Microfluidic Technology for Low-input Epigenomic Analysis

Yan Zhu

Dissertation submitted to the faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Chemical Engineering

Chang Lu

William A. Ducker

Aaron S. Goldstein

Liwu Li

April 20th 2018

Blacksburg, VA

Keywords: Chromatin immunoprecipitation (ChIP), Next generation sequencing (NGS), Epigenetics, Transcriptional regulations, DNA methylation, Histone modifications, Microfluidics, Circulating tumor cell (CTC)

Microfluidic Technology for Low-input Epigenomic Analysis

Yan Zhu

Abstract

Epigenetic modifications, such as DNA methylation and histone modifications, play important roles in gene expression and regulation, and are highly involved in cellular processes such as stem cell pluripotency/differentiation and tumorigenesis. Chromatin immunoprecipitation (ChIP) is the technique of choice for examining in vivo DNA-protein interactions and has been a great tool for studying epigenetic mechanisms. However, conventional ChIP assays require millions of cells for tests and are not practical for examination of samples from lab animals and patients. Automated microfluidic chips offer the advantage to handle small sample sizes and facilitate rapid reaction. They also eliminate cumbersome manual handling.

In this report, I will talk about three different projects that utilized microfluidic immunoprecipitation followed by next genereation sequencing technologies to enable low input and high through epigenomics profiling. First, I examined RNA polymerase II transcriptional regulation with microfluidic chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) assays. Second, I probed the temporal dynamics in the DNA methylome during cancer development using a transgenic mouse model with microfluidic methylated DNA

immunoprecipitation followed by next generation sequencing (MeDIP-seq) assays. Third, I explored negative enrichment of circulating tumor cells (CTCs) followed by microfluidic ChIP-seq technology for studying temporal dynamic histone modification (H3K4me3) of patient-derived tumor xenograft on an immunodeficient mouse model during the course of cancer metastasis.

In the first study, I adapted microfluidic ChIP-seq devices to achieve ultrahigh sensitivity to study Pol2 transcriptional regulation from scarce cell samples. I dramatically increased the assay sensitivity to an unprecedented level (~50 K cells for pol2 ChIP-seq). Importantly, this is three orders of magnitude more sensitive than the prevailing pol2 ChIP-seq assays. I showed that MNase digestion provided better ChIP-seq signal than sonication, and two-steps fixation with MNase digestion provided the best ChIP-seq quality followed by one-step fixation with MNase digestion, and lastly, no fixation with MNase digestion.

In the second study, I probed dynamic epigenomic changes during tumorigenesis using mice model with a tiny quantity of tissue samples. Conventional epigenomic tests do not support such analysis due to the large amount of materials required by these assays. In this study, I developed an ultrasensitive microfluidics-based methylated DNA immunoprecipitation followed by next-generation sequencing (MeDIP-seq) technology for profiling methylomes using as little as 0.5 ng DNA (or ~100 cells) with 1.5 h on-chip process for immunoprecipitation. This technology enabled me to examine genome-wide DNA methylation in a C3(1)/SV40 T-antigen transgenic mouse model during different stages of mammary cancer development. Using this data, I identified differentially methylated regions and their associated genes in different periods of

cancer development. Interestingly, the results showed that methylomic features are dynamic and change with tumor developmental stage.

In the last study, I developed a negative enrichment of CTCs followed by ultrasensitive microfluidic ChIP-seq technology for profiling histone modification (H3K4Me3) of CTCs to resolve the technical challenges associated with CTC isolation and difficulties related with tools for profiling whole genome histone modification on tiny cell samples.

Microfluidic Technology for Low-input Epigenomic Analysis

Yan Zhu

General Audience Abstract

The human genome has been sequenced and completed over a decade ago. The information provided by the genomic map inspired numerous studies on genetic variations and their roles in diseases. However, genomic information alone is not always sufficient to explain important biological processes. Gene activation and expression are not only associated with alteration in the DNA sequence, but also affected by other changes to DNA and histones. Epigenetics refers to the molecular mechanisms that affect gene expression and phenotypes without involving changes in the DNA sequence. For example, the DNA can get methylated, the histone protein that is wrapped around by DNA can also get methylated or acetylatied, and transcription factors can bind to different part of DNA. All of these can affect gene expression without alter the DNA sequences. Epigenetic changes occur throughout all stages of cell development or in response to environmental cues. They change transcription patterns in a tissue/cell-specific fashion.

For example, transcriptional silencing of tumor-suppressor genes by DNA methylation plays an important role in cancer development. Therefore, understanding of epigenetic regulations will help to improve various aspects of biomedicine. For instance, personalized medicine can be

tailored based on epigenetic profile of certain patient to specifically control gene expression in the disease treatment. However, the technology for profiling epigenetic modifications, i.e. Chromatin Immunoprecipitation (ChIP), suffers from serious limitations. The key limitation is the sensitivity of the assay. Conventional assay requires a large number of cells ($>10^6$ cells per ChIP). This is feasible when using cell lines. However, such requirement has become a major challenge when primary cells are used because very limited amounts of samples can be generated from lab animals or patients. Population heterogeneity information may also be lost when a large cell number is used.

In this project, we developed an automated ultrasensitive microfluidic chromatin/DNA immunoprecipitation followed by next-generation sequencing (ChIP/MeDIP-Seq) technology for profiling epigenetic modifications (e.g., histone modifications, transcriptional regulations, and DNA methylation). We extensively optimized design parameters for each and every step of ChIP/MeDIP (e.g. sonication/crosslinking time, antibody concentration, washing conditions) in order to reach highest sensitivity of 0.1 ng DNA (or ~50-100 cells) as starting material for IP, which is roughly 4-5 orders of magnitude higher than the prevailing protocol and 2-3 orders of magnitude higher than the-state-of-the-art(~50 ng). With such sensitivity, we were able to study temporal dynamics in the DNA methylomes during the various stages of mammary cancer development from a transgenic mouse mode. We were able to investigate transcriptional regulation of RNA polymerase II from scarce cell samples. We were also able to study histone modification (H3K4Me3) of circulating tumor cells during cancer metastasis.

Acknowledgements

First of all, I would like to thank my advisor Prof. Chang Lu for the continuous support of my Ph.D. study and research, for his openness, patience, and motivation. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and exciting. I am so thankful that he has tremendous trust and confidence in my abilities to let me be the first person to start on bioinformatics and mice work in the lab. I leanred to take initives, to be independent, and to reach the goal from bigger picuture while keep my pace fast. His advice and guidance not only helped me to become a better researcher but also a better person.

I would also like to thank my committee member Prof. William Ducker, Prof. Aaron Goldstein, and Prof. Liwu Li for their insightful comments and suggestions for my research work and dissertation, but also for the hard questions which helped me to widen my research from various perspectives.

I would also like to extent my appreciation to Jennifer Jenrette and Megan Frair from VBI, Dr. Nicole Lindstrom from Veterinarian school, Dana Reynolds from Vivarium, and my previous and current lab members for their help on my experiments and stimulating discussions.

I would especially like to thank all the members from Ph.D. Work Out Club. They have been helping me to get through all the frustrations and stresses from my Ph.D. research.

Finally, I would like to thank my family and friends for their love and support in all my endeavors.

Table of Content

List of Figures

List of Tables

1. Background

1.1 Chromatin Structure

In eukaryotic cells, genomic DNA is packaged with supporting proteins (i.e., histones) to form protein/DNA complexes called chromatin. The nucleosome is the basic unit of chromatin, which is composed of 146 base pairs (bp) of DNA wrapping around a histone octamer. This octamer consists of two H2A-H2B dimmers and an H3-H4 tetramer[1], where H2A, H2B, H3, and H4 are four core histones tightly packed in globular regions[2-4]. Amino-terminal tails extended from the globular regions make histones more accessible to histone modifying enzymes. Thus, they are subjected to various post-translational modifications (PTMs) including acetylation, methylation, phosphorylation, and ubiquitination[5, 6]. Another protein, linker histone H1, sits on top of nucleosome, and binds to the linker DNA region between nucleosomes to compact chromatin into higher-order structures that make up chromosomes. Generally, nucleosomes fold up to produce a 30-nanometer chromatin fiber, and the fiber forms loops that have 300 nm in average length. The loops are compressed and folded to produce a 250 nm wide fiber, which is tightly coiled into the chromatid of a chromosome[7, 8].

Chromatin structure not only helps to package DNA into tiny nucleus but also regulates the accessibility of DNA for transcription, recombination, DNA repair and replication. There are two functional chromatin states: euchromatin and heterochromatin. Euchromatin has relaxed nucleosome arrangement that makes DNA is accessible for transcription, whereas heterochromatin has condensed nucleosome arrangement that blocks bindings of transcription

factors or other chromatin-associated proteins. Therefore, euchromatin is related to gene activation, but heterochromatin is associated with gene silencing[5, 9-13].

1.2 Epigenetic Changes

The human genome has been sequenced and completed over a decade ago. The information provided by the genomic map inspired numerous studies on genetic variations and their roles in diseases. However, genomic information alone is not always sufficient to explain important biological processes. Gene activation and expression are not only associated with alteration in the DNA sequence, but also affected by other changes to DNA and histones. Epigenetics refers to the molecular mechanisms that affect gene expression and phenotypes without involving changes in the DNA sequence (**Figure 1.1**). Epigenetic changes occur throughout all stages of cell development or in response to environmental cues. They affect gene accessibility of regulatory proteins such as methyl-CpG-binding proteins, transcription factors, RNA polymerase II, and other components of the transcriptional machinery, eventually changing transcription patterns in a tissue/cell-specific fashion[14-19]. Epigenetics is the field that examines the environmental factors and determine how they alter gene regulations and what roles they play in disease development. The study of epigenetics at the genome scale is referred to as epigenomics[20]. With a large epigenomic database, one can formulate hypotheses about the causalities between epigenomic changes and phenotypical changes, link chromatin modifications with aberrant gene expressions in diseases, and define core regulatory mechanisms that occur in different tissues and in different developmental stages[21, 22].

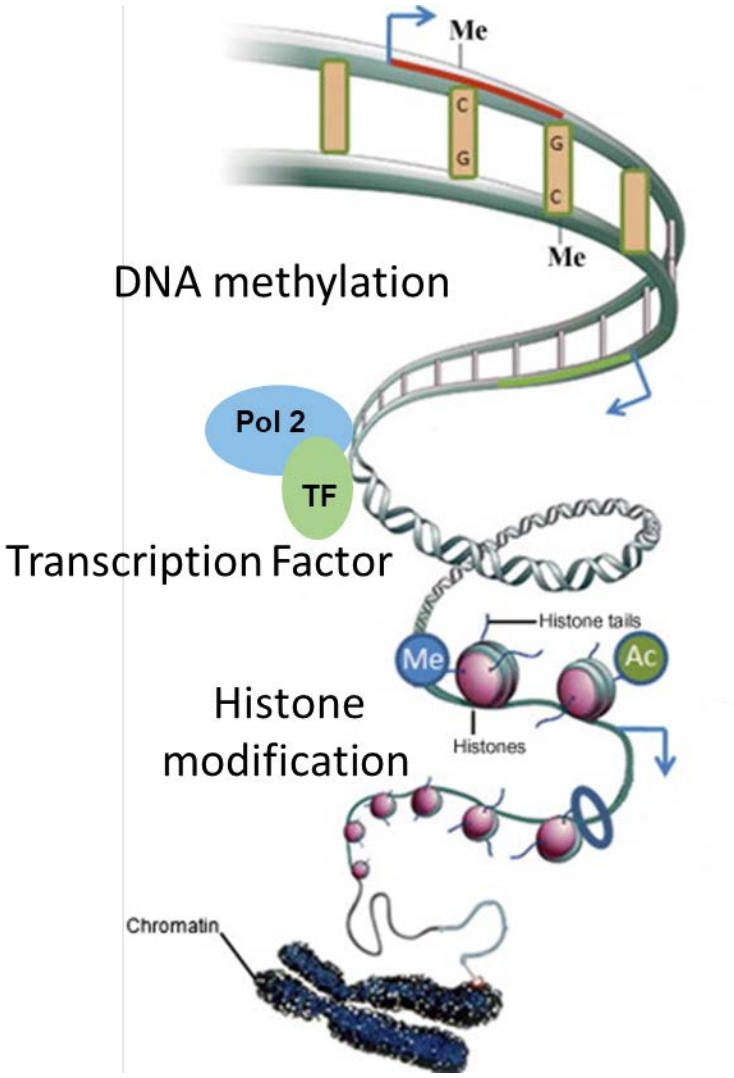


Figure 1.1 Epigenetic Changes – DNA methylation, transcriptional regulations, and histion modifications (adapted from Rajender et al. *Mutat. Res.* 727 (2011) 64)[23].

