UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR
NEURO- UND BIOINFORMATIK

Quantitative epigenetische Analyse von Histonmodifikationen in Wildmäusen

Quantitative epigenetic analysis of histone modifications in wild-caught mice

Masterarbeit im Rahmen des Studiengangs Mathematik in Medizin und Lebenswissenschaften der Universität zu Lübeck vorgelegt von Linda Krause.

Die Masterarbeit wurde ausgegeben und betreut von Prof. Dr. Bernhard Haubold mit Unterstützung von Dr. Angelika Börsch-Haubold.

Lübeck, den 17.12.2013. Formal überarbeitet am 28.01.2014.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Lübeck, 17.12.2013 _____

Summary

Epigenetic modifications of nuclear proteins play a crucial role in regulating gene activity. Studies in inbred laboratory animals help to classify activating and silencing marks. They also demonstrate that environmental factors such as social stress change epigenetic settings. We set out to assess the degree of epigenetic difference in mice with a heterozygous genome under similar or different environmental conditions. First, we use genome-wide sequencing data from chromatin immunoprecipitations for an activating histone mark (H3K4me3) in order to quantify inter-individual differences from two males of a wild-caught house mouse population. Second, we investigate whether social stress has an impact on epigenetic settings in liver tissue. The experimental model are mice that live in a semi-natural enclosure and form social hierarchies with dominant and subordinate males. We compare two active histone marks by immunoprecipitation and quantitative polymerase chain reaction from healthy males that live in family groups and from males that bear signs of social defeat. Statistical analyses show changes in epigenetic marks that cluster with the social background of the mice.

Zusammenfassung

Epigenetische Modifikationen nuklearer Proteine spielen eine entscheidende Rolle in der Regulation der Genaktivität. Studien an ingezüchteten Labormäusen helfen die aktivierenden von den repressiven Markierungen zu unterscheiden. Sie zeigen auch, dass Umwelteinflüsse wie sozialer Stress das epigenetische Muster verändern. Unser Ziel ist es, die Größe epigenetischer Unterschiede in Mäusen mit einem heterozygoten Genom unter ähnlichen und verschiedenen Umweltbedingungen zu erfassen. Zunächst verwenden wir genomweite Sequenzierungsdaten aus Chromatin Immunopräzipitationen einer aktiven Histonmarkierung (H3K4me3), um die individuellen Unterschiede zwischen zwei Wildmäusen einer Mauspopulation zu quantifizieren. Als Nächstes untersuchen wir, inwieweit sozialer Stress einen Einfluss auf epigenetische Muster in Lebergewebe hat. Wildmäuse, die in Freilaufgehegen leben und soziale Hierarchien mit dominanten und untergeordneten Männchen ausbilden, dienen als experimentelles Modell. Wir vergleichen zwei aktive Histonmarkierungen durch Immunopräzipitation gefolgt von quantitativer Polymerasekettenreaktion von gesunden Männchen, die in Familiengruppen leben und von Männchen, die Anzeichen sozialer Ausgrenzung aufweisen. Statistische Analysen zeigen Veränderungen der epigenetischen Markierungen, die mit dem sozialen Hintergrund der Mäuse zusammenfallen.

Danksagung

Mein besonderer Dank gilt Dr. Angelika Börsch-Haubold, die mir mit diesem interessanten Thema die Möglichkeit gab, über den Tellerrand hinauszuschauen und einen Blick in die spannende Welt der Epigenetik zu werfen.

Ich möchte mich bei Prof. Dr. Bernhard Haubold dafür bedanken, dass er mir stets bei allen Problemen und Fragen seinen Rat und seine Zeit schenkte.

Cornelia Burghardt möchte ich für ihre Geduld, die sie aufbrachte, um mir die Techniken im Labor zu erläutern, und ihre praktischen Tipps danken.

Weiterer Dank gilt meinen Lektorinnen und Lektoren, die mir bei Formulierungsschwierigkeiten weiterhelfen konnten und mich auf kleinere und größere Fehler aufmerksam machten.

Die Arbeit wurde am Max-Planck-Institut für Evolutionsbiologie in Plön durchgeführt und von der Max-Planck-Gesellschaft finanziert.

Inhaltsverzeichnis

1 Einleitung

Die Genomsequenzen von Maus und Mensch sind seit mehr als zehn Jahren vollständig bekannt [42] [8]. Wie aus einer DNA Sequenz, einem Genotypen, viele verschiedene Zellarten mit teilweise sehr spezialisierten Aufgaben entstehen können, beschäftigt Wissenschaftler auf dem Feld der Epigenetik schon seit mehreren Jahrzehnten. Epigenetik wurde deshalb ursprünglich definiert als alle Mechanismen, die den Zustand der Genexpression einer Zelle stabil durch die Mitose an Tochterzellen weitergeben können, ohne die zu Grunde liegende DNA Sequenz zu verändern [9]. Mittlerweile erweiterte sich diese Definition um Mechanismen, die die Genexpression regulieren, deren Vererbung aber noch diskutiert wird. Epigenetische Markierungen sind, anders als genetische Polymorphismen, potentiell reversibel, deshalb ist es interessant, den Zusammenhang zwischen Epigenetik und Krankheiten zu erforschen. Das Angelman-Syndrom, eine Störung der Entwicklung des Nervensystems, beruht zum Beispiel auf der Mutation oder Deletion des mütterlichen Allels des Gens Ube3a, während das intakte väterliche Allel epigenetisch inaktiviert ist. Durch eine Reaktivierung des väterlichen Allels mit Hilfe von Medikamenten, die Topoisomerasehemmer beinhalten und in Mäusen bereits Wirkung gezeigt haben, könnte die Krankheit therapiert werden [24].

Menschen unterscheiden sich in ihrer DNA Sequenz an ca. jeder 1300. Position [41]. Eineiige Zwillinge, die die gleiche DNA Sequenz haben, sind sich in ihren epigenetischen Mustern ähnlicher als zweieiige Zwillinge, die genetisch verschieden sind [9]. Die Abfolge der Nukleotide der DNA bestimmt demnach zunächst primär den epigenetischen Zustand. Viele Forschungsergebnisse beruhen auf Untersuchen an ingezüchteten Labormäusen, die alle den gleichen genetischen Hintergrund, die gleiche DNA Sequenz besitzen. Wildmäuse hingegen weisen ca. alle 500 Positionen einen Unterschied in der Abfolge der DNA Basen auf. Da mit Hilfe von Wildmäusen epigenetische Veränderungen vor dem Hintergrund genetischer Vielfalt untersucht werden können, wurde am Max-Planck-Institut für Evolutionsbiologie (MPI) in Plön ein Freilaufexperiment mit Wildmäusen durchgeführt und deren Epigenetik untersucht.

Die wohl am besten verstandene epigenetische Markierung ist die DNA Methylierung. Dabei wird eine Methylgruppe kovalent an die 5' Position im Nukleotid Cytosin gebunden. DNA Methylierung tritt in Säugetieren fast ausschließlich an Cytosin-Guanin Dinukleotiden (abgekürzt CpG) auf und ist über das gesamte Genom verteilt. Das Methylierungsmuster wird nicht von der DNA Replikationsmaschinerie reproduziert, sondern von unabhängigen Enzymen, die die mitotische Vererbung der DNA Methylierung sicher stellen, den DNA Methyltransferasen (DNMT) [47] [60]. Dadurch, dass die DNA Methylierung eine kovalente Modifikation der DNA ist, ist sie eine sehr stabile epigenetische Markierung.

Liegen Promotorregionen von Genen methyliert vor, wird die Genexpression hauptsächlich durch zwei Mechanismen inhibiert [33]. Zunächst kann die Methylierung des Cytosin dazu führen, dass Transkriptionsfaktoren nicht mehr an die Promotorregion binden können, da sich die Bindungsstellen durch die Methylierung getarnt haben. Ohne Transkriptionsfaktoren kann es keine Genexpression geben und das Gen ist still gelegt. Ab einer bestimmten Dichte an DNA Methylierung binden Proteine mit Affinität zu methylierter DNA und erzeugen eine geschlossene Chromatinstruktur, Heterochromatin genannt [47].

Der Anteil der Basen Guanin und Cytosin (GC-Gehalt) beträgt 38% im menschlichen Genom. Unter der Annahme einer zufälligen Verteilung der Basen würde das Dinukleotid CpG mit einer Häufigkeit von 3,6% vorliegen. Im gesamten menschlichen Genom beträgt der Anteil an CpG Dinukleotiden aber nur 0,9%. Der Grund dafür ist die spontane Desaminierung von methyliertem Cytosin zu Thymin, eine anderen Base der DNA, die deshalb von den DNA Reparaturenzymen nicht als Fehler erkannt wird [58].

Trotzdem tauchen im Genom immer wieder Bereiche auf, in denen der Anteil an CpG Dinukleotiden (CpG-Gehalt) und der GC-Gehalt hoch sind und die oft unmethyliert vorliegen, die CpG-Inseln. Es ist noch nicht aufgeklärt, wie die CpG-Inseln während der frühen Entwicklungsphase, in der die globe *de novo* Methylierung gelegt wird, unmethyliert bleiben [27]. Auffällig ist, dass die ca 1000 Basenpaar (Bp) langen CpG-Inseln mit ungefähr 60 bis 70% der Promotorregionen von menschlichen Genen überlappen [44].

Aufgrund dieser Beobachtungen wurden zwei Klassen von Promotoren beschrieben [58]. Promotoren mit einem hohen CpG-Gehalt (HCP, high CpG-content promotor), dazu zählen 72%, regulieren Gene, die in fast allen Geweben exprimiert werden. Werden Gene hingegen nur in wenigen Zelltypen exprimiert oder kurzfristig hoch bzw. herunterreguliert, fallen ihre Promotoren oft in die Gruppe mit einem niedrigen CpG-Gehalt (LCP, low CpG-content promotor) [58].

Abgesehen von der DNA Methylierung gibt es noch einen weiteren epigenetischen Mechanismus, der die Chromatinstruktur verändert: die Modifikation von Histonen. Mit ihnen beschäftigen wir uns in dieser Arbeit. Chromatin ist ein Komplex aus DNA mit daran gebundenen oder damit interagierenden Proteinen, deren kleinste Wiederholungseinheit das Nukleosom ist. Ein Nukleosom besteht aus ungefähr 150 Bp DNA, die um ein Oktamer aus vier verschiedenen Histonproteinen gewickelt sind. Die N-Termini der Histone unterliegen vielfältigen posttranslationalen Veränderungen. In dieser Masterarbeit wurden zwei Histonmodifikationen untersucht, die mit aktiver Transkription assoziiert werden. Die Trimethylierung des Lysin 4 am Histon H3 (H3K4me3) und die Acetylierung des Lysin 27 am Histon H3 (H3K27ac) führen durch die Neutralisierung der positiv geladenen Aminosäure Lysin zu einer offenen Chromatinstrukur (Euchromatin), so dass die Transkriptionsmaschinerie leichter binden kann, um die DNA als RNA abzulesen [38] [55].

Dies sind zwei von einer Vielzahl von inzwischen bekannten Histonmodifikationen. Zum Beispiel kann die Dimethylierung des Lysin 4 am Histon H3 (H3K4me2) sowohl an aktiven als auch an inaktiven Genen vorliegen. H3K4me3 hingegen markiert nur aktive Gene [57]; insbe-

sondere markiert sie Positionen der Transkriptionsinitiation [17]. H3K27ac markiert aktive Gene und insbesondere auch deren Enhancer Regionen. Es ist möglich, dass sie deshalb an aktiven Genen vorliegt, da die Trimethylierung des Lysin 27 am Histon H3 (H3K27me3), die mit inaktiven Genen assoziiert ist, nicht gleichzeitig vorliegen kann [68] [38].

Karlić et al. untersuchten die Korrelation zwischen Genexpression und Histonmodifikationen [38]. Dabei trennten sie LCP und HCP und stellten fest, dass die H3K4me3 Markierung für LCP und H3K27ac für HCP besonders gut zur Vorhersage geeignet ist. Darüber hinaus zeigten sie, dass beide Markierungen in einem ca. 1000 Bp großen Fenster um Transkriptionsstartpunkte (TSS, transcription start site) angereichert sind. An Nukleosomen in HCP liegt fast immer H3K4me3 vor, wohingegen in LCP diese Markierung nur auftritt, wenn das Gen exprimiert wird [48].

Der Zusammenhang zwischen H3K4me3 Markierung, DNA Methylierung und RNA Expression ist noch nicht vollständig geklärt. Meist wird die DNA Methylierung mit unmodifiziertem H3K4 assoziiert [7]. Aktive Histonmarkierungen, wie H3K4me3, können die Ausbildung von DNA Methylierung in Genpromotoren *in vitro* inhibieren [51] [4]. Angstkonditionierungsstudien an Ratten zeigten hingegen, dass die H3K4me3 Markierung am Promotor des Gens Egr1 zunimmt, während gleichzeitig die DNA Methylierung und die RNA Expression gesteigert wird [18]. Ein Gen kann demnach aktiv transkribiert werden und gleichzeitig methyliert vorliegen. Der Zusammenhang könnte zell- oder genspezifisch verschieden sein.

Um die DNA Abschnitte, die um modifizierte Histone gewunden sind, experimentell zu bestimmen, ist es zunächst wichtig, diese DNA chemisch mit den Histonen zu verbinden. Bereits in den 60iger Jahren wurde Formaldehyd verwendet, um Proteine mit RNA oder DNA zu verknüpfen [5] [28]. Seit 1981 ist bekannt, dass die Behandlung mit Formaldehyd die Chromatinstrukur erhält [29] [30]. Nach dem Vernetzen von DNA mit Histonen werden die Histone immunopräzipitiert, indem Antikörper, die spezifisch für die Histonmodifikation von Interesse sind, zugegeben werden. Mitte der 90iger Jahre gab es die ersten Antikörper gegen Histonacetylierungen und seitdem werden diese immer spezifischer in Bezug auf die Position der Acetylierung im Histonmolekül [21]. Im Anschluss an die Immunopräzipitation ist die chemische Verbindung zwischen DNA und Histonproteinen zu lösen, um die co-präzipitierte DNA zu quantifizieren [40].

Die quantitative Polymerasekettenreaktion (qPCR) ist eine Möglichkeit dafür. Dabei wird mit Hilfe von fluoreszierenden Stoffen die Menge des PCR Produkts bestimmt [22]. Noch 2007 wird die qPCR als die beste Methode beschrieben, um Immunopräzipitate zu analysieren [19]. Im selben Jahr stellten Robertson et al. ihre neu entwickelte Methode vor: eine Kombination aus Chromatin Immunopräzipitation und massiv-paralleler Sequenzierung (ChIP-Seq). Zunächst zur Bestimmung exakter Bindungsstellen von Transkriptionsfaktoren im Genom verwendet [56], wurde die Methode auch bald eingesetzt, um Bereiche im Genom mit bestimmten Histonmarkierungen zu finden [3] [48]. Robertson et al. untersuchten 2008 den Zusammenhang zwischen H3K4me1 bzw. H3K4me3 und Bindungsstellen von Transkriptionsfaktoren [55]. Die dafür ent-

wickelten Verfahren und die von Johnson et al. 2007 [32] beschriebenen Techniken sind die Grundlage für die aktuellen Kits zur Chromatin Immunopräzipitation. Mittlerweile ist ChIP-Seq die wichtigste Methode zur genomweiten *in vivo* Untersuchung von Protein-DNA Wechselwirkungen [6].

Jedes ChIP-Seq Experiment ergibt viele Millionen kurzer Nukleotidsequenzen, die nicht homogen über das Genom verteilt sind, sondern in bestimmten Bereichen angereichert vorliegen [16]. Diese Anreicherungen sind zu quantifizieren und zu visualisieren, um sie auszuwerten.

Das am MPI in Plön durchgeführte Freilaufexperiment mit Wildmäusen war ähnlich dem von Montero et al. 2013 beschriebenen [49]. Die ersten ausführlichen Beobachtungen zu Wildmäusen beschrieb Crowcroft 1955 im Zuge seiner Studien, die das Ziel hatten, die Ausbreitung von Mäusen in Kornspeichern besser zu kontrollieren [10]. In seinen Freilaufgehegen, die mit Nistboxen ausgestattet waren, führten Kämpfe zwischen den Männchen dazu, dass diese bestimmte Gebiete für sich vereinnahmten und gegen Eindringlinge verteidigten. Innerhalb ihrer Territorien lebten die Männchen mit einem oder mehreren Weibchen und den gemeinsamen Nachkommen zusammen. Bereits in den ersten Experimenten beobachtete Crowcroft, dass es neben den dominanten auch untergeordnete Männchen gab. Er beschrieb die Anhäufung dieser Männchen in einer Nistbox, in der außerdem keine Weibchen anzutreffen waren. Diese Entstehung sozialer Hierarchien bemerkte Crowcroft auch in späteren Experimenten immer wieder [11]. Die Hierarchie wirkte sich auf das Leben der untergeordneten Männchen stark aus. Zum Einen konnten sie keine Nachkommen zeugen, da sie keine Weibchen fanden. Zum Anderen wurden sie beim Fressen beeinträchtigt, da sie von den dominanten Männchen aus deren Territorien gejagt wurden. Das führte zu untypischem Verhalten dieser Mäuse, denn die eigentlich nachtaktiven Tiere liefen bei Tageslicht auf der Suche nach Fressen durch die Freigehege und standen unter solchem Stress, dass teilweise ihre Angst vor Menschen verloren ging.

Auch im Freilaufexperiment am MPI in Plön bildeten die Wildmäusen in dem Freigehege, in denen kleine Häuser zu Verfügung standen, schnell Paare und Nester, wie schon 1955 von Crowcroft beobachtet wurde. Bekamen die Paare Nachwuchs, lebten die Familien weiterhin zusammen in den Häusern [10]. Mit zunehmender Populationsgröße entwickelte sich eine soziale Hierarchie, die auch an den unterschiedlichen äußeren Erscheinungen der Mäuse beobachtet werden konnte. Die Männchen, die in Häusern mit ihrer Familie zusammen lebten, hatten gesundes Fell und machten insgesamt einen guten Eindruck. Sie werden im Weiteren als sozialisierte Mäuse bezeichnet. Die frei herumlaufenden Männchen ohne Hauszugehörigkeit hatten zu einem großen Anteil ein höheres Gewicht, das mit einem höheren Lebensalter einhergeht [12], und wiesen Bisswunden und eine schlechte Fellbeschaffenheit auf, was auf Kämpfe hindeutete. Es war offensichtlich, dass diese Mäuse sozial ausgegrenzt wurden und unter einem höheren Stressfaktor lebten [Angelika Börsch-Haubold, persönliche Mitteilung]. Diese Mäuse wurden bei den regelmäßigen Raumkontrollen aus dem Experiment entfernt [49].

In der Literatur gibt es weitere Beispiele für im Labor simulierte Stresssituationen. Indem Tiere eine Erinnerung an stressige Situationen ausbilden, lernen sie aus diesen und können auf ihre Erfahrungen zurückgreifen, wenn sie mit einem ähnlichen Ereignis konfrontiert werden. Epigenetische Mechanismen in Neuronen des zentralen Nervensystems spielen eine Rolle beim komplexen Aufbau dieser Erinnerungen [60]. Nachkommen von Männchen, die chronisch sozialem Stress ausgeliefert sind, weisen erhöht depressive und ängstliche Verhaltensmuster auf [14]. Ein Vergleich natürlicher und mit *in vitro*-Fertilisation gezeugter Nachkommen dieser gestressten Männchen gibt Hinweise auf eine epigenetische Regulation der Vererbung des Angst-Phänotyps [14].

Crowcroft beschreibt aggressives Verhalten in Wildmäusen insbesondere zwischen Männchen, aber auch bei trächtigen oder säugenden Weibchen [11]. Ungefähr im Alter von drei Monaten zeigt sich aggressives Verhalten innerhalb der männlichen Nachkommen, die bereits im Alter von ca. eineinhalb Monaten geschlechtsreif werden. Das Verhalten führt zur Ausbildung sozialer Hierarchien und dient zur Etablierung und Verteidigung von Territorien. Aggressives Verhalten und die Fähigkeit Kämpfe zu gewinnen ist in Mäusen erblich, die Anzahl und das Geschlechterverhältnis der Geschwister haben dabei kaum Auswirkungen, aber Epigenetik könnte eine Rolle spielen [12].

Stress und soziale Benachteiligung führen nach Crowcroft dazu, dass längere Wege während der Futtersuche zurückgelegt werden müssen und die Nahrung zu falschen Zeiten aufgenommen wird [11]. Die Leber ist das wichtigste metabolische Organ in Säugetieren, das unregelmäßige Fressen könnte sich demnach dort manifestieren. Außerdem ist die Leber im Vergleich zu beispielsweise dem Gehirn ein relativ homogenes Organ, so dass der Vergleich von Leberproben aus Leberlappen-Sezierschnitten einfacher durchzuführen ist, als der Vergleich von Hirn-Proben, bei denen eine akkurate Trennung verschiedener funktioneller Bereiche vorgenommen werden müsste. Deshalb stellen wir die Frage, wie sich epigenetische Folgen von sozialem Stress durch Fressverhalten in der Leber zeigen.

In dieser Masterarbeit werden die H3K4me3 ChIP-Seq Datensätze zweier Wildmäuse der Population des MPI in Plön analysiert, mit dem Ziel, grundlegende Unterschiede, die zwischen zwei Individuen auftauchen, zu beschreiben. Der Zusammenhang zwischen genetischen Polymorphismen, die in Wildmäusen anders als in Bl6 Labormäusen auftreten, und epigenetischer Variabilität, ist noch nicht geklärt. Deshalb ist es wichtig, zunächst die mögliche Variabilität der gewählten epigenetischen Markierung zu untersuchen.

Um unserer Frage nachzugehen, was die epigenetischen Folgen von sozialem Stress in der Leber sind, untersuchen wir Lebergewebe von 24 Wildmäusen aus dem Freilaufexperiment mit unterschiedlichem sozialen Status an 15 metabolischen Loci auf H3K4me3 und H3K27ac Anreicherung. Wir zeigen, dass die epigenetischen Muster der Mäuse sich so stark verändert haben, dass sie sich in statistischen Analysen nach ihrem sozialen Status trennen lassen.

2 Material und Methoden

2.1 Herkunft der untersuchten Wildmäuse

Am Max-Planck-Institut für Evolutionsbiologie in Plön werden Wildmäuse, die ursprünglich in der Region Köln-Bonn in Ställen gesammelt wurden, unter einem „outbreeding scheme“ gezüchtet [26]. Die in der vorliegenden Arbeit verwendeten Leberproben stammen von Nachkommen aus dieser Zucht.

2.1.1 ZH1 und ZH6

Um das Ausmaß an Unterschieden zwischen zwei Wildmaus-Individuen abzuschätzen, wurden H3K4me3 ChIP-Seq Daten von zwei Männchen erhoben, die aus Käfig-Zuchten stammen (Tabelle 2.1). Sie wuchsen zwar im Käfig unter gleichen äußeren Bedingungen auf, waren aber in ihrer frühen Entwicklung leichten sozialen Unterschieden ausgesetzt: die Maus ZH1 hatte vier Schwestern und nur einen Bruder; die Maus ZH6 hatte nur eine Schwester und fünf Brüder. Früher Konkurrenzkampf unter den Brüdern könnte dazu geführt haben, dass die Maus ZH6 nur noch einen halben Schwanz hatte. Die Maus ZH1 wiegt relativ wenig (22,30 g) und ist klein, während die Maus ZH6 über 25% mehr wiegt (28,00 g) und Fettpolster am Testis aufweist.

2.1.2 Wildmäuse des ChIP qPCR Experiments

Die untersuchten Wildmäuse aus dem Freilaufexperiment sind ausschließlich Männchen, die etwa sieben Monate alt waren. Je acht der analysierten Mäuse stammen aus zwei Nestern, den Häusern 11 und 18. Weitere acht Mäuse gehören zu der Gruppe Mäuse ohne Hauszugehörigkeit, die beim Monitoring frei durch den Raum liefen und Fellwunden aufwiesen. Das Durchschnittsgewicht der Gruppe aus Haus 11 beträgt 23,44 g, aus dem Haus 18 22,10 g und die Mäuse ohne Hauszugehörigkeit wogen im Schnitt 24,21 g (Tabelle 2.2).

Tab. 2.1: Die Wildmäuse ZH1 und ZH6. (w=weiblich, m=männlich)

ID	Geschwister	Gewicht	Zustand
ZH1	4xw, 1xm	22,30 g	Kleine Maus mit kleiner Milz.
ZH6	1xw, 5xm	28,00 g	Halber Schwanz, Fettpolster am Testis.

Tab. 2.2: Die 24 mit ChIP qPCR analysierten Wildmäuse.

MouseID	Haus	Körpergewicht	Zustand
H11_50	H11	22,41 g	gut
H11_51	H11	21,51 g	gut
H11_52	H11	23,02 g	gut
H11_54	H11	21,99 g	gut
H11_55	H11	23,07 g	gut
H11_56	H11	28,26 g	gut
H11_62	H11	22,77 g	gut
H11_63	H11	24,51 g	gut
H18_50	H18	20,88 g	gut
H18_51	H18	22,22 g	gut
H18_52	H18	22,35 g	gut
H18_54	H18	20,87 g	gut
H18_55	H18	22,54 g	gut
H18_56	H18	21,01 g	gut
H18_57	H18	23,42 g	gut
H18_66	H18	23,48 g	gut
2frei_41	frei laufend	25,14 g	zerbissen
2frei_45	frei laufend	20,75 g	zerbissen
2frei_47	frei laufend	23,07 g	zerbissen
2frei_48	frei laufend	24,90 g	zerbissen
2frei_49_2	frei laufend	25,38 g	zerbissen
2frei_49_3	frei laufend	22,21 g	zerbissen
2frei_49_4	frei laufend	23,67 g	zerbissen
2frei_49_5	frei laufend	28,52 g	zerbissen

2.2 Auswertung von ChIP-Seq Datensätzen

ChIP-Seq ist eine moderne Methode um genomweit epigenetische Daten zu erfassen. Die aus der Sequenzierung erhaltenen Daten waren Dateien im `fastq` Format. Eine Einheit einer `fastq` Datei sieht folgendermaßen aus:

```
@HWUSI-EAS1615R_0087:1:1:1327:1333#0/1
TCTGTTGGGTATCATCTGACATGACTCTGTGGGGTAT
+HWUSI-EAS1615R_0087:1:1:1327:1333#0/1
ggggggggggggdggffgggggggfgggggggeff
```

Diese vier Zeilen bilden einen *Read*. Ein Read beschreibt ein kleines Stück DNA durch die Angabe der Basenabfolge. In der ersten Zeile folgt auf @ der Name des Reads. Die beiden mehrstelligen Ziffern vor dem # Zeichen identifizieren den Read innerhalb eines Experiments eindeutig. In der zweiten Zeile steht die Sequenz von 37 Nukleotiden. Die dritte Zeile beginnt mit einem + Zeichen, das gefolgt wird von der Bezeichnung des Reads. In der vierten Zeile stehen 37 Qualitätskennzahlen für die Sequenz in der zweiten Zeile. Die Qualitätskennzahlen sind ASCII codiert, der Buchstabe g entspricht beispielsweise der Dezimalzahl 103. Um aus einem Buchstaben-Wert b die Wahrscheinlichkeit P zu berechnen, dass das entsprechende Nukleotid falsch ist, gilt

$$P = 10^{-(b-64)/10}$$

In unserem Beispiel ist also P=$10^{-3,9}$=0,0001.

2.2.1 Weiterverarbeitung von `fastq` Dateien

Die Datensätzen ZH1 und ZH6 bestehen aus über 40 Millionen Reads. Das Programm `bowtie` [43] dient dazu, diese auf das Referenzgenom zu platzieren oder zu *mappen*.

Wir verwenden folgenden Aufruf, um die Reads in der Datei `zh1.fastq` auf das Referenzmausgenom mm9 zu mappen.

```
bowtie --phred64-quals --sam --best -m 1 -p 29 index/
   m_musculus_ncbi37 zh1.fastq > zh1.sam
```

Bei diesen Einstellungen versucht `bowtie`, jeden Read jeweils mit höchstens zwei nicht passenden Positionen (*Mismatches*) in den ersten 28 Stellen des Nukleotiden auf die Referenz zu mappen. An den Positionen 29-37 sind weitere Mismatches erlaubt, solange die Summe der Qualitätskennzahlen an Positionen, die einen Mismatch aufweisen, nicht größer als 70 ist. Zu beachten ist, dass `bowtie` die Qualitätskennzahlen stets auf die nächste Zehnerstelle rundet und das Maximum bei 30 erreicht ist.

Die Option `--phred64-quals` bedeutet, dass die Qualitätskennzahlen unserer Reads in einem Format vorliegen, bei dem nach einer Umwandlung der ASCII Zeichen in Dezimalzahlen 64 abgezogen werden muss, um im Bereich von 0 bis 40 zu landen. Die Ausgabe des Aufrufs soll eine Datei im `sam` Format sein, deshalb wählen wir die Option `--sam`. Die `--best` Einstellung

bei `bowtie` garantiert, dass das Alignment optimal in Hinsicht auf die Mismatches und die Qualitätskennzahl an diesen Positionen ist. Da wir nur Reads in unserem Mapping haben wollen, die eindeutig auf die Referenz passen, verwenden wir die Option `-m 1`. Zur Beschleunigung der Berechnung kann mit Hilfe der Einstellung `-p 29` die Anzahl an Prozessoren, auf denen gleichzeitig gerechnet werden kann, angegeben werden. Nach der Angabe aller Optionen folgt der Pfad zum `bowtie`-Index des Referenzgenoms und abschließend die `fastq` Datei.

Um einen Überblick zu erhalten, wo die Reads gemappt haben, kann mit dem Programm `samtools` [45] gearbeitet werden. Dabei wird zunächst die `sam` Datei in eine binäre `bam` Datei formatiert, anschließend wird sie sortiert und ein Index erstellt. Das Kommando `samtools mpileup` berechnet dann, wo und wieviele Reads gemappt haben und gibt eine Tabelle mit Chromosom, Position und der *Coverage* aus. Die Coverage beschreibt, wie viele Reads auf genau diese Position im Referenzgenom von `bowtie` gemappt wurden. Unser Aufruf zur Berechnung der Coverage von ZH1:

```
samtools view -b -S zh1.sam > zh1.bam
samtools sort zh1.bam zh1_sort
samtools index zh1_sort.bam
samtools view -b zh1_sort.bam | samtools mpileup - >
   coverage_zh1.txt
```

Wenn Datensätze zu vergleichen sind, die auf unterschiedlichen Geräten oder mit ungleichen Methoden erhoben wurden, empfiehlt sich eine technische Normalisierung auf die Anzahl gemappter Reads. Es ist zu beachten, dass diese Normalisierung vor jeder weiteren Form der Filterung durchzuführen ist.

Um das Hintergrundrauschen in den Daten zu verringern, wenden wir eine Filterung nach zwei Kriterien an: die Coverage an jeder Position muss mindestens die Höhe zwei haben und insgesamt muss ein Peak auf die Höhe acht anwachsen. Ein Beispiel für das Ergebnis einer Anwendung dieser Filtermethode ist in Tabelle 2.3 angegeben. Wenn Datensätze verglichen werden sollen, zwischen denen eine technische Normalisierung notwendig ist, ist dieser Normalisierungsfaktor auch auf die 2/8er Regel anzuwenden.

Tab. 2.3: Beispiel der 2/8er Regel.

Chromosom	Position	Coverage	Filterergebnis
chr1	1	2	2
chr1	2	10	10
chr1	9	2	-
chr1	10	1	-
chr1	11	4	4
chr1	12	8	8

Um später besser mit den Datensätzen zu arbeiten, werden die Coverage Tabellen in eine relationale Datenbank eingepflegt. Die Tabellendefinition erfolgt mit dem Skript `createTable.sql`:

```
DROP TABLE if EXISTS coverage_zh1;
CREATE TABLE coverage_zh1(
       chr varchar(2),
       pos int(10),
       cov int(10),
       primary key(chr,pos)
)ENGINE = MyISAM DEFAULT CHARSET=latin1;
CREATE INDEX coverage_zh1 USING btree ON coverage_zh1(
   chr,pos);
```

Die Datentabelle wird mit folgendem Aufruf importiert:

```
mysql -D <DB-Name> < createTable.sql
mysqlimport <DB-Name> -L coverage_zh1.txt
```

Zur Visualisierung werden aus den Coverage Tabellen *Custom Tracks* auf dem UCSC Genom-Browser (mm9) erstellt [39].

2.2.2 Definition von Peaks

Die Analyse hat das Ziel, die Histonmarkierungen an bekannten Genen zu quantifizieren und zu vergleichen. Darum berechnen wir in einem 2.000 Bp großen symmetrischen Fenster um bekannte TSS die Summe der Reads, die dorthin gemappt wurden. Diese Summe bezeichnen wir im Weiteren als Größe des TSS Peaks. Als Referenz für TSS verwenden wir die Annotationen von ENSEMBL (`ensGene` auf mm9) [25]. Zum weiteren Arbeiten werden die TSS Peaks ebenfalls in der relationalen Datenbank unter dem Tabellennamen `geneCov` eingepflegt, wobei die ENSMUST Kennziffer als Identifikation der TSS verwendet wird.

