285 (2014).
312. Bogdanović, O., Fernández-Miñán, A., Tena, J. J., de la Calle-Mustienes, E. & Gómez-Skarmeta, J. L. The developmental epigenomics toolbox: ChIP-seq and MethylCap-seq profiling of early zebrafish embryos. *Methods* **62,** 207–15 (2013).
313. Anders, L. *et al.* Genome-wide localization of small molecules. *Nat. Biotechnol.* **32,** 92–6 (2014).
314. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39,** 2010–2012 (2011).
315. Radermacher, P. T. *et al.* O-GlcNAc reports ambient temperature and confers heat resistance on ectotherm development. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 5592–7 (2014).
316. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8,** R24 (2007).
317. Graveley, B., Brooks, A. & Carlson, J. The developmental transcriptome of Drosophila melanogaster. *Nature* **471,** 473–479 (2011).
318. Rafiee, M. R., Girardot, C., Sigismondo, G. & Krijgsveld, J. Expanding the Circuitry of Pluripotency by Selective Isolation of Chromatin-Associated Proteins. *Mol. Cell* **64,** 624–635 (2016).
319. Baldini, S. F. *et al.* Glucokinase expression is regulated by glucose through O-GlcNAc glycosylation. *Biochem. Biophys. Res. Commun.* **478,** 942–948 (2016).
320. Pourfarzad, F. *et al.* Locus-Specific Proteomics by TChP: Targeted Chromatin Purification. *Cell Rep.* **4,** 589–600 (2013).
321. Byrum, S. D., Raman, A., Taverna, S. D. & Tackett, A. J. ChAP-MS: A Method for Identification of Proteins and Histone Posttranslational Modifications at a Single Genomic Locus. *Cell Rep.* **2,** 198–205 (2012).
322. Déjardin, J. & Kingston, R. E. Purification of Proteins Associated with Specific Genomic Loci. *Cell* **136,** 175–186 (2009).
323. Kennedy-Darling, J. *et al.* Discovery of Chromatin-Associated Proteins via Sequence-Specific Capture and Mass Spectrometric Protein Identification in Saccharomyces cerevisiae. *J. Proteome Res.* **13,** 3810–3825 (2014).
324. Ide, S. & Dejardin, J. End-targeting proteomics of isolated chromatin segments of a mammalian ribosomal RNA gene promoter. *Nat. Commun.* **6,** 6674 (2015).
325. Butala, M., Busby, S. J. W. & Lee, D. J. DNA sampling: a method for probing protein binding at specific loci on bacterial chromosomes. *Nucleic Acids Res.* **37,** e37–e37 (2009).
326. Waldrip, Z. J. *et al.* A CRISPR-based approach for proteomic analysis of a single genomic locus. *Epigenetics* **9,** 1207–1211 (2014).