1.2.1 Histone Modifications

Histone modifications, known as PTMs, can alter the chromatin structure and regulate gene expressions. They are critically involved in processes, such as transcription, recombination, DNA repair and replication[2, 4, 6, 24]. Of these PTMs mentioned above, histone (H3 and H4)

acetylation and methylation are the best understood. In histone acetylation, there are positively charged lysine residues of amino-terminal tails that can tightly bind to the negatively charged DNA to compact nucleosomes and form a condensed chromatin structure, which disables the binding of transcription factors. However, acetylation on those lysine residues removes positive charges and decreases the affinity between histone and DNA[6, 9, 25]. Thus closed chromatin structure is opened up into relaxed structure that allows binding of transcription factor and increases gene expression[26].

Histone methylation, on the other hand, does not affect the charges of amino-terminal tails. It is less likely to be able to modify the chromatin structure. Therefore, histone methylation can either increase or repress gene expression[27]. Histone methylation mostly occurs on different lysine residues, but it can also occur on arginine. Arginine can be methylated once or twice, whereas lysine can be methylated once, twice, or three times (mono-, di-, or tri-methylation). Methylation on arginine is related to gene activation. Tri-methylation on 4th lysine residue (K4) of Histone H3 (H3K4me3) is generally associated with gene activation[11, 24, 28-30]. But tri-methylation on 9th lysine residue (K9) and 27th lysine residue (K27) of histone H3 (H3K9me3 & H3K27me3) is generally associated with gene repression[5, 6, 31].

1.2.2 DNA Methylation

In mammals, DNA methylation typically refers to the addition of a methyl group at the carbon-5 position of cytosine residues within CpG dinucleotides, forming 5-methylcytosine (5mC). This process is catalyzed by enzymes called DNA methyltransferases. Approximately 70–80% of CpG dinucleotides are methylated, primarily in heterochromatic regions. However, CpG-rich

sequences with low levels of DNA methylation are referred to as CpG islands (CGIs). The human genome contains 0.7 % clusters of CpG islands and those CpG sites contain 7 % of all CpG dinucleotides. CpG islands are mostly enriched at gene promoters. About 60% of all mammalian gene promoters are CpG rich[32]. CpG islands at promoter regions typically stay unmethylated in normal cells. This opens up space on nucleosome for DNA to be easily transcribed and is associated with active gene expression during differentiation. On the other hand, methylated CGIs are associated with gene repression[33, 34] **(Figure 1.2)**.

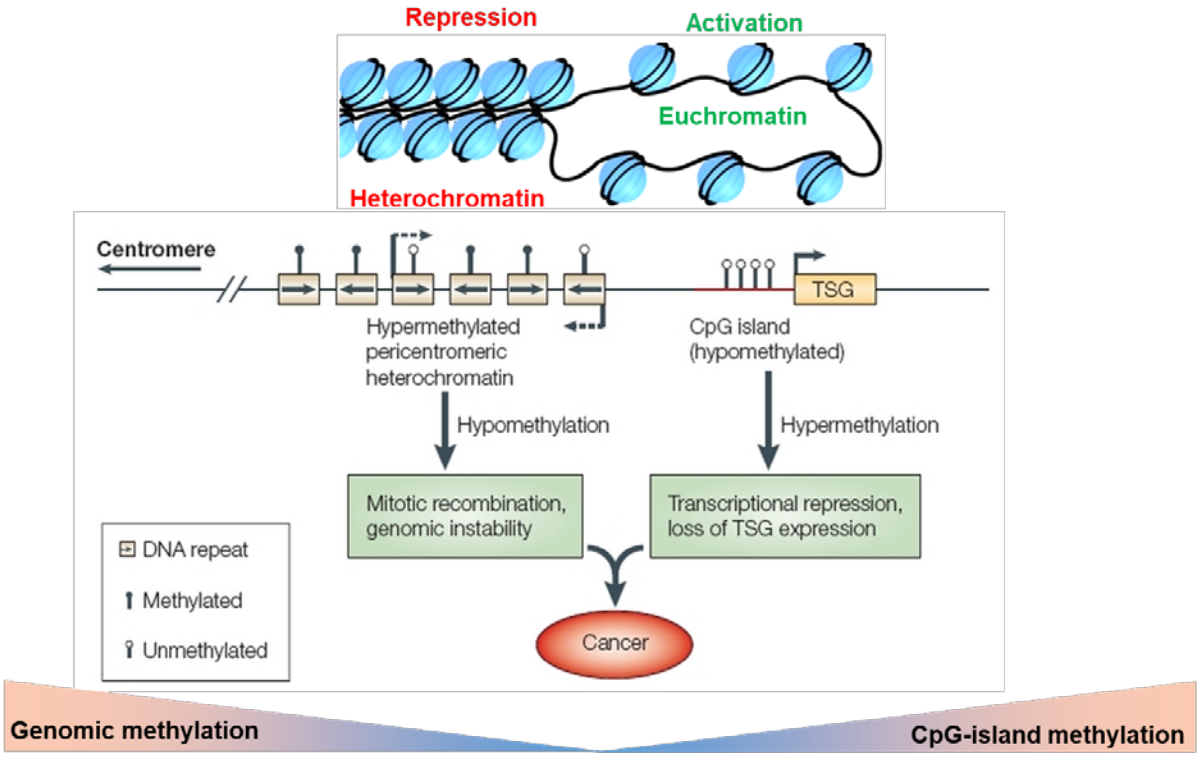


Figure 1.2 DNA methylation and cancer (adapted from Robertson *Nat Rev Genet. 6* (2005) 598 and Esteller *N. Engl. J. Med.* 358 (2008) 1150)[35, 36].

DNA hypermethylation and hypomethylation compared to normal tissues have been associated with a large number of human diseases. For example, transcriptional silencing of tumor suppressor genes by CGI-promoter hypermethylation plays an important role in cancer

development[37-39]. Genes that are critically involved in cancer biology, including the cell-cycle inhibitor p16-INK4a and the DNA-repair genes MLH1 and BRCA1, have been shown to undergo methylation-associated silencing in tumor cells[40, 41]. Typically, global hypomethylation leads to activation of oncogenes and local hypermethylation of CGIs at the promoter region represses tumor suppressor genes during cancer development[42] (**Figure 1.3**).

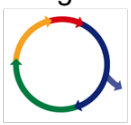
Gene function	Gene name	Cancer type
Cell-cycle regulation 	RB1	Retinoblastoma
	CDKN2A (INK4A transcript)	Colon, lung, many others
	CDKN2A (ARF transcript)	Colon
Tumor suppressor 	CDH1	Breast, gastric, thyroid, leukemia, liver
	CDH13	Lung, ovarian, pancreatic
	TIMP3	Brain, kidney
	VHL	Renal cell
	HIC1	Breast, many others
DNA repair / detoxification 	MLH1	Colon, endometrial, gastric
	MGMT	Brain, colon, lung, breast
	BRCA1	Breast, ovarian
	GSTP1	Prostate, liver, colon, breast, kidney

Figure 1.3 Examples of Genes Silenced by Aberrant DNA Hypermethylation in Cancer.

1.2.3 Transcription factor

There are a number of processes that affect gene expressions without changing DNA sequence. Transcriptional regulation is one of the most important mechanisms that affect gene activity in cell proliferation and differentiation. It is widely accepted that the molecular requirements for human cancer include loss of growth inhibition by ligands, self-sufficiency in growth stimulation, limitless replicative potential, and avoidance of apoptosis[43]. All of these characteristics have

close ties to dysregulated transcription. Transcription of DNA to RNA is carried out by RNA polymerase II binding with other general transcription factors (e.g., TFIIA, TFIIB, etc.) and mediators near the site of transcription initiation. These factors can either recruit other factors that modify chromatin structure or directly interact with transcription machinery. Both ways end up recruiting transcription machinery to a core promoter[44]. The core promoter helps position RNA polymerase II to its preinitiation complex state. However, this complex of RNA polymerase II with other general transcription factors at the promoter is not in an active conformation to begin transcription. Initiation of transcription begins when the open complex forms, where 11-15 base pairs of DNA near the transcription start site are unwound and the template strand of promoter is positioned within the active site cleft of RNA polymerase, and Ser5 is phosphorylated. Subsequently, Ser2 phosphorylation is needed for transcription elongation to occur[45].

It is well established that the differential gene expression patterns in cancer cells correlate with the recruitment of specific and general transcription factors on relevant promoters. Intracellular cell signaling pathways typically rely on activated transcription factors or proteins that activate transcription factors for relaying the signal into the nucleus. These signaling proteins may experience mutation or overexpression and these changes alter spatiotemporal transcription patterns. As a result, a number of transcription factors (e.g., STATs, NF-κB, β-catenin, and NICD) are well known to have increased activity in most human cancers and prevent apoptosis of tumor cells in these cases[46-48]. In tumor cells, for example, transcription factor NF-κB is normally sequestered in the cytoplasm through the association with an I-κB protein. When the cell is exposed to activation signals such as tumor necrosis factor-α (TNF-α) binding to cell

surface receptors, the I-κB protein is phosphorylated on serines 32 and 36, then ubiquitinated, and broken down in proteasomes. After being freed from its association with I-κB, the NF- κB complex moves to the nucleus where it binds to specific sequences in the promoter/enhancer regions of genes to regulate transcription[49, 50].

1.3 Next Generation Sequencing

DNA sequencing is the process to determine the precise order of the four bases (adenine, guanine, cytosine, and thymine) in a strand of DNA. In 1977, Fred Sanger and Alan R. Coulson came up a chain termination method for fast determination of DNA sequences, known as Sanger sequencing. It paved the road for deciphering complete genes and later entire genomes. The human genome took 13 years to sequence based on this method, and it was finally finished in 2003. Next-generation sequencing (NGS) refers to non-Sanger-based sequencing technologies. The major difference between Sanger and next-generation sequencing is that next-generation sequencing allows millions or billions of DNA strands to be sequenced in parallel, resulting in high-throughput[51, 52].

There are different kinds of next-generation technologies, mainly distinguished by their template preparation, sequencing and imaging methods[53]. In template preparation, the template is attached or immobilized to a solid surface or support to allow millions or billions of sequencing reactions to be performed in parallel. There are two template preparation methods: Clonally amplified templates, which can be branched into emulsion PCR (Roche/454, Life/APG, Polonator)[54-57], solid-phase amplification (Illumina/Solexa)[58], and Single-molecule templates (Helicos BioSciences, Pacific Biosciences, Life/Visigen, LI-COR

Biosciences)[59, 60]. In sequencing and imaging, methods include cyclic reversible termination (CRT) (Illumina/Solexa, Helicos BioSciences)[61-65], single-nucleotide addition (SNA) (e.g., Roche/454 — Pyrosequencing)[66, 67], real-time sequencing (Pacific Biosciences)[60, 68], and sequencing by ligation (SBL) (e.g., Life/APG — SOLiD)[69-71]. Here, solid-phase amplification template preparation method and cyclic reversible termination sequencing method will be discussed since Illumina provides industry-leading sequencing technology in respects of data quality, throughput and cost, and all of our preliminary studies were conducted by Illumina sequencing.

Solid-phase amplification is used to generate randomly distributed, clonally amplified clusters from fragments on a glass slide. Forward and reverse primers are covalently bonded to the slide. The surface density of the amplified clusters can be determined by the ratio of the primers to the template on the support. Solid-phase amplification can produce 100–200 million spatially separated template clusters[53]. It provides free ends for universal sequencing primer to hybridize and initiate the NGS reaction.

Cyclic reversible termination uses reversible terminators in a cyclic method that is comprised of nucleotide incorporation, fluorescence imaging and cleavage. In the first step, a DNA polymerase is bounded to the primed template. It adds or incorporates one fluorescently modified nucleotide, which represents the complement of the template base. After incorporation, the remaining unincorporated nucleotides are washed away. Imaging is then used to determine the identity of the incorporated nucleotide. Next, the cleavage step removes the terminating/inhibiting group and the fluorescent dye. Another washing is performed before

starting the next incorporation step. The important feature of the CRT method is the reversible terminator. There are two types of terminators: 3' blocked and 3'unblocked. 3'-blocked terminators require break of two chemical bonds to remove the fluorophore from the nucleobase and restore the 3'-OH group, whereas 3'unblocked only require break of one chemical bonds to remove the fluorophore from the nucleobase. Unblocked 3'-OH group is the natural substrate for incorporating the next incoming nucleotide, so 3'unblocked terminators are more efficient on restoring the nucleotide for the next CRT cycle than 3'-blocked terminators. However, 3' blocked terminators, such as 3'-O-azidomethyl have been successfully used in CRT by Illumina/Solexa[61-65].