Neben den TSS Peaks können im UCSC Genom-Browser auch die genauen Grenzen der Histonmarkierung bestimmt werden. Die Summe der Reads, die auf diesen Bereich gemappt wurden, entspricht dann der Größe des exakten Peaks. Im Genom gibt es Gebiete, in denen nur ein Datensatz eine Histonmarkierung aufweist. Wir sprechen dann von exklusiven Peaks.

2.2.3 Analyse der Anreicherung in funktionellen Einheiten

Um herauszufinden, ob Gruppen von Genen in einer bestimmten funktionellen Einheit angereichert vorliegen, verwenden wir das Programm `WebGestalt` [62]. Wir betrachten Anreicherungen unterschiedlich markierter Gene zwischen ZH1 und ZH6 gegenüber allen bekannten Genen der Maus und allen in unseren Datensätzen in der Leber markierten Genen. Die Auswahl der funktionellen Einheiten erfolgte durch einen hypergeometrischen Test mit einer Benjamin-Hochberg Korrektur für multiples Testen. Mindestens zwei Gene müssen in einer funkionellen Einheit enthalten sein, damit diese für die weitere Anreicherungsanalyse berücksichtigt wird. Die betrachteten funktionellen Einheiten sind *KEGG Pathways*. KEGG ist die Abkürzung für die Kyoto Enzyklopädie von Genen und Genomen, eine Datenbank mit deren Hilfe molekulare Daten in einen biologischen Zusammenhang gestellt werden können [34] [35] [36].

2.2.4 Normalisierung

Wie oben bereits erwähnt, ist eine technische Normalisierung über die Summe aller gemappten Reads sinnvoll, wenn Datensätze aus unterschiedlichen Experimenten betrachtet werden. Des Weiteren wird von Vandesompele et al. [61] empfohlen über mehrere der *Housekeeping-Gene* zu normalisieren. Housekepping-Gene sind Gene, von denen erwartet wird, dass sie in allen Geweben gleichmäßig exprimiert und markiert sind. Wir wählen die in Tabelle 2.4 genannten Gene zur biologischen Normalisierung aus.

Wenn sich durch die biologische Normalisierung kein gleichmäßiges Bild ergibt, kann auch über die Summe aller Peaks, die in TSS Peaks landen, normalisiert werden. Das entspricht der Summe über die Coverages in der entsprechenden `geneCov` Tabelle. Zu beachten ist, dass einige Gene mehrere annotierte TSS haben und somit mehrmals in diesem Normalisierungsfaktor berücksichtigt werden. Um das zu vermeiden, kann pro Gen eine TSS ausgewählt werden.

Tab. 2.4: Gene der biologischen Normalisierung in mm9.

Gen	Position der exakten Peaks von ZH1 und ZH6
Gapdh	chr6: 125.114.128 - 125.115.683
Sdha	chr13: 74.486.750 - 74.488.295
Tbp	chr17: 15.636.545 - 15.637.510
Ubc	chr5: 125.868.558 - 125.870.697
YWHAZ	chr15: 36.721.910 - 36.724.387
Actb	chr5: 143.667.104 - 143.668.526
B2m	chr2: 121.973.324 - 121.975.043
Rpl13a	chr7: 52.382.913 - 52.385.009
Hmbs	chr9: 44.151.210 - 44.152.372

2.2.5 Differenzmaße

Um Peakgrößen miteinander ins Verhältnis zu setzen, brauchen wir ein Differenzmaß. In Abbildung 2.1 sind vier verschiedene Differenzmaße gegenübergestellt, die in den Gleichungen 2.1 bis 2.4 definiert sind.

Differenz

$$\Delta_D(X,Y) = X - Y \tag{2.1}$$

Quotient

$$\Delta_Q(X,Y) = \frac{X}{Y} \tag{2.2}$$

Produkt aus Differenz und Quotient

$$\Delta_P(X,Y) = \begin{cases} (X-Y)\cdot\frac{X}{Y}, & \text{falls } X > Y \\ (X-Y)\cdot\frac{Y}{X}, & \text{falls } X < Y \end{cases} \tag{2.3}$$

Normierte Differenz

$$\Delta_N(X,Y) = \frac{X-Y}{\sqrt{X+Y}} \tag{2.4}$$

Die Verwendung der Differenz führt dazu, dass kleine Werte unterrepräsentiert und große Werte überrepräsentiert werden (Abbildung 2.1 A). Beim Quotienten hingegen liegt der umgekehrte Fall vor. Wie in Abbildung 2.1 B deutlich zu erkennen ist, liegen außerhalb der grünen Linien, die das 95% Quantil markieren, fast ausschließlich kleine Peakgrößen. Deshalb ist dieses Unterscheidungsmaß, das häufig in der Auswertung von Transkriptanalysen verwendet wird, für unsere Zwecke ungeeignet.

Die beiden weiteren Differenzmaße verhalten sich besser. Es werden weder kleine noch große Peakhöhen überbewertet. Das Problem des Produkts aus Differenz und Quotient ist, dass sobald einer der Werte Null ist, es nicht definiert ist und somit exklusive Peaks eine getrennte Analyse erfordern. Es wurde gezeigt, dass die Normierte Differenz zur Identifikation von ChIP-Seq Peaks besser geeignet ist als vergleichbare Differenzmaße, die nur mit Summen und Differenzen arbeiten [52]. Die Normierte Differenz wird auch erfolgreich in einem Softwarepaket verwendet, um schwache H3K4me3 Peaks in ChIP-Seq Daten zu detektieren [66].

Der Zähler der Normierten Differenz ist eine Schätzung der Standardabweichung. Das folgt aus folgender Überlegung: Seien X und Y zwei Messwerte der Verteilung M. Ihr Mittelwert, der als Schätzer des Erwartungswerts der Verteilung M dient, ist $\frac{1}{2}\cdot(X+Y)$. Die Varianz von M kann geschätzt werden durch:

$$\begin{aligned} Var(M) &= (X - E(M))^2 \cdot P(M = X) + (Y - E(M))^2 \cdot P(M = Y) \\ &= (X - \frac{1}{2}\cdot(X+Y))^2 \cdot \frac{1}{2} + (Y - \frac{1}{2}\cdot(X+Y))^2 \cdot \frac{1}{2} \\ &= \frac{1}{8}X^2 - \frac{1}{4}XY + \frac{1}{8}Y^2 + \frac{1}{8}Y^2 - \frac{1}{4}XY + \frac{1}{8}X^2 = \frac{1}{4}(X-Y)^2 \end{aligned}$$

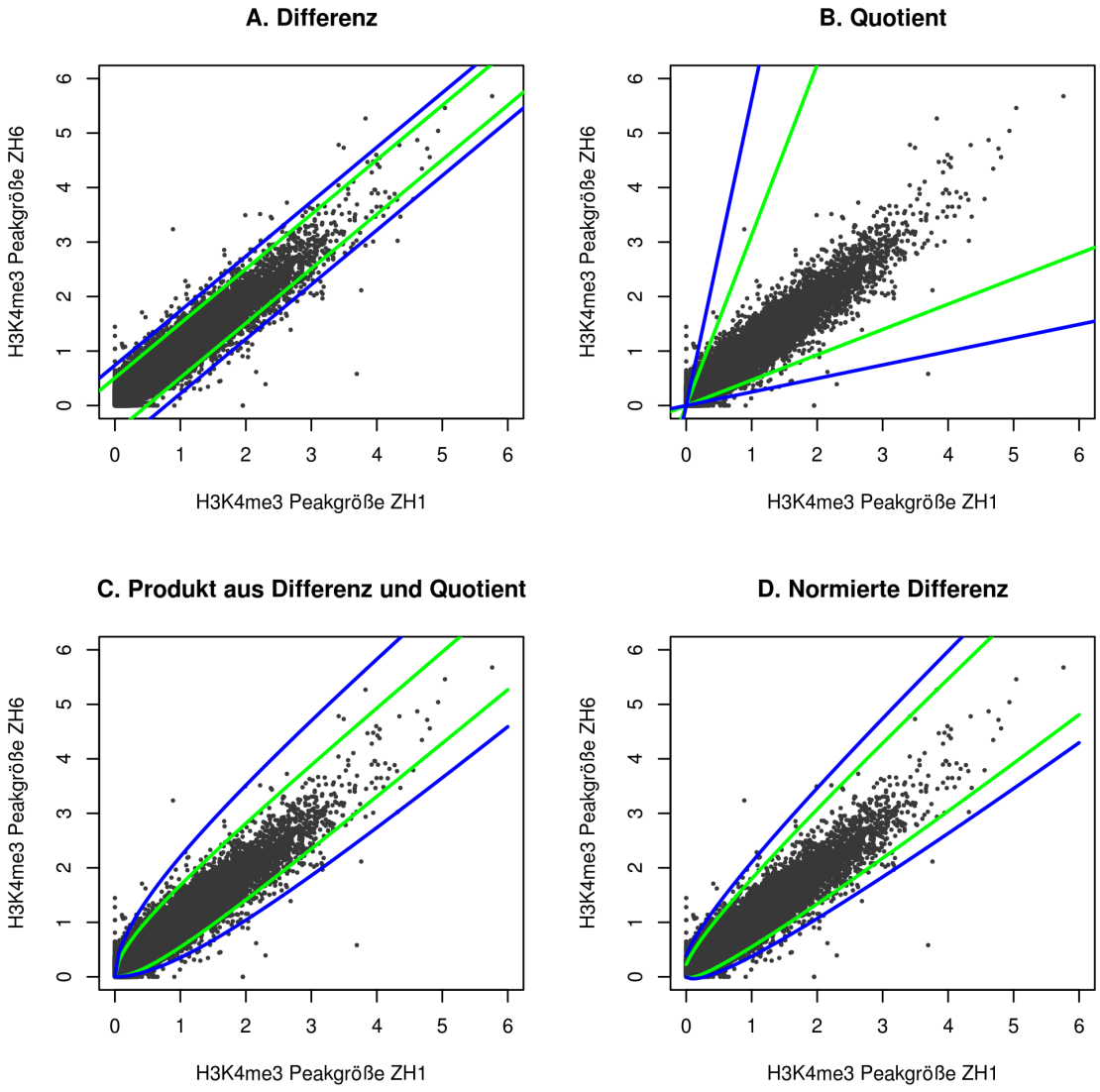

Abb. 2.1: Visualisierung der Grenzlinien verschiedener Differenzmaße. Die grünen Linien kennzeichnen jeweils das 2,5% und 97,5% Quantil, die blauen das 0,5% und das 99,5% Quantil. Neun Gene mit Peakgrößen größer als sechs sind zur besseren Übersicht weggelassen.

Die Standardabweichung ergibt sich als Wurzel aus der Varianz zu $\frac{1}{2}(X - Y)$. Da eine Skalierung bei einem Differenzmaß keinen Unterschied macht, kann der Faktor weggelassen werden und es folgt der Zähler der Normierten Differenz.

Der Nenner der Normierten Differenz dient zur Normalisierung. Es wird nicht (X+Y) gewählt, denn das würde die gleiche Rangordnung wie bei einer einfachen Quotient Berechnung (X/Y) bewirken. Das zeigt folgende Rechnung: Seien X_1, X_2, Y_1, Y_2 größer eins und beliebig, aber fest. Dann folgt:

$$\begin{aligned}
&\frac{X_1 - Y_1}{X_1 + Y_1} > \frac{X_2 - Y_2}{X_2 + Y_2} \\
\Leftrightarrow\ &(X_1 - Y_1) \cdot (X_2 + Y_2) > (X_2 - Y_2) \cdot (X_1 + Y_1) \\
\Leftrightarrow\ &X_1X_2 + X_1Y_2 - X_2Y_1 - Y_1Y_2 > X_1X_2 + X_2Y_1 - X_1Y_2 - Y_1Y_2 \\
\Leftrightarrow\ &X_1Y_2 - X_2Y_1 > X_2Y_1 - X_1Y_2 \\
\Leftrightarrow\ &2 \cdot X_1Y_2 > 2 \cdot X_2Y_1 \\
\Leftrightarrow\ &X_1Y_2 > X_2Y_1 \\
\Leftrightarrow\ &\frac{X_1}{Y_1} > \frac{X_2}{Y_2}
\end{aligned}$$

Um den Unterschied zwischen Histonmarkierungen zu quantifizieren, verwenden wir deshalb die Normierte Differenz nach Gleichung 2.4.

2.2.6 Finden von SNPs

Mit den Ergebnissen von Sequenzierexperimenten können auch Einzelnukleotidpolymorphismen (SNP, single nucleotide polymorphism) detektiert werden. Dazu verwenden wir `bcftools`, ein Programmpaket von `samtools` [45], und das `variant call format` [13] mit folgendem Aufruf.

```
samtools faidx index/m_musculus_ncbi37.fasta
samtools mpileup -E -C50 -uf index/m_musculus_ncbi37.
   fasta zh1_sort.bam |
bcftools view -bvgc - |
bcftools view - |
vcfutils.pl varFilter -D100 > zh1.vcf
```

Für diesen Befehl ist eine `fasta` Datei des Referenzgenoms notwendig, die mit `faidx` zu indizieren ist. Die `fasta` Datei kann entweder aus dem Internet heruntergeladen, oder aus dem `bowtie` Index mit Hilfe von `bowtie-inspect` extrahiert werden.

2.3 Chromatin Immunopräzipitation gefolgt von quantitativer PCR

2.3.1 Chromatinpräparation

Zur Durchführung der Chromatin Immunopräzipitation wird das SimpleChIP™ Enzymatic Chromatin IP Kit von Cell Signaling (Beverly, MA) verwendet.

Der erste Schritt der Chromatinpräparation umfasst eine chemische Vernetzungsreaktion, wobei die DNA mit den Histonproteinen verbunden wird. 70 bis 220 mg tiefgefrorenes Lebergewebe werden in einem 2 ml Reaktionsgefäß durch Zugabe von 1 ml einer 1%igen Formaldehydlösung in Phosphatgepufferter Salzlösung (PBS) fixiert. Dabei wird das Gewebe mit einer kleinen Schere eine Minute lang zerkleinert und liegt danach neun Minuten bei Raumtemperatur unberührt. Nach kurzer Zentrifugation (30 Sekunden bei 4820 rpm) wird der Überstand verworfen. Die Gewebestücke werden in 1 ml 125 mM Glycin in PBS mindestens fünf Minuten auf einem rotierenden Rad inkubiert, um die Vernetzungsreaktion zu stoppen.

Nach kurzer Zentrifugation (30 Sekunden bei 4820 rpm) wird der Überstand verworfen und die Proben werden auf Eis zweimal mit je 1 ml eiskalten PBS/PMSF (1 mM) gewaschen. Anschließend wird ein Lysepuffer hinzugegeben (Puffer A aus dem SimpleChIP™ Enzymatic Chromatin IP Kit, Cell Signaling, Beverly, MA), und die Proben werden mit Hilfe eines Dounce Homogenisators auf Eis homogenisiert. Dabei werden zehnmal Pistill A und zehnmal Pistill B langsam rauf- und runtergefahren. Zur weiteren Zelllyse bleiben die Proben mindestens weitere zehn Minuten auf Eis. Zur Isolation der Zellkerne werden die Proben jeweils zweimal zunächst in einer auf 4°C gekühlten Mikrozentrifuge für vier Minuten bei 3000 rpm zentrifugiert. Der Überstand wird entfernt und das Pellet anschließend im Puffer B (SimpleChIP™ Enzymatic Chromatin IP Kit, Cell Signaling, Beverly, MA) resuspendiert. Nach diesem Schritt können die Proben bei Bedarf auf Trockeneis schockgefroren und bei -80°C gelagert werden.

Die Zelllysate werden mit 5 μl Nuclease pro $4{\cdot}10^7$ Zellen versetzt und bei 37°C 20 Minuten inkubiert. Alle drei bis fünf Minuten werden die Reaktionsgefäße zur Durchmischung des Inhalts gedreht. Durch eine Zugabe von 100 μl EDTA (0,5 M) wird die Reaktion gestoppt. Die Proben werden eine Minute bei 13.000 rpm zentrifugiert und der Überstand wird entfernt. Das Pellet aus Zellkernen wird in 500 μl ChIP Puffer (SimpleChIP™ Enzymatic Chromatin IP Kit, Cell Signaling, Beverly, MA) gelöst und ruht zehn Minuten auf Eis. Die anschließende Ultraschallbehandlung (*Duty Cycle* 80% und *Output Control* 2) der Proben erfolgt in 1,5 ml Reaktionsgefäßen je vier bis zehnmal 20 Sekunden lang mit einer Minute Kühlzeit dazwischen. Eine abschließende zehnminütige Zentrifugation bei 4°C und 10.000 rmp liefert als Überstand gelöstes Chromatin, welches bei -80°C gelagert wird. Davor werden Aliquots zur Überprüfung der Länge der DNA Stücke (25μl) und zur Feststellung der DNA Konzentration (1μl) entnommen.

2.3.2 Kontrolle des präparierten Chromatins

Bestimmung der DNA Konzentration im Chromatin

Wir verwenden das Quant-iT^{TM} dsDNA Broad-Range Assay Kit (Life Technologies, Carlsbad, CA), um die DNA Konzentration im Chromatin zu bestimmen. Das Chromatin wird dabei eins zu zwanzig mit der Farbreagenz gemischt und vor Licht geschützt fünf Minuten inkubiert. Die Auswertung erfolgt mit dem Fluoreszensspektrometer *Nanodrop 3300 Fluorospectrometer*.

Überprüfung der Länge der DNA Stücke

Um zu überprüfen, ob die Nuklease das Chromatin erfolgreich in Fragmente bestehend aus einem bis fünf Nukleosomen verdaut hat (150 bis 900 Bp), wird ein 25 μl Chromatin-Aliquot bei 37°C für 30 Minuten mit 1 μl RNAse (10 mg/ml; Thermo Scientific Fermentas, Waltham, MA) inkubiert. Der Verdau wird durch Zugabe von 1 μl Proteinase K bei 65°C weitere zwei Stunden lang fortgesetzt. Ein Aliquot dieses Verdaus (6μl) wird mit 3μl des Laufpuffers (6X Thermo Specific MassRuler Loading Dye, Thermo Scientific) gemischt und auf ein 1% Agarosegel aufgetragen. Als Markierung wird eine niedrigmolekulare DNA Leiter verwendet, die DNA Längen von 50, 200, 400, 850 und 1500 Bp markiert. Jede Chromatinprobe wird in eine Tasche des Gels eingefüllt und wandert 35 Minuten bei 120 V.

2.3.3 Ergebnisse der Kontrollen der Chromatinpräparation

Verhältnis von Gewebeeinwaage zu DNA Konzentration im Chromatin

Abbildung 2.2 zeigt den Zusammenhang zwischen der Gewebeeinwaage zu Beginn des Experiments und der in Abschnitt 2.3.2 beschriebenen mit dem *Nanodrop* gemessenen DNA Konzentration am Ende der Chromatinpräparation. Das Ziel der Beschallung ist das Auftrennen der Zellkernmembran. Eine Faustregel ist, dass die Lösung klar wird, wenn die Zellkernmembran aufgebrochen wurde. Da dies insbesondere bei den Proben mit einer hohen Gewebeeinwaage nicht der Fall war, lösten wir im weiteren Verlauf des Experiments die Zellkerne dieser Proben in der doppelten Menge des zur Beschallung verwendeten ChIP Puffers (rote Punkte in Abbildung 2.2). Entsprechend ist das Ergebnis der DNA Konzentrationsmessung bei diesen verdünnten Proben geringer und muss zur Berechnung der Korrelation zwischen Einwaage und DNA Konzentration angepasst werden. Im rechten Teil von Abbildung 2.2 ist die DNA Konzentration der roten Datenpunkte (der y-Wert) verdoppelt worden. Das führt zu einem Korrelationskoeffizienten (nach Spearman) von 0,69 für alle Datenpunkte. Nur die blauen Punkte für sich betrachtet haben eine Korrelation von 0,89. Es gibt somit eine gute Korrelation zwischen der Gewebeeinwaage zu Beginn des Experiments und der DNA Konzentration nach der Chromatinpräparation.

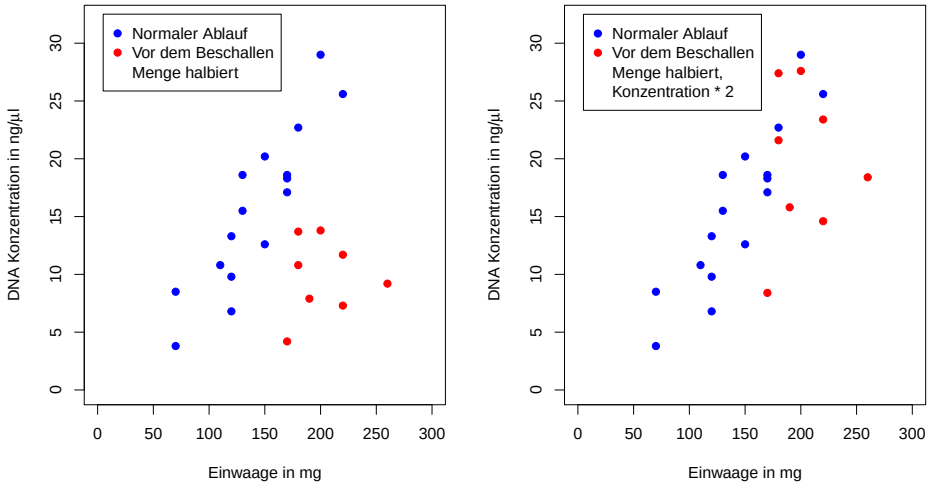

Abb. 2.2: DNA Konzentrationen nach der Beschallung aufgetragen gegen die Gewebeeinwaage zu Beginn der Präparation. In rot markiert sind diejenigen Proben, die wegen hoher Chromatinkonzentration vor der Beschallung 1:1 verdünnt wurden.

Überprüfung der Länge der DNA Stücke

Das Ergebnis der in Abschnitt 2.3.2 beschriebenen Gelelektrophorese ist in Abbildung 2.3 zu sehen. Pro Spur ist eine DNA-Bande sichtbar, die zwischen den 50 Bp und 200 Bp Banden der DNA Leiter liegt. Somit war das Auftrennen der Kernmembran durch die Beschallung und das Zerkleinern der DNA durch die Nuklease erfolgreich. Die Unterschiede in der Intensität der Banden spiegelt einerseits die schwankende DNA Konzentration im Chromatin wieder (Abbildung 2.2), andererseits könnte nicht exaktes Pipettieren der geringen Probemenge in die Geltasche die Ursache sein.

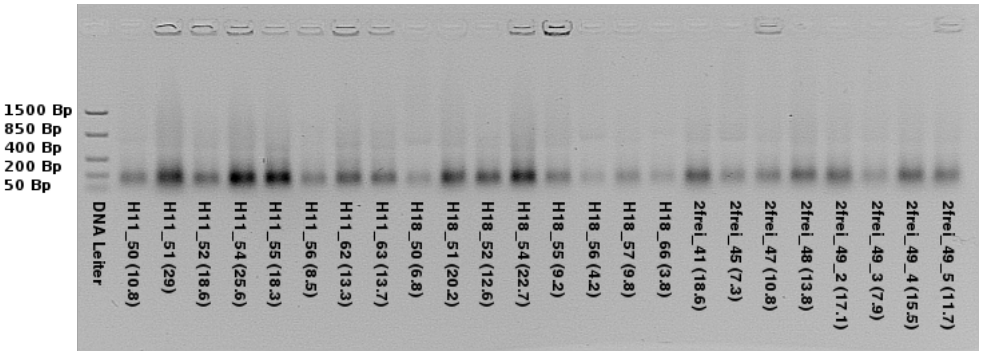

Abb. 2.3: Ergebnis des 1% Agarosegels zur Überprüfung der Länge der DNA Stücke. Die DNA Leiter beschreibt die Stufen: 50, 200, 400, 850 und 1500 Bp. Hinter der Maus ID ist jeweils in Klammern die gemessene DNA Konzentration in ng/μl angegeben.

2.3.4 Immunopräzipitation

Aus der DNA Konzentrationsmessung (Schritt 2.3.2) wird für jede Probe das Volumen an Chromatin berechnet, das 1 μg DNA pro Immunopräzipitation bzw. Blindprobe entspricht. Das Chromatin wird in ChIP (1x) Puffer auf ein Volumen von 300 μl pro Immunopräzipitation verdünnt. Die Proben werden eine Stunde lang im Kühlraum bei 4°C mit einer kleinen Menge Beads vorgereinigt. Bevor die Antikörper zugegeben werden, wird 2% der Menge (6 μl) des vorgereinigten Chromatins als *Input* Probe entnommen und bei -20°C gelagert. Das übrige vorgereinigte Chromatin wird in drei Reaktionsgefäße zu je 300 μl aufgeteilt. Die erste Immunopräzipitation erfolgt mit 2 μl des H3K4me3 Antikörpers (#9751, Cell Signaling), die zweite mit 2 μl des H3K27ac Antikörpers (#4353, Cell Signaling) und die dritte mit 0,5 μl des unspezifischen *Rabbit IgG* Antikörpers (#2729, Cell Signaling). Über Nacht rotieren alle Proben im Kühlraum bei 4°C, damit die Antikörper an die entsprechenden Histonproteine binden können.

Die Immunopräzipitate werden an Agarose Beads (#9007, Cell Signaling) gebunden, indem die Proben mindestens zwei Stunden im Kühlraum bei 4°C auf einem Rad rotieren. Es folgt ein dreimaliges Waschen der Immunopräzipitate mit je 1 ml eiskaltem ChIP (1x) Waschpuffer mit niedrigem Salzgehalt und einmaliges Waschen mit 1 ml ChIP (1x) Waschpuffer mit hohem Salzgehalt. Nach dem letzten Waschen wird vorsichtig die gesamte Waschflüssigkeit entfernt und zu den Beads wird 150 μl Elutionspuffer (ChIP 1x; SimpleChIP^{TM} Enzymatic Chromatin IP Kit, Cell Signaling, Beverly, MA) gegeben. Die Proben werden anschließend 30 Minuten bei 65°C und leichtem Schütteln (1.200 rmp) inkubiert, um die Verbindung zwischen DNA und Histonproteinen zu lösen. Anschließend wird der Überstand mit 6 μl NaCl (5M) und 2 μl Proteinase K zwei Stunden lang bei 65°C inkubiert, damit die Histonproteine und die Antikörper abgebaut werden können.

Parallel werden die Input Proben ebenfalls mit 150 μl Elutionspuffer, 6 μl NaCl (5M) und 2 μl Proteinase K zwei Stunden lang bei 65°C inkubiert.

Die abschließende Aufreinigung der DNA erfolgt mit Spinsäulen (MinElute PCR Purification Kit, Quiagen, Hilden, Deutschland). Die aufgereinigten Proben, 70 μl pro Antikörper und Input Probe, werden bei -20°C bis zur Quantifizierung gelagert.

2.3.5 Quantitative Polymerasekettenreaktion

Zur vergleichenden Bestimmung der immunopräzipitierten DNA Konzentrationen wird eine quantitative Polymerasekettenreaktion (qPCR) durchgeführt. Die aufgereinigten ChIP-DNA Proben aus Schritt 2.3.4 werden auf Eis aufgetaut. Zu den 70 μl ChIP-DNA werden je 130 μl HPLC Wasser und 250 μl *Fast SYBR Green Master Mix* (#4385612, Applied Biosystems, Foster City, CA) gegeben. Die Proben sind vor Licht zu schützen, weil in dieser Mischung der lichtempfindliche Farbstoff SYBR Green I enthalten ist, der an doppelsträngige DNA bindet. Die qPCR wird auf 96 Well-Platten durchgeführt, und zwar für jeden Primer und jede Probe ein Triplikat. In jedes 10 μl Well wird 1 μl Primer vorgelegt und 9 μl der oben beschriebenen Mischung zugefügt. Dabei ist in jeder Zeile zwölf Mal derselbe Primer nebeneinander aufgetragen und jeweils in den ersten drei Spalten wird die Input Probe zugegeben, in den Spalten vier bis sechs die mit dem Antikörper H3K4me3 präzipitierte Probe, in den Spalten sieben bis neun folgt die mit dem Antikörper H3K27ac versehene Probe und in den letzten drei Spalten die Blindprobe mit dem Rabbit IgG Antikörper. Somit besteht eine Reaktionsmischung aus 1 μl Primer, 1 μl ChIP-DNA, 3 μl Wasser und 5 μl Fast SYBR. Die qPCR wurde auf dem Gerät *7900HT Fast Real-Time PCR System* von Applied Biosystems (Singapur) unter Verwendung des Programms `SDS 2.3` durchgeführt. Die in der qPCR verwendeten Primer sind in Tabelle 2.5 aufgelistet. Die funktionelle Klasse und die molekulare Funktion der Gene, die untersucht werden, sind in Tabelle 2.6 angegeben.

In einer PCR wird eine DNA Sequenz vervielfältigt, indem ein dreischrittiger Zyklus mehrfach durchlaufen wird. Zunächst wird die DNA denaturiert, das heißt ihre doppelsträngige Struktur wird durch hohe Temperaturen aufgebrochen. Das erfolgt zu Beginn 20 Sekunden lang bei 95°C. Anschließend wird die Temperatur für 20 Sekunden auf 60°C erniedrigt, um die Anlagerung der Primer zu ermöglichen. Haben die Primer ihre komplementären Stellen in der vorliegenden DNA gefunden, beginnt die Elongation, bei der die DNA Polymerase am 3' Ende der Primer ansetzt und die DNA wieder zu einem Doppelstrang vervollständigt. Es schließt sich eine ein Sekunden lange Steigerung auf 95°C an, damit die Amplifikate wieder denaturiert werden. Die letzten beiden Schritte werden 40 Mal wiederholt.

Bei einer quantitativen PCR wird nach jedem Zyklus die Fluoreszenz gemessen, die durch die Bindung des Farbstoffs SBYR Green I an doppelsträngiger DNA entsteht. Mit steigender DNA Amplifikation wird das Fluoreszenzsignal immer stärker. Das qPCR Gerät hat eine eigene Software, den `SDS RQ Manager`, der nach Abschluss des Experiments den *Ct-Wert* bestimmt [22]. Der Ct-Wert (kurz für *Cycle Threshold*) benennt denjenigen Zyklus, bei dem das Fluoreszenzsignal signifikant über das Hintergrundsignal ansteigt.

Tab. 2.5: Die qPCR Primer.

Locus	Sequenz des Vorwärtsprimers 5'-3'	Sequenz des Rückwärtsprimers 5'-3'	Amplicon-länge	Position im Mausgenom
Gapdh	GCACCAGCATCCCTAGACC	GTGCAGTGCCAGGTGAAAAT	103	chr6:125.115.479-125.115.581
CD36	ACTTGTGGCAAACAGGGCTGGAG	AGTCCGATGAGAGAGTGCAAGGCC	102	chr5:17.355.497-17.355.598
Slc27a5	AGCCACATCTTTATGCAGCCAGCG	CTTGTGCTGCTTGGCTTGGCATTG	150	chr7:13.583.139-13.583.288
Ppara	GTCTGGAGACCCACAGCCACT	AAACAGCTGCGAACACCAATGT	145	chr15:85.566.202-85.566.346
Pparg	GGCTTGTGGGGTCTGGATCTGACT	GAGTGTGGCTTTCCAGCCCGTATC	147	chr6:115.311.404-115.311.550
Acox2	CCTTTGTCTGCCTGCCTTCCTGTG	CTGACCAGGAGAGTTGGGGAGCAG	114	chr14:9.090.867-9.090.982
Cyp4a14	TGCATGGGAAATGCTGGAGGGTCT	TCTCTGGGTTCTTCCAATGGGCCT	117	chr4:115.168.559-115.168.675
Fasn	ACACTTGACCTGGCCTGACCCTAC	AGTGCCCCAGACCCTGTTTCTTGA	151	chr11:120.684.384-120.684.534
Nr3c1	AGCCAGATAAACAAGTCGGCGTGC	TGGAAAAGAGGGGGCGACTGTTGA	83	chr18:39.649.175-39.649.257
Pck1	AGGTTCCCAGGGTGCATGAAAGGT	GCTTTGCAGCTCAGGTCGCCATTA	102	chr2:172.979.596-172.979.698
Insig2	CTGCGGAGGGCAGGCTGAAGAA	CCTCCCCTCTTCCCTCCACTCATCC	115	chr1:123.228.131-123.228.245
Plin5	GCAGGTCGCCTACCAAAGGCTGC	TGCAGGTCACCCAAACAGCCCCC	111	chr17:56.255.950-56.256.060
Igfbp2	CTGTGAGCACTAGTTTTGGGCTTG	CTCTCTTTTCCCACGTGAGAGTCA	88	chr1:72.871.549-72.871.636
Sqle	TGGTGACAAAAGACCGTTTACAGC	CAACGGCTCCTGATTACACACTTC	134	chr15:59.146.990-59.147.123
Serpina6	AAAAATCTCAAGCAAGCAGGCACT	AGACGTGAGTCACCCCTGACAGTA	111	chr12:104.895.140-104.895.250

Tab. 2.6: Gene der qPCR Analyse und ihre Funktion. Informationen wurden entnommen aus den Datenbanken *NCBI* und *UniProt*.