In general, Illumina sequencing starts with a flow cell, which is a glass slide with lanes (**Figure 1.4**). Each lane is a channel coated with a lawn composed of two types of oligonucleotides. Hybridization is enabled by the first of the two types of oligos on the surface. This oligo is complementary to the adapter region on one of the fragments strands. A polymerase creates a complement of the hybridized fragment. The double-stranded molecule is denatured and the original template is washed away. The strands are clonally amplified through bridge amplification[72]. In this process, the strand folds over, and the adapter region hybridizes to the second type of oligo on the flow cell. Polymerases generate the complementary strand forming a double-stranded bridge. This bridge is denatured resulting in two single-stranded copies of the molecule that are tethered to the flow cell. The process is then repeated over and over and occurs simultaneously for millions of clusters, resulting in clonal amplification of all the fragments. After bridge amplification, the reverse strands are cleaved and washed off, leaving only the forward strands. The 3' ends are blocked to prevent unwanted priming. Sequencing begins with

the extension of the first sequencing primer to produce the first read. With each cycle, fluorescently tagged nucleotides compete for addition to the growing chain, and only one is incorporated based on the sequence of the template. After the addition of each nucleotide, the clusters are excited by a light source and a characteristic fluorescent signal is emitted. This proprietary process is called sequencing by synthesis. The number of cycles determines the length of the read. The emission wavelengths along with the signal intensity determine the base call. For a given cluster, all identical strands are read simultaneously. Hundreds of millions of clusters are sequenced in a massively parallel process. After the completion of the first read, the read product is washed away. The index one read primer is then introduced and hybridized to the template. The read is generated similar to the first read. After completion of the index read, the read product is washed off and the 3' end of the template is deprotected. The template now folds over and binds the second oligos on the flow cell. Index two is read in the same manner as index one. Index two read product is washed off at the completion of this step. Polymerases extend the second flow cell oligos forming a double-stranded bridge. This double stranded DNA is then linearized and the 3' ends are blocked. The original forward strand is cleaved off and washed away, leaving the reverse strand. Read two begins with the introduction of the read two sequencing primer. As with read one, the sequencing steps are repeated until the desired read length is achieved. The read two product is then washed away. This entire process generates millions of reads, representing all the fragments. Sequences from pooled sample libraries are separated based on the unique indices introduced during the sample preparation. For each sample, reads with similar stretches of base calls are locally clustered. Forward and reverse reads are paired creating contiguous sequences. These contiguous sequences are aligned back to the reference genome for variant identification[53, 65].

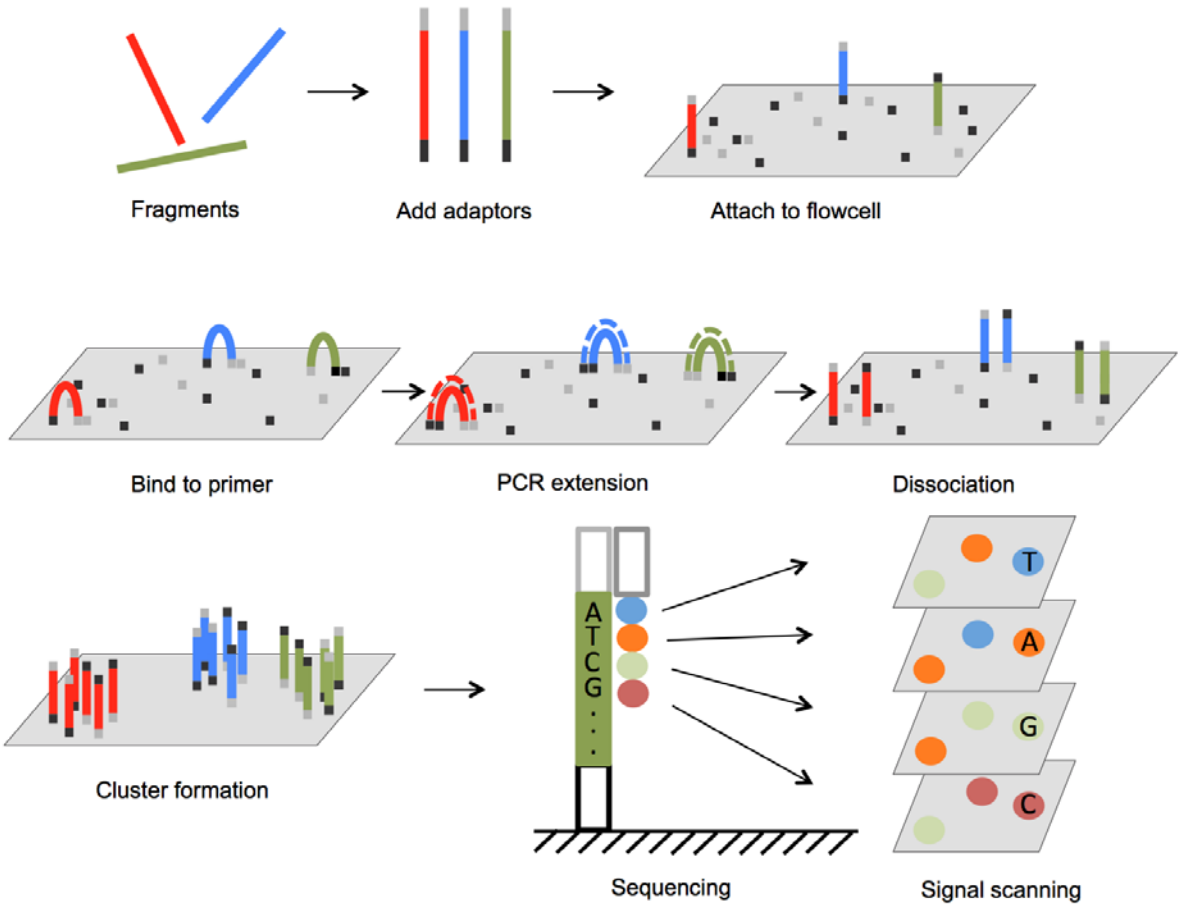


Figure 1.4 Sequencing by synthesis (adapted from Yuan Lu, Yingjia Shen, Wesley Warren and Ronald Walter (2016). Next Generation Sequencing in Aquatic Models, Next Generation Sequencing - Advances, Applications and Challenges, Dr. Jerzy Kulski (Ed.), InTech, DOI: 10.5772/61657. Available from: https://www.intechopen.com/books/next-generation-sequencing-advances-applications-and-challenges/next-generation-sequencing-in-aquatic-models).

1.4 ChIP-seq

Epigenetic regulations of gene expressions play pivotal roles in normal and disease development. Understanding of epigenetic regulations will improve various aspects of biomedicine, including strategies for manipulating cell fate for regenerative medicine, novel epigenetics-based disease markers and biomarkers, and novel therapeutics. Chromatin immunoprecipitation followed by sequencing analysis (ChIP-Seq) is by far the powerful tool to study genome-wide distributions and binding sites of chromatin-binding proteins and histone modifications in any genome. It reveals insights about how gene regulations are involved in various diseases and biological pathways at epigenomic level[73].

1.4.1 Chromatin Immunoprecipitation

To date, chromatin immunoprecipitation (ChIP) assay is the gold standard for investigating in vivo epigenetic modifications and transcription factor binding at the genome scale. ChIP assays involve a number of steps (**Figure 1.5**). First, DNA is either cross-linked (Cross-linking chromatin immunoprecipitation, X-ChIP)[74] with protein by cross-linking agents (e.g., formaldehyde) in order to freeze protein–DNA and protein–protein interactions or not cross-linked (Native chromatin immunoprecipitation, N-ChIP)[75] depending on the purpose of the experiment. Subsequently, cells are lysed in order to release chromatin. Then, chromatin is either sonicated or MNase digested also depending on the purpose of the experiment to yield fragments of protein-bound DNA that are typically 200–600 base pairs long. In general, for histone modifications, MNase digestion of native chromatin into mononucleosome-sized particles without cross-liking is the preferred method because as the histone is already wrapped around by

DNA, there is no need to cross link them together. Also it provides high-resolution data for nucleosome modifications and eliminates artifactual signals caused by cross-linking with other genomic regions. However, this method may suffer from a loss of signal or bias due to unstable nucleosomes[76, 77]. For mapping binding sites of transcription factors, sonication of formaldehyde cross-linked chromatin may be the preferred method because some transcription factors are not tightly bound or indirectly bound to DNA, cross-linking here helps to enhance the bindings of transcription factors and DNA. Also if MNase is used, it will degrade linker DNA, where transcription factors tend to bind. However, mechanical shearing is harsher then MNase digestion and may cause damage to epitopes of interest[78]. In the next step, these fragments are then immunoprecipitated onto the surface of antibody-coated beads. The antibody here specifically targets a transcription factor, a specific modified form of histone, or a cytosine methyl group. Finally, the immunoprecipitated fraction is isolated. The cross-linking is reversed and the released DNA fragments are assayed to determine their sequences[79]. The identification of the DNA sequences can be done by qPCR if there are known gene candidates. Alternatively, these binding sites can be mapped at the genome scale by microarray (ChIP-chip) or sequencing (ChIP-seq) using high-throughput sequencing technology (e.g., Illumina HiSeq 4000 System) [80, 81]. ChIP–seq offers many advantages over ChIP–chip. First, ChIP–seq provides single nucleotide resolution comparing to 30-100bp offered by ChIP-chip. Second, ChIP-seq coverage is not limited by sequences on the array, which enables the studies of repetitive sequences. Third, the signal intensity measured on ChIP-chip arrays might not be linear over its entire range, and its dynamic range is limited below and above saturation points. Moreover, ChIP–seq requires less ChIP DNA (10-50ng) compared to few microgram required by ChIP–chip. Last but not least, ChIP–seq can sequence multiple samples at the same time, referred as multiplexing, which

becomes important for cost effectiveness. However, the main disadvantage with ChIP–seq is its current cost and availability. Depending on size of the genome and sequencing depth needed, ChIP-seq still could be cheaper than ChIP-chip when sequence large genome like human.

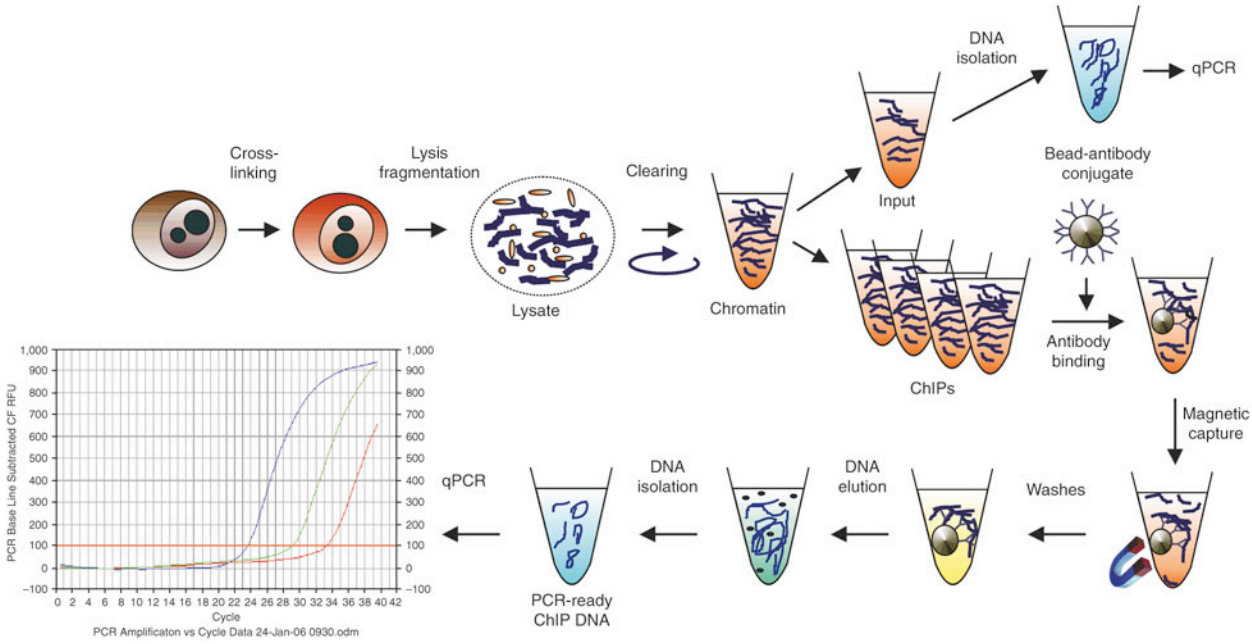


Figure 1.5 Convention ChIP work flow(adapted from Dahl et al. Nat. Protocols 3 (2008) 1033)[82].