Locus	Genname	Funktionelle Klasse	Molekulare Funktion
Gapdh	Glyceraldehyde-3-phosphate dehydrogenase	Metabolisches Enzym	Glykolyse, Kernfunktion
CD36	CD36 antigen	Zellmembranrezeptor	Transport von Lipoproteinen
Slc27a5	Solute carrier family 27 (fatty acid transporter), member 5	Zellmembranrezeptor	Fettsäuretransport
Ppara	Peroxisome proliferator activated receptor alpha	Transkriptionsfaktor	Fettsäuremetabolismus, Gallensäurenhomöostase
Pparg	Peroxisome proliferator activated receptor gamma	Transkriptionsfaktor	Fettsäureoxidation, Entzündungsreaktion
Acox2	Acyl-Coenzyme A oxidase 2, branched chain	Enzym: Oxidoreductase	Fettsäurebetaoxidation
Cyp4a14	Cytochrome P450, family 4, subfamily a, polypeptide 14	Enzym: Oxidoreductase	Fettsäureabbau
Fasn	Fatty acid synthase	Enzym: Acyltransferase	Fettsäurebiosynthese
Nr3c1	Glucocorticoid receptor	Kernrezeptor	Regt die Transkription von Target Genen an
Pck1	Phosphoenolpyruvate carboxykinase 1, cytosolic	Enzym: Kinaseaktivität	Glukoneogenese, Antwort auf einen Insulinstimulus
Insig2	Insulin induced gene 2	ER Membranprotein	Cholesterolbiosynthese
Plin5	Perilipin 5 (lipid storage droplet protein 5)	Bestandteil von Lipidpartikeln	Regulation des Lipidmetabolismus
Igfbp2	Insulin-like growth factor binding protein 2	Transportprotein für Igf1	Regulation von Zellwachstum
Sqle	Squalene epoxidase	Enzym: Oxidoreductase	Cholesterolmetabolismus
Serpina6	Serine peptidase inhibitor, clade A, member 6	Alpha-Globulin	Transport von Glucocorticoiden im Blut

2.4 Bestimmung der Primereffizienz

Um eine qPCR richtig auszuwerten, ist es wichtig, die Effizienz der verwendeten Primerpaare zu kennen. Diese Effizienz kann experimentell bestimmt werden. Prinzipiell gilt, dass sich bei einer zweifachen Verdünnung der Ct-Wert um eine Einheit vergrößern sollte. Um dies zu überprüfen, wird empfohlen, eine Verdünnungsreihe in 10er-Schritten der zur Kontrolle verwendeten Input DNA zu erstellen (1:1 bis zu 1:1000) und die verwendeten Primer in einer qPCR zu testen. Anschließend wird die Steigung der qPCR Standardkurve berechnet [22]. Eine qPCR Standardkurve ist ein Graph, in der Ct-Werte gegen den dekadischen Logarithmus der eingesetzen DNA Konzentration aufgetragen werden. Beispiele für eine qPCR Standardkurve sind im Abschnitt 3.2.2 in Abbildung 3.12 zu sehen. Die Steigung wird mit Gleichung 2.5 berechnet. Dabei sind x_1 und x_2 die DNA Konzentrationen in ihrer nicht-logarithmisierten Form und es gilt $x_1 < x_2$.

$$\text{Steigung} = \frac{Ct_2 - Ct_1}{log_{10}(x_2) - log_{10}(x_1)} \tag{2.5}$$

Die Primereffizienz E wird dann mit Gleichung 2.6 bestimmt [53].

$$E = 10^{-\frac{1}{\text{Steigung}}} \qquad \text{bzw.} \qquad E_{\%} = (10^{-\frac{1}{\text{Steigung}}} - 1) \cdot 100 \tag{2.6}$$

Die Primereffizienz sollte idealerweise bei zwei liegen. Das entspricht einer Steigung von -3.32 in der qPCR Standardkurve. $E_{\%}$ gibt in Prozent an, wie effizient der Primer ist; dabei ergeben sich Werte zwischen 0 und 100 [54]. Eine Steigung von -3.32 in der qPCR Standardkurve führt zu einem $E_{\%}$ von 100. Wenn die Steigung in der qPCR Standardkurve betragsmäßig größer wird, wird die Effizienz geringer. Wird die Kurve flacher, was einer betragsmäßig kleineren Steigung entspricht, wird die Effizienz größer.

Ein direkter Zusammenhang zwischen der Steigung in der qPCR Standardkurve und der Effizienz des Primers ergibt sich, wenn die Formeln für die Steigung aus Gleichung 2.5 direkt in Gleichung 2.6 einsetzt werden:

$$E = 10^{-\frac{1}{\text{Steigung}}} = 10^{-\frac{log_{10}(x_2) - log_{10}(x_1)}{Ct_2 - Ct_1}} = \frac{10^{\frac{log_{10}(x_1)}{Ct_2 - Ct_1}}}{10^{\frac{log_{10}(x_2)}{Ct_2 - Ct_1}}} = \frac{x_1^{\frac{1}{Ct_2 - Ct_1}}}{x_2^{\frac{1}{Ct_2 - Ct_1}}} = \left(\frac{x_1}{x_2}\right)^{\frac{1}{Ct_2 - Ct_1}}$$

Hier zeigt sich, dass die Basis des Logarithmus in Gleichung 2.5 der Basis in Gleichung 2.6 entspricht und gleichzeitig der Schrittgröße der verwendeten Verdünnungsreihe.

Experimentelles Vorgehen zur Bestimmung der Primereffizienz

In unserem ChIP qPCR Experiment verwenden wir eine DNA Input Menge von 2% verglichen mit der Menge DNA, die zur Immunopräzipitation benötigt wird. Für die Verdünnungsreihe benutzen wir deshalb eine 20%tige, 2%tige und 0,2%ige Konzentration. Die Präparation des Chromatins und die Aufreinigung der DNA erfolgt wie in den Abschnitten 2.3.1 und 2.3.4 beschrieben. Es wird ebenfalls eine quantitative PCR durchgeführt, allerdings werden in jeder Zeile

neunmal derselbe Primer aufgetragen und in den ersten drei Spalten die 20% Input Probe, in den Spalten vier bis sechs die 2% Input Probe und in den Spalten sieben bis neun die 0,2% Input Probe dazu gegeben.

Um zu überprüfen, ob sich die Primereffizienz unabhängig von der genetischen Grundlage der Mäuse verhält, verwenden wir zu ihrer Bestimmung zwei Proben von Bl6 Labormäusen (MausID: 5904 und 6313) und zwei Proben von Wildmäusen (MausID: H11_50 und H18_51).

2.5 Vorgehen bei der Auswertung des ChIP qPCR Experiments

2.5.1 Interpretation der qPCR Daten

Pro Datensatz wird über die Software des qPCR-Messgeräts ein Grenzwert bestimmt, der ausschlaggebend für die exakten Ct-Werte dieses Datensatzes ist. Um Vergleiche zwischen unterschiedlichen Datensätzen möglich zu machen, wird deshalb eine Normalisierung benötigt. Zunächst verwenden wir die %Input Methode [50]. Dafür ist der angepasste Input Ct-Wert nach Gleichung 2.7 zu berechnen. Da für die qPCR Reaktion mit der Input DNA nur 2% der Menge verwendet wurde, aus der immunopräzipitiert wurde, ist von dem gemessenen Ct-Wert der Logarithmus von 50 zur Basis 2 abzuziehen. Das ergibt sich daraus, dass von 2% zu 100% eine Verfünfzigfachung der Menge nötig ist und in jedem Schritt der qPCR, das bedeutet pro Ct-Wert, eine Verdoppelung der Menge erreicht wird. %Input Werte werden nach Gleichung 2.8 berechnet. Ct(Immunopräzipitiert) kann sowohl für die Ct-Werte der Immunopräzipitation mit H3K4me3, H3K27ac oder auch dem unspezifischen Antikörper stehen.

$$\text{Ct(Angepasster Input)} = \text{Ct(Input)} - \log_2(50) \tag{2.7}$$

$$\Delta\text{Ct} = \text{Ct(Immunopräzipitiert)} - \text{Ct(Angepasster Input)}$$

$$\%\text{Input} = 100 \cdot 2^{-\Delta\text{Ct}} \tag{2.8}$$

Das Ergebnis von Gleichung 2.8 ist der Anteil, der aus dem Chromatin präzipitiert wurde. Abbildung 2.4 gibt einen Überblick über die Größenordnungen der %Input Werte. In Abbildung 2.4 A sind die Werte nach den Antikörpern aufgeteilt. Die y-Achse gibt auf einer logarithmischen Skala die %Input Werte gemittelt über alle 24 Wildmausdatensätze an. Es fällt auf, dass die Werte des unspezifischen Antikörpers zwar die niedrigsten sind, aber die Werte des H3K27ac Antikörpers nur einen halben bis einen Prozentpunkt höher liegen. Die %Input Werte des Primers CD36 liegen unabhängig von dem verwendeten Antikörper zwischen 0,5 und eins. Anders verhält es sich bei Gapdh: beim unspezifischen Antikörper ist der %Input Wert der niedrigste, bei H3K27ac liegt er unterhalb des ersten Quartils, und bei H3K4me3 zeigt er den höchsten Wert. Acox2 bewegt sich dazu komplementär: die %Input Werte nehmen beim unspezifischen Antikörper und bei H3K27ac die höchsten Werte an, aber bei H3K4me3 liegt der Wert unterhalb des ersten Quartils.

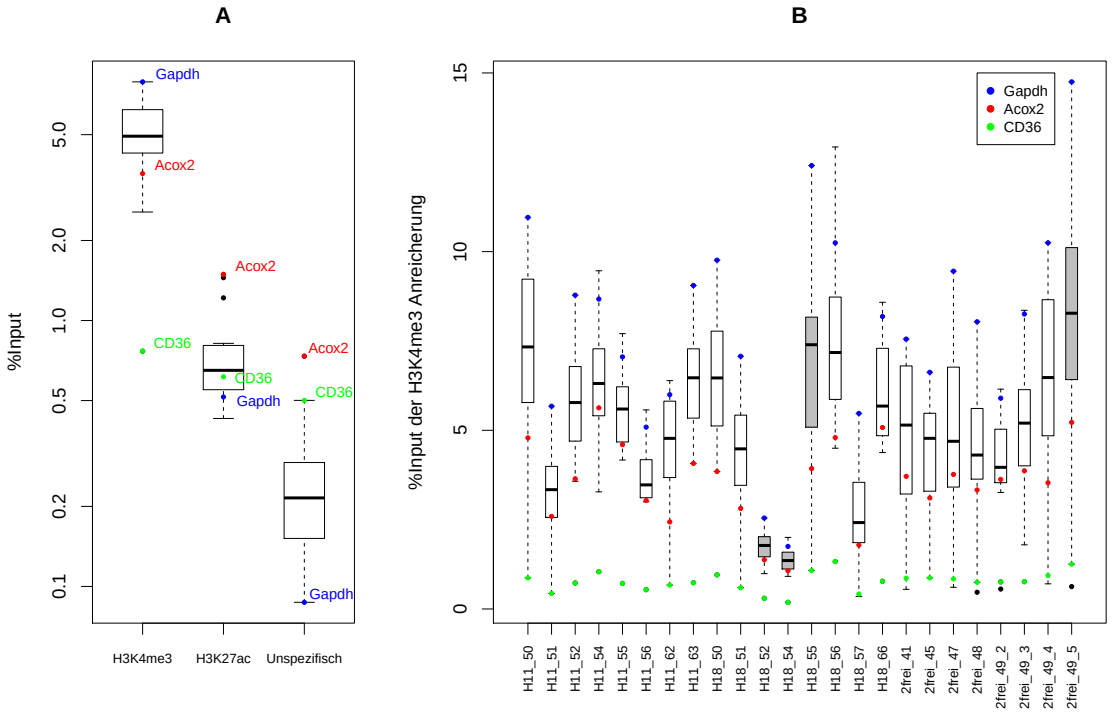

Abb. 2.4: Die %Input Werte. A: Boxplots der %Input Werte pro Immunopräzipitation, gemittelt über die 24 Wildmausdatensätze. B: Boxplots der %Input Werte der H3K4me3 Anreicherung pro untersuchter Wildmaus. Farblich markiert sind auffällige Primer. Grau markiert sind die Mäuse, bei denen eine Wiederholungsmessung durchgeführt wurde.

Wiederholungsmessungen

In Abbildung 2.4 B werden die H3K4me3 Anreicherungen der 24 untersuchten Wildmäuse einzeln betrachtet. Auffällig sind die unterschiedlichen Varianzen zwischen den Individuen. Gapdh gehört zu den am höchsten markierten Genen, während CD36 konsistent bei den nierigsten Werten liegt. Acox2 liegt stets unterhalb des ersten Quartils. Die Rangreihenfolge der Markierungen an den jeweiligen Genen ist in den %Input Werten zwischen den Tieren sehr ähnlich.

Auffällig ist, dass die %Input Werte für zwei Datensätze besonders niedrig sind: H18_52 und H18_54. Es könnte sein, dass hier die Immunopräzipitation nicht richtig funktioniert hat. Deshalb wurden diese beiden Messungen wiederholt. Da der Versuchsablauf auf vier Mäuse ausgelegt ist, wurden zu den beiden Mäusen mit sehr geringen %Input Werten noch zwei Mäuse für die Wiederholungsmessungen ausgewählt, deren %Input Werte eine besonders große Varianz aufwiesen (H18_55 und 2frei_49_5).

Das Ergebnis der Wiederholungsmessungen wird in Abbildung 2.5 deutlich. Ein Boxplot der ersten %Input Ergebnisse der H3K4me3 Anreicherung ist jeweils links neben dem der entsprechenden Wiederholungsmessung dargestellt. Markiert sind wie in Abbildung 2.4 die Resultate der

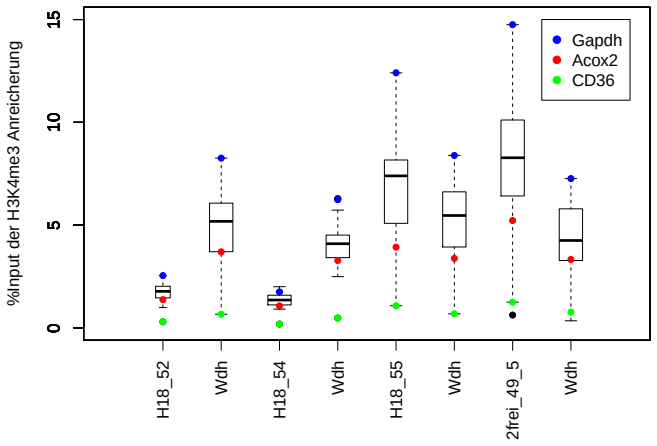

Abb. 2.5: %Input Werte der Wiederholungsmessungen.

Primer Gapdh, Acox2 und CD36. Auch hier führt CD36 meist zu den niedrigsten %Input Werten, die entsprechenden Werte von Acox2 sind kleiner oder gleich dem ersten Quartil und Gapdh hat die größten Werte. Die Wiederholungsmessungen der Proben H18_52 und H18_54 führen zu %Input Werten, die vergleichbar zu den Messungen der anderen Wildmäuse sind. Deshalb werden wir im Weiteren die neuen Messungen verwenden. Die Wiederholungsmessungen der Mäuse H18_55 und 2frei_49_5 zeigen eine geringere Streuung. Da die Ergebnisse ähnlich zur ersten Messung sind, verwenden wir für diese beiden Datensätze ab jetzt den Mittelwert aus der ersten und zweiten Messung.

2.5.2 Normalisierung

Um qPCR Datensätze miteinander zu vergleichen, ist eine weitergehende Normalisierung als die %Input Methode erforderlich. Vandesompele et al. beschreiben verschiedene Housekeeping-Gene, die zur Normalisierung bei Transkriptionsstudien verwendet werden sollten, darunter auch Gapdh [61]. Dabei wird empfohlen, in einer Reverse Transkriptase-Polymerasekettenreaktion (RT-PCR) mindestens drei Gene als Referenz zu verwenden. Wir binden in unsere Experimente nur das Gen Gapdh in die Analyse mit ein (siehe Tabelle der verwendeten Primer 2.5), da die Anzahl an messbaren Loci mit 15 begrenzt ist.

Eine weitere Möglichkeit der Normalisierung besteht darin, pro Maus den Mittelwert der %Input Werte aller untersuchten Loci zu bestimmen und dadurch zu dividieren. Somit erhält jede Maus an jedem Locus im Schnitt den Wert eins und Variationen in der Anreicherung führen zu entsprechend höheren oder niedrigeren Werten.

2.5.3 Hauptkomponentenanalyse

Eine Hauptkomponentenanalyse (PCA, principal component analysis) dient dazu, einen Datensatz zu strukturieren, indem das Koordinatensystem der Analyse verändert wird. Mathematisch gesehen wird eine orthogonale Tranformation durchgeführt, wobei die Koordinatenachsen (Variablen) in eine neue Menge unkorrelierter Variablen überführt werden. Diese neuen Variablen heißen Hauptkomponenten und sind eine Linearkombination der alten Variablen. Dabei soll die erste Hauptkomponente in die Richtung der größten Varianz des Datensatzes zeigen. Jede weitere k. Hauptkomponente soll orthogonal zu den vorherigen k-1 Hauptkomponenten sein und die Richtung der k. größten Varianz aufweisen. Das wird erreicht, indem eine Eigenwertzerlegung der Kovarianzmatrix, bzw. der Korrelationsmatrix, durchgeführt wird. Die erste Hauptkomponente entspricht dann dem Eigenvektor zum größten Eigenwert der Kovarianz- bzw. Korrelationsmatrix, die k. Hauptkomponente dem Eigenvektor zum k. größten Eigenwert. Für eine Herleitung dieses Zusammenhangs siehe [65].

Wir verwenden folgenden Aufruf in `R` um eine Hauptkomponentenanalyse durchzuführen:

```
pca.1 <- prcomp(data[,2:16], center=TRUE, scale=TRUE)
```

In dem `data.frame` „data“ steht in der ersten Spalte die Identifikationsnummer der Maus und in den Spalten zwei bis 16 folgen die mittelwertnormalisierten qPCR Anreicherungswerte an den 15 gemessenen Loci. In unserem Fall entsprechen die untersuchten Loci den Variablen, die wir transformieren wollen. Die Option `center=TRUE` besagt, dass die Daten noch nicht zentriert um Null vorliegen, sondern pro Variable der Mittelwert über diese Dimension subtrahiert werden muss. Da wir mittelwertnormalisierte Werte haben, ist der Mittelwert, der abzuziehen ist, eins. Wir verwenden außerdem die Option `scale=TRUE`, mit der die Varianz aller Variablen vor der weiteren Berechnung auf eins skaliert wird. Damit haben alle Variablen das gleiche Gewicht und es können Skalierungsprobleme umgangen werden. Das entspricht in der oben genannten Berechnung der Hauptkomponenten einer Verwendung der Korrelationsmatrix anstelle der Kovarianzmatrix.

Die Berechnung mit der in `R` integrierten Funktion `prcomp` kann nachgestellt werden, indem zunächst die Korrelationsmatrix des Datensatzes berechnet und anschließend eine Eigenwertzerlegung vorgenommen wird. Die Ergebnisse von `pca.1` und `pca.2` sind identisch.

```
pca.2 <- eigen(cor(data[,2:16]))
```

Um die Ergebnisse zu veranschaulichen, sind die Datenpunkte im neuen, transformierten Koordinatensystem darzustellen. Da wir die Variablen vor der Analyse zentriert und skaliert haben, müssen wir diese Transformation auch mit den Datenpunkten durchführen, bevor wir sie visualisieren können. In `R` funktioniert das mit Hilfe der Funktion `predict`. Der folgende Aufruf gibt die Datenpunkte projiziert auf die ersten beiden Hauptkomponenten an:

```
pc.1 <- predict(pca.1)[,1:2]
```

Dabei vollzieht die `predict` Funktion zunächst eine Skalierung und anschließend wird das Skalarprodukt zwischen den skalierten Datenpunkten und den Eigenvektoren der Korrela-

tionsmatrix berechnet. Folgender Aufruf führt zu demselben Ergebnis wie die Verwendung der `predict` Funktion:

```
sc.data <- scale(as.matrix(data[,2:16]), center=sapply(
   data[,2:16], mean), scale=sapply(data[,2:16], sd))
pc.2 <- sc.data %*% pca.2$vectors[,1:2]
```

2.5.4 Statistische Auswertung der qPCR Ergebnisse

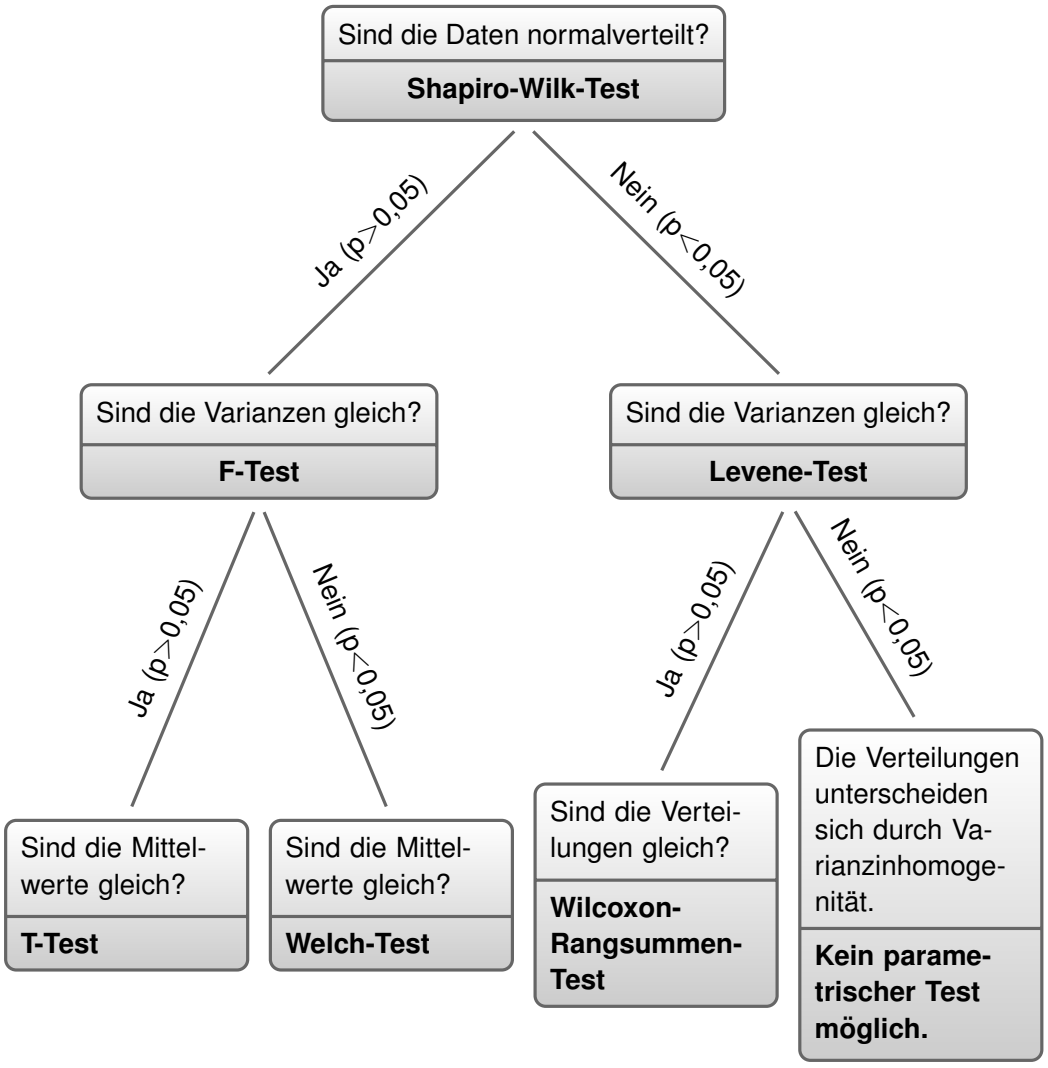

Abb. 2.6: Überblick über das Vorgehen beim statistischen Testen der qPCR Daten.

Ziel der statistischen Auswertung der qPCR Ergebnisse ist herauszufinden, ob sich die Mäuse aus den Häusern H11 und H18 signifikant in ihren H3K4me3 oder H3K27ac Anreicherungen an bestimmten Genen voneinander und/oder von den Mäusen ohne Hauszugehörigkeit unterscheiden. Einen Überblick über die benötigten statistischen Tests gibt Abbildung 2.6.

Zuerst ist zu überprüfen, ob die gemessenen Daten aus einer normalverteilten Grundgesamtheit stammen. Dafür kann der Shapiro-Wilk-Test jeweils pro Mausgruppe pro Primer angewendet werden. Die Nullhypothese dieses Tests ist, dass die Daten normalverteilt sind. Deshalb dürfen wir bei einem p-Wert größer als 0,05 davon ausgehen, dass die Daten normalverteilt sind. Ist der p-Wert hingegen kleiner als 0,05, dürfen im Weiteren nur Tests angewendet werden, die keine normalverteilten Daten voraussetzen.

Ergibt der Shapiro-Wilk-Test keine Ablehnung der Nullhypothese, kann als nächstes mit dem F-Test die Gleichheit der Varianzen in den zu vergleichenden Gruppen getestet werden. Bei einem p-Wert des F-Tests kleiner als 0,05 ist die Ungleichheit der Varianzen anzunehmen und es dürfen keine Tests verwendet werden, die von einer Gleichheit der Varianzen ausgehen.

Folgt aus dem F-Test keine Ablehnung der Nullhypothese, darf zum Testen auf Unterschiede in den Mittelwerten der Gruppen der Zweistichproben-T-Test verwendet werden. Der T-Test testet die Nullhypothese der Gleichheit der Mittelwerte unter den Voraussetzungen, dass die Daten normalverteilt sind und die Varianzen in den Gruppen gleich sind. Bei einem p-Wert kleiner als 0,05 kann die Nullhypothese auf einem Signifikanznivenau von 5% abgelehnt werden und es kann ein signifikanter Unterschied in den Mittelwerten der beiden untersuchten Gruppen festgestellt werden.

Wird mit Hilfe des F-Tests festgestellt, dass die Varianzen in den Gruppen ungleich sind, ist zum Testen auf Unterschiede in den Mittelwerten der Welch-Test zu verwenden. Dieser schätzt pro Gruppe eine Varianz, die zur Berechnung der Teststatistik benötigt wird. Ergibt der Welch-Test einen p-Wert kleiner als 0,05, kann auf einem Signifikanzniveau von 5% die Nullhypothese abgelehnt werden und geschlossen werden, dass sich die Mittelwerte der Gruppen signifikant unterscheiden.

Ergibt der Shapiro-Wilk-Test hingegen eine Ablehnung der Nullhypothese, dürfen weder F- noch T-Test angewendet werden. Eine Lösung bietet der Levene-Test, der Unterschiede in der Varianz von zwei Gruppen untersucht ohne normalverteilte Daten vorauszusetzen. Die Nullhypothese des Levene-Tests ist die Gleichheit der Varianzen in den untersuchten Gruppen. Ergibt dieser Test einen p-Wert kleiner als 0,05, kann auf einem Signifikanzniveau von 5% festgestellt werden, dass sich die untersuchten Gruppen in ihren Varianzen unterscheiden.

Führt der Levene-Test zu einem p-Wert größer als 0,05, kann mit Hilfe des Wilcoxon-Rangsummen-Tests bestimmt werden, ob die Verteilungen der beiden Datensätze aus der gleichen Grundgesamtheit stammen. Da die Nullhypothese des Tests die Gleichheit der Grundgesamtheiten beschreibt, darf bei einem p-Wert kleiner als 0,05 geschlossen werden, dass die beiden zu vergleichenen Datensätze aus unterschiedlichen Verteilungen stammen.

3 Ergebnisse

3.1 Ergebnisse der ChIP-Seq Analyse

Der Vergleich der ChIP-Seq Daten von ZH1 und ZH6 soll uns individuelle Unterschiede in H3K4me3 Markierungen zwischen zwei Wildmäusen aufzeigen.

3.1.1 Qualität der Reads

Die durchschnittlichen Qualitätskennzahlen der Reads aus den beiden Datensätzen ZH1 und ZH6 sind sehr ähnlich (3.1 A). Am 3'-Ende der Reads ist eine leichte Abnahme der Qualität zu beobachten, wie auch bereits in der Literatur beschrieben wurde [16].

3.1.2 Mapping der Reads

Die Ergebnisse des Mappings mit `bowtie` sind in Tabelle 3.1 aufgeführt. Jedes Mapping dauerte 32 bis 34 Minuten auf 29 CPUs. In beiden Datensätzen werden durch das Mappen auf mm10 über 100.000 Reads nicht mehr als nicht alignierbar klassifiziert, im Vergleich zum Mapping auf mm9. Diese in mm10 alignierbaren Reads tauchen aber nicht in unseren betrachteten Alignments auf, da sie mehrere Treffer im neuen mm10 Referenzgenom haben und somit als nicht eindeutig eingeordnet werden. Die Anzahl an Reads mit genau einem Alignment geht bei einem Mapping auf mm10 statt mm9 um mehrere tausend Reads zurück, obwohl mm10 3% mehr Basenpaare enthält (mm9: $2{,}65 \cdot 10^9$ Bp. mm10: $2{,}73 \cdot 10^9$ Bp.).

Tab. 3.1: Ergebnisse der Mappings der Datensätze ZH1 und ZH6 auf die Referenzmausgenome mm9 (von Juli 2007) und mm10 (von Dez. 2011).

Datensatz	Referenz	Anzahl Reads	Alignierbarkeit: eindeutig	nicht alignierbar	nicht eindeutig
ZH1	mm9	40.172.749	22.112.113	11.237.325	6.823.311
ZH1	mm10	40.172.749	22.108.714	11.122.548	6.941487
ZH6	mm9	40.372.729	23.073.724	10.164.899	7.134.106
ZH6	mm10	40.372.729	23.064.318	10.030.715	7.277.696

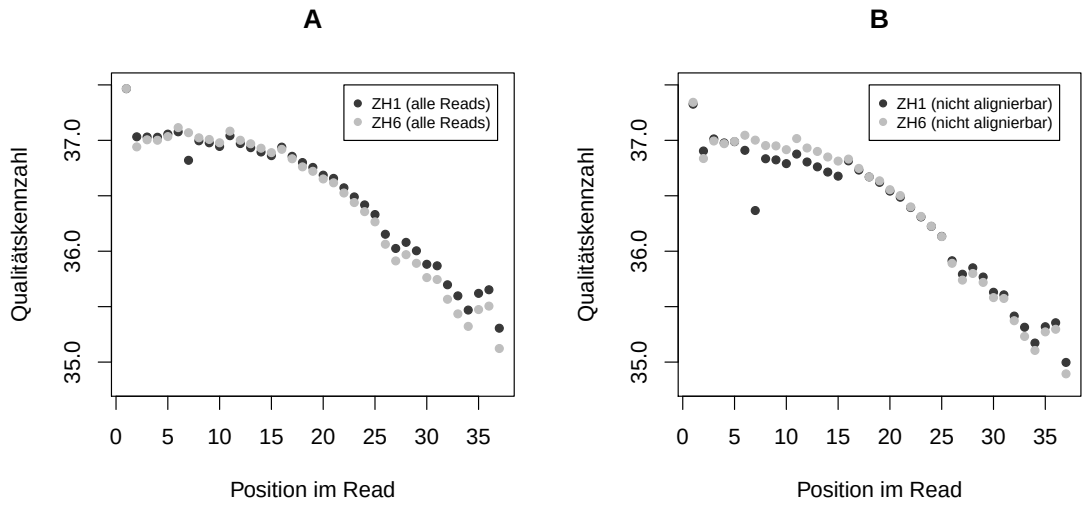

Abb. 3.1: Mittlere Qualitätskennzahl pro Position im Read für die beiden Datensätze ZH1 und ZH6. A: Alle Reads. B: Die von `bowtie` als nicht alignierbar klassifizierten Reads.

Untersuchung der nicht alignierbaren Reads

Über ein Viertel der Reads aus den Datensätze ZH1 und ZH6 wurden von `bowtie` als nicht alignierbar eingeordnet. Ihre Qualitätskennzahlen sind sehr ähnlich (Abbildung 3.1 B) im Vergleich zur Gesamtheit aller Reads (Abbildung 3.1 A). Eine geringe Anzahl Reads weisen eine durchweg schlechte Qualität auf, wie man beim Auftragen der Summe der Qualitätskennzahlen über alle 37 Positionen in jedem Read gegen die Häufigkeit erkennen kann (Abbildung 3.2). Die maximale Qualitätskennzahl liegt bei 40, multipliziert mit 37 Positionen ergibt sich ein Maximum von 1480. Ein großer Teil der nicht alignierbaren Reads liegt in einem Bereich nahe dieses Maximums. Deshalb kann eine geringe Qualität nicht der alleinige Grund für das ungültige Alignment sein. Auch unbestimmte Positionen in den Reads (*N*) scheiden als Grund aus. Innerhalb der nicht alignierbaren Reads weisen 98,7% (ZH1) bzw. 99,4% (ZH6) an keiner Position im Read ein *N* auf. Der einzige verbleibende Grund für das Finden keines Alignments durch `bowtie` ist die zu große Anzahl an Mismatches. Bei ZH1 und ZH6 handelt es sich um Wildmäuse, deshalb ist eine bestimmte Anzahl an Mismatches zu erwarten. Im folgenden bestimmen wir deshalb zum Vergleich die Anzahl an SNPs der Wildmäuse.

3.1.3 SNP Detektion

In der Literatur sind SNPs von Wildmäusen des *Mus musculus domesticus* Stammes an 21 Loci von ca. 600 Bp Länge untersucht worden [20] [1]. Aus diesen Angaben ergibt sich, dass im Mittel 0,21% der Positionen einen Polymorphismus aufweisen. Aus den ChIP-Seq Daten (Abschnitt 2.2.6) ergaben sich für ZH1 316.356 polymorphe Positionen von 151.177.788 Stellen auf die insgesamt gemappt wurde. Das entspricht einer Häufigkeit von 0,21%. Bei ZH6 ergeben

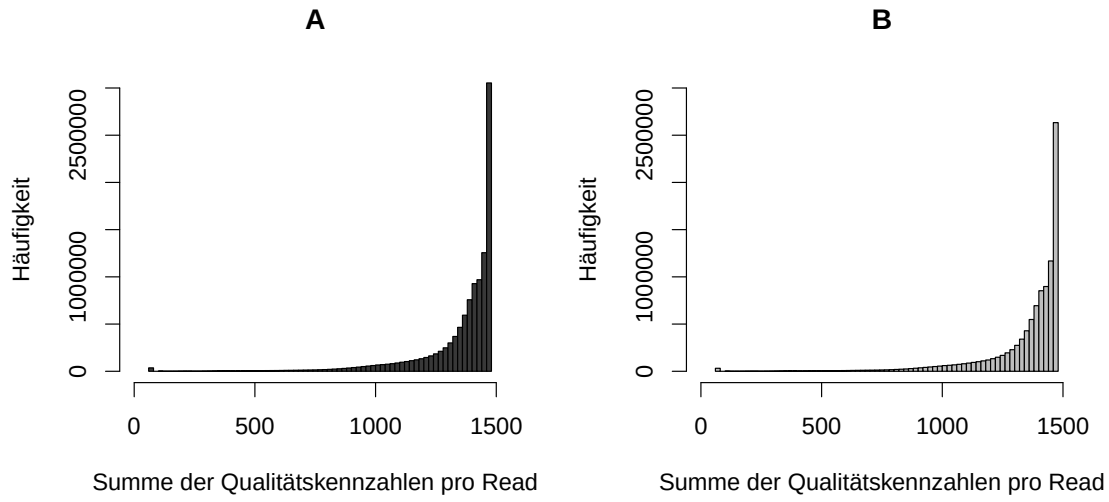

Abb. 3.2: Histogramm der Summe Qualitätskennzahlen der unmappbaren Reads. Das Maximum liegt bei 37 · 40=1480. A: ZH1. B: ZH6.

sich 334.412 SNPs auf 161.925.114 Positionen und damit ebenfalls eine Häufigkeit von 0,21%. Die SNPs verteilen sich auf die Chromosomen proportional zur Länge der Chromosomen. Der Korrelationskoeffizient (Spearman) zwischen Länge der Chromosomen und Anzahl an SNPs ergibt für ZH1 ρ=0,88 und für ZH6 ρ=0,87.

3.1.4 Ergebnisse der Normalisierung

Bei den in Abschnitt 2.2.4 vorgestellten Möglichkeiten der Normalisierung von ChIP-Seq Datensätzen ergeben sich für ZH1 und ZH6 die in Tabelle 3.2 gegenüber gestellten Werte. Die Anzahl aller TSS, an denen entweder ZH1 oder ZH6 einen Wert größer als Null haben, ist 33.882. Wenn pro Gen nur eine TSS betrachtet wird, schrumpft diese Zahl auf 12.007. Wir wählen folgendes Verfahren, um pro Gen nur eine TSS auszuwählen: Zunächst wird der Mittelwert aller TSS Peaks berechnet (MW(ZH1)= 5498,91 und MW(ZH6)=6098,33) und die Peaks werden mittelwertnormalisiert. Daraufhin wird die Normierte Differenz aller Peaks bestimmt und pro Gen wird die TSS ausgewählt, die den größten Unterschied repräsentiert.

Die Ergebnisse in Tabelle 3.2 zeigen deutlich, dass diese Auswahl der TSS zu einem Normalisierungsfaktor führt der dem der biologischen Normalisierung sehr nahe kommt. Da die Normalisierungsfaktoren aber insgesamt alle ungefähr bei eins liegen, werden wir, außer der Normalisierung auf jeweils den Mittelwert, zwischen den Datensätzen keine weitere Normalisierung vornehmen.

Tab. 3.2: Normalisierungsfaktoren zwischen ZH1 und ZH6.

Methode	Wert ZH1	Wert ZH6	Norm.faktor
Technische Normalisierung			
# Reads	40.172.749	40.372.729	0,995
# gemappter Reads	22.112.113	23.073.724	0,958
Summe Coverage (vor 2/8)	818.148.181	853.727.788	0,958
Summe Coverage (nach 2/8)	72.274.650	80.596.801	0,897
# Positionen mit Cov.$\neq$0 (vor 2/8)	559.593.961	569.991.090	0,982
# Positionen mit Cov.$\neq$0 (nach 2/8)	9.073.853	9.877.136	0,919
Normalisierung mit `geneCov`			
Alle ENSMUST	186.313.939	206.623.624	0,902
1 ENSMUST pro Gen	10.988	11.224,2	0,979
Biologische Normalisierung			
Gapdh	14.644	14.470	1,012
Sdha	12.547	13.326	0,942
Tbp	5.326	6.142	0,867
Ubc	23.371	24.469	0,955
YWHAZ	14.106	13.362	1,056
Actb	12.446	13.068	0,952
B2m	17.036	16.651	1,023
Rpl13a	18.178	13.286	1,368
Hmbs	8.316	6.411	1,297
Mittelwert			1,052

3.1.5 Normierte Differenz der Datensätze

Einige Eigenschaften der Normierten Differenz angewendet auf die Datensätzen ZH1 und ZH6 sind in Abbildung 3.3 dargestellt. In Abbildung 3.3 A und C werden alle Peaks für die Berechnungen berücksichtigt (Anzahl: 12.005), in Abbildung 3.3 B und D nur die Peaks, an denen beide Datensätze größer Null sind (Anzahl: 10.229). In den beiden QQ-Plots der Abbildung (A und B) sind zur besseren Übersicht die Peakgrößen bei den Genen AL928915.2 (Norm.Diff.=2,22) und Cwc22 (Norm.Diff.=7,10) entfernt worden. Eine Annäherung an die Normalverteilung ist für die Datenverteilung inklusive aller Peakgrößen deutlich stärker sichtbar, weil nur 66 Genpeaks außerhalb des Bereichs zwischen -0,70 und +0,60 liegen. Dieser Bereich ist nach dem QQ-Plot derjenige, in dem die Verteilung nicht mehr gut durch eine Normalverteilung angenähert werden kann. Durch ein Weglassen der exklusiven Peaks stimmt die Verteilung der Normierten Differenz nicht mehr so gut mit einer Normalverteilung überein; 599 Gene liegen außerhalb des Bereiches in dem die Quantilen übereinstimmen.

3.1.6 Vergleich ZH1 mit ZH6

3.1.6.1 Allgemeine Beobachtungen

In Abbildung 3.4 sind die H3K4me3 Peakgrößen von ZH1 gegen die von ZH6 aufgetragen. Es wird deutlich, dass außerhalb der 95% bzw. 99% Quantilen der Normierten Differenz zwischen den Peakgrößen viele exklusive Markierungen liegen (in rot markiert in Abbildung 3.4). Ihr Anteil im Bereich außerhalb des 95% Quantils (grüne Linie) beträgt 38,7% (233 zu 369 Peaks) und außerhalb des 99% Quantils (blaue Linien) 43,4% (53 zu 69). Dabei gilt, dass die Peakgrößen der exklusiven Peaks im Vergleich zu allen Peakgrößen gering sind, was durch die Boxplots in Abbildung 3.5 deutlich zu erkennen ist. Deshalb werden wir zunächst die exklusiven Peaks (siehe Abschnitt 3.1.6.2) und anschließend die übrigen Peaks (siehe Abschnitt 3.1.6.3) betrachten.

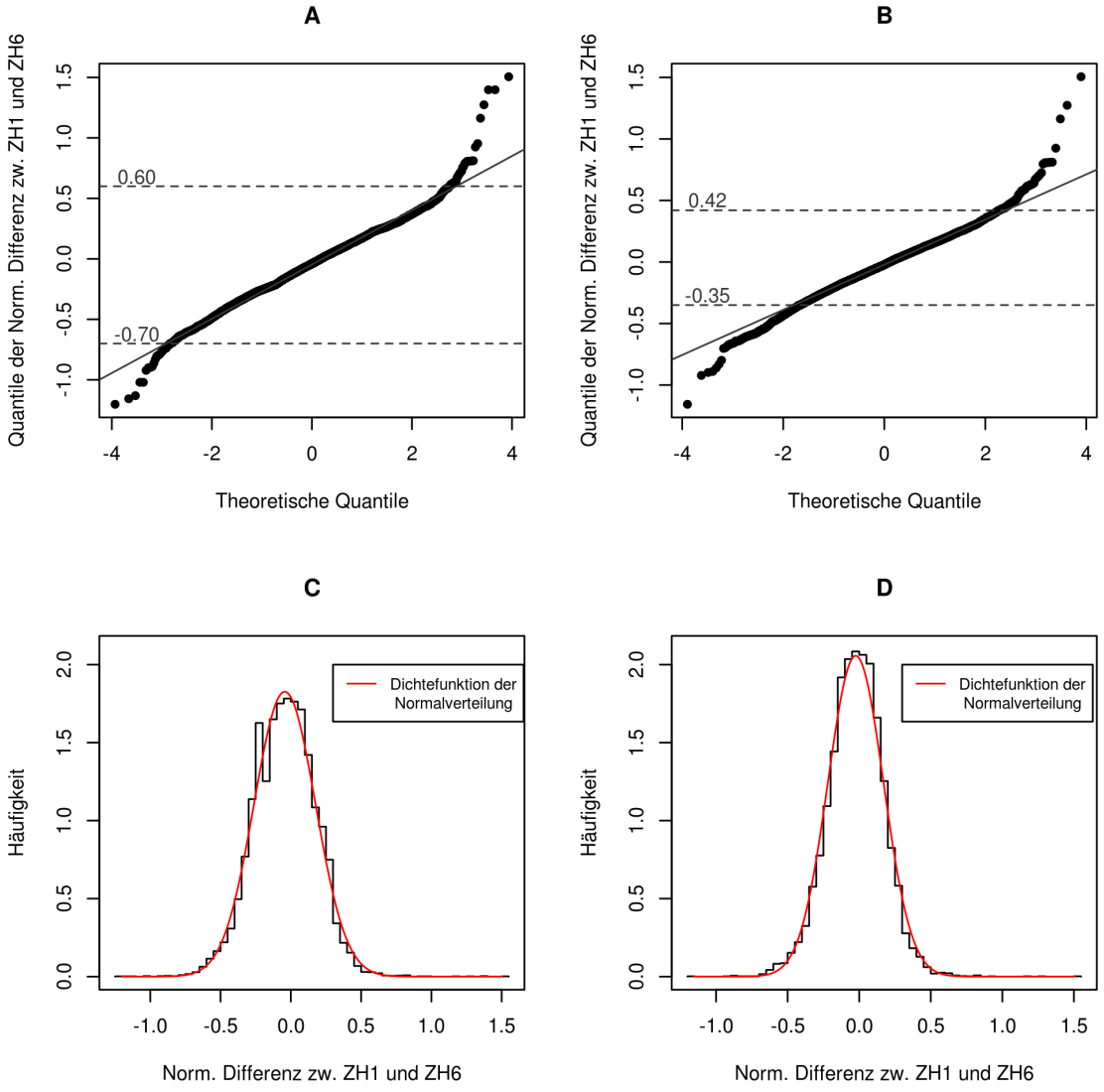

Abb. 3.3: Normierte Differenz zwischen ZH1 und ZH6. A und C: Alle Peaks. B und D: Ohne exklusive Peaks. A und B: QQ-Plots der Normierten Differenz zwischen ZH1 und ZH6 gegen eine Normalverteilung. C und D: Verteilung der Normierten Differenz und Dichtefunktion einer passenden Normalverteilung.

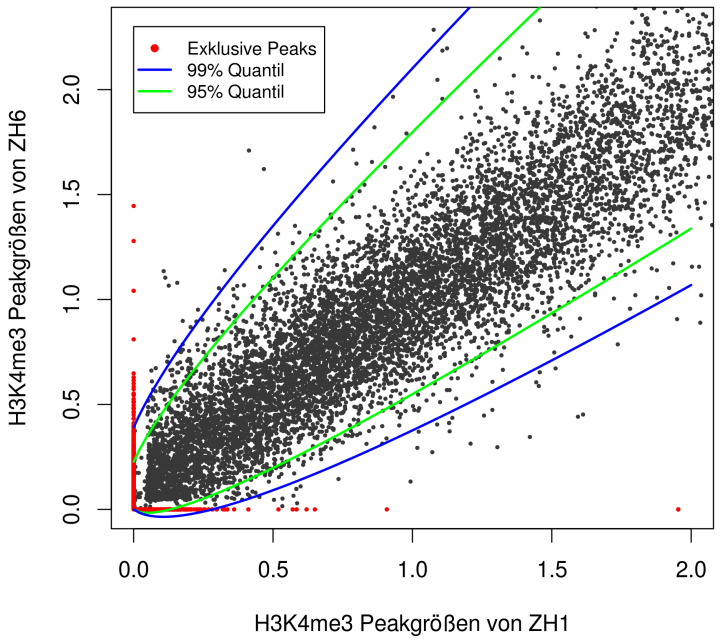

Abb. 3.4: Ausschnitt der Verteilung der H3K4me3 ChIP-Seq Peakgrößen. Rot markiert sind die exklusiven Peaks. Die Peaks sind normalisiert auf den Mittelwert aller H3K4me3 markierten Gene. Bei Genen mit mehreren TSS-Annotationen wurde in dieser Abbildung diejenige gewählt, bei der der Wert der Normierten Differenz am größten ist.

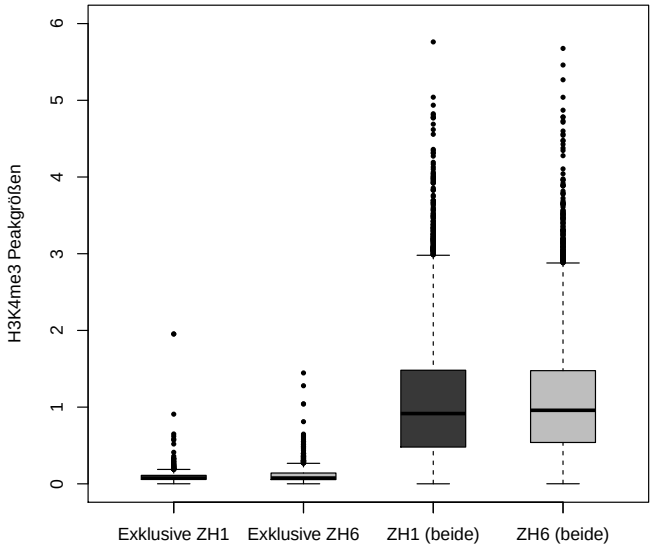

Abb. 3.5: Verteilung der H3K4me3 Peakgrößen zwischen den exklusiven Peaks und den Peaks, an denen beide Datensätze Markierungen aufweisen.

3.1.6.2 Exklusive Peaks

Exklusive Peaks zeichnen sich dadurch aus, dass ein Datensatz an diesem Gen markiert ist und der andere nicht. In Abbildung 3.6 sind die Markierungen an allen exklusiven Peaks dargestellt. Insgesamt gibt es 485 exklusive ZH1 Peaks und 1291 exklusive ZH6 Peaks. Diesen exklusiven Peaks stehen 10.231 Peaks gegenüber, an denen beide Datensätze Markierungen aufweisen. Jeweils die extremen 5% bzw. 1% sind in der Abbildung grün bzw. blau markiert. Da in ZH6 fast dreimal so viele Gene exklusiv markiert sind wie in ZH1, liegen auch in den extremen 5% bzw. 1% knapp dreimal mehr Gene.

Eine KEGG Analyse mit `WebGestalt` der Gene mit einer mittleren Peakgröße in den oberen 5% der Verteilung der exklusiven ZH1 Peaks ergibt im Vergleich zu allen Genen im Mausgenom zwei signifikant (korrigiert: $p<0{,}01$) angereicherte funktionelle Einheiten: den PPAR Signaltransduktionsweg und die neuroaktive Ligand-Rezeptor Interaktion. Ein ähnliches Ergebnis erhält man, wenn die Anreicherung gegen die in der Leber markierten Gene analysiert wird. Wenn die oberen 1% betrachtet werden, ergibt sich noch der PPAR Signaltransduktionsweg als signifikant angereichert im Vergleich zu allen Genen (korrigiert: $p < 10^{-5}$) und im Vergleich zu den ausgewählten Lebergenen (korrigiert: $p < 10^{-4}$). Cyp4a14 und Scd1 sind diejenigen Gene in den oberen 1%, die Teil des PPAR Signalwegs sind. Nach Pdf sind das die zwei höchsten exklusiven ZH1 Markierungen (in Abbildung 3.6 mit den Gennamen markiert und in Tabelle 3.3 aufgelistet). Pdf ist der größte exklusive Peak im gesamten Datensatz. Abbildung 3.7 zeigt diesen exklusiven Peak im UCSC Genome Browser und es wird deutlich, dass in diesem Bereich des Genoms eine CpG-Insel liegt (grün in Abbildung 3.7). An den Markierungen von Cyp4a14 und Scd1 liegen keine CpG-Inseln vor.

Eine KEGG Analyse der exklusiven ZH6 Peaks ergibt keine signifikant angereicherten funktionellen Einheiten. Vier exklusive Peaks des Datensatzes ZH6 ragen heraus (Abbildung 3.6): Scara5, Npr1, Tmem51 und Slc39a4 (Tabelle 3.3). Darunter liegen zwei exklusive Markierungen parallel zu annotierten CpG-Inseln (Npr1 und Tmem51).

Tab. 3.3: Die größten exklusiven Markierungen. Informationen stammen aus *UniProt* und *NCBI*.

Datensatz	Gen	Genbezeichnung	Funktion	Besonderheit
ZH1	Pdf	Peptiddeformylase	Entfernung von Aldehydgruppen am Methionin	CpG-Insel
	Cyp4a14	Cytochrome P450 4A14	Häm-Thiolat Monooxygenase	
	Scd1	Stearoyl-CoA Desaturase 1	Fettsäure-Desaturase	
ZH6	Scara5	Scavenger receptor class A member 5	Ferritin-Rezeptor	
	Npr1	Atrial natriuretic peptide receptor 1	Hormonrezeptor	CpG-Insel
	Tmem51	Transmembrane protein 51	Membranprotein	CpG-Insel
	Slc39a4	Zinc transporter ZIP4	Zink-Transportprotein	

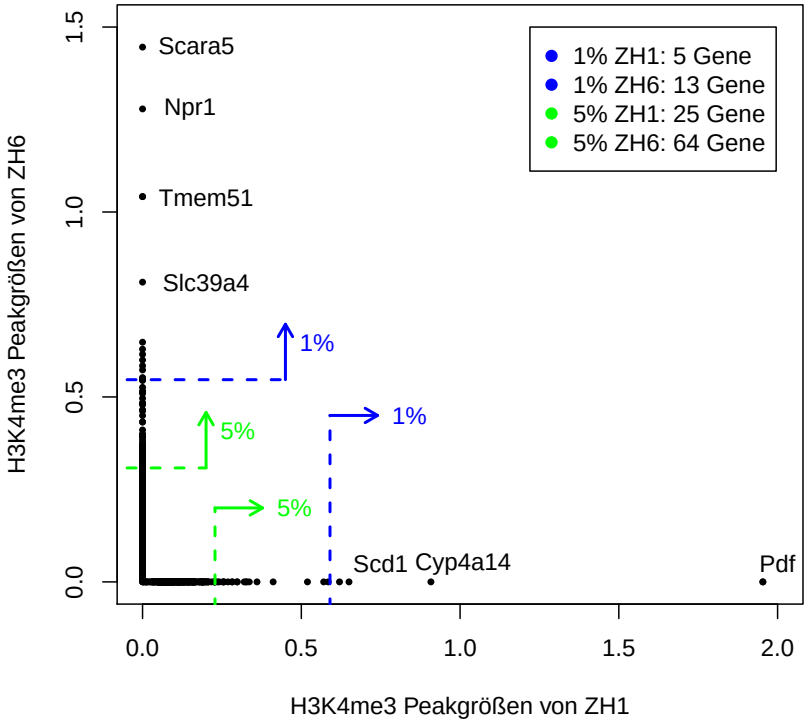

Abb. 3.6: Überblick über die exklusiven Markierungen in den Datensätzen ZH1 und ZH6.

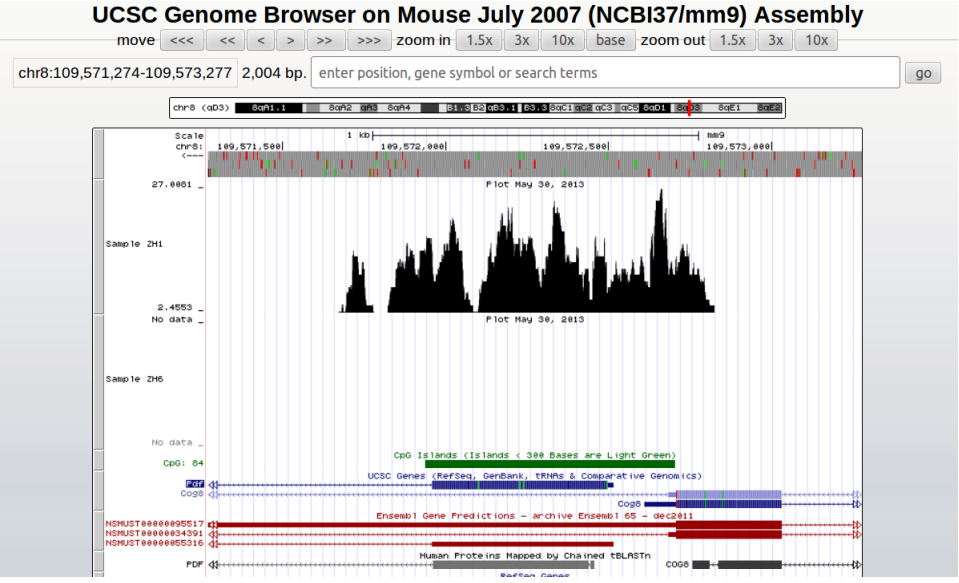

Abb. 3.7: Aufnahme der exklusiven ZH1 Markierung am Gen Pdf im UCSC Genome Browser in einem Fenster der Größe 2004 Bp zentriert um den TSS ENSMUST00000055316.

3.1.6.3 Alle Peaks ohne exklusive Peaks

Abbildung 3.8 zeigt ein Streudiagramm der H3K4me3 ChIP-Seq Peakgrößen derjenigen Gene, die sowohl in ZH1 als auch in ZH6 markiert sind. Dabei handelt es sich um 10.231 Gene. Außerhalb der grünen Linien liegen die 5% der Datenpunkte, bei denen die Normierte Differenz zwischen ZH1 und ZH6 größer als die 2,5% bzw. 97,5% Quantile ist. Eine KEGG Analyse dieser Gene gegen alle Mausgene liefert elf signifikant (korrigiert: $p<0{,}001$) angereicherte funktionelle Einheiten. Wenn als Referenz alle in der Leber markierten Gene verwendet werden, gibt es auf einem Signifikanzniveau von 5% keine Anreicherung. Die beiden funktionellen Einheiten mit den kleinsten p-Werten sind *Cell Adhesion Molecules* (korrigiert: p=0,0642) und *Circadian Rhythm* (korrigiert: p=0,0856).

Eine KEGG Analyse der Gene, die außerhalb des 99% Quantils in Abbildung 3.8 liegen, gegen alle Gene ergibt zehn funktionelle Einheiten, die signifikant (korrigiert: $p<0{,}05$) angereichert sind. Eine Änderung der Referenz auf alle in der Leber markierten Werte führt wiederum dazu, dass keine funktionelle Einheit signifikant angereichert ist.

Werden Gene, die in ZH1 stärker markiert sind als in ZH6, in einer KEGG Analyse mit allen in der Leber markierten Genen verglichen, ergibt sich eine signifikant angereicherte funktionelle Einheit (korrigiert: p=0,0031): *Cell Adhesion Molecules*. Bei der Analyse wurden nur Gene berücksichtigt, deren Normierte Differenz größer als das 97.5% Quantil sind. Wenn umgekehrt nur diejenigen Gene betrachtet werden, die in ZH6 stärker markiert sind als in ZH1, ergeben sich bei der KEGG Analyse gegenüber allen in der Leber markierten Genen drei signifikant angereicherte funktionelle Einheiten (korrigiert: $p<0{,}05$). Alle drei funktionellen Einheiten gehören zu den Stoffwechselwegen: *Proteasome*, *Pyrimidine metabolism* und *Fructose and mannose metabolism*.

Die größten Unterschiede zwischen ZH1 und ZH6 bestehen an den in Tabellen 3.4 und 3.5 gelisteten Loci. Die Funktionen der Gene, entnommen aus den Datenbanken *NCBI*, *MGI* und *UniProt*, sind vielfältig. Genomische Informationen beziehen sich auf Besonderheiten der untersuchten Wildmäuse in Variation der Genkopiezahl (CNV, Copy Number Variations) und bekannten genetischen Unterschieden verschiedener Labormaus-Stämme (RFLP, PCRV).

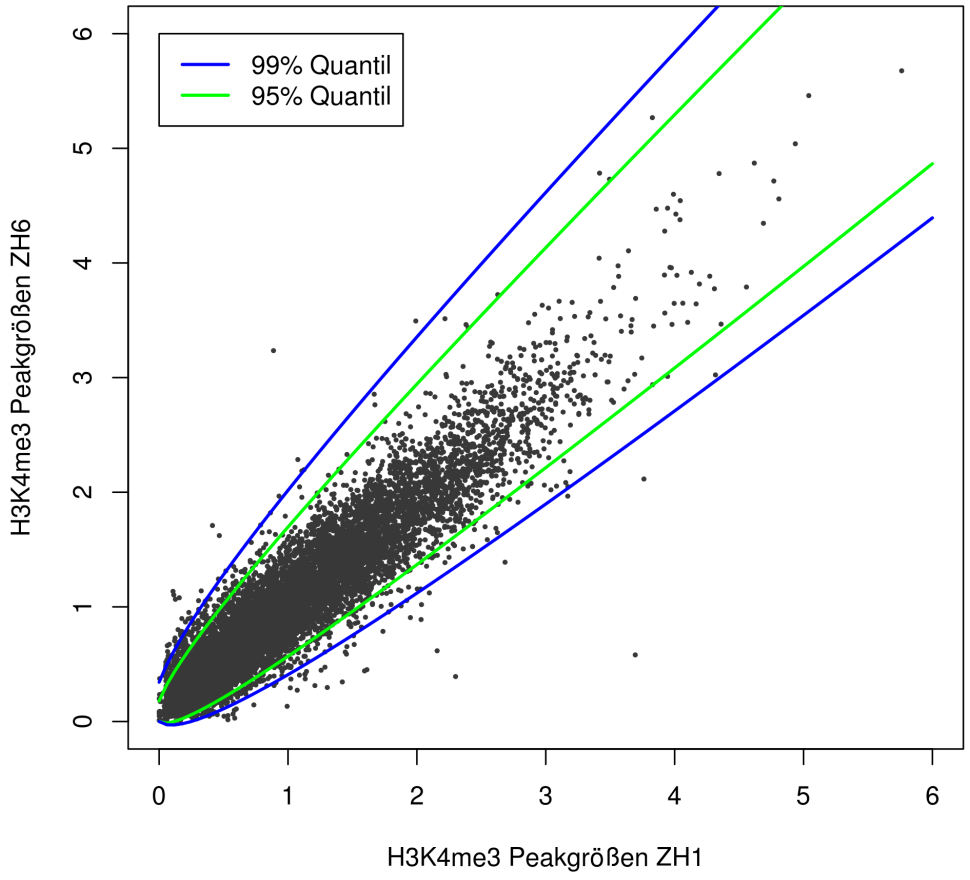

Abb. 3.8: Verteilung aller nicht exklusiven H3K4me3 ChIP-Seq Peakgrößen. Innerhalb der blauen bzw. grünen Linien liegen 99% bzw. 95% der Datenpunkte. In der Abbildung fehlen neun Gene, deren Peakgrößen größer als sechs sind (AL837506.1, AL837506.2, AL928915.2, Asl, Cwc22, Hjurp, mmu-mir-715, SSU_rRNA_5, Trpm8).

Tab. 3.4: Gene, an denen ZH1 stärker markiert als ZH6. CNV: Copy Number Variation. RFPL: Restriction Fragment Length Polymorphism.

Gen	Norm. Differenz	Funktion	Genetische Informationen
Cwc22	7,10	mRNA Prozessierung	CNV
AL928915.2	2,22	keine Annotation	CNV
AL928791.1	1,51	keine Annotation	CNV
Asl	1,27	Arginin Biosynthese	RFLP(1)
Cyp26a1	1,16	Retinsäuremetabolismus	CpG-Insel
Slc16a1	0,92	Membran-Transportprotein	CpG-Insel
Id1	0,81	Transkriptionsfaktor	CpG-Insel
Pnpla7	0,81	Serin Hydrolase	
Per2	0,81	Zirkadianer Rhythmus	CpG-Insel
Akap13	0,81	Bindet Proteinkinase A	
Tpst1	0,80	Tyrosine Sulfation	CpG-Insel

Tab. 3.5: Gene, an denen ZH6 stärker markiert als ZH1. CNV: Copy Number Variation. RFPL: Restriction Fragment Length Polymorphism. PCRV: Polymerase Chain Reaction Variation.

Gen	Norm. Differenz	Funktion	Genetische Informationen
Nr1d1	-1,16	Steroidhormonrezeptor	
Cpne8	-0,92	evtl. Membrantransportprotein	CpG-Insel
Mt2	-0,90	Bindung von Schwermetallen, Entgiftung	PCR(4), RFLP(1), CpG-Insel
Ctsa	-0,86	Proteolyse	RFLP(1), CpG-Insel
Ptk2b	-0,86	Proteinphosphorylierung	
Il1r1	-0,83	Regulation von Entzündungsreaktionen	PCR(1), RFLP(3)
Neurl2	-0,80	Intrazelluläre Signaltransduktion	CpG-Insel

3.1.6.4 Alle H3K4me3 Peaks im Vergleich

Wir wollen nun alle H3K4me3 Peaks, egal ob sie exklusiv sind oder nicht, gemeinsam betrachten. Dabei unterscheiden wir zwischen Genen, an denen ZH1 stärker markiert ist als ZH6 und dem umgekehrten Fall. Abbildung 3.9 zeigt ein Histogramm der Normierten Differenzen zwischen ZH1 und ZH6. In grün markiert sind die Bereiche, in denen die 5% größten Unterschiede liegen. Da die Normierte Differenz ein unsymmetrisches Differenzmaß ist, finden sich die Gene, an denen ZH1 stärker markiert ist als ZH6, am rechten Ende der Verteilung. Die Gene, die in ZH1 weniger stark markiert sind als in ZH6, liegen am linken Rand.

Die Ergebnisse einer KEGG Analyse der Gene, an denen ZH1 hoch markiert ist, sind in Tabelle 3.6 aufgezeigt. Dabei werden nur Gene betrachtet, die im oberen 2,5% Bereich der Verteilung der Normierten Differenz liegen (Norm.Diff $>$ 0,36). Es fällt auf, dass in diesem untersuchten Datensatz viele Gene liegen, die in metabolischen Einheiten auftauchen. Dabei sind innerhalb der angereicherten funktionellen Einheiten immer wieder dieselben Gene zu finden, die zu dieser funktionellen Einheit und gleichzeitig zu den in ZH1 höher markierten Genen gehören. Besonders große Überlappungen gibt es zwischen dem Pyrimidin- und dem Purinmetabolismus. Sieben der 99 bzw. 168 Gene dieser funktionellen Zusammenhänge tauchen in dem untersuchten Datensatz auf, davon stimmen fünf miteinander überein. Auch in den funktionellen Einheiten bei Krebs und dem TGF-beta Signaltransduktionsweg gibt es zwei überlappende Gene: Myc und Tgfbr2.