1.4.2 Genome Mapping and Peak Calling

Next-generation sequencing normally generate short reads (50bp or 100bp). The short reads do not have genomic position information. So those reads need to be aligned or mapped to the reference genome with a mapping algorithm (e.g., BOWTIE[83, 84]). This mapping algorithm tries to locate read sequences that match to reference sequences, while tolerating a certain amount of mismatch. Sometimes, reads can be mapped to multiple locations in the genome (e.g., repetitive sequences). In this case, only uniquely mapped reads will be used in downstream analysis.

After sequenced reads are aligned to the genome, the next step is to identify regions that are enriched in the ChIP sample relative to the control with statistical significance. Peak calling algorithms (e.g., MACS[85], SPP[86, 87]) are used to identify the enriched regions. Briefly, ChIP fragments are sequenced from 5' ends on both the positive strand and negative strand of DNA (**Figure 1.6**). Mapping them to the genome forms two distributions with consistent distant between the peaks of the distributions. Extending those two distributions of mapped reads with an estimated binding fragments size results in a combined peak profile that tells where the DNA-protein binding site is[73, 88]. Then peaks are scored (based on P-value) to determine their statistical significance. If one only looks at fold enrichment of ChIP sample relative to the control sample, the information is not sufficient to draw a conclusion about statistical significance of enriched regions. The reason is that an enrichment of 10 estimated from 100 and 10 tags (from the ChIP and control experiments, respectively) has a different statistical significance to the same enrichment estimated from 1000 and 100 tags[89]. In this case, a Poisson model is used to account for both enrichment and tag numbers[90, 91]. However, if sequencing depth is increased, the number of significant peaks will also increase since peaks with small enrichment become statistically significant as more tags accumulate. An important factor that affects the sequencing depth that is required for a ChIP–seq experiment is whether the protein (or chromatin modification) is a point-source factor, a broad-source factor or a mixed-source factor[92, 93]. Point-source factors are sequence-specific transcription factors (e.g., CTCF, MYC) and chromatin marks (e.g., H3K4me3) that bind to specific locations in the genome and generate narrow peaks around 100s of base pairs. Broad-source factors include many chromatin marks (e.g., H3K27me3, H3K9me3, and H3K36me3) that generate broad peaks around 100s of kilobases. Mixed-source factors (e.g.,

RNA Pol II, SUZ12) generate combined point and broad peaks[89]. According to The Encyclopedia of DNA Elements (ENCODE) project guidelines for ChIP–seq experiments, 20 million reads[92] per factor should be used for point-source factor experiments, and 40 million reads[92] per factor should be used for broad-source or mixed-source factors. In MeDIP–seq, the sequencing depth depends on number of CpG dinucleotides in the genome. In this case ENCODE suggests that 60 million reads[92, 94] are enough to detect the majority of methylated CpG in the human genome[95].

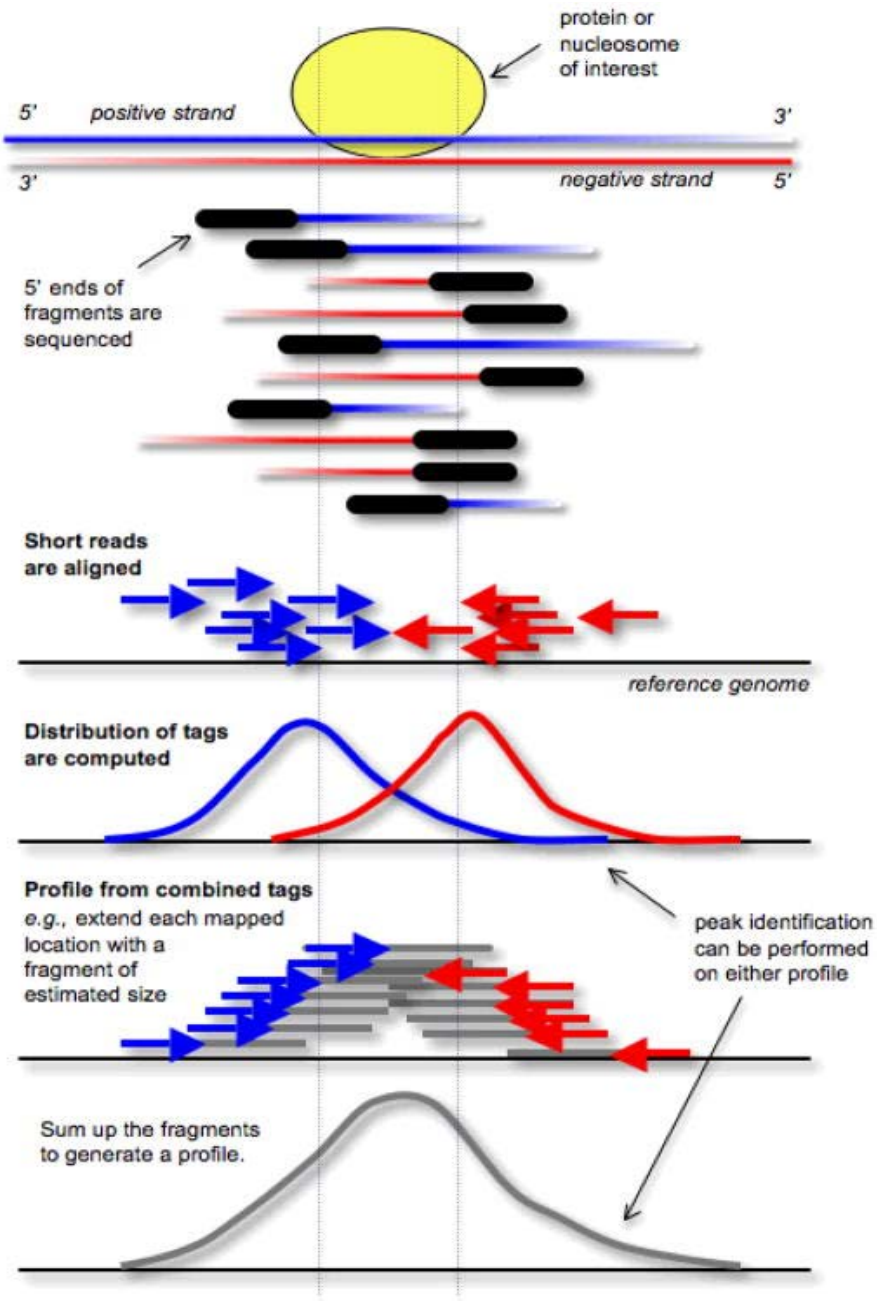


Figure 1.6 Strand-specific profile at enriched sites (adapted from Park *Nat. Rev. Genet.* 10 (2009) 674)[80].

MACS finds a global peak shift size from significant fold changes (mfold) regions. Then all forward and reverse strand tags are shifted toward to the center by this estimated shift size. Next, MACS calls peaks on the combined tags using a Poisson distribution through sliding windows. In Poisson distribution, lambda is number of occurrence of expectation, and it is equal to mean and variation. Lambda in peak calling represents number of tag counts/hits. Each lambda value is associated with different Poisson density function. The density function has y-axis as tag density/probability, and x-axis as tag counts/hits. MACS uses input sample to generate dynamic parameter lambda local, which is estimated from the maximum tag counts/hits of 1 kb, 5 kb or 10 kb window centered at the peak location. ChIP-seq sample is observation in this case, for number of observed occurrence of tag counts/hits (k), one can calculate the probability of observed k counts/hits for given lambda local expected counts/hits. If observed k counts/hits from ChIP-seq has way lager value than local lambda from input, which will provide a really small probability (e.g. expectation for overflow floods occur once every 100 years, lambda=1, what would be the probability of having 3 overflow floods in 100 years, k=3, it will be really small $P = \frac{\lambda^k e^{-\lambda}}{k!}$). Therefore it relates to a very small p-value (the peak is statistically significant). The taller the peak, the smaller the p-value. The ratio between the ChIP-Seq tag count and λlocal is reported as the fold_enrichment. The empirical FDR in MACS is defined as number of control peaks divided by the number of ChIP peaks. MACS is not good at identifying sharp peaks in terms of inconsistent peak numbers.

SPP is a better option for calling sharp peaks. Unlike MACS, SPP finds global peak shift size from a cross-correlation analysis. The estimated shift size is the one that provides the highest linear Pearson correlation of all forward tag counts and reverse tag counts. SPP then uses this shift size as a window size to slide through genome to call the peak. In peak calling, SPP calls a

peak based on a locally maximized binding score, which is defined as twice of the difference between the geometric mean and the arithmetic mean of the forward tag counts in upstream window and the reverse tag counts in downstream window. Next, binding scores in the signal window are compared with control data to generate FDR (SPP is not based on any statistical models). If no control data presents, SPP will randomly generate a data set from signal data.

1.5 MeDIP-seq

DNA methylomes, similar to other epigenetic information, are specific to cell and tissue types, the disease condition and its developmental stage. Profiling DNA methylomes contributes to various aspects of cancer intervention in the context of personalized medicine. First, methylomic profiles may serve as highly-sensitive cancer cell markers because CGI hypermethylation of tumor-suppressor genes occurs early in the tumorigenesis and the number of genes undergoing this epigenetic alteration increases during cancer development. Second, methylomic profiles can provide valuable information for prognosis because certain patterns of methylation-based silencing are associated with tumor aggressivity and angiogenesis[96]. Third, methylomes have also been used to predict responses to a therapeutic procedure or reagent[97]. Finally, methylomic marks have been used as therapeutic targets. DNA-demethylating agents have been explored as promising drugs for reversing promoter DNA hypermethylation and associated gene silencing[98]. Methylated DNA immunoprecipitation followed by sequencing analysis (MeDIP-seq) provides an unbiased tool for probing DNA methylation events throughout the entire genome.

Methylomic profiles (like genomic hypomethylation and CpG hypermethylation) may serve as highly-sensitive cancer cell markers for cancer diagnosis and prognosis. Profiling DNA

methylomes contributes to cancer intervention in the context of personalized medicine. One can use epigenetic information for personalized medicine (**Figure 1.7**). One can profile DNA methylomes in a normal tissue and compare it to temporal dynamic methylomic profiles from a disease tissue in disease development. Then one can tell which gene is activated and which gene is silenced and annotate the function of those genes. At the same time patient's methylomic profiles are extracted, and compare with established profiles from lab (with annotated gene functions) for disease identification. Finally one can apply that information to develop targeting-specific drugs for reversing DNA hypermethylation and associated gene silencing in the individual patient.

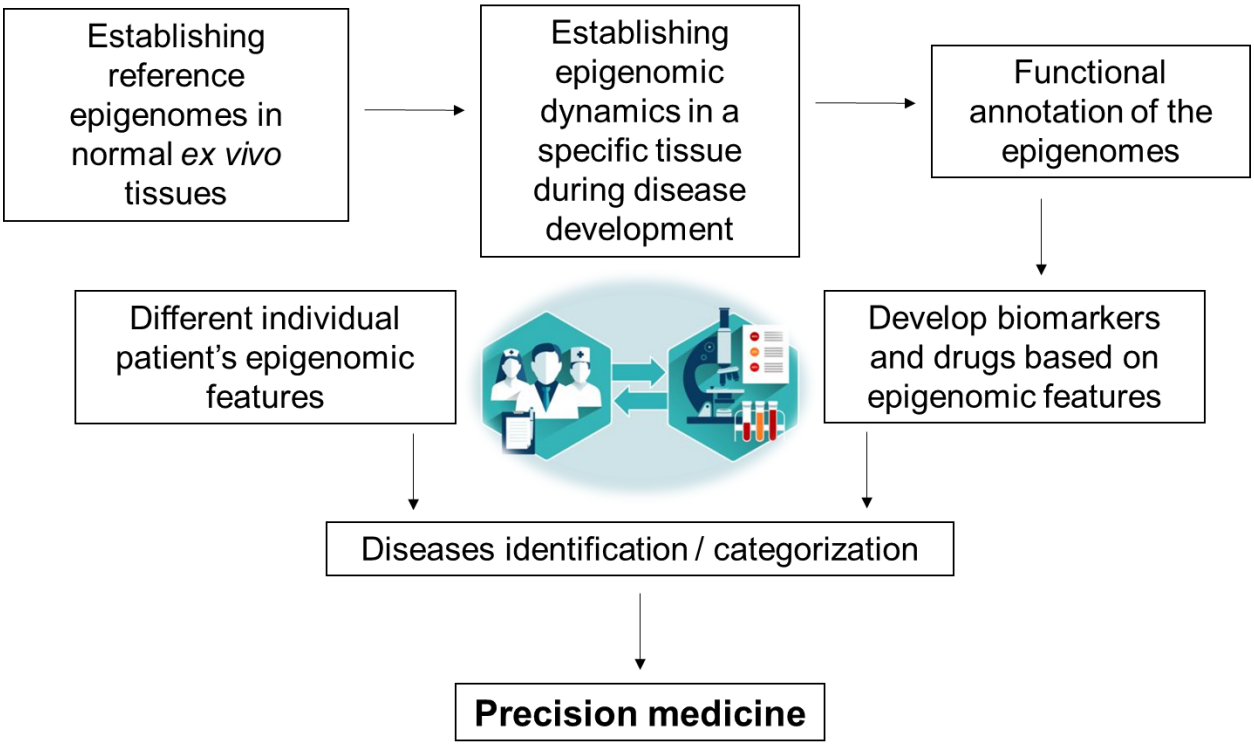


Figure 1.7 From the Bench to the Bedside: Use Epigenomics Information for Precision Medicine.