Gene, die in ZH6 stärker markiert sind als in ZH1, liegen nicht in metabolischen Zusammenhängen. In den Ergebnissen der KEGG Analyse in Tabelle 3.7 tauchen hingegen funktionelle Einheiten auf, die an der Immunantwort einer Zelle beteiligt sind oder die den Zellzyklus regulieren.

Insgesamt ist zu beachten, dass die Anzahl an Genen, die in den genannten funktionellen Einheiten und gleichzeitig im betrachteten Datensatz liegen, gering ist im Vergleich zu allen Genen dieser Kategorien. Es gibt keine funktionelle Einheit, in dem ein Großteil der Gene zwischen den Mäusen ZH1 und ZH6 unterschiedlich markiert ist. Trotzdem hilft die Betrachtung der funktionellen Einheiten dabei, die ungleich markierten Gene in einen Zusammenhang zu stellen.

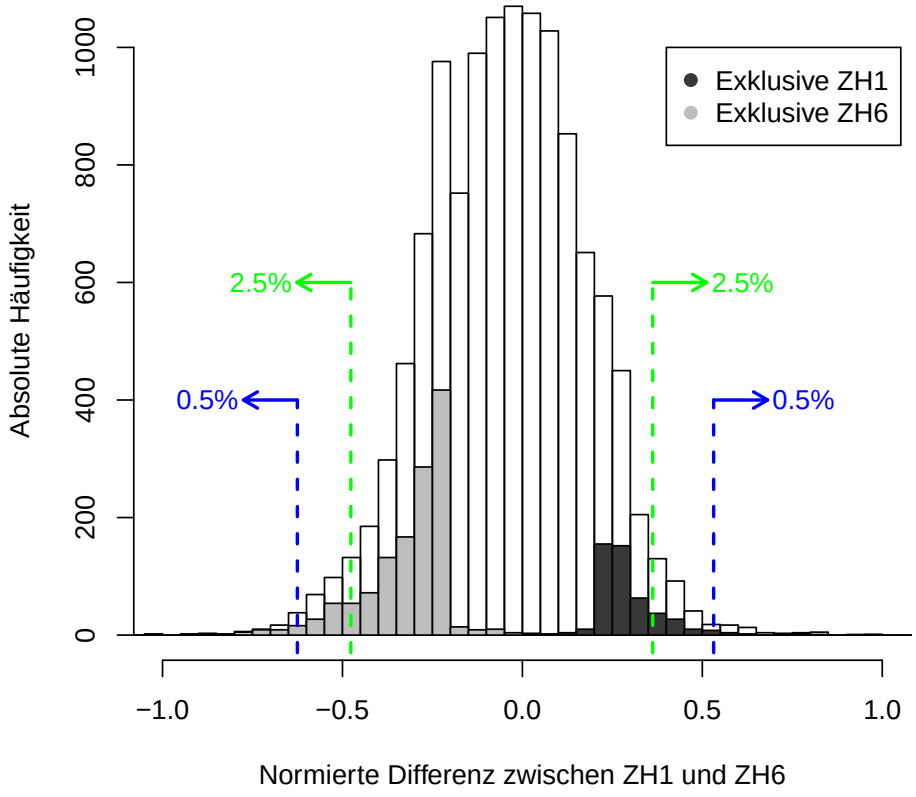

Abb. 3.9: Histogramm der Normierten Differenz zwischen ZH1 und ZH6. Markiert sind die exklusiven Peaks: in grau ZH1 und in schwarz ZH6. Außerdem sind diejenigen Bereiche markiert, in denen 95% (grün) bzw. 99% (blau) der Werte liegen.

Tab. 3.6: KEGG Analyse der Gene, an denen ZH1 stärker markiert ist als ZH6 (obere 2,5%). F: Anzahl Gene in der funktionellen Einheit. G: Anzahl Gene mit signifikant höherer Markierung in ZH1 als in ZH6 und gleichzeitig in der funktionellen Einheit.

Funktionelle Einheit	F	G	Gene aus G	korrigierter p-Wert
Stoffwechselwege	1184	32	u.a. Adssl1, Cda, Cyp7a1, Cyp4a14, Fuk, Nme2, Nt5c2, Nt5e, Papss2, Pola2, Tpi1, Znrd1	$1{,}87 \cdot 10^{-14}$
Pyrimidinmetabolismus	99	7	Cda, Nme2, Nt5c2, Nt5e, Pola2, Tk1, Znrd1	$1{,}18 \cdot 10^{-5}$
Fruktose - und Mannosemetabolismus	37	5	Fuk, Pfkfb1, Pfkfb3, Pfkfb4, Tpi1	$1{,}61 \cdot 10^{-5}$
Purinmetabolismus	168	7	Adssl1, Nme2, Nt5c2, Nt5e, Pola2, Papss2, Znrd1	0,0002
TGF-beta Signaltransduktionsweg	85	5	Id1, Id3, Id4, Myc, Tgfbr2	0,0005
PPAR Signaltransduktionsweg	80	5	Cyp7a1, Cyp4a14, Ppard, Scd1, Slc27a2	0,0005
Zirkadianer Rhythmus (Säugetier)	22	3	Npas2, Per2, Rorc	0,001
Proteinverarbeitung im ER	169	6	Dnaja1, Hsp90ab1, Preb, Ube2e1, Ubqln1, Uggt2	0,001
Funktionelle Einheiten bei Krebs	325	8	Egln3, Hsp90ab1, Myc, Pik3r1, Ppard, Prkcb, Tgfbr2, Wnt5a	0,001

Tab. 3.7: KEGG Analyse der Gene, an denen ZH6 stärker markiert ist als ZH1 (untere 2,5%). F: Anzahl Gene in der funktionellen Einheit. G: Anzahl Gene mit signifikant höherer Markierung in ZH6 als in ZH1 und gleichzeitig in der funktionellen Einheit.

Funktionelle Einheit	F	G	Gene aus G	korrigierter p-Wert
Hepatitis C	137	6	Irf7, Irf9, Oas1b, Scarb1, Socs3, Traf3	0,0017
Endozytose	220	6	Ehd4, Git1, Grk4, H2-T24, Pld2, Rab11fip4	0,0059
Ubiquitin gesteuerte Proteolyse	140	5	Anapc4, Fbxo4, Socs3, Ube2e2, Xiap	0,0059
Apoptose	86	4	Capn1, Il1r1, Myd88, Xiap	0,0059
Jak-STAT Signaltransduktionsweg	153	5	Il7, Il13ra1, Irf9, Jak2, Socs3	0,0059
Toll-like Rezeptor Signaltransduktionsweg	101	4	Irf7, Lbp, Myd88, Traf3	0,009
Chemokine Signaltransduktionsweg	185	5	Jak2, Grk4, Prkcd, Ptk2b, Rasgrp2	0,0098

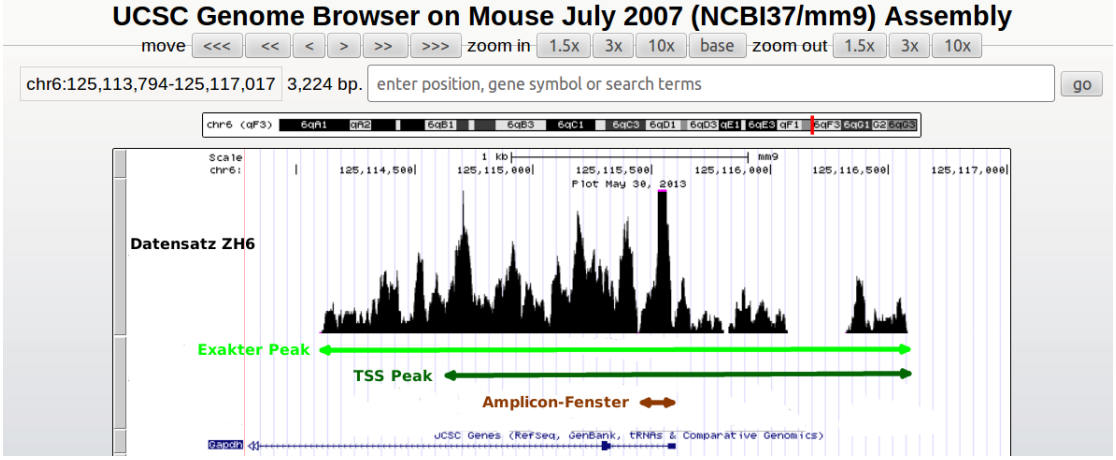

Abb. 3.10: Ausschnitt aus dem UCSC Browser Track des Datensatzes ZH6 am Gen Gapdh. In hellgrün ist der 2532 Bp lange exakte Peak markiert, in dunkelgrün der 2000 Bp lange TSS Peak und in braun das 103 Bp lange Amplicon-Fenster.

3.2 Validierung der ChIP qPCR Methode für Wildmäuse

3.2.1 Normalisierung

Die Normalisierung in einem qPCR Experiment ist kritisch. In der Literatur werden verschiedene Möglichkeiten der Normalisierung beschrieben (siehe Abschnitt 2.5.2). Laut [61] sind mindestens drei Housekeping-Gene genügend für eine akkurate Normalisierung. Da die Anzahl an messbaren Loci in unserem qPCR Experiment begrenzt ist (15), verwenden wir als Housekeeping-Gen nur Gapdh. Da dies nach [61] nicht ausreicht, kann auch der Mittelwert aus allen 15 Messungen benutzt werden. Um zu testen, ob dieses sinnvoll ist, vergleichen wir eine Normalisierung auf das Housekeeping-Gen Gapdh mit einer Normalisierung über den Mittelwert aller 15 Loci mit Hilfe von ChIP-Seq Daten.

In der qPCR wird nur ein begrenzter Bereich von etwa 80 bis 160 Basenpaaren amplifiziert, das Amplicon-Fenster. Nur über diesen Bereich im Genom kann eine Aussage getroffen werden. Hingegen geben ChIP-Seq Daten die Möglichkeit, die Verteilung der Histonmarkierungen über das gesamte Genom zu betrachten und damit einen exakten Peak pro Gen zu bestimmen. Die exakten Peaks haben für die hier verwendeten Gene eine Länge von etwa 1000 bis 4000 Basenpaare. Da es schwierig ist für alle Gene einen exakten Peak zu bestimmen, betrachten wir in der automatisierten Analyse der ChIP-Seq Daten 2000 Bp große symmetrische Fenster um TSS, die TSS Peaks. Ein Beispiel für den Zusammenhang zwischen einem exakten Peak, einem TSS Peak und einem Amplicon-Fenster ist in Abbildung 3.10 für das Gen Gapdh dargestellt.

Um zu überprüfen, inwieweit die in der qPCR amplifizierten Bereiche des Peaks mit dem exakten Peak übereinstimmen, vergleichen wir in Abbildung 3.11 die Auswirkungen der Normalisierung auf das Housekeeping-Gen Gapdh (3.11 A) mit der Normalisierung über den Mittelwert

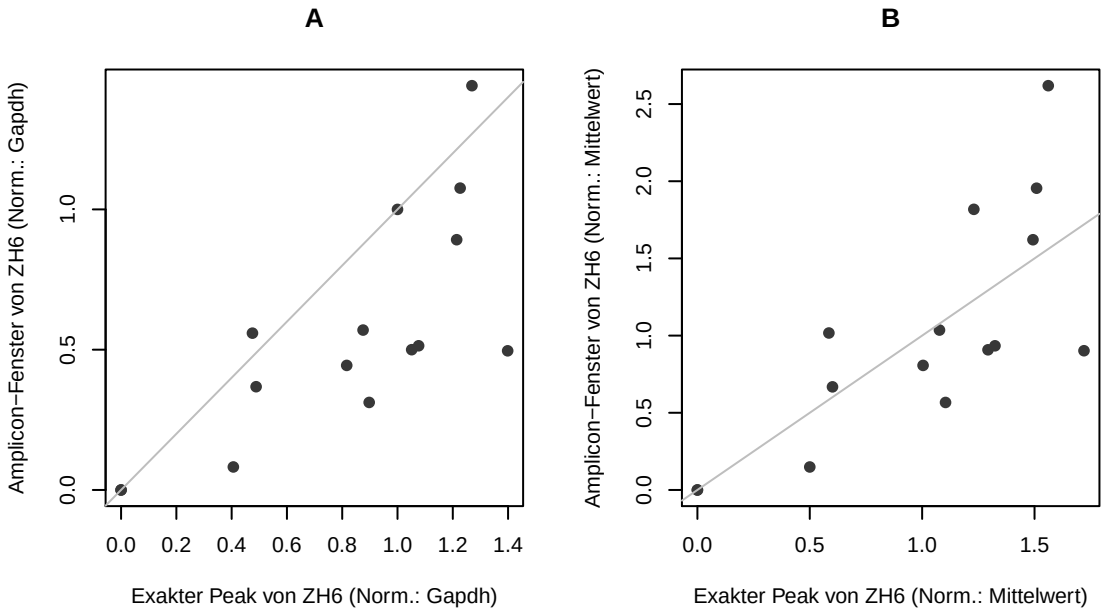

Abb. 3.11: Zusammenhang zwischen dem ChIP-Seq Signal im Amplicon-Fenster der 15 in der qPCR untersuchten Gene und dem exakten TSS Peak im ChIP-Seq Signal dieser Gene am Beispiel des H3K4me3 Datensatzes von ZH6. A: Normalisierung Gapdh. B: Normalisierung Mittelwert.

der 15 in der qPCR untersuchten Gene (3.11 B). Die x-Achse gibt jeweils die normalisierten Werte im Bezug auf den exakten Peak an und die y-Achse in Bezug auf das Amplicon-Fenster. Die graue Linie entspricht der Ursprungshalbgeraden mit Steigung eins. Lägen alle Punkte auf dieser Linie, gäbe es eine perfekte positive Korrelation zwischen dem Wert im Amplicon-Fenster und der Größe des exakten Peaks.

Ein unterschiedliche Normalisierung verändert die Korrelation zwischen den x- und y-Werten nicht, da der Korrelationskoeffizient invariant gegenüber Skalierungen ist. Dieser ist für die Berechnung nach Pearson bei ρ =0,76 für den Datensatz ZH6 und bei ρ =0,73 für den Datensatz ZH1 (nicht gezeigt). Hingegen ändert sich die Steigung einer linearen Regressionsgeraden durch die Punkte: beim Datensatz ZH6 von 0,68 (Gapdh) auf 1,01 (Mittelwert) und im Datensatz ZH1 von 0,80 (Gapdh) auf 0,99 (Mittelwert).

Da sich die Steigung einer linearen Regressionsgeraden durch eine Normalisierung über den Mittelwert näher an eins bewegt und sich dadurch die Werte, die eine Analyse mit einer qPCR im Amplicon-Fenster ergeben, leichter mit einem TSS Peak vergleichen lassen, verwenden wir im Weiteren diese Normalisierung.

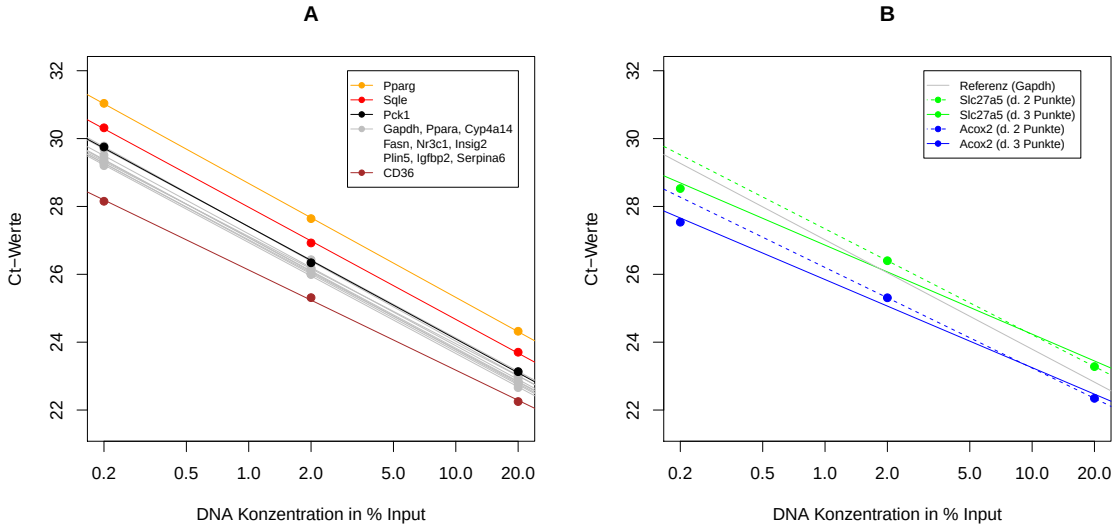

Abb. 3.12: QPCR Standardkurven für die 15 verwendeten Primerpaare jeweils aus den Mittelwerten der vier untersuchten Mäuse berechnet (zwei Bl6 und zwei Wildmäuse).

3.2.2 Ergebnisse der experimentellen Bestimmung der Primereffizienz

In Abbildung 3.12 ist das Ergebnis der experimentellen Bestimmung der Primereffizienz dargestellt. Abgebildet sind qPCR Standardkurven, die den Zusammenhang zwischen dem dekadischen Logarithmus der eingesetzten DNA Konzentration auf der x-Achse und den gemessenen Ct-Werten auf der y-Achse beschreiben. Durch die drei Messpunkte wurde mit Hilfe eines linearen Modells eine Regressionsgerade gelegt. Bei Primern, die eine konstante Effizienz von zwei über einen großen Konzentrationsbereich haben, liegen alle gemessenen Punkte im einfach logarithmischen Graph auf einer Geraden mit einer Steigung von -3,32.

Dies zeigt sich bei allen Primern in Abbildung 3.12 A. Bei ihnen wurde mit Hilfe der Gleichungen 2.5 und 2.6 eine Primereffizienz von ungefähr zwei berechnet (1,979 bis 2,182). Unterschiede der Regressionsgeraden kann man in den y-Abschnitten entdecken. Sie sind parallel zu höheren oder niedrigeren Ct-Werten verschoben. Das hat keinen Einfluss auf die Effizienz des Primers. Da wir mit %Input Werten weiterrechnen, wirkt sich der Unterschied zwischen den zum Beispiel konstant höheren Ct-Werten von Pparg im Vergleich zu CD36 nicht mehr aus.

Ist die Steigung in einer qPCR Standardkurve hingegen größer als -3,32, ist die Effizienz des Primers größer als zwei. Abbildung 3.12 B zeigt diese Kurven der beiden Primer, deren Effizienzen von zwei abweichen. Die Regressionsgeraden durch die drei Messwerte sind deutlich flacher als die der Referenz Gapdh, die in grau abgebildet ist. Diese Steigungen würden zu Primereffizienzen von 2,405 für Slc27a5 bzw. 2,428 für Acox2 führen. Außerdem liegen die drei Messwerte nicht exakt auf den Regressionsgeraden, wie es für die Primer in Abbildung 3.12 A der Fall ist. Werden hingegen nur die Messwerte 20% und 2% verwendet, um eine Regres-

sionsgerade zu zeichnen (gestrichelte Linien in Abbildung 3.12 B), ergeben sich Steigungen, die näher an der Referenzsteigung von Gapdh sind. Diese Steigungen entsprechen Primereffizienzen von 2,092 für Slc27a5 und von 2,175 für Acox2. Bei größeren Verdünnungen scheint in diesem Fall die Messung nicht linear zu verlaufen, was wir auch bei der Datenerhebung beachten müssen.

Es gibt keine Unterschiede zwischen den Primereffizienzen für die Bl6 Mäuse im Vergleich zu den untersuchten Wildmäusen. Im Weiteren verwenden wir deshalb zur Berechnung der Anreicherung über den Input eine Primereffizienz von zwei für alle Primer. Da bei Slc27a5 und Acox2 unsere experimentellen Ct-Werte zwischen den 20% und 2% Werten des Primereffizienz Tests liegen, rechnen wir auch für diese Primer mit Effizienzen von zwei.

3.2.3 Ergebnisse der Gelelektrophorese

Kontrolle der Amplifikationsreaktion der qPCR

Zur Überprüfung der qPCR Reaktion wurde eine Gelelektrophorese des PCR Produkts durchgeführt. Dazu wurden Input Proben verwendet. Es gibt einen deutlichen Zusammenhang zwischen der jeweiligen Ampliconlänge, die sich aus der Reaktion der jeweiligen Primerpaare ergibt, und der im Gel zurückgelegten Strecke, wie anhand der Wildmaus H11_50 gezeigt wird (Abbildung 3.13). In der Kontrollprobe (vorletzte Spur) ist wie erwartet keine Bande zu erkennen. Das Ergebnis einer Bl6 Maus ist vergleichbar (nicht gezeigt). Die Helligkeit der Banden hat keine Aussagekraft, da ungleichmäßig pipettiert wurde.

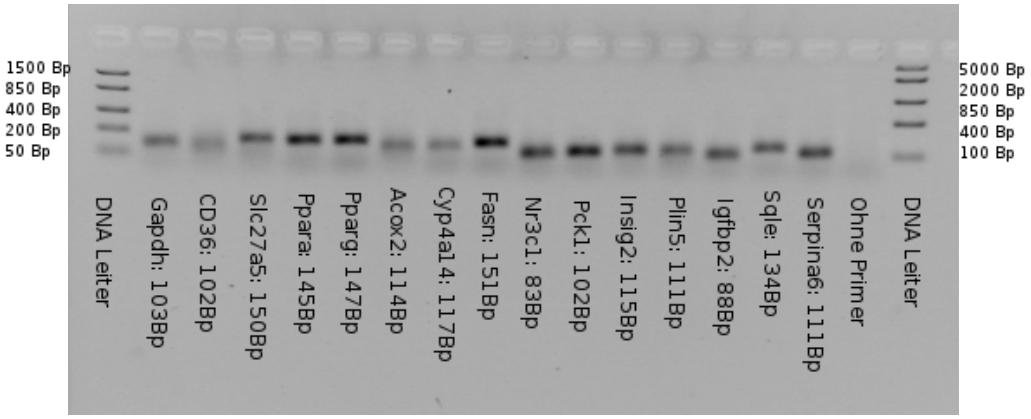

Abb. 3.13: 2% Gel des qPCR Produkt der Input Probe der Wildmaus H11_50.

Die qPCR Platte einer Wildmaus

Abbildung 3.14 zeigt die Gelelektrophorese (2% Agarosegel) der qPCR Produkte von Probe H11_50. Pro verwendetem Primerpaar sind in vier nebeneinander liegenden Taschen jeweils die 2% Input Probe gefolgt von der Immunopräzipitation mit den Antikörpern H3K4m3 und H3K27ac aufgetragen worden. Die jeweils vierte Spur zeigt das Ergebnis der Amplifikation mit dem unspezifischen Antikörper. Am rechten und linken Rand ist eine DNA Leiter aufgetragen mit den Stufen 50, 200, 400, 850 und 1500 Bp. Alle Proben zeigen einen Schleier bzw. einen Schatten im niedrigmolekularen Bereich ($<$50 Bp). Da dieser Schleier unabhängig von der Probe und auch bei der Kontrollmessung auftritt, ist davon auszugehen, dass er unsere weitere Analyse nicht beeinflusst. Bei einigen Primerpaaren ist eine deutliche Bande bei allen vier Spuren zu erkennen, beispielsweise bei Gapdh, Fasn, Nr3c1 oder Serpina6. Die zum unspezifischen Antikörper gehörende Spur, die jeweils vierte Spur pro Primer, ist manchmal schwächer ausgeprägt, das heißt, es liegt eine geringere Anzahl an amplifizierten DNA Stücken vor. Beispiele hierfür sind Insig2 und Plin5. Prinzipiell schwächere Banden zeigen CD36 und Cyp4a14. Bei Slc27a5 ist besonders auffällig, dass die Probe des unspezifischen Antikörpers eine zweite niedrigmolekulare Bande aufweist, die stärker ist als die Bande, die auf der selben Höhe wie die der 2% Input Probe liegt. Bei Acox2 ist eine zweite Bande bei allen vier Spuren deutlich zu erkennen. Obwohl dieses Primerpaar bei der *in silico* PCR ein eindeutiges Produkt liefert, gibt es bei Acox2 nach der qPCR ein zweites Amplifikat. Das Experiment mit den qPCR Produkten einer Bl6 Maus lieferte vergleichbare Ergebnisse (nicht gezeigt).

Abbildung 3.15 zeigt noch einmal die Amplifikationsprodukte von Acox2 im Vergleich mit Gapdh, Slc27a5, Igfbp2 und Serpina6. Das zweite Amplifikat von Acox2 tritt konsistent auf. Bei Slc27a5 ist ein ähnlicher Effekt zu beobachten, allerdings deutlich schwächer ausgeprägt.

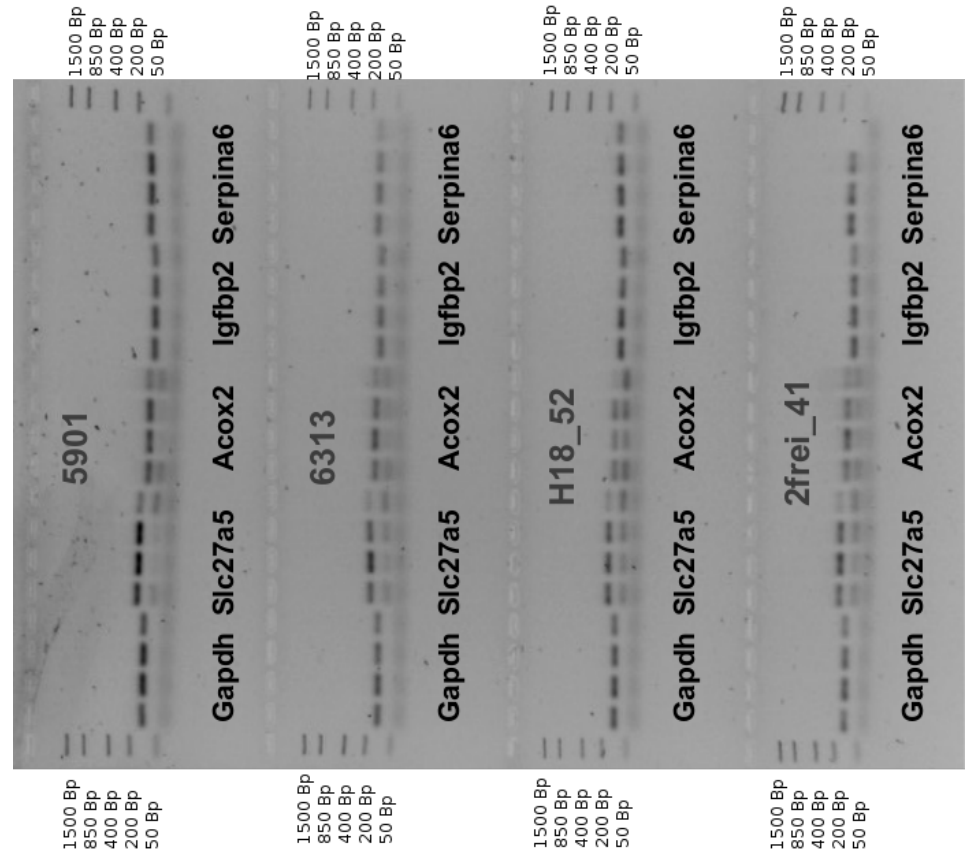

Abb. 3.15: 2% Gel ausgewählter Primer und Proben.

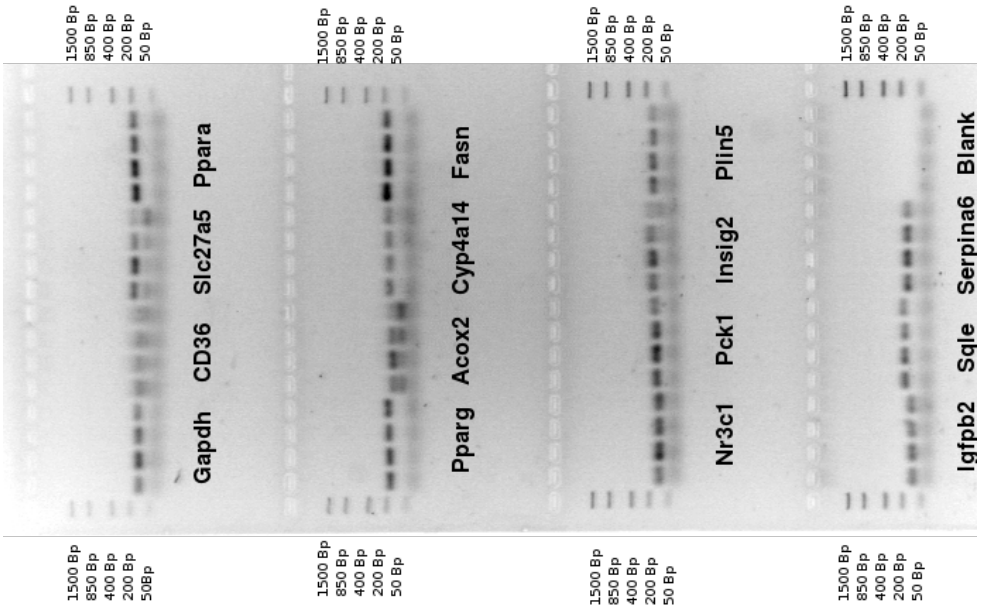

Abb. 3.14: 2% Gel der qPCR Platte von H11_50.

3.2.4 Die Triplikate in der qPCR Analyse

In der qPCR messen wir jeweils in Triplikaten an 15 Loci die Ct-Werte von vier unterschiedlich behandelten Proben einer Maus. In Abbildung 3.16 sind alle gemessenen Ct-Werte der Maus H18_66 aufgetragen. Der Input (grau) ist die Messung des Chromatins vor der Immunopräzipitation, H3K4me3 und H3K27ac sind die immunopräzipitierten Proben und der unspezifische Antikörper ist eine Blindprobe. Die Varianz innerhalb der Triplikate ist insgesamt sehr gering.

Niedrige Ct-Werte bedeuten eine hohe DNA Anreicherung. Die niedrigsten Werte gehören zu den Triplikaten der Immunopräzipitation mit dem Antikörper H3K4me3 mit Ausnahme des Locus CD36. Pparg und Sqle zeigen hier ebensfalls eine schwächere Anreicherung.

Die Ct-Werte der Immunopräzipitation mit dem Antikörper H3K27ac liegen deutlich höher, und zwar über den Input Werten, aber unter den Messwerten des unspezifischen Antikörpers. Die Anreicherung ist insbesondere bei Pparg und Sqle sehr schwach, weil die Ct-Werte sehr nahe an denen des unspezifischen Antikörpers liegen. Die beste Anreicherung finden wir bei Acox2. Bei diesem Primer ist auch der Wert des unspezifischen Antikörpers sehr niedrig.

Die Ct-Werte der Immunopräzipitation mit dem unspezifischen Antikörper liegen deutlich höher als die der DNA enthaltenden Proben. Vergleichsweise niedrige Werte ergeben sich bei Acox2 und CD36. Die Varianz der Triplikate der Blindprobe ist bei manchen Primern vergleichsweise hoch. Der mit dem Pfeil gekennzeichnete Messwert bei Slc27a5 ist ein Beispiel für einen seltenen Ausreißer; solche werden bei den weiteren Berechnungen ausgeschlossen.

Die Relationen der Ct-Werte untereinander sind in den anderen Datensätzen ähnlich zu Abbildung 3.16. Einen Überblick über alle Ct-Werte gibt Abbildung 3.17, bei der der Mittelwert der Ct-Werte pro Primer über alle 24 analysierten Wildmäuse berechnet wurde. Pro Immunopräzipitation und für den Input ist ein Boxplot über die Mittelwerte der 15 Primer abgebildet. Die Primer Acox2, Sqle, Pparg und CD36 sind jeweils hervorgehoben. Auch hier ist zu erkennen, dass der niedrigste Ct-Wert meist zur H3K4me3 Anreicherung gehört, die Ct-Werte des Inputs folgen und schließlich H3K27ac und der unspezifische Antikörper die höchsten Ct-Werte liefern. Das bedeutet, dass in der H3K4me3 Immunopräzipitation eine deutliche Anreicherung gegenüber der genomischen DNA stattfindet, während der H3K27ac Antikörper im Vergleich zur Blindprobe nur schwach anreichert. Am Locus CD36 liegt der Input unter dem H3K4me3 Wert. Im Vergleich zu den anderen Loci scheint die Histonmarkierung hier sehr gering zu sein. Theoretisch sollte der Input Wert pro Präparation für alle Primer gleich sein. Die geringen Schwankungen zeigen die technische Variabilität der Methode. Deshalb verrechnen wir bei der Auswertung die Ct-Werte der Immunopräzipitationen mit den jeweiligen Input Ct-Werten durch die %Input Methode.