1.5.1 Methylated DNA Immunoprecipitation

In MeDIP, genomic DNA is extracted from cells without cross-linking (**Figure 1.8**). Then, genomic DNA is either sonicated or MNase digested to produce 200-600 base pair long fragments. The DNA fragments are denatured to produce single stranded DNA. Those single stranded DNA are then immunoprecipitated onto the surface of antibody-coated beads. The antibody here specifically targets 5-methylcytosine (5mc) on single stranded DNA. Finally, the immunoprecipitated fraction is isolated[99]. Similar, the identification of the DNA sequences can be done by qPCR if there are known gene candidates. Alternatively, these binding sites can be mapped at the genome scale by high-throughput sequencing.

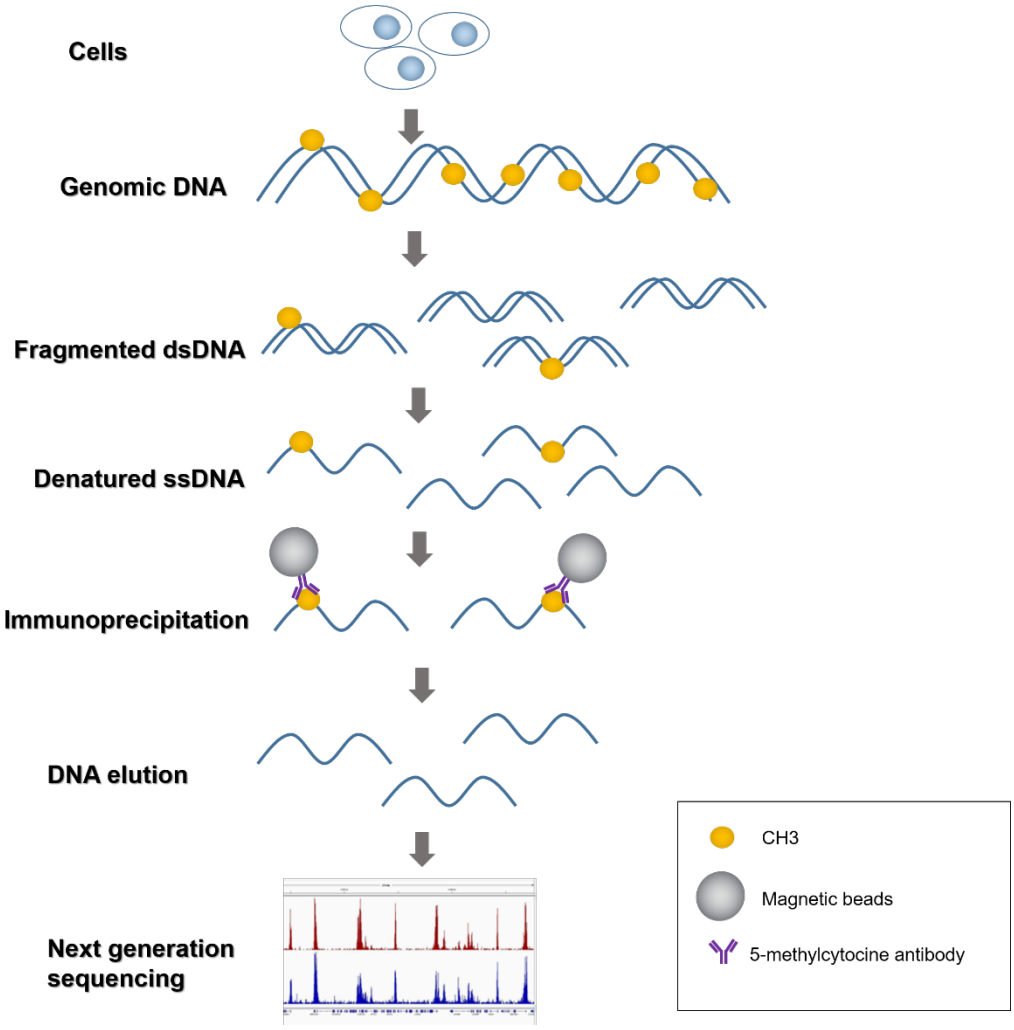


Figure 1.8 MeDIP-seq workflow.

1.5.2 Differentially Methylated Regions (DMRs)

Differentially methylated regions (DMRs) are regions of DNA that have different methylation patterns compared to other samples. For example, DMRs between cancer and normal samples show the aberrant methylation in cancers.

To identify DMRs, the genome is first divided into bins of equal size. The raw MeDIP-seq reads in each bin can be transformed into reads per million (rpm) format to represent coverage profile at each bin. It has been shown that MeDIP signals scale with local densities of CpGs[100, 101]. Therefore, it is necessary to normalize MeDIP signals with CpG density when comparing two biological samples. Here, a coupling factor which contains local CpG densities has to be calculated. For each genomic bin at position b, the density of surrounding CpGs is defined within range of [b-d, b, b+d], where d is maximal distance to be defined. One way to calculate coupling factor is to count the number of CpGs within the maximal distance d around a genomic bin at position b. Another way is to apply a weighting function to weight each CpG by its distance to the current genomic bin at position b. For example, CpGs close to the genomic bin will receive higher weights, whereas CpGs farther away from the current genomic bin will receive smaller weights. Once coupling factor is calculated, MeDIP-seq signal can be then normalized to relative methylation score (RMS) based on CpGs density, and ratio of mean rms values between the control sample and the treatment sample can be obtained as well. To assess differential methylation, it is of interest to pre-define regions of interest (ROIs) like CpG islands, promoter, etc., or any given genome wide length. Statistical tests (e.g., Wilcoxon rank test, Ttest) then can be used in order to rate whether the RMS data series of the genomic bins within any ROI significantly differs in the control sample compared to the treatment sample. Finally, ROIs are filtered with given p-value, false discovery rate (FDR), or ratio of mean rms values between the control sample and the treatment sample. The remaining ROIs are considered as candidate genomic regions where events of differential methylation can be deduced from the data in a statistically significant way[102].

A frequency vs reads count can be described as normal distribution (bell-shaped curve). If one were to compare two normal samples (not much variation), the P-value would be large (> 0.05) because the two samples overlap. Very rarely, would one get two samples that do not overlay. When this happens, the P-value will be <0.05. This is called a false positive (or Type I error), because the small P-value suggests that the samples are from two types of condition (two separate distribution), and this is false. Normally, false positives are rare. 95% of the time the two samples will overlap. This will result in P-values greater than 0.05. Five percent of the time they do not. This will result in P-values of less than 0.05. But human and mouse cells have at least 10,000 transcribed genes. If we took two samples from the same condition and compared all 10,000 genes, 5% of 10,000 tests could statistically give rise to 500 false positives. That would be 500 genes that appear interesting, even when they are not. The false discovery rate can control the number of false positives. Technically, the FDR is not a method to limit false positives, but the term is used interchangeably with the methods. In particular, it is used for the Benjamini-Hochberg method. A histogram of P-values is generated by comparing two normal conditions (same distribution). X-axis is possible values for P-values. Y-axis is number of P-values in each bin. The P-values are uniformly distributed, there is an equal probability that a test's P-values falls into any one of these bins. A histogram of P-values is generated by comparing two different conditions (e.g., normal vs tumor, two different distributions). Most of the P-values are <0.05, and are heavily skewed and closer to 0. The P-values greater than 0.05 are the false negatives from where the samples overlapped. If tumor (DNA methylation) affects 1000 genes, the measurements for these genes will come from two different distributions. The P-values are skewed. The remaining 9000 genes might not be affected by tumorigenesis. This means the measurements for most of genes will come from the same distribution. The P-values

are uniformly distributed. The histogram of P-values we obtained from all 10,000 genes is the sum of the two separate histograms. The uniformly distributed P-values come from the genes unaffected by the DNA methylation. The P-values on the left side are a mixture from genes affected and genes unaffected by DNA methylation. Visually, we can assess where the P-values are uniformly distributed and determine how many tests are in each bin. We can extend this line and use it as a cutoff to identify the true positives. One way to isolate the true positives from the false positives would be to only consider the smallest P-values (<0.05) above the dotted line. Benjamini and Hochberg developed an algorithm to covert this procedure by eye into a mathematic formula. It adjusts P-values in a way that limits the number of false positives that are reported as significant. Adjusts p-values refers that it makes them larger values. For example, before the FDR correction, P-values might be 0.04 (significant). After the FDR correction, P-values might be 0.06 (no longer significant). If cutoff for significant is FDR less than 0.05, then less than 5% of the significant results will be false positives. Not all of the true positive genes have adjusted FDR P-values <0.05, because not all true positive genes will have very small P-values.

1.6 Soft Lithography

Soft lithography is a fabrication technique that uses replica molding of soft elastomers to fabricate stamps and microfluidic channels. On can build microfluidic systems containing on-off valves, switching valves, and pumps completely out of elastomer. The most used elastomer Polydimethylsiloxane (PDMS) is a soft material with Young's modulus around 750 kPa, which is 5 orders of magnitude smaller than hard materials such as silicon and silicon nitride allowing

large deflections with small actuation pressures[103]. With such a soft material, device areas could be further reduced by more than 2 orders of magnitude compared with silicon-based devices. Other advantages of soft lithography include rapid prototyping, ease of fabrication, and biocompatibility[104-106].

To make microfluidic device with soft lithography, one has to use computer-aided design software to create a mask that has microfluidic channel features on it. Then, a light-sensitive material called photoresist is soft baked on top of silicon wafer. Mask, photoresist, and silicon wafer "sandwich" is developed/exposed under UV light with certain wavelength to make photoresist master as replica molding. Next, PDMS is cast on top of the photoresist master, and hardened by baking to form PDMS stamp. Fluidic and actuation channel PDMS stamps can be fabricated same way, but the fluid channel layer is several millimeters thick to allow reliable connections with macroscopic elements such as tubing and syringes. Finally, the fluidic layer is thermal bounded to the actuation layer, and the PDMS conjugate and substrate (e.g., glass) are oxidized with plasma and sealed[107, 108] (**Figure 1.9**).

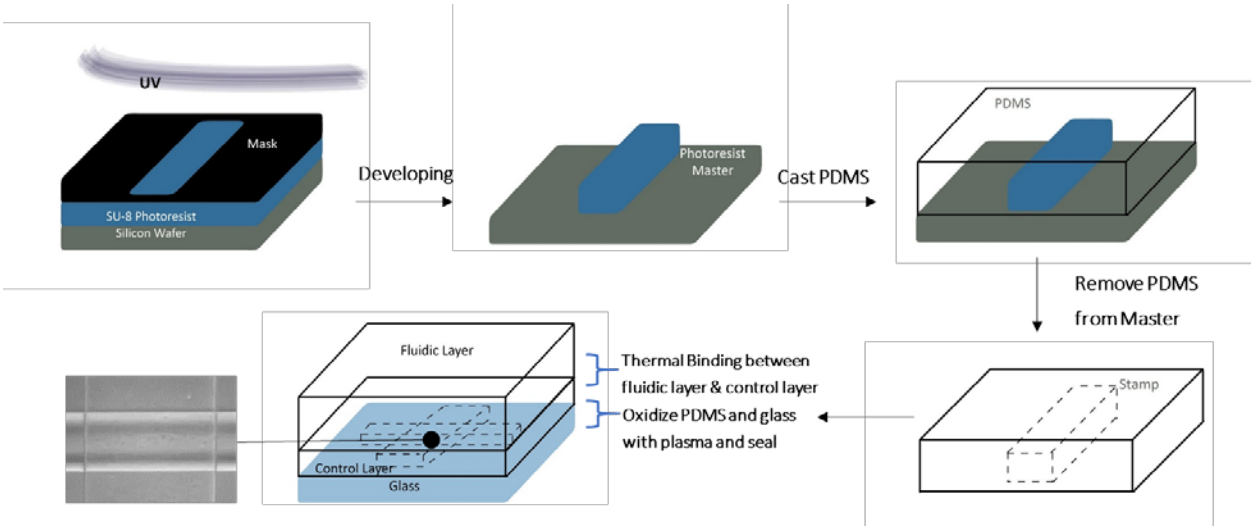


Figure 1.9 Multilayer soft lithography workflow.

There are positive and negative photoresists. When expose to light, positive photoresist becomes soluble to the photoresist developer. The portion of the photoresist that is unexposed remains insoluble to the photoresist developer. In contrast, when negative photoresist is exposed to light, it becomes insoluble to the photoresist developer. The unexposed portion of the photoresist can be dissolved by the photoresist developer (**Figure 1.10**). Most importantly, negative photoresist (e.g., SU-8) does not reflow at post exposure bake, it keeps rectangular channel shape. In this case, when pressure is applied to the valve, the rectangular channel will only be partially closed to stop particle flow while permitting liquid passage. Positive photoresist (e.g., AZ) will reflow at 140 °C (5 min), it obtains rounded channel shape (**Figure 1.11**). When pressure is applied to the valve, the rounded channel will be fully closed to stop both liquid and particles[109](**Figure 1.12**).

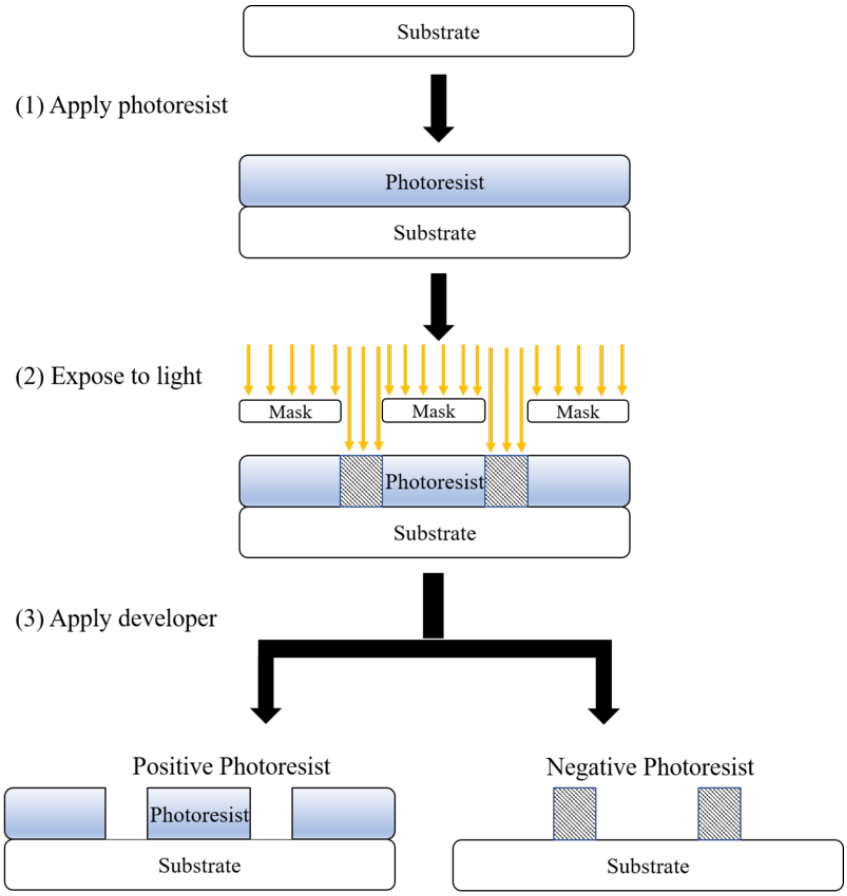


Figure 1.10 Positive vs negative photoresist (adapted from https://en.wikipedia.org/wiki/Photoresist).

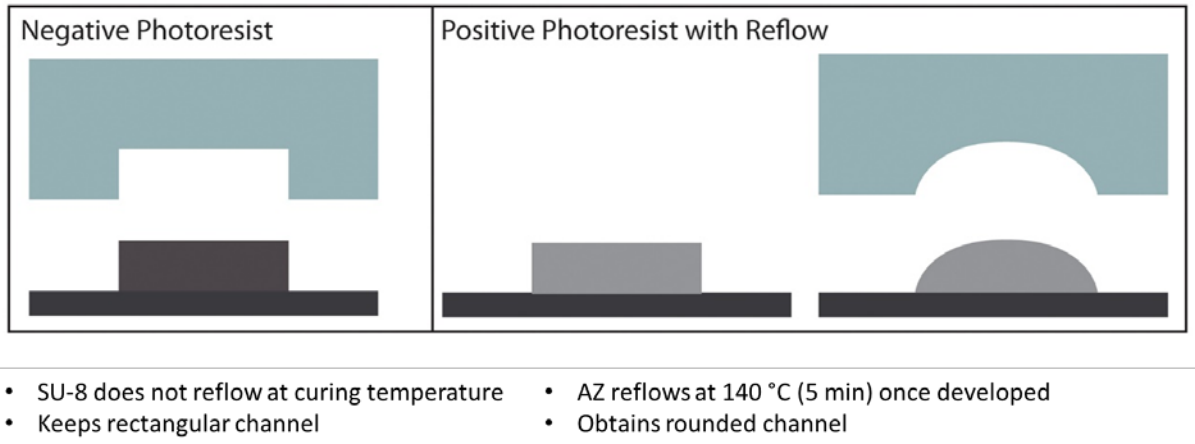


Figure 1.11 Rectangular and Round Channel (adapted from Studer et al. *J. Appl. Phys.* 95 (2004) 394)[110].

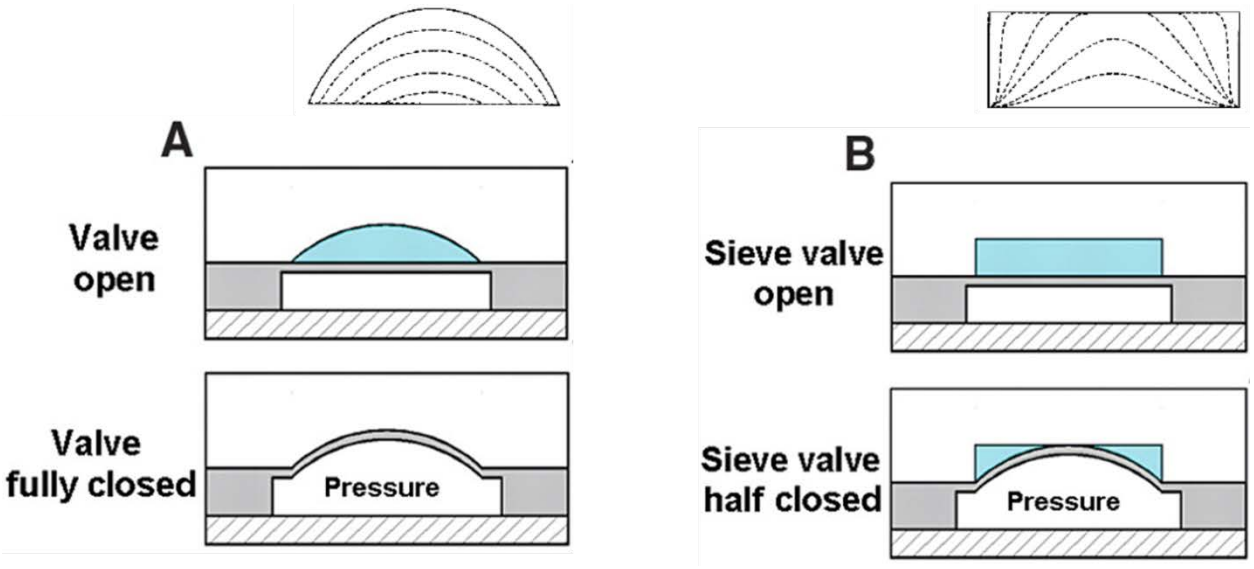


Figure 1.12 Fully closed and partially closed valve (adapted from Unger et al. *Science* 288 (2000) 114)[111].

When comes to microfluidic valve design, there are two different types of valve geometries: push-down and push-up[110, 111](**Figure 1.13**). A push-down valve has the actuation channel on top of the fluidic channel. When pressure is given, the membrane in between channels is pushed down to close the fluidic channel. However, actuation pressure highly depends on the thickness of fluidic channel membrane with this design. A push-up valve on the other hand, has the fluidic channel on top of the actuation channel. When pressure is given, the membrane in between channels is pushed up to close the fluidic channel. This design is advantageous because the actuation pressure is independent of thickness of fluidic channel membrane.

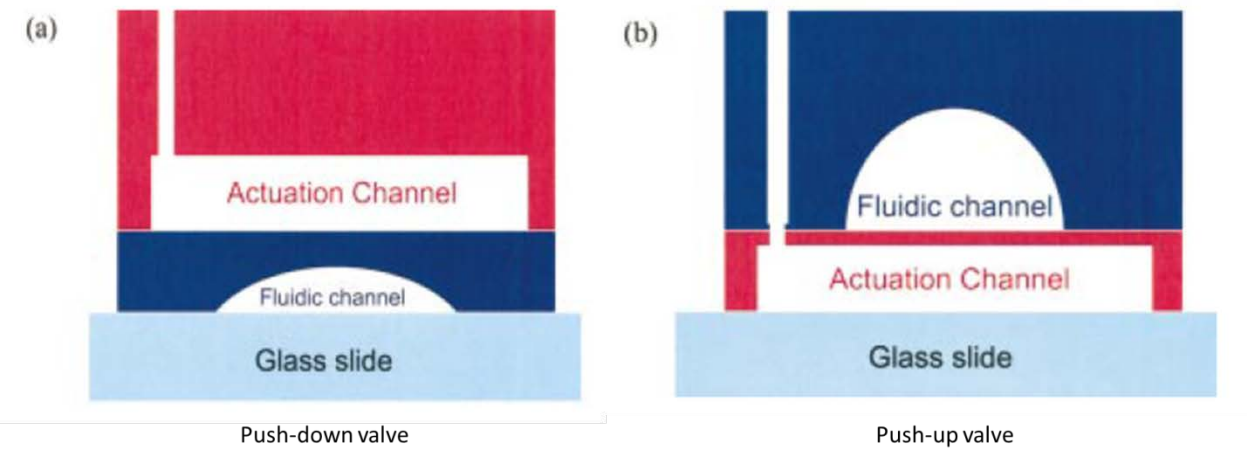


Figure 1.13 Push-down and Push-up Valve (adapted from Studer et al. *J. Appl. Phys.* 95 (2004) 394)[110].

1.7 Microfluidic ChIP Assay

Microfluidics enables controlling and transferring tiny quantities of liquids to allow biological assays to be integrated on a small chip. Microfluidic chips can be made from polydimethylsiloxane (PDMS) by soft lithography at low cost[111, 112]. Microfluidic chips offer reduction in sample amounts, high level of integration and automation, and high throughput. Microfluidics provide an automated platform for performing ChIP assays based on a low number of cells with a short assay time. Automation not only helps to standardize and simplify the ChIP process, but also eliminates errors from pipetting. Moreover, a tiny volume inside a microfluidic device tends to create a high concentration that benefits processes such as immunoprecipitation. Although a significant amount of effort has been directed toward DNA/RNA analysis using microfluidics[113-118], the reports on microfluidics-based ChIP assays (or epigenetic/epigenomic studies) have been scarce until recently.

1.8 Transgenic Mouse of Mammary Cancer

A transgenic mouse is a mouse which genome was modified by insertion of the modifying DNA (a trans-gene). Transgenic mice are good models for human disease since their tissues and organs are similar to human and they almost have same genes that operate in human. For example, transgenic mice that develop mammary tumors are good models to study human breast cancer. They have also been used to study diabetes, heart disease, obesity, arthritis, and Parkinson disease.