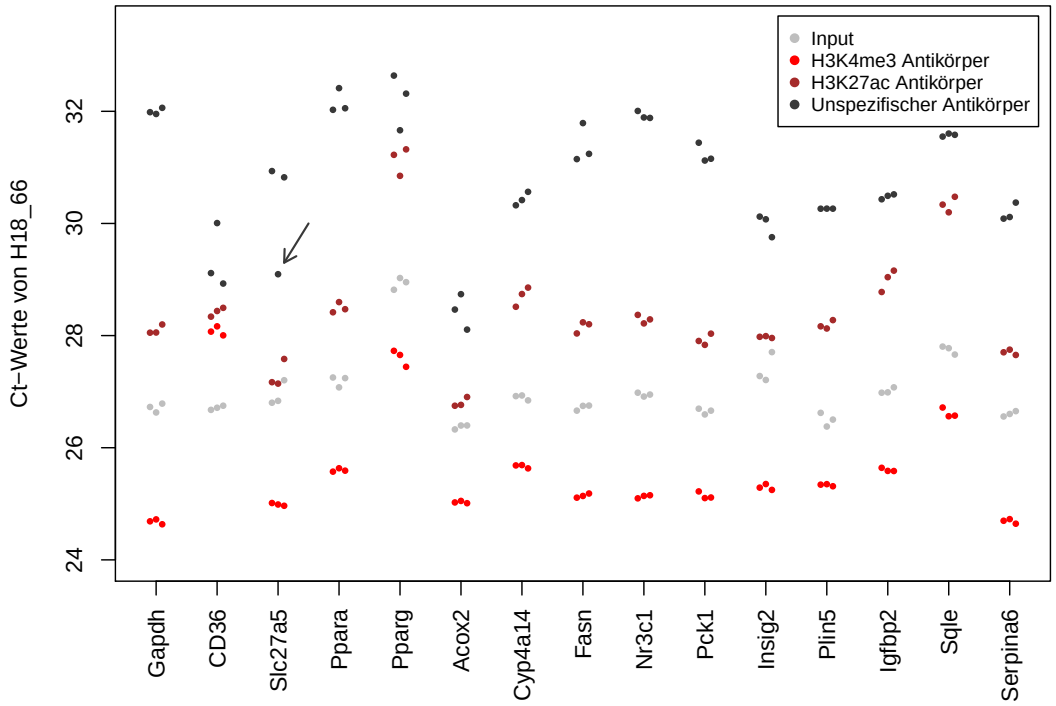

Abb. 3.16: Triplikate der qPCR Ergebnisse der Maus H18_66. Als unspezifischer Antikörper wurde Rabbit IgG verwendet. Der mit dem Pfeil gekennzeichnete Messwert ist ein Ausreißer im entsprechenden Triplikat.

3.2.5 Die Input Ct-Werte

Da der Vergleich zwischen den Mäusen mit Hilfe der %Input Werte ermöglicht wird, analysieren wir zur Methodenvalidierung auch die Input Ct-Werte der verschiedenen Messungen. Abbildung 3.18 A untersucht die Streuung der Input Ct-Werte vom mittleren Input Ct-Wert pro Maus für jedes Primerpaar. Pro Maus wurde ein mittlerer Input Ct-Wert aus den 15 verschiedenen Primerergebnissen bestimmt. Aufgetragen ist der Input Ct-Wert an einem bestimmten Primer minus diesen Mittelwert. In blau sind die Mäuse aus Haus 11, in grün die Mäuse aus Haus 18 und in rot die Mäuse ohne Hauszugehörigkeit markiert. Die drei Gruppen verhalten sich über alle Primer hinweg sehr ähnlich. Pparg hat sehr hohe Ct-Werte, sie liegen ein bis zwei Einheiten über dem Mittelwert der anderen Primer. Die Ct-Werte von Acox2 hingegen liegen knapp einen halben bis einen Ct-Wert unterhalb des Durchschnitts. Die Varianz innerhalb der Primer ist ähnlich, nur bei Slc27a5 und Acox2 sind Ausreißer zu beobachten. Insbesondere bei Slc27a5 fällt auf, dass drei Werte von Mäusen ohne Hauszugehörigkeit kleiner sind. Diese Werte sind aber immer noch in einer Größenordnung, die vergleichbar mit anderen Primern ist. Der untere Ausreißer der Mäusen ohne Hauszugehörigkeit bei Acox2 entspricht keinem der unteren Ausreißer bei Slc27a5.

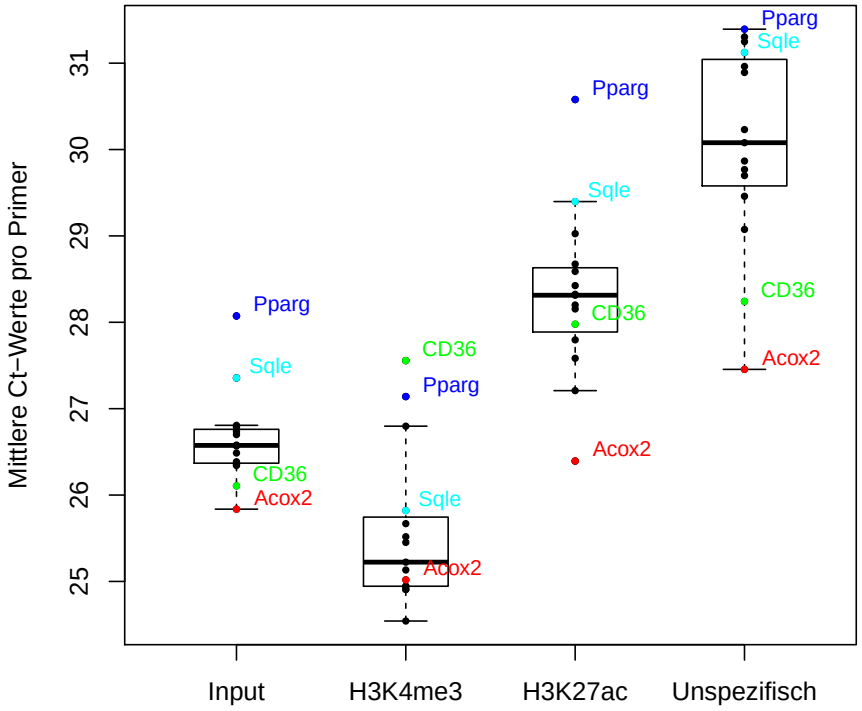

Abb. 3.17: Mittlere Ct-Werte über alle Messungen pro in der qPCR verwendeten Primer.

Die DNA Konzentration im Chromatin beeinflusst die Ct-Werte. Deshalb berechnen wir aus den DNA Konzentrationsmessungen (Abschnitt 2.3.3) ein Verdünnungsverhältnis, um im Experiment die DNA Konzentration auf einen konstanten Wert von 1μg pro Probe einstellen. Ungenauigkeiten in der Messung führen zu unterschiedenlichen Ct-Niveaus verschiedener Proben. Drei Beispiele dafür sind in Abbildung 3.18 B dargestellt. In dieser Grafik sind die unnormierten Input Ct-Werte pro Primer aufgetragen. Prinzipiell verlaufen die drei Kurven sehr ähnlich mit einem Maximum bei Pparg und einem Minimum bei Acox2, aber sie sind parallel zu höheren oder niedrigeren Ct-Werten verschoben.

In Abbildung 3.18 C ist die Größe des Amplicons pro Primer aufgetragen. Man könnte vermuten, dass ein größeres Amplicon-Fenster zu einer stärkeren Fluoreszenz führt, da sich mehr SYBR Green Moleküle einlagern könnten. Eine stärkere Fluoreszenz müsste dann zu niedrigeren Ct-Werten führen. Dieser Zusammenhang ist nicht gegeben. Nr3c1 hat das kleinste Amplicon-Fenster und Fasn das größte, und bei beiden liegen die Input Ct-Werte ziemlich genau im Mittel aller Input Ct-Werte. Unsere Messungen sind damit unabhängig von der Länge des Amplicons.

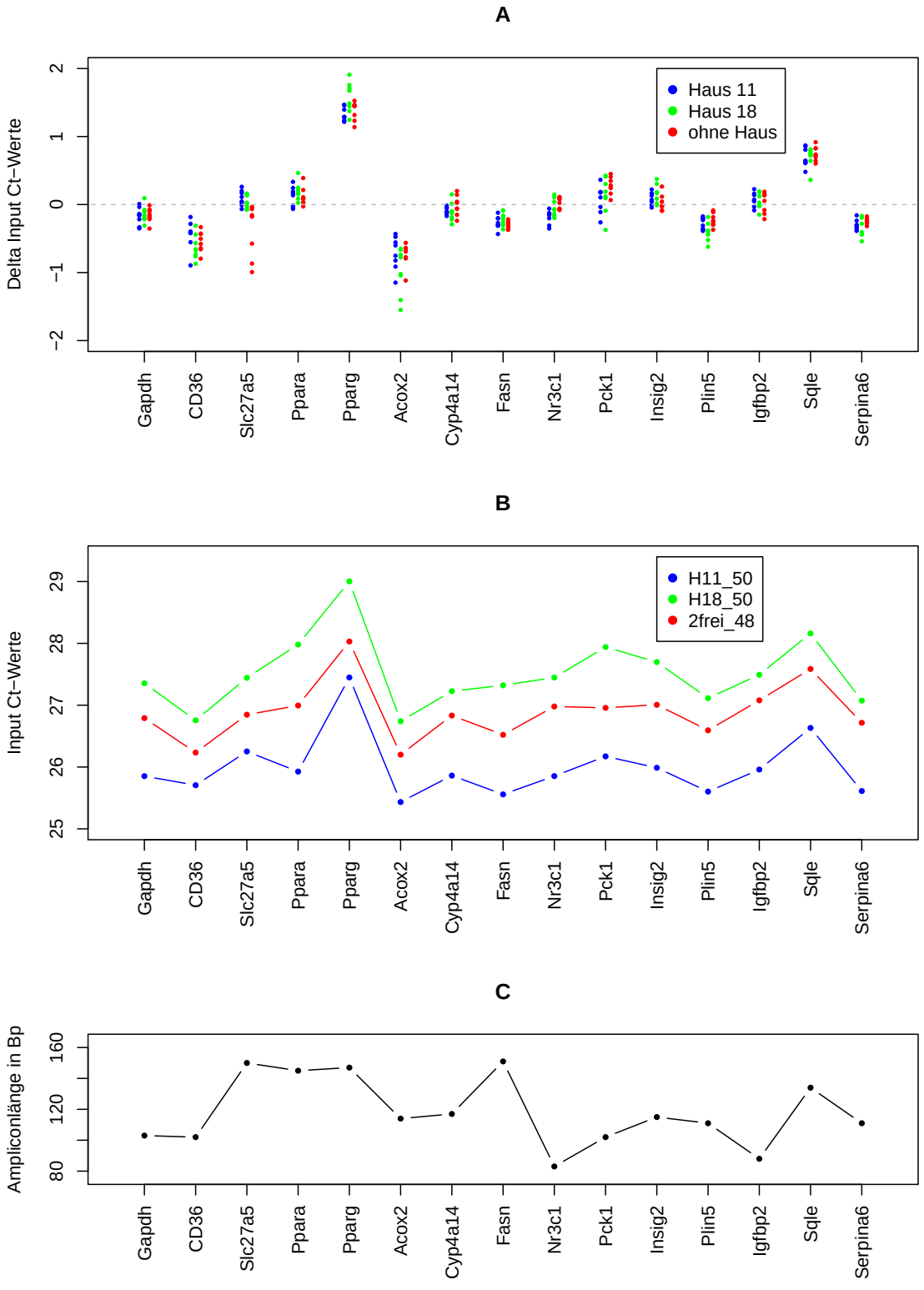

Abb. 3.18: Vergleich der Schwankungen der Input Messwerte mit dem Amplicon-Fenster. A: Input-Werte der einzelnen Datensätze an den 15 Primern jeweils im Vergleich zum Mittelwert aller Input-Werte einer Maus. B: Der Verlauf unnormierter Input Ct-Werte am Beispiel von drei Mäusen. C: Länge des Amplicons der verschiedenen Primer in Basenpaaren.

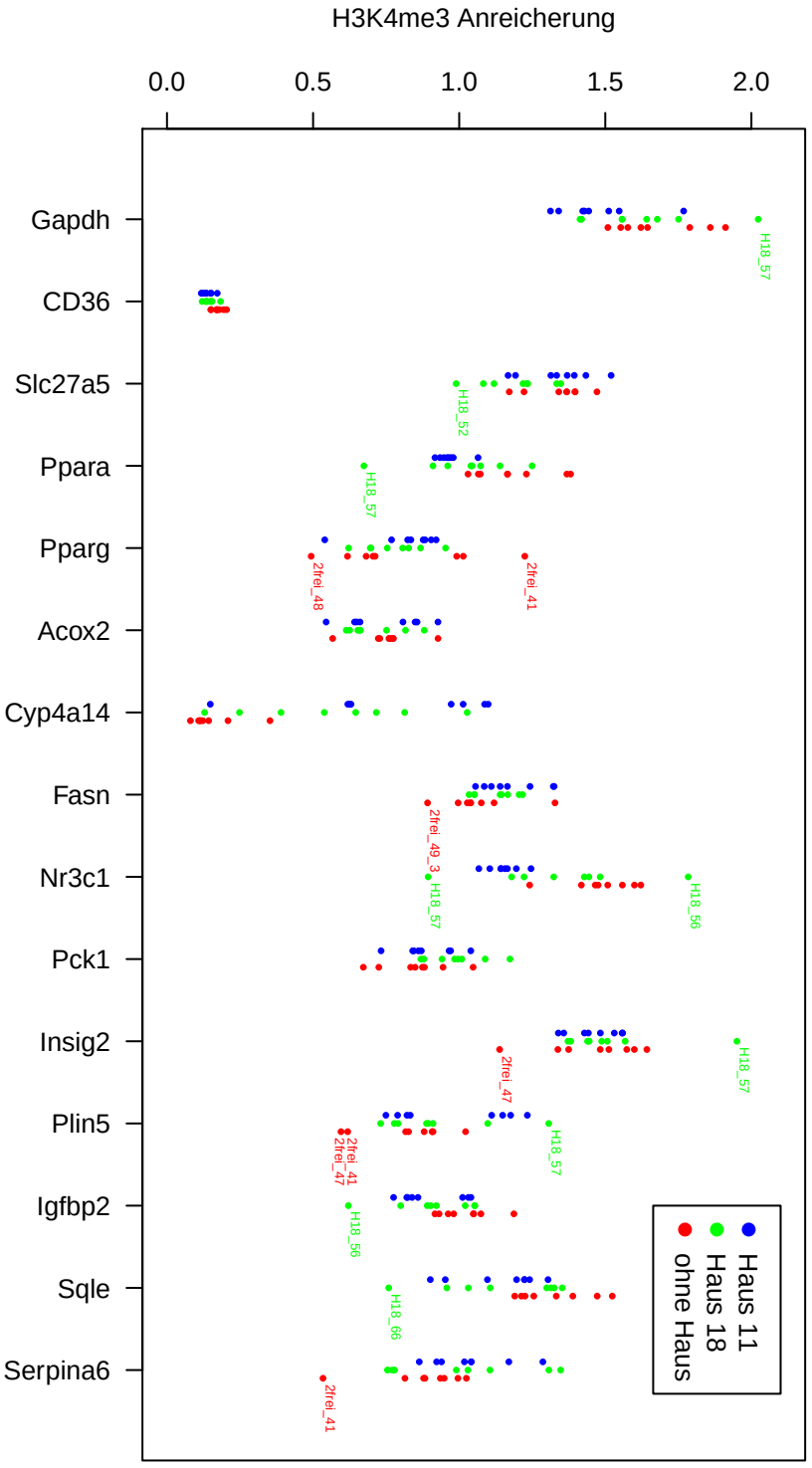

Abb. 3.19: H3K4me3 Anreicherung gegenüber dem Input Signal, normiert auf den Mittelwert pro Maus. Jede Maus ist pro Locus durch einen Punkt repräsentiert. Mäuse, die an einem oder mehreren Primern auffallen, sind mit ihrer ID gekennzeichnet.

3.3 Ergebnisse des ChIP qPCR Experiments

3.3.1 Grafische Darstellung der qPCR Ergebnisse

3.3.1.1 H3K4me3 Anreicherung

Abbildung 3.19 zeigt die Verteilung der H3K4me3 Anreicherung der 24 analysierten Mäuse an den 15 untersuchten Loci. Jede Maus wird pro Primer durch einen Punkt dargestellt. Die meisten Datenpunkte streuen mit einer Standardabweichung von 0,16 um die Mittelwerte. Ausreißer beobachten wir innerhalb von Haus 18 und innerhalb der Mäuse ohne Hauszugehörigkeit. Diese sind in der Abbildung mit ihrer Identifikationsnummer markiert. Zum Beispiel taucht die Maus H18_57 an fünf Loci als niedrigster oder höchster Wert auf. Das bedeutet, dass bei dieser Maus die untersuchten Genbereiche im Vergleich mit den anderen Mäusen sehr unterschiedlich markiert sind. In der Gruppe der Mäuse ohne Hauszugehörigkeit fällt die Maus 2frei_41 dreimal durch extreme Werte auf. Die anderen extremen Markierungen gehören zu unterschiedlichen Mäusen.

Als Erstes wollen wir die H3K4me3 Anreicherung aller 24 Datensätze gemeinsam betrachten. Da wir erwarten, dass die Mäuse, die im selben Haus lebten, näher miteinander verwandt sind, als mit den anderen Mäusen, vergleichen wir anschließend die drei Gruppen *Haus 11*, *Haus 18* und *ohne Hauszugehörigkeit*.

Einen hohen Mittelwert haben Gapdh (MW=1,60) und Insig2 (MW=1,48; Tabelle 3.8). Das bedeutet, Gapdh und Insig2 sind im Vergleich zu den anderen untersuchten Genen in dem verwendeten Amplicon-Fenster am stärksten mit H3K4me3 markiert. Den niedrigsten Mittelwert und die kleinste Varianz hat CD36 (MW=0,15; sd=0,02). Deshalb könnte CD36 bei allen Mäusen gar nicht markiert sein. Bei Cyp4a14 scheint es unmarkierte und markierte Proben zu geben. Dementsprechend liegt der Mittelwert von Cyp4a14 sehr niedrig (MW=0,50), aber die Varianz (sd=0,36) ist die größte.

Tab. 3.8: H3K4me3 Anreicherung: Extreme Werte der Mittelwerte (MW) und Standardabweichungen (sd).

	Größter Wert	Zweitgrößter Wert	Zweitkleinster Wert	Kleinster Wert
MW (alle 24)	Gapdh: 1,60	Insig2: 1,48	Cyp4a14: 0,50	CD36: 0,15
sd (alle 24)	Cyp4a14: 0,36	Nr3c1: 0,21	Acox2: 0,11	CD36: 0,02
MW: Haus 11	Gapdh: 1,47	Insig2: 1,46	Acox2: 0,74	CD36: 0,14
MW: Haus 18	Gapdh: 1,63	Insig2: 1,51	Acox2: 0,71	CD36: 0,15
MW: ohne Haus	Gapdh: 1,68	Nr3c1: 1,49	CD36: 0,17	Cyp4a14: 0,16
sd: Haus 11	Cyp4a14: 0,33	Plin5: 0,20	Ppara: 0,04	CD36: 0,02
sd: Haus 18	Cyp4a14: 0,30	Nr3c1: 0,26	Fasn: 0,07	CD36: 0,02
sd: ohne Haus	Pparg: 0,25	Insig2: 0,17	Cyp4a14: 0,09	CD36: 0,02

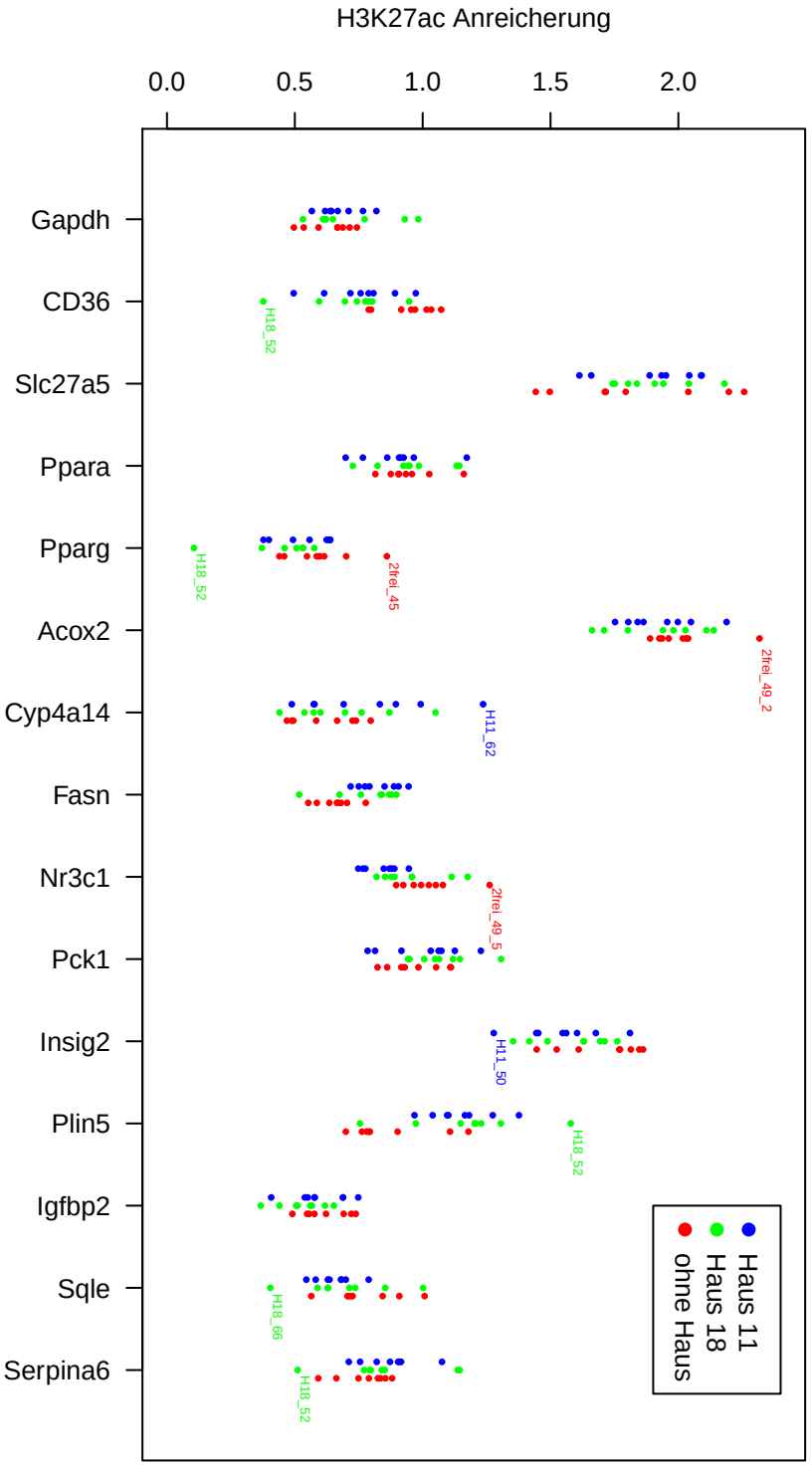

Abb. 3.20: Anreicherung der H3K27ac Markierung verglichen mit dem Input Wert, normalisiert auf den mittleren Wert pro Maus. Besonders hohe oder niedrige Werte sind mit der ID der entsprechenden Maus gekennzeichnet.

Bei Betrachtung der einzelnen Gruppen finden wir Unterschiede in der Rangfolge bei der zweithöchsten Markierung. Für die Mäuse aus Haus 11 und 18 ist diese Insig2, aber für die Mäuse ohne Hauszugehörigkeit ist diese Nr3c1 (exakte Werte siehe Tabelle 3.8). Auch die Markierungen bei den zweitkleinsten Werten clustern nach experimentellen Gruppen. Diese liegen für die Mäuse aus Haus 11 bei Acox2 (MW=0,74) und für die Mäuse aus Haus 18 und für die Mäuse ohne Hauszugehörigkeit bei Cyp4a14 (MW=0,56 bzw. MW=0,16).

Die Standardabweichung der Mäuse aus Haus 18 ist größer als diejenige der anderen beiden Gruppen. Sie liegt im Mittel über die 15 Loci bei 0,16 im Vergleich zu 0,12 bzw. 0,13 für die Mäuse aus Haus 11 bzw. die Mäuse ohne Hauszugehörigkeit. Allerdings führt ein Weglassen der Maus H18_57, die an fünf Loci einen extremen Wert annimmt, bei der Berechnung der Standardabweichung zu einem vergleichbaren Wert (sd=0,13).

Eine besonders kleine Varianz kann man für die Mäuse aus Haus 18 bei Fasn beobachten (sd=0,07). Eine ähnlich kleine Standardabweichung liegt bei den Mäusen ohne Hauszugehörigkeit beim Gen Slc27a5 (sd=0,09). Bei Ppara fällt auf, dass die Varianz innerhalb der Mäuse aus Haus 11 extrem gering ist (sd=0,04). Die größte Varianz haben die Mäuse aus Haus 11 und 18 bei Cyp4a14. Bei den Mäusen ohne Hauszugehörigkeit tritt diese bei Pparg auf (sd=0,25).

Die H3K4me3 Anreicherung am Gen Cyp4a14 unterscheidet sich zwischen den Gruppen. Die Mäuse ohne Hauszugehörigkeit sind nur sehr gering markiert (MW=0,16; sd=0,09). Die Mäuse aus den Häusern 11 und 18 weisen einen höheren Mittelwert (MW=0,77 bzw. MW=0,56) und eine größere Varianz (sd=0,33 bzw. sd=0,30) auf. Das bedeutet, dass die H3K4me3 Anreicherung an diesem Genbereich bei den Mäusen aus den Häusern 11 und 18 deutlich stärker ausgeprägt ist als bei den Mäusen ohne Hauszugehörigkeit.

3.3.1.2 H3K27ac Anreicherung

In Abbildung 3.20 sind die H3K27ac Anreicherungen dargestellt. Auch hier wird jede Maus pro Gen durch einen Punkt repräsentiert, dessen Farbe die Zugehörigkeit zu einer der Gruppen Haus 11, Haus 18 oder ohne Hauszugehörigkeit beschreibt. Punkte, die durch ihre Lage im oberen oder unteren Spektrum auffallen, sind mit der Indentifikationsnummer der zugehörigen Maus markiert. In vier Fällen ist Maus H18_52 ein Ausreißer. Ansonsten sind keine Auffälligkeiten festzustellen.

Zunächst betrachten wir alle 24 Datensätze gemeinsam, dann unterteilen wir die Mäuse nach ihrer Gruppenzugehörigkeit.

Die größten Markierungen über alle 24 Mäuse hinweg finden sich bei Acox2 (MW=1,96) und Slc27a5 (MW=1,88), die kleinsten bei Pparg (MW=0,53) und Igfbp2 (MW=0,58; Tabelle 3.9). Bei Igfbp2 ist die Variabilität am geringsten (sd=0,10). Ähnlich kompakt sind die Werte bei Gapdh (sd=0,12). Die größte Streuung zeigen Plin5 und Slc27a5 (sd=0,23 bzw. sd=0,22).

Wenn nun die Gruppenzugehörigkeiten mit betrachtet werden, ergeben sich keine Auffälligkeiten in den Mittelwerten. Bei der Variablität innerhalb der Gruppen ergibt sich ein anderes

Bild. Insgesamt haben die Mäuse aus Haus 18 im Mittel eine größere Varianz (sd=0,16) als die Mäuse aus Haus 11 und die Mäuse ohne Hauszugehörigkeit (beide sd=0,13). Der Grund für die erhöhte Varianz der Mäuse aus Haus 18 größer ist, dass Maus H18_52 an vier Genbereichen einen Ausreißer darstellt. Wenn Maus H18_52 bei Berechnung der gruppeninternen Standardabweichungen weggelassen wird, ergibt sich auch hier ein mittlerer Wert von 0,13. Die Mäuse aus Haus 11 zeigen bei Cyp4a14 die größte Standardabweichung (sd=0,25) und bei Slc27a5 die zweitgrößte (sd=0,18). Hingegen haben die Mäuse aus Haus 18 bei Plin5 die größte Varianz und bei Serpina6 die zweitgrößte (sd=0,24 bzw. sd=0,21). Die Mäuse ohne Hauszugehörigkeit sind an den Loci Slc27a5 und Plin5 am variabelsten (sd=0,30 bzw. sd=0,17). Die kleinste Varianz haben die Mäuse aus Haus 11 bei Nr3c1 (sd=0,07). Die Mäuse aus Haus 18 sind am Locus Igfbp2 am kompaktesten (sd=0,09). Fasn ist der Genbereich, in dem die Mäuse ohne Hauszugehörigkeit die kleinste Standardabweichung zeigen (sd=0,07).

Auffällig ist in diesem Datensatz insbesondere die hohe Varianz der Mäuse ohne Hauszugehörigkeit am Locus Slc27a5. Die Standardabweichung ist dort mit 0,30 fast doppelt so hoch wie bei dem Gen mit der zweitgrößten Variabilität (Plin5 mit sd=0,17).

Tab. 3.9: H3K27ac Anreicherung: Extreme Werte der Mittelwerte (MW) und Standardabweichungen (sd).

	Größter Wert	Zweitgrößter Wert	Zweitkleinster Wert	Kleinster Wert
MW (alle 24)	Acox2: 1,96	Slc27a5: 1,88	Igfbp2: 0,58	Pparg: 0,53
sd (alle 24)	Plin5: 0,23	Slc27a5: 0,22	Gapdh: 0,12	Igfbp2: 0,10
MW: Haus 11	Acox2: 1,93	Slc27a5: 1,91	Igfbp2: 0,60	Pparg: 0,55
MW: Haus 18	Acox2: 1,92	Slc27a5: 1,90	Igfbp2: 0,53	Pparg: 0,45
MW: ohne Haus	Acox2: 2,01	Slc27a5: 1,83	Igfbp2: 0,62	Pparg: 0,62
sd: Haus 11	Cyp4a14: 0,25	Slc27a5: 0,18	Sqle: 0,08	Nr3c1: 0,07
sd: Haus 18	Plin5: 0,24	Serpina6: 0,21	Pck1: 0,12	Igfbp2: 0,09
sd: ohne Haus	Slc27a5: 0,30	Plin5: 0,17	Gapdh: 0,09	Fasn: 0,07

3.3.1.3 Vergleich H3K4me3 mit H3K27ac Anreicherung

H3K4me3 und H3K27ac sind beides Markierungen, die mit aktiver Transkription assoziiert werden [63] [38]. Interessanterweise zeigen sich im Vergleich der Abbildungen 3.19 und 3.20 einige Unterschiede. Da beide Datensätze pro Maus über den Mittelwert aus allen 15 Loci normalisiert wurden, können wir die einzelnen Werte miteinander vergleichen.

Der Locus mit dem größten Mittelwert der H3K4me3 Anreicherung ist Gapdh (MW=1,60), während er bei H3K27ac unterdurchschnittlich markiert ist (MW=0,68). Dagegen sind Acox2 und Slc27a5 die höchsten H3K27ac Markierungen (MW=1,96 bzw. MW=1,88) und tauchen bei den H3K4me3 Markierungen nur im mittleren Bereich auf (MW=0,73 bzw. MW=1,29). Anders als bei der H3K4me3 Markierung fällt CD36 bei H3K27ac nicht als unmarkiert auf (MW=0,15 bzw. MW=0,81). Die Spanne der H3K27ac Anreicherung ist im Allgemeinen nach oben verschoben. Sie reicht von im Mittel 0,53 (Pparg) bis 1,96 (Acox2). Betrachtet man die experimentellen Gruppen, liegen die Standardabweichungen bei beiden Markierungen in einem sehr ähnlichen Bereich: die Mäuse aus Haus 18 zeigen in beiden Fällen eine leicht höhere Varianz (sd=0,16), während die anderen beiden Mausgruppen im Mittel in beiden Datensätzen nur eine Varianz von 0,13 haben. Die H3K4me3 Daten zeigen an zwei Genbereichen eine extreme Varianz: Cyp4a14 hat eine besonders große Varianz (sd=0,36) und CD36 eine besonders kleine (sd=0,02). Die größte bzw. kleinste Varianz in der H3K27ac Anreicherung passt mit einer Standardabweichung von 0,23 (Plin5) bzw. 0,10 (Igfbp2) in der Größenordnung besser zur zweitgrößten bzw. zweitkleinsten Standardabweichung im H3K4me3 Datensatz. Diese liegt bei 0,21 (Nr3c1) bzw. 0,11 (Acox2).

3.3.2 Statistische Auswertung

3.3.2.1 H3K4me3 Anreicherung

Der Shapiro-Wilk-Test angewendet auf die einzelnen Gruppen ergibt, dass bei Cyp4a14 die Werte der Mäuse aus Haus 11 und ohne Hauszugehörigkeit nicht normalverteilt sind (p=0,007 bzw. p=0,009). Alle anderen Werte ergeben keinen p-Wert kleiner als 0,05 und es kann somit die Nullhypothese auf Normalverteilung der Daten nicht abgelehnt werden.