There are two common approaches to create transgenic mouse. The first involves pronuclear injection of modifying DNA into a single cell of the mouse embryo[119, 120]. The second involves transforming embryonic stem cells growing in tissue culture with the modifying DNA[121]. Mice used in our study were created by first method. Briefly, fusion gene constructs in which a sequence of interest is placed upstream of a polyadenylation signal sequence and downstream of a cell-specific promoter, are used for targeting of transgene expression to different cell types. The fusion gene construct is then microinjected into pronuclei of a fertilized ovum. Embryos are re-implanted in the oviducts of pseudopregnant recipient mice. Homologous recombination (in which nucleotide sequences are exchanged between two similar or identical molecules of DNA) or transposition occurs in between a chromosome and inserted modifying sequence. Progenies then can be found that are either homozygous transgenic or heterozygous transgenic. In eukaryotic cells, there are two matching sets of chromosomes, namely diploid. If both alleles of a diploid are the same, then it is called homozygous at that locus. If they are different, it is called heterozygous at that locus.

The C3(1)/SV40 T-antigen transgenic mouse model used in our study contains a recombinant gene expressing the simian virus 40 early-region transforming sequences under the regulatory control of the rat prostatic steroid binding protein C3(1) gene. The expression of TAg in the mammary epithelium results in progressive lesions and tumor development in all female mice. In this mouse model, atypia of the mammary ductal epithelium develops at about 8 weeks of age, representing low-grade mammary intraepithelial neoplasia (MIN). The atypical lesions progress to high-grade MIN (resembling human DCIS (Ductal carcinoma in situ)) at about 12 weeks of age. Invasive carcinomas are finally observed at about 16 weeks of age[122, 123](**Figure 1.14**).

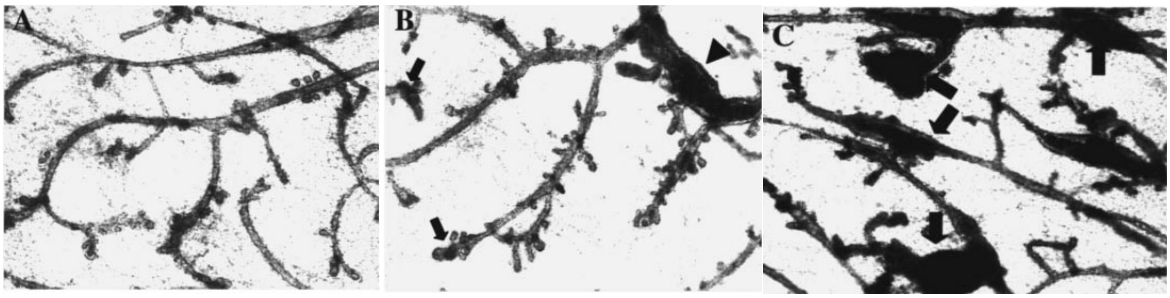


Figure 1.14 Tumorigenesis in the C3(1)/SV40 T-antigen transgenic mouse model (adapted from Green et al. *Oncogene* 19 (2000) 1024)[122].

1.8.1 Genotyping

Genotyping is the process to examine differences in the genetic make-up (genotype) from an individual to another individual's sequence or a reference sequence using biological assays. It reveals the alleles an individual has inherited from their parents. When genotyping transgenic mice, a single genomic region is needed to determine genotype.

There are different ways for collecting tissue samples for genetic (PCR) analysis of mice and rats. For example, DNA can be obtained from toe amputation, ear punch, tail snip, blood, hair and fecal samples, and oral or rectal swabs. Tail snip is a commonly used method for collecting tissue for DNA analysis. No anesthesia or analgesia is required for tail biopsies (<3mm) when mice are less than 25 days of age. General or local anesthesia is only considered when mice are more than 25 days of age to alleviate pain and distress.

1.8.2 Histology of Tumorigenesis

Histology is study of microanatomy of cells and tissues from animals or plants. Tissue samples from different stages of tumorigenesis are normally first fixed with 10% formalin to preserve tissue from degradation, and to maintain the structure of the cell and sub-cellular components. Alternatively, tissue samples also can be flash frozen with liquid nitrogen. Samples are then transferred through baths of progressively more concentrated ethanol to remove the water from the tissue. A hydrophobic clearing agent (e.g., xylene) is used to remove the alcohol. Finally an infiltration agent/embedding media (e.g., paraffin wax, epoxy resin) is used to replace the xylene, and to solidify the samples for sectioning. Next, the tissue samples are placed into molds with liquid embedding material for external embedding. Heating is applied to epoxy resins and cooling is applied to wax to further harden them. Flash frozen samples, are placed into molds with water-based embedding material (e.g., optimal cutting temperature compound) to be further frozen into hardened blocks. Formalin-fixed, paraffin-embedded (FFPE) tissues may be stored indefinitely at room temperature, and nucleic acids (both DNA and RNA) may be recovered from them decades after fixation. However, frozen samples have shorter storage time, typically 2

to 3 years at -80°C. Embedded tissue blocks are then placed on microtome to be sectioned into 4 micrometer thick, and mounted on a glass microscope slide to be stained. Frozen tissues can be sectioned using a microtome mounted in a refrigeration device known as a cryostat[124].

Since biological tissue has little inherent contrast in either the light or electron microscope, staining is employed to give both contrast to the tissue as well as highlighting particular features of interest. Hematoxylin eosin (HE) staining is the most common staining technique in histology. It is a bichromatic stain. The detailed topographic stain reveals the whole structure and morphology of cells in a tissue sample. Hematoxylin is a cationic stain (basic/positive stain) with an affinity for negatively charged cell compounds (DNA/RNA). It stains specifically the nuclei of cells in dark blue/purple. Eosin is an anionic stain (acidic/negative stain) with an affinity for positively charged cell compounds (Proteins). It stains the cytoplasm of cells pink, muscles dark red, collagen pale pink, and mitochondria pale pink[125].

2. Study of RNA polymerase II Transcriptional Regulation and Estrogen Receptor α binding with Microfluidic ChIP-seq Assays

2.1 Introduction

Higher organisms have evolved a sophisticated and complicated transcription mechanism. Briefly, transcription starts with recruitment of an RNA polymerase II (Pol2), other general transcription factors, and mediators to form preinitiation complex (PIC) (**Figure 2.1**). Transcription continues following steps of initiation, clearance of Pol2 from the site of initiation, promoter-proximal pausing of Pol2, escape from pause, elongation, and termination. Pol2 binding is the most important sign of transcription. Despite histone modifications and transcription factors binding, transcription would not happen without recruitment of Pol2. Studying ChIP-seq on Pol2 provides a more thorough understanding of the gene regulation mechanism on the actual process of transcription. For example, promoter-proximal pausing of Pol2 could affect chromatin structure at promoter to alter gene activity. Different patterns of promoter-proximal pausing (paused and transcribed, not paused and transcribed, paused and not transcribed, not paused and not transcribed) also reveal information on the mechanisms of transcription of each gene[126](**Figure 2.2**). Moreover, when histone modifications or transcription factor binding regulate gene expression, it is not clear that such modification or binding would lead to transcriptional activation or repression[127]. Comparing ChIP-seq data on histone modifications or transcription factors with Pol2 data would reveal the effects of histone modification or transcription factor binding on transcription of each bound gene[128](**Figure 2.3**). Furthermore, mammalian genes use different promoter regions to generate tissue-specific

expression of different protein isoforms. For example, the protein isoforms LEF1 generated from the two promoters perform opposing regulation functions on Wnt target genes. ChIP-seq on Pol2 provides useful information on where transcribe starts[128, 129]. Therefore, it reveals different physiological processes associated with normal and diseased states in different tissues.

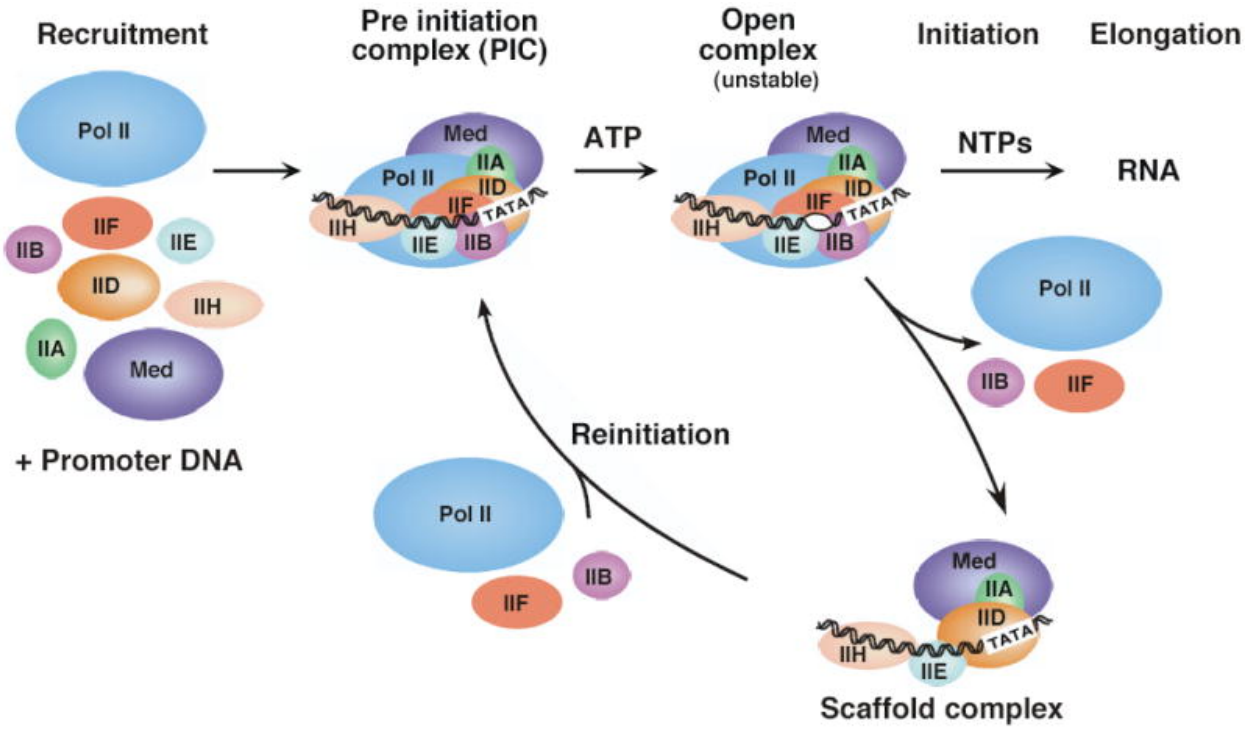


Figure 2.1 Transcriptional Regulation – RNA polymerase II Transcription Initiation Pathway (adapted from *Hahn Nat Struct Mol Biol.* 11 (204) 395)[130].

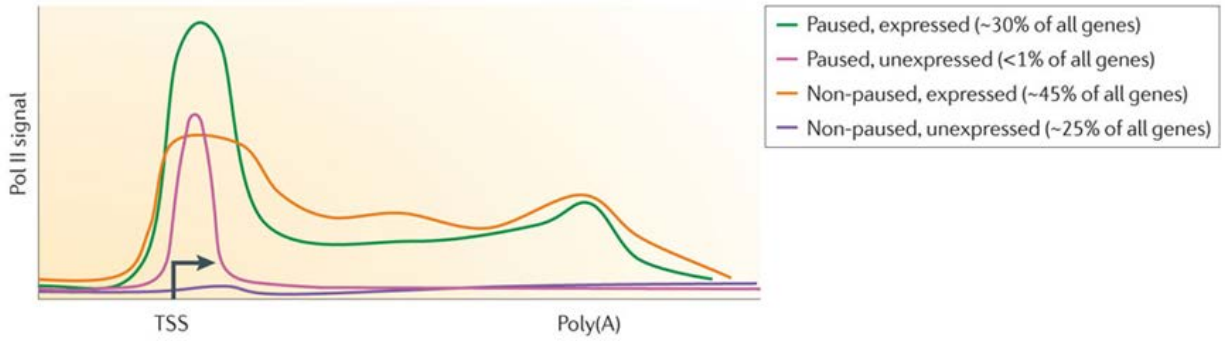


Figure 2.2 Pol2 pausing patterns. The different patterns provide information on the mechanisms of transcription of each gene (adapted from Adelman et al. *Nat Rev Genet.* 13 (2012) 724)[131].