Zwischen der H3K4me3 Anreicherung der Mäuse aus Haus 11 und Haus 18 gibt es nur einen statistisch signifikanten Unterschied in den Mittelwerten (Slc27a5). Das bedeutet, dass diese Mäuse sich in der H3K4me3 Markierung sehr ähnlich sind. Anders verhält es sich bei dem Vergleich zwischen den Mäusen aus Haus 11 und den Mäusen ohne Hauszugehörigkeit: an acht der 15 betrachteten Loci unterscheidet sich die H3K4me3 Anreicherung signifikant zwischen den beiden Gruppen. Am stärksten ist das Signal beim Locus Nr3c1 ($p=5{,}1 \cdot 10^{-6}$). Ebenso finden sich zwischen den gemessenen H3K4me3 Anreicherungen der Mäuse aus Haus 18 im Vergleich zu den Mäusen ohne Hauszugehörigkeit einige signifikante Unterschiede (5 Loci). An drei Loci zeigen beide Vergleiche zwischen Mäusen in Häusern und ohne Hauszugehörigkeit einen signifikanten Unterschied (CD36, Ppara und Cyp4a14).

Tab. 3.10: Ergebnisse der Tests auf Gleichheit der Varianzen (bei nicht normalverteilten Datenpunkten, markiert mit ‡) bzw. Gleichheit der Mittelwerte.

Gen	Haus 11 vs. Haus 18	Haus 11 vs. ohne Haus	Haus 18 vs. ohne Haus
Gapdh		p=0,012	
CD36		p=0,002	p=0,010
Slc27a5	p=0,030		p=0,020
Ppara		p=0,002	p=0,041
Cyp4a14		p=0,004 ‡	p=0,01 ‡
Nr3c1		$p=5{,}1 \cdot 10^{-6}$	
Pck1			p=0,024
Igfbp2		p=0,031	
Sqle		p=0,018	
Serpina6		p=0,048	

3.3.2.2 H3K27ac Anreicherung

Der Shapiro-Wilk-Test auf Normalverteilung ergibt bei den Daten der H3K27ac Anreicherung, dass die Werte von Pparg in den Gruppen Haus 11 und Haus 18 nicht normalverteilt sind (p=0,045 bzw. p=0,007). Der weitere Testablauf erfolgt wie in Abbildung 2.6 beschrieben und führt zu den Ergebnissen in Tabelle 3.11. Eine graphische Darstellung der hier verwendeten Daten ist in Abbildung 3.20.

Die H3K27ac Anreicherung unterscheidet sich zwischen den Mäusen aus Haus 11 und Haus 18 nur an einem untersuchten Locus signifikant (Nr3c1). Die sozialisierten Mäuse sind sich demnach in dieser epigenetischen Markierung sehr ähnlich, was den Ergebnissen der H3K4me3 Analyse entspricht. Die Mäuse aus Haus 11 unterscheiden sich von denen ohne Hauszugehörigkeit an den Loci CD36, Fasn, Nr3c1 und Plin5. Dabei ergibt sich der niedrigste p-Wert beim Test auf Gleichheit der Mittelwerte am Locus Fasn (p=0,0005). Der Vergleich der H3K27ac Anreicherung zwischen den Mäusen aus Haus 18 und den nicht sozialisierten Mäusen ergibt an den Genbereichen CD36, Pparg, Fasn und Plin5 signifikante Unterschiede. Im Vergleich zwischen den sozialisierten Mäusen und den Mäusen ohne Hauszugehörigkeit bringt die H3K27ac Analyse demnach jeweils vier signifikant unterschiedliche Loci hervor, von denen drei übereinstimmen: CD36, Fasn und Plin5.

Tab. 3.11: Ergebnisse der Tests auf Gleichheit der Mittelwerte.

Gen	Haus 11 vs. Haus 18	Haus 11 vs. ohne Haus	Haus 18 vs. ohne Haus
CD36		p=0,012	p=0,006
Pparg			p=0,046
Fasn		p=0,0005	p=0,031
Nr3c1	p=0,040	p=0,002	
Plin5		p=0,003	p=0,013

3.3.3 Hauptkomponentenanalyse

Die qPCR Analyse der 24 Wildmäuse hat 15 Dimensionen, die den 15 untersuchten Genen entsprechen. Eine PCA projiziert Daten auf einen niedrig dimensionalen Raum, der einen größtmöglichen Anteil der Varianz erklärt. Bei den vorhandenen H3K4me3 Datensätzen trennen sich die Mäuse ohne Hauszugehörigkeit von den Mäusen aus Häusern 11 und 18 entlang der ersten Hauptkomponente (siehe Abbildung 3.21 A). Bei dieser Trennung werden eine Maus aus Haus 18 und eine aus Haus 11 falsch zugeordnet. Die erste Hauptkomponente kann bei dieser Zerlegung 26,5% der Varianz in dem Datensatz erklären, die zweite 18,3%. Das ist eine kumulierte Varianz von 44,8%. Bei einer Hauptkomponentenanalyse auf Basis der H3K27ac Anreicherung ergibt sich ein noch deutlicheres Bild: projiziert auf die ersten beiden Hauptkomponenten können die Datenpunkte entlang der zweiten Hauptkomponente fast fehlerfrei getrennt werden (siehe Abbildung 3.21 B). Nur eine der Mäuse ohne Hauszugehörigkeit wird falsch klassifiziert. Dabei erklärt die erste Hauptkomponente 25,9% der Varianz und die zweite 18,5%.

Kontrolle des Ergebnisses

Um zu überprüfen, ob die Trennung zwischen den Mäusen ohne Hauszugehörigkeit und denen aus den Häusern 11 und 18 nicht auf Anomalien in den Daten zurückzuführen ist, führten wir zwei Kontrollen durch. In Abbildung 3.22 A ist das Ergebnis der Projektion auf die ersten beiden Hauptkomponenten der Input Ct-Werte dargestellt. Vor der Anwendung der PCA wurden die Input Ct-Werte jeweils auf den mittleren Input Ct-Wert pro Maus normalisiert. Es ist klar ersichtlich, dass die Datenpunkte nicht entlang der ersten oder zweiten Hauptkomponente getrennt werden können, die jeweils 26,8% bzw. 18,5% der Varianz in den Daten erklären. Ähnlich verhält es sich bei den Ergebnissen der PCA mit den Datensätzen der Anreicherung mit dem unspezifischen Antikörper (Abbildung 3.22 B). Die ersten beiden Hauptkomponenten können jeweils 35,2% bzw. 15% der Varianz erklären, aber es gibt keine Möglichkeit, die Daten in dieser Projektion klar zu trennen. Daraus schließen wir, dass die Trennung der Gruppen in Abbildung 3.21 nicht zufällig ist, sondern die untersuchten Gene in den Gruppen tatsächlich unterschiedlich markiert sind. Um festzustellen, welche Gene einen besonderen Einfluss auf die Trennbarkeit zwischen Mäusen mit und ohne Hauszugehörigkeit haben, ist es sinnvoll, die entsprechenden Hauptkomponenten näher zu betrachten.

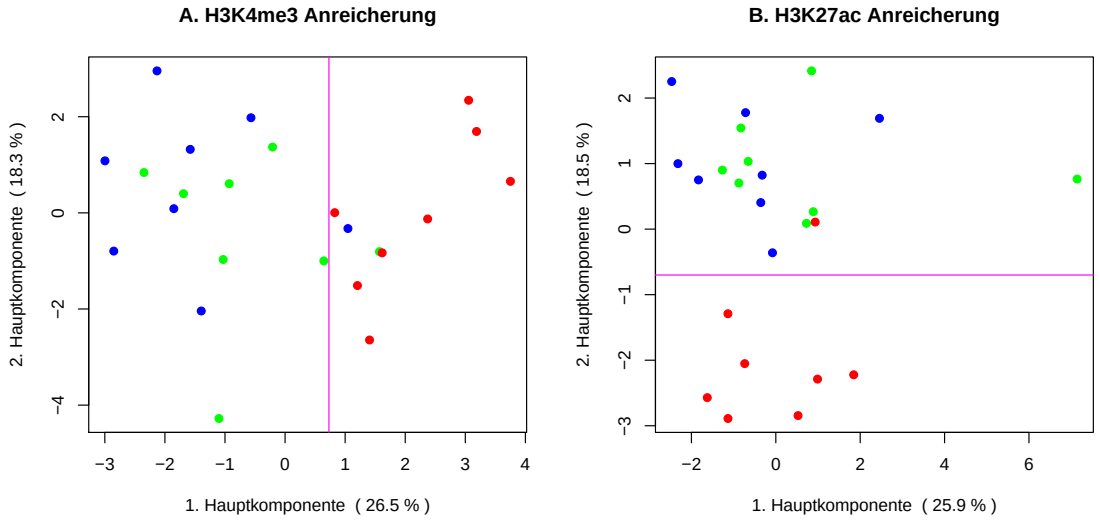

Abb. 3.21: Ergebnis der Hauptkomponentenanalyse der 24 Wildmäuse. In blau sind die Mäuse aus Haus 11, in grün die aus Haus 18 und in rot die Wildmäuse ohne Hauszugehörigkeit aufgetragen. Die magentafarbene Linie zeigt eine mögliche Trennung der Gruppen auf.

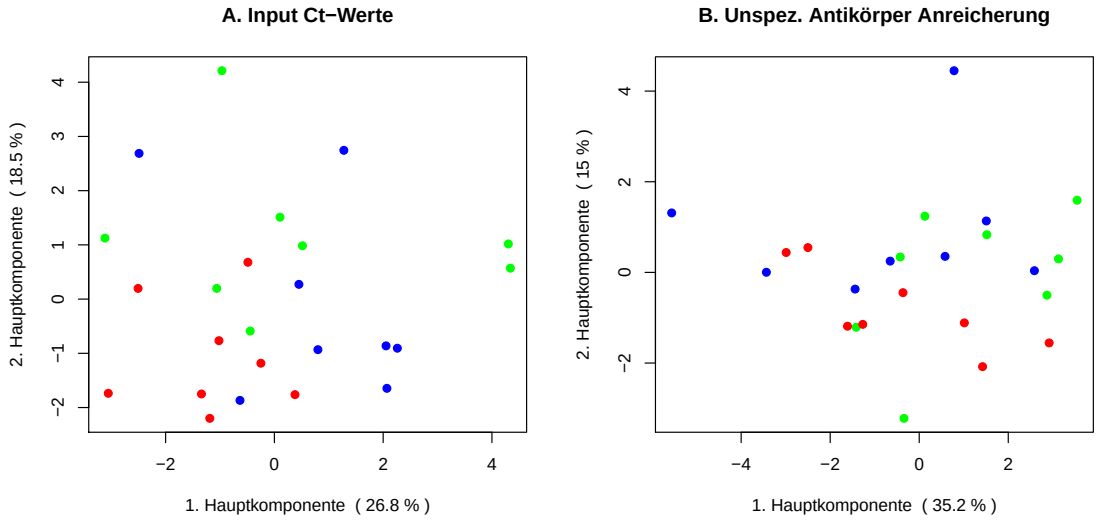

Abb. 3.22: PCA der Input Ct-Werte und der Anreicherung mit dem unspezifischen Antikörper. In blau die Mäuse aus Haus 11, in grün die aus Haus 18 und in rot die Wildmäuse ohne Hauszugehörigkeit.

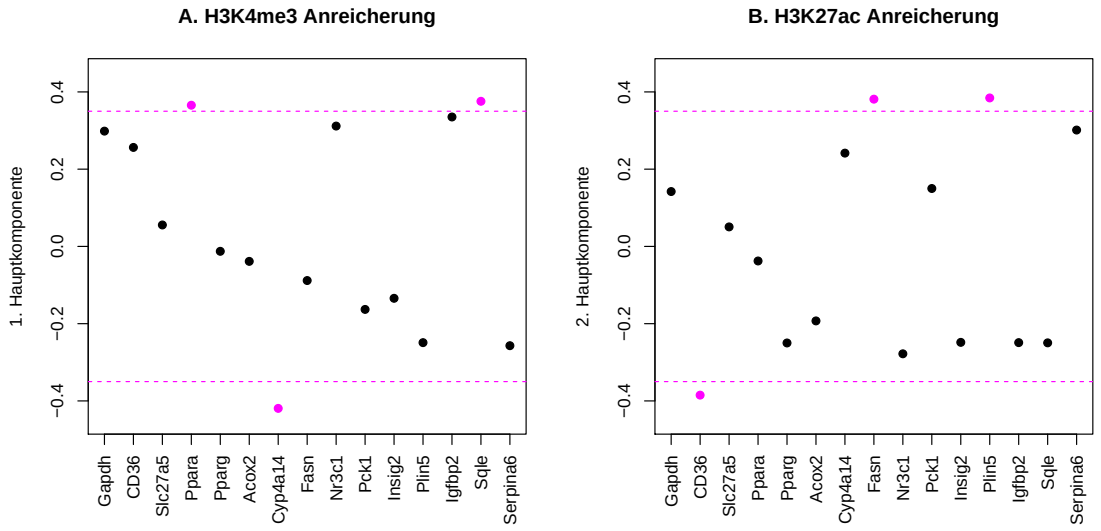

Abb. 3.23: Visualisierung der ersten bzw. zweiten Hauptkomponente der PCA angewandt auf die H3K4me3 (A) bzw. HK27ac Anreicherung (B). Variablen, deren absoluter Wert größer als 0,35 ist, sind farblich markiert.

Vergleich zur statistischen Analyse der H3K4me3 Anreicherung

Da die 24 Wildmäuse bei Verwendung der H3K4me3 Anreicherung entlang der ersten Hauptkomponente getrennt werden können, ist in Abbildung 3.23 A diese erste Hauptkomponente grafisch dargestellt. Die Parameter der Hauptkomponente, die den 15 Variablen des Datensatzes entsprechen, sind auf der x-Achse in der Abbildung dargestellt. Die Gene, die den größten Einfluss in einer Hauptkomponente haben, sind diejenigen mit den größten absoluten Werten. In der Abbildung sind diese farblich markiert. Bei der H3K4me3 Anreicherung spielen die Gene Ppara, Cyp4a14 und Sqle die größte Rolle in der ersten Hauptkomponente. In der statistischen Analyse der Daten ergab sich ein signifikanter Unterschied jeweils zwischen den sozialisierten Mäusen in den Häusern 11 und 18 und den nicht sozialisierten Mäusen unter anderem an den Loci Ppara und Cyp4a14 (siehe Tabelle 3.8). Auch an Sqle fand sich ein signifikanter Unterschied zwischen den Mäusen aus Haus 11 und denen ohne Hauszugehörigkeit. Somit führen die beiden unterschiedlichen Analyseverfahren zu einer vergleichbaren Aussage.

Vergleich zur statistischen Analyse der H3K27ac Anreicherung

In der Projektion auf die ersten beiden Hauptkomponenten konnten die Datensätze der H3K27ac Anreicherung entlang der zweiten Hauptkomponente getrennt werden, die in Abbildung 3.23 B dargestellt ist. Auch hier fallen drei Gene durch einen hohen absoluten Wert auf: CD36, Fasn und Plin5. Das sind genau diejenigen Loci, an denen die statistische Analyse einen signifikanten Unterschied jeweils im Vergleich der Mäuse aus Haus 11 und 18 und denen ohne Hauszugehörigkeit aufgedeckt hat (siehe Tabelle 3.9). Beide Auswertungen des H3K27ac Datensatzes ergeben demnach dieselben Ergebnisse.

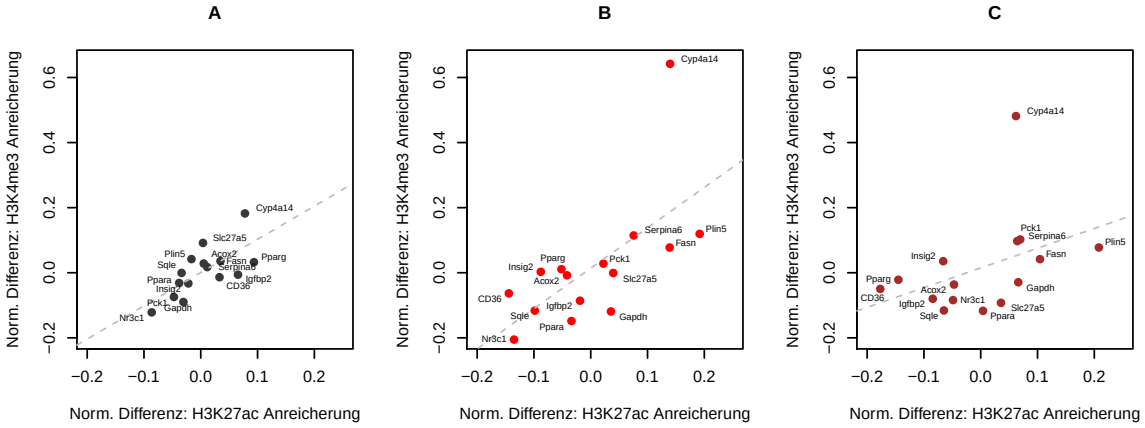

Abb. 3.24: Vergleich der Normierten Differenz der H3K4me3 und H3K27ac Anreicherung der 15 in der qPCR untersuchten Loci. A: Vergleich Haus 11 mit Haus 18. B: Vergleich Haus 11 mit ohne Haus. C: Vergleich Haus 18 mit ohne Haus.

3.3.4 Vergleich der H3K4me3 mit der H3K27ac Anreicherung

Neben der graphischen Darstellung (siehe Abschnitt 3.3.1.3) ist es auch möglich, die H3K4me3 Anreicherung mit der von H3K27ac zu vergleichen, in dem Zusammenhänge zwischen den Normierten Differenzen betrachtet werden. In Abbildung 3.24 sind jeweils die Normierten Differenzen der Vergleiche zwischen den Gruppen aufgetragen und zwar als Unterschiede in der H3K27ac Anreicherung gegen die der H3K4me3 Anreicherung. Die gestrichelten Linien sind die Ergebnisse eines linearen Modells durch die gegebenen Punkte.

Die Werte des Differenzmaßes korrelieren miteinander: in A beträgt der Korrelationskoeffizient nach Spearman 0,67, in B 0,68 und in C 0,51. Es kann von einer mittleren Korrelation gesprochen werden [15]. Da beide Histonmarkierungen mit Genen assoziiert werden, die aktiv transkribiert werden, war diese Korrelation zu erwarten. Bemerkenswert ist die positive Korrelation trotzdem, da in Abbildung 2.4 A deutlich wird, dass die Immunopräzipitiation mit dem Antikörper gegen H3K27ac entweder weniger effizient ist als mit denjenigen gegen H3K4me3, oder dass weniger H3K27ac Markierungen vorliegen.

3.3.5 Zusammenhang zwischen der Normierten Differenz und den p-Werten der statistischen Analyse

Die in Gleichung 2.4 definierte Normierte Differenz kann auf die Ergebnisse des ChIP qPCR Experiments angewendet werden, indem jeweils die Mittelwerte innerhalb der betrachteten Gruppen (Haus 11, Haus 18 und ohne Hauszugehörigkeit) berechnet werden und diese Werte in Gleichung 2.4 eingesetzt werden. Abbildung 3.25 zeigt ein Streudiagramm, in dem pro Primer die Normierte Differenz gegen die in der statistischen Analyse der Daten berechneten p-Werte aufgetragen ist. In Abbildung 3.25 A wird der Unterschied in der H3K4me3 Anreicherung zwischen den Gruppen Haus 11 und ohne Hauszugehörigkeit dargestellt, in C der Unterschied in der H3K27ac Anreicherung zwischen diesen Gruppen. Abbildung 3.25 B und D zeigen die entsprechenden Grafiken für den Vergleich zwischen Haus 18 und ohne Hauszugehörigkeit.

Die Loci mit einer Normierten Differenz von ungefähr Null entsprechen denjenigen, die in der statistischen Analyse keine signifikanten p-Werte ergeben. Pro Primer kann mit Hilfe der Formel des T-Tests ein Grenzwert berechnet werden, ab dem die Normierte Differenz gleichbedeutend ist mit einem statistisch signifikanten Unterschied der Mittelwerte. Dieser Grenzwert liegt im Mittel aller Vergleiche und Primer bei $\pm$0,1. Abbildung 3.25 macht aber bereits deutlich, dass dieser Grenzwert nur eine Orientierung darstellt und Abweichungen in beide Richtungen vorkommen. Zum Beispiel hat Cyp4a14 in Abbildung 3.25 C eine Normierte Differenz von 0,14 ($>$0,1), aber der p-Wert liegt bei 0,12 ($>$0,05). CD36 hingegen hat in Abbildung 3.25 B einen p-Wert von 0,01 ($<$0,05) bei einer Normierten Differenz von $-$ 0,05 ($> -$ 0,1).

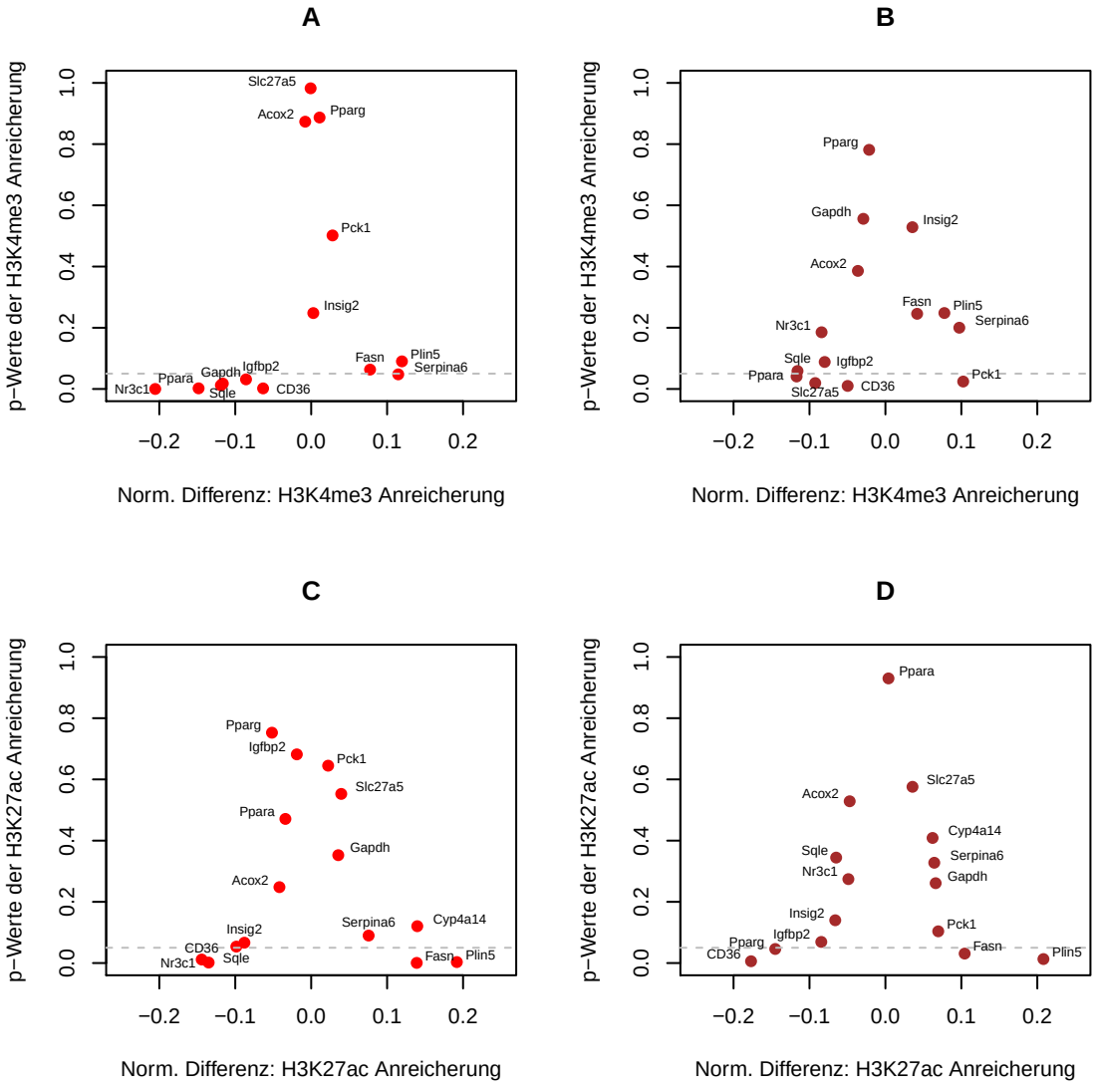

Abb. 3.25: Vergleich der Normierten Differenz der Mittelwerte der H3K4me3 (A, B) und H3K27ac (C, D) Anreicherungen zu den p-Werten der statistischen Tests auf Unterschiede in den Mittelwerten. A, C: Vergleich Haus 11 mit ohne Haus. B, D: Vergleich Haus 18 mit ohne Haus. Der Locus Cyp4a14 ist in A und B nicht abgebildet, er liegt bei $p<0{,}05$ und einer Normierten Differenz von 0,6.

3.3.6 Vergleich zu Bl6 Mäusen

Neben den Wildmäusen aus dem Freilaufexperiment wurden im Labor auch ChIP qPCR Daten von acht Bl6 Mäusen erhoben. Diese stammen aus einem anderen Experiment, wurden im Käfig aufgezogen und lebten unter anderen Umweltbedingungen als die Wildmäuse. Es handelt sich ebenfalls um adulte Männchen. Die Interpretation eines Vergleichs dieser beiden Mäusegruppen ist beschränkt, weil wir nicht zwischen genetischen und experimentellen Einflüssen unterscheiden können. Allerdings zeigen sich in der Analyse einige interessante Details:

Eine Hauptkomponentenanalyse der Datensätze der H3K4me3 und H3K27ac Anreicherungen in den Wildmäusen aus Haus 11 und 18 und den Bl6 Mäusen wurde wie in Abschnitt 2.5.3 beschrieben durchgeführt. Die ersten beiden Hauptkomponenten sind jeweils in Abbildung 3.26 aufgezeigt. In Abbildung 3.26 A wird deutlich, dass im Datensatz der H3K4me3 Anreicherung die ersten beiden Hauptkomponenten, die gemeinsam 47,6% der Varianz erklären, benötigt werden, um die Wildmäuse von den Bl6 Mäusen fehlerfrei zu trennen. Bei Verwendung der H3K27ac Anreicherung genügt es, die erste Hauptkomponente zu betrachten, die 44,1% der Varianz erklärt, um die Mäuse aus dem Freilaufexperiment von denen aus dem Käfig abzugrenzen (Abbildung 3.26 B). Das bedeutet, dass die Unterschiede zwischen den beiden Mausgruppen bei der H3K27ac Anreicherung stärker ausgeprägt sind als bei der H3K4me3 Anreicherung.

In den ersten beiden Hauptkomponenten der Transformation des H3K4me3 Datensatzes fallen insbesondere die Loci Nr3c1 und Igfbp2 mit hohen absoluten Werten auf. Im Datensatz der H3K27ac Anreicherung treten die Gene Pparg und wieder Igfbp2 durch hohe absolute Werte in der ersten Hauptkomponente hervor (nicht gezeigt).

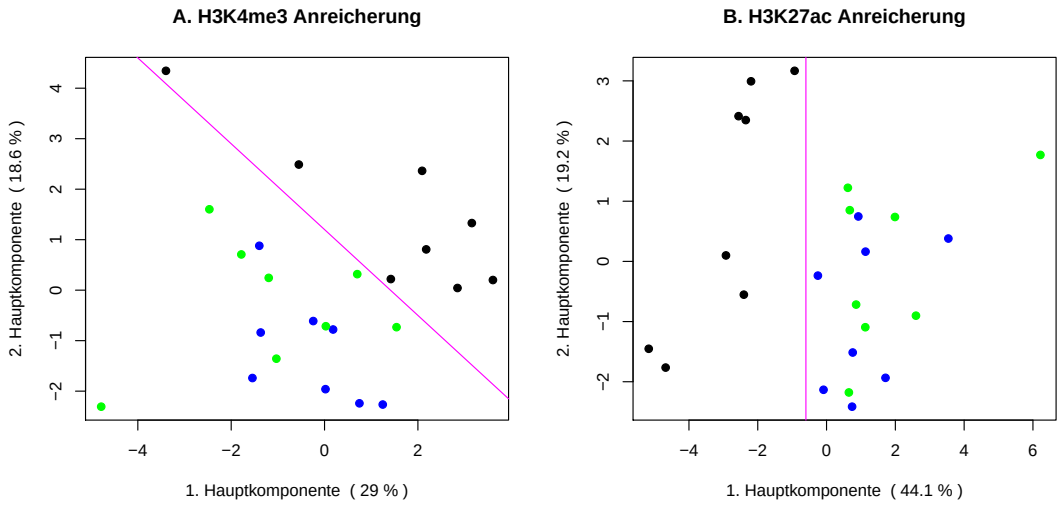

Abb. 3.26: Ergebnis der Hauptkomponentenanalyse der 16 in Häusern lebenden Wildmäuse und der acht Bl6 Mäuse. In blau die Wildmäuse aus Haus 11, in grün die aus Haus 18 und in schwarz die Bl6 Mäuse. Die magentafarbene Linie zeigt eine mögliche Trennung.

4 Diskussion

Epigenetische Mechanismen regulieren die Expression von Genen. Eine abnormale Regulation kann zu schwerwiegenden Krankheiten wie Krebs führen [37]. Die meisten Untersuchungen zur Epigenetik beruhen auf Messungen an Zelllinien oder ingezüchteten Labortieren. Um Epigenetik an echtem Gewebe auf dem Hintergrund natürlicher genetischer Variabilität zu untersuchen, haben wir Histonmodifikationen in Lebergewebe von Wildmäusen gemessen. Dabei betrachten wir genomweite ChIP-Seq Daten von zwei Männchen aus unserer Population und ChIP qPCR Datensätze von 24 Wildmäusen. Untersuchungen zu epigenetischen Auswirkungen von Stress werden fast auschließlich im Gehirn nach von außen induziertem Stress durchgeführt [60]. In dem Freilaufexperiment aus dem die mit ChIP qPCR analysierten Wildmäuse stammen, wurde der untersuchte Phänotyp, hoher oder niedriger sozialer Status, unter natürlichen Bedingungen erzeugt.

In der genomweiten ChIP-Seq Analyse der Wildmäuse ZH1 und ZH6 suchten wir nach epigenetische Unterschieden. Wir nehmen an, dass diese zum größten Teil auf Grund genetischer Variabilität entstehen. Der Anteil polymorpher Nukleotide pro Position (s) kann auch in ChIP-Seq Daten erforscht werden. Da wir Reads nach einer Immunopräzipitation gegen H3K4me3 auf Polymorphismen untersuchen, berechnen wir s in Promotorregionen von aktiv transkribierten Genen. Unser Ergebnis (s=0,0021) ist überraschenderweise identisch mit dem veröffentlichten Wert [20] [1]. Diese Autoren bestimmten die Anzahl SNPs in ca. 600 Bp langen Bereichen in Promotorregionen oder Introns. Da wir mit Mäusen aus der gleichen Population arbeiten und SNPs in Promotorregionen messen, ist die Übereinstimmung zwischen unserem Ergebnis und dem in der Literatur angegebenen Wert nachvollziehbar [1].

Wir erwarten durch s=0,0021 circa alle 500 Bp einen Polymorphismus. Somit können die SNPs nicht die 25 bzw. 29 Prozent der Reads erklären, die `bowtie` nicht auf das Referenzmausgenom platzieren konnte (Tabelle 3.1). Wahrscheinlich führten Messfehler des Sequenziergeräts zu falsch bestimmten Nukleotiden und damit unmappbaren Reads.

In den ersten ChIP-Seq Experimenten wurden keine Kontrollen mitgeführt [56]. Zhang et el. zeigten 2008 jedoch, dass offenes Chromatin zu einer höheren Coverage führt [67]. Eine Chromatin Input Kontrolle kann für den Chromatinzustand korrigieren [6]. Xu et al. stellten 2010 einen Algorithmus vor, der Peaks in ChIP-Seq Daten findet, wenn eine Blindprobe, die durch einen unspezifischen Antikörper oder Input DNA gewonnen wird, angegeben wird. Nur 35,6% der mit diesem Algorithmus bestimmten Peaks liegen in TSS. Ohne negative Kontrolle können wir keinen der vorhandenen Algorithmen zum *de novo* Finden von Peaks anwenden. Wir definieren unsere Peaks deshalb in einem Fenster von 2000 Bp symmetrisch um bekannte TSS.

Eine geeignete Normalisierung von ChIP-Seq Daten ist entscheidend für die richtige Interpretation einer quantitativen Analyse. Durch falsch gewählte Faktoren könnten vorhandene Unterschiede nivelliert werden oder künstliche entstehen. In einer Gegenüberstellung verschiedener Normalisierungsfaktoren ergibt sich für unsere Daten, dass am wenigsten Verzerrungen entstehen, wenn die Datensätze jeweils auf die Anzahl aller Reads, die in TSS Peaks landen, normalisiert werden. Zur quantitativen Analyse ist neben einer geeigneten Normalisierung ein Differenzmaß, das die Unterschiede bemisst, wichtig. In dieser Arbeit wird die Normierte Differenz verwendet, die von Nix et al. 2008 bereits zur Detektion von ChIP-Seq Peaks verwendet wurde [52].

ZH1 und ZH6 enthalten sieben exklusive Peaks, die in Abbildung 3.6 durch ihre großen Peakgrößen auffallen. Die größte exklusive Markierung von ZH1 gehört zu dem Gen, das für das Protein Peptiddeformylase (Pdf) kodiert. Pdf entfernt in Mitochondrien die Aldehydgruppe (Formylgruppe) des Methionins am Beginn von neu translatierten Peptiden. Das Protein wird im Zellkern kodiert und in allen Geweben exprimiert [59]. Im Custom Track des UCSC Browser ist erkenntlich, dass die exklusive H3K4me3 Markierung in ZH1 mit einer CpG-Insel überlappt (grün in Abbildung 3.7).

Weitere exklusive ZH1 Markierungen liegen an Cytochrome P450 4A14 (Cyp4a14) und Stearoyl-CoA Desaturase (Scd1) vor. Cyp4a14 kodiert für ein Enzym, das in Lebermikrosomen am Abbau von Fettsäuren beteiligt ist [59]. Cyp4a14 wird hauptsächlich in der Leber, aber auch in der Niere und in der Milz exprimiert [23]. Scd1 ist ebenfalls ein zentrales Enzym des Fettsäuremetabolismus. Cyp4a14 und Scd1 werden vom PPAR Signaltransduktionsweg gesteuert: Cyp4a14 als Teil der Fettsäure-Oxidation und Scd1 als Teil der Lipogenese.

Der größte exklusive Peak von ZH6 gehört zu Scara5. Das Gen kodiert für einen Ferritin-Rezeptor, der eine Belieferung mit Eisen regelt, die unabhängig von Transferrin ist [46]. Das Gen wird im Testis, in der Trachea, in der Lunge, in der Blase, im Dünndarm und insbesondere in Epithelzellen, die mit Schleimhäuten in Verbindung stehen, exprimiert [31]. Eine Expression in der Leber ist bisher nicht bekannt.

Interessant ist nun, dass die in ZH1 exklusiv markierten Gene, Pdf, Cyp4a14 und Scd1, für Proteine kodieren, die in ihrem aktiven Zentrum Eisen binden. Der größte exklusive Peak von ZH6 gehört zu dem Eisenrezeptor Scara5, der den Eisentransport reguliert und dessen Expression in der Leber bisher noch nicht beschrieben wurde. Es wäre interessant als nächstes die Expression von Scara5 in Leber direkt zu untersuchen. Möglich ist, dass exklusive Peaks eventuell dort auftauchen, wo durch Enzymfamilien eine Übernahme der Funktion durch ein anderes Enzym gewährleistet ist. Zum Beispiel hat Cyp414 einen exklusiven Peak in ZH1, aber Cyp4a32, auch eine Häm-Thiolat Monooxygenase, ist in beiden Datensätzen ähnlich markiert.

Neben exklusiven Markierungen betrachten wir auch quantitative Unterschiede in der H3K4me3 Anreicherung. Bei den Genen, die in ZH1 stärker markiert sind als in ZH6, ist der PPAR Signaltransduktionsweg signifikant angereichert (siehe Tabelle 3.6). Diese Anreicherung ist auch zu beobachten, wenn die exklusiven ZH1 Markierungen analysiert werden. PPAR steht für *Per-*

oxisome proliferator-activated receptors und es gibt drei Untergruppen dieser Rezeptoren: α, β/δ und γ. Es handelt sich um nukleare Hormonrezeptoren, die von Fettsäuren und deren Derivaten aktiviert werden. Eine Aufgabe von PPARα ist der Metabolimus von Lipiden durch die Regulierung der daran beteiligten Gene. PPARβ/δ spielt eine Rolle in der Lipid-Oxidation und in der Proliferation von Fettzellen. PPARγ unterstützt die Differenzierung von Fettzellen und ist ein wichtiger Regulator in Glukosehomöostase [36].

Die größten Unterschiede zwischen ZH1 und ZH6 (Tabellen 3.4 und 3.5) liegen fast ausschließlich an Genen, die bei den im Plöner Labor gehaltenen Wildmäusen bereits durch Variation der Genkopiezahlen aufgefallen sind [Angelika Börsch-Haubold, persönliche Mitteilung] oder die andere genetische Besonderheiten wie CpG-Inseln und bekannte genetische Unterschiede in Labormaus-Stämmen (RFLP, PCRV) aufweisen. Auch KEGG Analysen der unterschiedlich markierten Gene ergeben stets nur eine geringe Anzahl an Genen, die differentiell markiert und in einer funktionellen Einheit vorliegen (Tabellen 3.6 und 3.7). Deshalb sind die Unterschiede zwischen den Wildmäusen ZH1 und ZH6 punktuell und betreffen keine ganze funktionelle Einheit.

Durch die ChIP-Seq Analyse zweier Wildmäuse unserer Population haben wir einen Einblick in die epigenetische Variabilität der H3K4me3 Markierung erhalten und Gene detektiert, die auffällig sein können, ohne dass ein bestimmter Phänotyp vorliegt. Diese Gene liegen meist in metabolischen Netzwerken, was zu erwarten ist, da wir mit Lebergewebe arbeiten.

Große Unterschiede zwischen den H3K4me3 Markierungen der Wildmäuse ZH1 und ZH6 überlappen im Genom oft mit CpG-Inseln, unabhängig davon, ob exklusive Peaks oder quantitative Abweichungen betrachtet werden. Ein großer Anteil nicht exprimierter Gene wies in einer Studien zu Histonmodifikationen und DNA Methylierung in Darmkrebs H3K4me3 Markierungen auf [2]. An 95 bis 99% dieser Gene wurden gleichzeitig eine CpG-Insel und weitere Histonmarkierungen, die mit der Hemmung von Genexpression assoziiert werden, gefunden [2]. Bereits 2007 stellten Mikkelsen et al. fest, dass in embryonalen Stammzellen 99% der Promotoren mit einem hohen CpG-Gehalt eine signifikante Anreicherung der Histonmodifikation H3K4me3 aufweisen. Obwohl nicht alle diese Promotoren aktiv waren, zeigte sich eine starke Korrelation zwischen der Intensität der Markierung und der Genexpression [48]. Promotoren mit einem niedrigen CpG-Gehalt tragen nur dann eine H3K4me3 Markierung, wenn sie exprimiert werden.

Deshalb bedeuten unsere Markierungsunterschiede zwischen den Wildmäusen ZH1 und ZH6 nicht unbedingt, dass die Gene differentiell exprimiert werden. Das gleichzeitige Messen einer repressiven Histonmarkierung oder von mRNA wären die nächsten Schritte, das aufzuklären.

Wir verwendeten die qPCR, um eine quantitative Aussage über die Anreicherung zweier Histonmodifikationen (H3K4me3 und H3K27ac) zu treffen. Da wir mit Wildmausleber arbeiteten, validierten wir zunächst die Methode bevor wir die Daten interpretierten.

Haring et al. stellten 2007 ein robustes ChIP Protokoll vor und empfehlen darin verschiedene Normalisierungen und Kontrollen [19]. Wir führten den dort vorgeschlagenen Primertest durch (siehe Abschnitt 3.2.2) und banden eine Kontrollmessung in unsere qPCR Reaktion ein. Außerdem verwendeten wir zur Normalisierung unserer qPCR Daten die %Input Methode und berechneten keine Anreicherung über der Blindprobe (Rabbit IgG), die starken Schwankungen unterliegen kann. Durch die Mittelwertnormalisierung pro Datensatz machten wir diese miteinander vergleichbar.

Der verwendete Primer Acox2 lieferte sowohl in mm9 als auch in mm10 in der *in silico* PCR ein eindeutiges Amplifikationsprodukt. Wir fanden jedoch sowohl in unseren Wildmäusen als auch in Bl6 Mäusen ein zweites Amplifikat (Abbildung 3.15). Das ist vermutlich durch eine Duplikation verursacht, die in der aktuellen Genomversion fehlt. Die Schatten im niedrigmolekularen Bereich der Ergebnisse der Gelelektrophorese (Abbildung 3.14) könnten RNA entsprechen, da wir in der ChIP Präparation keinen RNAse Verdau durchführen.

Die Ct-Werte der Input Proben (Abbildung 3.18) schwankten stark zwischen den verwendeten Primern: Pparg ergab beispielsweise sehr hohe Input Ct-Werte. Dabei spielte die Länge des Amplicon-Fensters keine Rolle. In ChIP-Seq Datensätzen verschiedener Wildmäuse gab es immer wieder Bereiche einer Länge von ca 150 bis 200 Bp in der Nähe von TSS, in denen keine Reads platziert wurden (Abbildung 3.10). Durch die aktive Transkription des Gens könnte an diesen Positionen die DNA ohne bindende Histone vorliegen und die Immunopräzipitation würde diese Bereiche nicht erfassen. Eine weitere Möglichkeit ist, dass die Nuclease während der Chromatinpräparation genau in diesem Bereich schneidet. Im Amplicon-Fenster von Pparg liegt so ein Bereich, der dadurch die hohen Input Ct-Werte erklären könnte. Insgesamt könnten die Input Ct-Werte unter anderem davon abhängig, wo die Nuklease schneidet und ob die DNA an Histone gebunden vorliegt.

Ppara ist in den Mäusen ohne Hauszugehörigkeit stärker mit H3K4me3 markiert als in den anderen Mäusen und zeigt in der statistischen Analyse einen signifikaten Unterschied (Tabelle 3.10). Es gibt auch einen signifikanten Unterschied zwischen den H3K4me3 Markierungen am Locus CD36 zwischen den Gruppen. Da dieser Locus insgesamt nur sehr schwach mit H3K4me3 markiert ist, vernachlässigen wir diesen. Cyp4a14 ist ein sehr variabler Locus, der in den Mäusen ohne Hauszugehörigkeit kompakter vorliegt als in den Vergleichsgruppen. Die Datensätze aus Haus 18 weisen insgesamt eine höhere Varianz auf, sodass nur fünf Gene einen statistischen Unterschied zwischen diesen Mäusen und denen ohne Hauszugehörigkeit zeigen, wohingegen die Markierungen der Mäuse aus Haus 11 acht signifikant verschiedene Loci zeigen. Die H3K4me3 Anreicherungen der sozialisierten Mäuse weisen nur einen signifikaten Unterschied auf (Slc27a5). Das Gen Nr3c1 weist eine Besonderheit auf: während an anderen Loci meist ein Unterschied zwischen den sozialisierten Mäusen und denen ohne Hauszugehörigkeit zu erkennen ist, haben hier die Mäuse aus Haus 11 die niedrigste Markierung, die Mäuse aus Haus 18 sind höher markiert, aber wegen der großen Streuung der Daten gibt es kein signifikantes Signal und im Vergleich der Markierung an diesem Locus zwischen Haus 11 und ohne Hauszugehörigkeit detektieren wir den größten Unterschied des H3K4me3 Datensatzes.

Nach einer PCA des H3K4me3 Datensatzes können die Mäuse aus den Häusern 11 und 18 entlang der ersten Hauptkomponente mit zwei Ausnahmen von den Mäusen ohne Hauszugehörigkeit getrennt werden (Abbildung 3.21 A). In der ersten Hauptkomponente sind Ppara, Cyp4a14 und Sqle wichtige Faktoren (Abbildung 3.23 A), die teilweise bereits in der statistischen Analyse signifikante Unterschiede zeigen.

Neben H3K4me3 haben wir auch die Histonmodifikation H3K27ac untersucht. Die beiden korrelieren an den 15 qPCR Loci (Korrelationskoeffizient zwischen 0,51 und 0,68), obwohl die Anreicherung über den Input sehr verschieden war (Abbildung 3.17). Histonmodifikationen sind dynamisch und unterliegen unterschiedlichen Turnoverraten [64]. Ein schwaches ChIP Signal, wie bei uns bei H3K27ac, könnte auf eine hohe Turnoverrate dieser Modifikation hindeuten [19].

Im H3K27ac Datensatz waren trotz der Korrelation zu den H3K4me3 Daten andere Gene differentiell markiert. CD36 ist hier insgesamt stärker markiert und in den Mäusen ohne Hauszugehörigkeit stärker als in den anderen. An den Loci Fasn und Plin5 hingegen sind die sozialisierten Mäuse signifikant höher markiert. Das stärkste Signal hat der Vergleich am Locus Fasn zwischen den Mäusen aus Haus 11 und denen ohne Hauszugehörigkeit. Die einzige unterschiedliche Markierung zwischen den Mäusen aus Haus 11 und Haus 18 liegt am Gen Nr3c1. Interessanterweise liegt bei den H3K27ac Daten aus Haus 11, wie bei den H3K4me3 Daten, eine niedrigere Markierung vor als bei den anderen beiden Gruppen.

Nr3c1 kodiert für den Glucocorticoidrezeptor (GR), dessen Cytosin-Methylierungsmuster im Hippocampus bei Menschen und Ratten durch Stress beeinflussbar ist (zusammengefasst in [60]). Wir untersuchen keine Neuronen, sehen trotzdem Unterschiede in der H3K4me3 und der H3K27ac Markierung des GR bei mehr oder weniger gestressten Mäusen.

Die PCA der H3K27ac Anreicherungen führt zu einer Trennung der Mäuse ohne Hauszugehörigkeit von den sozialisierten Mäusen entlang der zweiten Hauptkomponente (Abbildung 3.21 B). CD36, Fasn und Plin5 sind dort die entscheidenden Variablen (Abbildung 3.23 B), die auch schon durch die statistischen Tests identifiziert wurden.

Crowcroft beobachtete, dass Mäuse mit einem niedrigen sozialen Status ungeregelmäßig und zu ungewöhnlichen Zeiten fressen müssen, da sie von den dominanten Männchen gejagt werden [11]. Das veränderte Fressverhalten könnte zu Unterschieden in der Expression von metabolischen Genen führen. Die in der qPCR untersuchten Loci (Tabelle 2.6) stehen alle im Zusammenhang mit Fettmetabolismus. Wir detektieren in der ChIP qPCR Analyse Unterschiede in epigenetischen Markierungen an diesen Loci zwischen den Gruppen der sozialisierten Mäuse und den Mäusen ohne Hauszugehörigkeit und damit einem niedrigen sozialen Status, was auf das veränderte Fressverhalten dieser Mäuse zurückzuführen sein könnte. Der Vergleich zwischen sozialisierten Mäusen aus unterschiedlichen Nestern ergibt dagegen keine Unterschiede in den epigenetischen Markierungen.

Cyp4a14 ist sowohl in den ChIP-Seq Daten als auch in den ChIP qPCR unterschiedlich mit H3K4me3 markiert. Im ChIP-Seq Vergleich der Wildmäuse ZH1 und ZH6 zeigt es sich als ex-

klusive Markierung bei ZH1. In den ChIP qPCR Untersuchungen der 24 Wildmäuse hat die Cyp4a14 Markierung eine große Streuung (Abbildung 3.19): Die Werte der Mäuse ohne Hauszugehörigkeit clustern bei niedrigen Anreicherungswerten. Im selben Wertebereich liegen auch eine Maus aus Haus 11 und drei aus Haus 18. Alle weiteren Wildmäuse (13 Stück) weisen im ChIP qPCR Experiment höhere H3K4me3 Markierungen an Cyp4a14 auf.

Wir vermuten daher, dass an Cyp4a14 ein epigenetischer Polymorphismus in unserer Population vorliegt, der unabhängig vom beobachteten Phänotyp (sozialisiert oder ohne Hauszugehörigkeit) sein könnte. Die grundlegende Variabilität am Locus Cyp4a14 könnte zufällig auf die untersuchten Gruppen so aufgeteilt sein, dass anscheinend signifikante Unterschiede auftreten.

An diesem Gen wären weitergehende Untersuchungen interessant. Zum Einen könnten epigenetische Studien mit anderen Geweben klären, ob ein epigenetischer Polymorphismus vorliegt. Zum Anderen könnte man erforschen, ob und inwieweit genetische Unterschiede epigenetische Marker beeinflussen. Dazu müssten größere Bereiche um den Lokus herum sequenziert und vorhandene SNPs mit epigenetischen Phänotypen verglichen werden. Außerdem schließen wir aus den Beobachtungen zu Cyp4a14, dass bei phänotypisch motivierten Studien an Wildmäusen bedacht werden sollte, dass Kandidatengene in der Population prinzipiell unterschiedlich markiert vorliegen könnten.

In dieser Arbeit wurde am Beispiel von zwei Wildmäusen (ZH1 und ZH6) gezeigt, inwieweit epigenetische Muster zwischen Individuen variieren, deren Umwelt vergleichbar war. Es wäre wichtig, nicht nur zwei Mäuse miteinander zu vergleichen, sondern eine größere Anzahl, um die tatsächliche Variabilität besser abzuschätzen. Ebenso wäre es interessant einen Vergleich mit der Referenzmaus Bl6 zu haben, um zu sehen, inwieweit genomische Vielfalt die epigenetiche Vielfalt beeinflusst.

Nach Einwirkung von reellen unterschiedlichen Umweltbedingungen in Form ungleicher Sozialisation der Mäuse haben sich die epigenetischen Muster so verschoben, dass die Gruppen durch statistische Analysen getrennt werden konnten. Es wäre interessant von diesen Mäuse genomweite ChIP-Seq Daten zu erheben, um die Änderung unabhängig von der Vorauswahl der 15 metabolischen Loci untersuchen zu können.

Abkürzungsverzeichnis

Bp	Basenpaare
ChIP	Chromatin Immunopräzipitation
ChIP-Seq	Chromatin Immunopräzipitation gefolgt von massiv-paralleler Sequenzierung
ChIP qPCR	Chromatin Immunopräzipitation gefolgt quantitativer Polymeraseketten-reaktion
CpG	Cytosin-Guanin Dinukleotid
CpG-Gehalt	Anteil an CpG Dinukleotiden
CNV	Copy number variation (Variation der Genkopiezahl)
DNA	Deoxyribonucleic acid (Desoxyribonukleinsäure)
DNMT	DNA Methyltransferase
GC-Gehalt	Anteil der Basen Guanin und Cytosin
H3K4me2	Dimethylierung des Lysin 4 am Histon H3
H3K4me3	Trimethylierung des Lysin 4 am Histon H3
H3K27ac	Acetylierung des Lysin 27 am Histon H3
H3K27me3	Trimethylierung des Lysin 27 am Histon H3
HCP	High CpG-content promoter (Promotoren mit einem hohen CpG-Gehalt)
ID	Identifikationsnummer
LCP	Low CpG-content promoter (Promotoren mit einem niedrigen CpG-Gehalt)
KEGG	Kyoto encyclopedia of genes and genomes (Kyoto Enzyklopädie von Genen und Genomen)
MGI	Mouse Genome Informatics (Mausgenomdatenbank)
MPI	Max-Planck-Institut
MW	Mittelwert

NCBI	National Center for Biotechnology Information (Nationales Zentrum für Biotechnologie)
PBS	Phosphate buffered saline (Phosphatgepufferte Salzlösung)
PCA	Principal component analysis (Hauptkomponentenanalyse)
PCR	Polymerase chain reaction (Polymerasekettenreaktion)
PCRV	Polymerase chain reaction variation (Polymerasekettenreaktionvariation)
qPCR	Quantitative polymerase chain reaction (quantitative Polymerasekettenreaktion)
RFPL	Restriction fragment length polymorphism (Restriktionsfragmentlängenpolymorphismen)
rpm	Revolutions per minute (Umdrehungen pro Minute)
RT-PCR	Reverse transcription polymerase chain reaction (Reverse Transkriptase-Polymerasekettenreaktion)
sd	Standard deviation (Standardabweichung)
SNP	Single nucleotide polymorphism (Einzelnukleotidpolymorphismus)
TSS	Transcription start site (Transkriptionsstartpunkt)
UniProt	Universal protein resource (Proteindatenbank)
UCSC	University of California, Santa Cruz (Universität Californien, Santa Cruz)
WebGestalt	WEB-based GEne SeT AnaLysis Toolkit (Programm zur Berechnung von Anreicherungen in funktionellen Einheiten)

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] J. F. Baines and B. Harr. Reduced x-linked diversity in derived populations of house mice. *Genetics*, 175(4):1911–1921, Apr. 2007. PMID: 17287527.

[2] D. Balasubramanian, B. Akhtar-Zaidi, L. Song, C. F. Bartels, M. Veigl, L. Beard, L. Myeroff, K. Guda, J. Lutterbaugh, J. Willis, G. E. Crawford, S. D. Markowitz, and P. C. Scacheri. H3K4me3 inversely correlates with DNA methylation at a large class of non-CpG-island-containing start sites. *Genome Medicine*, 4(5):47, May 2012. PMID: 22640407.

[3] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007.

[4] A. B. Brinkman, S. W. C. Pennings, G. G. Braliou, L. E. G. Rietveld, and H. G. Stunnenberg. DNA methylation immediately adjacent to active histone marking does not silence transcription. *Nucleic Acids Research*, 35(3):801–811, Feb. 2007. PMID: 17202157 PMCID: PMC1807972.

[5] D. Brutlag, C. Schlehuber, and J. Bonner. Properties of formaldehyde-treated nucleohistone. *Biochemistry*, 8(8):3214–3218, Aug. 1969. PMID: 5809221.

[6] Y. Chen, N. Negre, Q. Li, J. O. Mieczkowska, M. Slattery, T. Liu, Y. Zhang, T.-K. Kim, H. H. He, J. Zieba, Y. Ruan, P. J. Bickel, R. M. Myers, B. J. Wold, K. P. White, J. D. Lieb, and X. S. Liu. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, 9(6):609–614, June 2012.

[7] X. Cheng and R. M. Blumenthal. Coordinated chromatin control: Structural and functional linkage of DNA and histone methylation. *Biochemistry*, 49(14):2999–3008, Apr. 2010.

[8] A. T. Chinwalla, L. L. Cook, K. D. Delehaunty, G. A. Fewell, L. A. Fulton, R. S. Fulton, T. A. Graves, L. W. Hillier, E. R. Mardis, J. D. McPherson, and et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, Dec. 2002.

[9] V. K. Cortessis, D. C. Thomas, A. J. Levine, C. V. Breton, T. M. Mack, K. D. Siegmund, R. W. Haile, and P. W. Laird. Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships. *Human Genetics*, 131(10):1565–1589, Oct. 2012.

[10] P. Crowcroft. Territoriality in wild house mice, mus musculus l. *Journal of Mammalogy*, 36(2):299–301, May 1955. ArticleType: research-article / Full publication date: May, 1955 / Copyright © 1955 American Society of Mammalogists.

[11] P. Crowcroft and F. P. Rowe. Social organization and territorial behaviour in the. *Proceedings of the Zoological Society of London*, 140(3):517–531, 1963.

[12] C. B. Cunningham, J. S. Ruff, K. Chase, W. K. Potts, and D. R. Carrier. Competitive ability in male house mice (mus musculus): Genetic influences. *Behavior Genetics*, 43(2):151–160, Mar. 2013.

[13] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, Aug. 2011. PMID: 21653522.

[14] D. M. Dietz, Q. LaPlant, E. L. Watts, G. E. Hodes, S. J. Russo, J. Feng, R. S. Oosting, V. Vialou, and E. J. Nestler. Paternal transmission of stress-induced pathologies. *Biological Psychiatry*, 70(5):408–414, Sept. 2011.

[15] L. Fahrmeir. *Statistik: der Weg zur Datenanalyse*. Springer, Berlin [u.a.], 2011.

[16] T. S. Furey. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 13(12):840–852, Dec. 2012.

[17] M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, July 2007.

[18] S. Gupta, S. Y. Kim, S. Artis, D. L. Molfese, A. Schumacher, J. D. Sweatt, R. E. Paylor, and F. D. Lubin. Histone methylation regulates memory formation. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 30(10):3589–3599, Mar. 2010. PMID: 20219993.

[19] M. Haring, S. Offermann, T. Danker, I. Horst, C. Peterhansel, and M. Stam. Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. *Plant Methods*, 3(1):11, Sept. 2007. PMID: 17892552.

[20] B. Harr. Genomic islands of differentiation between house mouse subspecies. *Genome Research*, 16(6):730–737, June 2006. PMID: 16687734 PMCID: PMC1473184.

[21] T. R. Hebbes, A. L. Clayton, A. W. Thorne, and C. Crane-Robinson. Core histone hyperacetylation co-maps with generalized DNase i sensitivity in the chicken beta-globin chromosomal domain. *The EMBO Journal*, 13(8):1823–1830, Apr. 1994. PMID: 8168481 PMCID: PMC395022.

[22] C. A. Heid, J. Stevens, K. J. Livak, and P. M. Williams. Real time quantitative PCR. *Genome Research*, 6(10):986–994, Oct. 1996. PMID: 8908518.

[23] Y. M. Heng, C. S. Kuo, P. S. Jones, R. Savory, R. M. Schulz, S. R. Tomlinson, T. J. Gray, and D. R. Bell. A novel murine p-450 gene, cyp4a14, is part of a cluster of cyp4a and cyp4b, but not of CYP4F, genes in mouse and humans. *The Biochemical journal*, 325 (Pt 3):741–749, Aug. 1997. PMID: 9271096.

[24] H.-S. Huang, J. A. Allen, A. M. Mabb, I. F. King, J. Miriyala, B. Taylor-Blake, N. Sciaky, J. W. Dutton, H.-M. Lee, X. Chen, J. Jin, A. S. Bridges, M. J. Zylka, B. L. Roth, and B. D. Philpot. Topoisomerase inhibitors unsilence the dormant allele of ube3a in neurons. *Nature*, 481(7380):185–189, Jan. 2012.

[25] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, Jan. 2002. PMID: 11752248.

[26] S. Ihle, I. Ravaoarimanana, M. Thomas, and D. Tautz. An analysis of signatures of selective sweeps in natural populations of the house mouse. *Molecular biology and evolution*, 23(4):790–797, Apr. 2006. PMID: 16421176.

[27] R. S. Illingworth and A. P. Bird. CpG islands – 'A rough guide'. *FEBS Letters*, 583(11):1713–1720, June 2009.

[28] Y. V. Ilyin and G. P. Georgiev. Heterogeneity of deoxynucleoprotein particles as evidencec by ultracentrifugation of cesium chloride density gradient. *Journal of molecular biology*, 41(2):299–303, Apr. 1969. PMID: 5805452.

[29] V. Jackson and R. Chalkley. A new method for the isolation of replicative chromatin: Selective deposition of histone on both new and old DNA. *Cell*, 23(1):121–134, Jan. 1981.

[30] V. Jackson and R. Chalkley. Use of whole-cell fixation to visualize replicating and maturing simian virus 40: identification of new viral gene product. *Proceedings of the National Academy of Sciences of the United States of America*, 78(10):6081–6085, Oct. 1981. PMID: 6273846 PMCID: PMC348981.

[31] Y. Jiang, P. Oliver, K. E. Davies, and N. Platt. Identification and characterization of murine SCARA5, a novel class A scavenger receptor that is expressed by populations of epithelial cells. *The Journal of biological chemistry*, 281(17):11834–11845, Apr. 2006. PMID: 16407294.

[32] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, June 2007. PMID: 17540862.

[33] P. A. Jones. The DNA methylation paradox. *Trends in Genetics*, 15(1):34–37, Jan. 1999.

[34] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, Jan. 2000. PMID: 10592173.

[35] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(Database issue):D109–114, Jan. 2012. PMID: 22080510.

[36] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*, Nov. 2013. PMID: 24214961.

[37] R. Kanwal and S. Gupta. Epigenetic modifications in cancer. *Clinical genetics*, 81(4):303–311, Apr. 2012. PMID: 22082348.

[38] R. Karlić, H.-R. Chung, J. Lasserre, K. Vlahoviček, and M. Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931, Feb. 2010. PMID: 20133639.

[39] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, June 2002. PMID: 12045153.

[40] M.-H. Kuo and C. Allis. In vivo cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment. *Methods*, 19(3):425–433, Nov. 1999.

[41] E. S. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197, Feb. 2011.

[42] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, and et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001. PMID: 11237011.

[43] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, Mar. 2009. PMID: 19261174.

[44] F. Larsen, G. Gundersen, R. Lopez, and H. Prydz. CpG islands as gene markers in the human genome. *Genomics*, 13(4):1095–1107, Aug. 1992. PMID: 1505946.

[45] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug. 2009. PMID: 19505943.

[46] J. Y. Li, N. Paragas, R. M. Ned, A. Qiu, M. Viltard, T. Leete, I. R. Drexler, X. Chen, S. Sanna-Cherchi, F. Mohammed, D. Williams, C. S. Lin, K. M. Schmidt-Ott, N. C. Andrews, and J. Barasch. Scara5 is a ferritin receptor mediating non-transferrin iron delivery. *Developmental cell*, 16(1):35–46, Jan. 2009. PMID: 19154717.

[47] P. O. McGowan and M. Szyf. Environmental epigenomics: understanding the effects of parental care on the epigenome. *Essays in Biochemistry*, 48(1):275–287, Sept. 2010.

[48] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, Aug. 2007.

[49] I. Montero, M. Teschke, and D. Tautz. Paternal imprinting of mating preferences between natural populations of house mice (mus musculus domesticus). *Molecular ecology*, 22(9):2549–2562, May 2013. PMID: 23506395.

[50] K. Nagaki, P. B. Talbert, C. X. Zhong, R. K. Dawe, S. Henikoff, and J. Jiang. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of arabidopsis thaliana centromeres. *Genetics*, 163(3):1221–1225, Mar. 2003. PMID: 12663558 PMCID: PMC1462492.

[51] H. H. Ng, F. Robert, R. A. Young, and K. Struhl. Targeted recruitment of set1 histone methylase by elongating pol II provides a localized mark and memory of recent transcriptional activity. *Molecular Cell*, 11(3):709–719, Mar. 2003.

[52] D. A. Nix, S. J. Courdy, and K. M. Boucher. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*, 9(1):523, Dec. 2008. PMID: 19061503.

[53] M. W. Pfaffl, G. W. Horgan, and L. Dempfle. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Research*, 30(9):e36–e36, May 2002. PMID: 11972351.

[54] A. Radonić, S. Thulke, I. M. Mackay, O. Landt, W. Siegert, and A. Nitsche. Guideline to reference gene selection for quantitative real-time PCR. *Biochemical and Biophysical Research Communications*, 313(4):856–862, Jan. 2004.

[55] A. G. Robertson, M. Bilenky, A. Tam, Y. Zhao, T. Zeng, N. Thiessen, T. Cezard, A. P. Fejes, E. D. Wederell, R. Cullum, G. Euskirchen, M. Krzywinski, I. Birol, M. Snyder, P. A. Hoodless, M. Hirst, M. A. Marra, and S. J. M. Jones. Genome-wide relationship between histone h3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Research*, 18(12):1906–1917, Dec. 2008. PMID: 18787082.

[56] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8):651–657, Aug. 2007.

[57] H. Santos-Rosa, R. Schneider, A. J. Bannister, J. Sherriff, B. E. Bernstein, N. C. T. Emre, S. L. Schreiber, J. Mellor, and T. Kouzarides. Active genes are tri-methylated at k4 of histone h3. *Nature*, 419(6905):407–411, Sept. 2002.

[58] S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5):1412–1417, Jan. 2006. PMID: 16432200.

[59] The UniProt Consortium. Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Research*, 41(D1):D43–D47, Nov. 2012.

[60] A. F. Trollope, M. Gutièrrez-Mecinas, K. R. Mifsud, A. Collins, E. A. Saunderson, and J. M. Reul. Stress, epigenetic control of gene expression and memory formation. *Experimental Neurology*, 233(1):3–11, Jan. 2012.

[61] J. Vandesompele, K. D. Preter, F. Pattyn, B. Poppe, N. V. Roy, A. D. Paepe, and F. Speleman. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, 3(7):research0034, June 2002. PMID: 12184808.

[62] J. Wang, D. Duncan, Z. Shi, and B. Zhang. WEB-based GEne SeT AnaLysis toolkit (WebGestalt): update 2013. *Nucleic Acids Research*, 41(W1):W77–W83, May 2013.

[63] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903, July 2008.

[64] J. H. Waterborg. Dynamics of histone acetylation in vivo. a function for acetylation turnover? *Biochemistry and Cell Biology*, 80(3):363–378, June 2002.

[65] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3):37–52, Aug. 1987.

[66] H. Xu, L. Handoko, X. Wei, C. Ye, J. Sheng, C.-L. Wei, F. Lin, and W.-K. Sung. A signal–noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, 26(9):1199–1204, May 2010. PMID: 20371496.

[67] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, Sept. 2008. PMID: 18798982.

[68] V. W. Zhou, A. Goren, and B. E. Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 12(1):7–18, Jan. 2011.