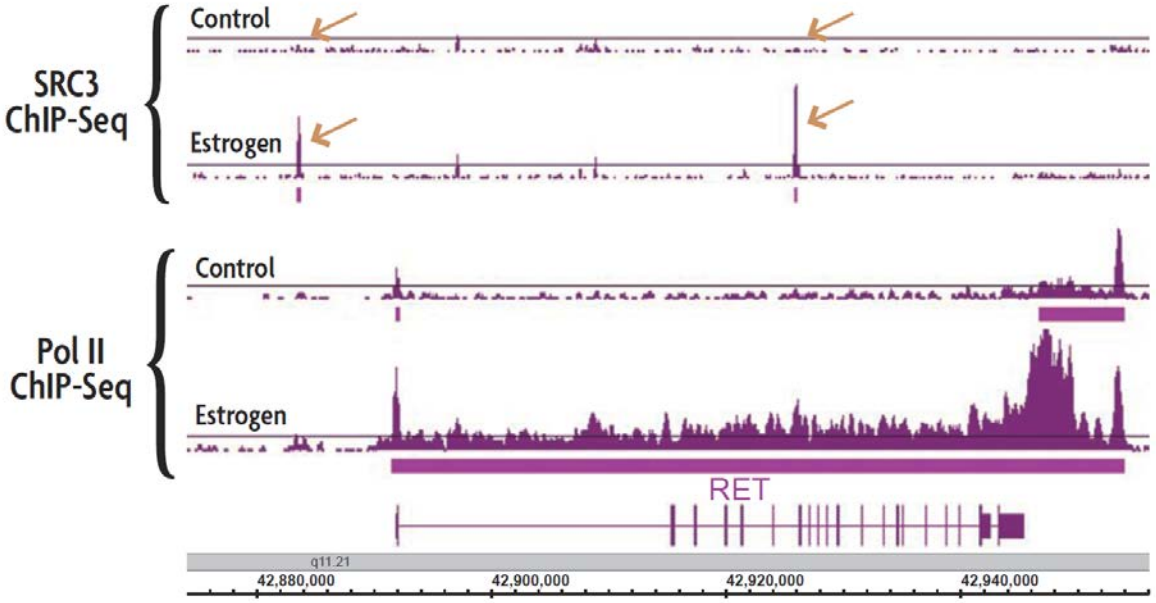


Figure 2.3 Combining transcription factor ChIP-seq data with RNA Pol II ChIP-seq reveal functional consequences of transcription factor binding (adapted from Mokry et al. *Nucleic Acids Res.* 40 (2012) 155)[132].

About 10% of the human genome codes for transcription factors and there are about 2600 known transcription factors (TFs)[133]. Many transcription factors turn genes on or off in response to stimuli. One of the many mechanisms in which this happens is ligand-activation. Estrogen Receptor α (ERα) is a classic example of ligand-activated transcription factor which is activated by estrogen. It contains a DNA Binding Domain (by which it binds to DNA), activation domain (by which it controls transcription), and ligand-binding domain (where the ligand estrogen binds). ERα rarely binds at promoters (< 5%) but mostly bind at enhancers located in introns and distal intergenic regions[134]. There are two mechanisms by which ERα mediates transcription: classical and nonclassical[135]. In the classical mechanism, ERα regulates genes such as TFF1, EBAG9, CASP7, and GREB1 by directly binding to DNA at estrogen response elements (EREs). In the non-classical mechanism, ERα regulates gene expression through protein-protein interactions with other direct DNA binding transcription factors such as Sp1, Ap1, CEBP, and Pitx1. Once activated by estrogen, ERα translocates into the nucleus and regulates transcription by one of these two mechanisms. One of its main functions is to regulate the activity of genes related to sexual maturation and gestation. In humans, ERα is mostly found in endometrium, mammary cells, ovarian stromal cells and the hypothalamus[136]. The clinical significance of Estrogen receptors is that it is over-expressed in around 80% of breast cancer patients, referred to as "ER-positive" tumours[136, 137]. Estrogen binding to the ERα receptors stimulates proliferation of the mammary cells which leads to increased DNA replication thus increasing the probability of mutations leading to breast cancer. The identification of ERα binding signatures on a genome-wide level via ChIP-Seq is therefore greatly necessary to understand the underlying biology in breast cancer.

Understanding the interaction between Pol2/ ERα and its binding sites in the genome yields important insights into its effectiveness and the transcriptional regulatory mechanisms. There have been a number of molecular techniques for identifying the interaction between regulatory proteins and DNA in general. However, in vitro methods such as electrophoresis mobility shift assays (EMSA) and DNase 1 protection assay cannot properly validate the in vivo relevance[138]. In comparison, chromatin immunoprecipitation (ChIP) assay has become the technique of choice for examining in vivo DNA-protein interactions over the years[80, 139-141]. Common ChIP assays confirm potential interactions between Pol2/ ERα and promoters of interest by conducting locus-specific qPCR analysis. On the other hand, unbiased and genome-wide mapping of Pol2/ ERα binding is also possible with techniques such as ChIP-ChIP and ChIP-seq[140]. Such examination provides a snapshot of the dynamic processes of Pol2 recruitment and regulation, without overexpressing any component.

In a typical ChIP experiment, DNA is first cross-linked with protein by cross-linking agents (e.g., formaldehyde), in order to freeze protein–DNA interactions. Subsequently, cells are lysed in order to release chromatin. Then, chromatin is sonicated/MNase digested to yield fragments of protein-bound DNA that are typically 200–600 base pair long. These fragments are then immunoprecipitated onto the surface of antibody-coated beads. The antibody here specifically targets a transcription factor, a specific modified form of histone, or a cytosine methyl group. Finally, the immunoprecipitated fraction is isolated. The cross-linking is reversed and the released DNA is assayed to determine the sequences[142]. The identification of the DNA sequences can be done by qPCR if there are known candidates. Alternatively, these binding sites can be

mapped at the genome scale by sequencing (ChIP-seq) using high-throughput sequencing technology (e.g., Illumina HiSeq 2500 System)[81, 89].

Although ChIP-qPCR/seq has been tremendously useful for epigenetic/ epigenomic studies, the technology suffers from serious limitations. First, the key limitation is that it requires a large number of cells (>10^6 cells per IP for ChIP-qPCR and 10^7-10^8 cells for ChIP-seq)[141, 143, 144]. This is feasible when using cell lines. However, such requirement has become a major challenge when primary cells are used because very limited amounts of samples can be obtained from animals or patients. For example, there are around 10,000 naturally occurring T regulatory cells per murine spleen, and ~5000 per ml peripheral blood. For metastatic cancer patients, there are only about 1–10 circulating tumor cells per ml of whole blood. Furthermore, primary samples are typically mixture of different cell types. Isolating single cell types from a mixture brings further loss in the sample amount. Second, the outcome of a ChIP assay can be affected when a large cell number is used. Population heterogeneity may contribute to the large standard deviations among trials. Finally, most ChIP assays involve lengthy manual handling and they normally take 3–4 days or longer to finish. These cumbersome procedures not only create loss of materials but also introduce human/ technical errors that lead to large variations between replicates[145, 146]. In general, ChIP assays with ultrahigh sensitivity and high degree of automation and integration are highly desirable.

There have been several attempts to reduce the quantity of starting material required for ERα ChIP-seq. A carrier molecule (mRNA and Histone H2B mix) was used to enable ChIP-Seq from

10,000 cells[134]. How the carrier molecule exactly works to enhance ChIP-Seq signal is not clearly known but it is hypothesized that the bulky material of the carrier helps to retain the small amounts of relevant chromatin throughout the procedure and the carrier also reduces nonspecific binding by acting as a competitor for nonspecific binding sites on the magnetic beads and elsewhere. Single-tube linear DNA amplification (LinDA) was developed to study ERα binding for as low as 5000 cells[147]. In this method, DNA was ligated with poly T and and in vitro transcribed to RNA. The RNA was then reverse transcribed and amplified using the T7 promoter-BpmI-oligo(dA)15 primer. Moreover, other groups have successfully performed ChIP-Seq of histone modifications in microfluidics devices to further reduce input requirement. A microfluidic device with sensitivity of 1000 cells was used to study histone modifications of mouse early embryonic cells[148]. Microfluidic oscillatory washing based ChIP-seq (MOWChIP-seq) with sensitivity of a 100 cells was developed to study histone modification in GM12878 cells and fetal liver cells[149]. However, not much progress has been made in the field of low-input pol2/transcription factor ChIP-seq with microfluidic devices.

In this project, we developed a microfluidic ChIP-seq device with ultrahigh sensitivity to study Pol2 transcriptional regulation and ERα binding from scarce cell samples. By conducting IP at microscale, we dramatically increased the assay sensitivity to an unprecedented level (~50 K cells for pol2 ChIP-seq, and 2.5 K for ERα ChIP-seq). Such sensitivities are 3 orders of magnitude higher than the prevailing pol2 ChIP-seq assays. By conducting the immunoprecipitation (IP) in a tiny microfluidic chamber and due to high degree of automation, we drastically shortened the time for IP to less than 1.5 hours (compared to overnight in most ChIP protocols). Thus we decreased the assay time of ChIP-qPCR to less than 6 h. These

unprecedented sensitivities will permit some transformative applications of ChIP technology to primary samples. We reasoned that mechanical sonication shearing was harsher then MNase digestion and might cause damage to epitopes of interest with our ChIP-qPCR data. Therefore, to get high quality chromatin, we performed MNase digestion to mainly fragmentize chromatin and followed by a short sonication to further break nuclei membrane. We also optimized the technology to reach ultrasensitive with a lymphoblastic cell line GM12878. We found that the combined use of a packed bed of beads for ChIP and effective oscillatory washing for removing nonspecific adsorption/trapping is key to extremely high yield of highly-enriched ChIP DNA.

Our microfluidic technology is superior for conventional ChIP for serveral reasons. First, high concentrations from trace amounts of molecules can be built up inside the tiny volumes of the microfluidic chamber. Adsorption kinetics and completeness was facilitated by such high concentration. Second, IP beads occupy a large fraction of the tiny volume so that the surface area/volume ratio (15-40%) is tremendously improved when compared to 5% in the conventional ChIP[150]. The close proximity among beads greatly increases the efficiency and rate for chromatin adsorption to the bead surface due to the short diffusion lengths involved. The adsorption of a chromatin molecule among beads was rapid given that travel time $\tau_D \sim w^2/D$, where w is diffusion distance between two beads, and D is diffusivity. Third, by using microfluidic technique uniquely suited for bead manipulation at the microscale, we effectively remove nonspecific adsorption after high efficiency adsorption using microfluidic oscillatory washing. This is critical for producing high quality ChIP DNA that preserves desired biological information. Finally, the microfluidic device integrates various steps and minimizes material loss among steps.

The new capability of our technology will allow establishing genome-wide Pol2 binding profile with 5×10^4 cells and ERα binding profile with 2500 cells. Our technology dramatically widens the sample range for Pol2/ ERα binding profiling to include primary cell samples from scarce sources. Transcriptional regulation and its mechanism during disease development from tiny cell samples of a patient that used to be not accessible to the researchers and clinicians due to the technological limitation now could become attainable. With such information, one can build epigenomic signatures for disease diagnosis and prognosis in the context of personalized medicine.

2.2 Methods and materials

Fabrication of the microfluidic ChIP device

The microfluidic ChIP device is composed of a microfluidic chamber (~800 nl), connecting channels, and a micromechanical vale that can be partially closed to stop magnetic beads while allowing liquid to pass. The main chamber is in elliptic shape with a major axis of 6 mm, a minor axis of 3 mm, and a depth of 40 μm. 27 micro-pillars are spotted inside the main chamber to prevent collapsing of PDMS.

Multilayer soft lithography was adopted to fabricate microfluidic ChIP device. Briefly, two photomasks (one for fluidic layer, and one for control layer) that had desired microscale patterns were designed with computer aided design software FreeHand MX (Macromedia) and printed on

high-resolution (5,080 d.p.i.) transparencies. To make fluidic layer master (~40 μm thick), photoresist (SU-8 2025, Microchem) was spun on a 3-inch silicon wafer (978, University Wafer) at 500 rpm for 10s and 2500 rpm for 30s followed by soft bake at 65°C for 1 min and 95°C for 7 min. To make the control layer master (~50 μm thick), SU-8 was spun at 500 rpm for 10s and 1500 rpm for 30s, followed by the same soft bake condition. Each master covered with its photomask was UV exposed for 17s at 580 mW exposure intensity and followed by a post exposure bake at 65°C for 1 min and 95°C for 7 min. Masters were then developed in SU-8 developer for 2-3 min, rinsed with Isopropyl alcohol (IPA) and air blown to dry. To make fluidic layer stamp, PDMS (General Electric silicone RTV 615, MG chemicals) with a mass ratio of A:B = 5:1 was thoroughly mixed and incubated under vacuum for 1 hr. It was then poured onto the fluidic layer master in a Petri dish to a height ~5 mm thick. To make the control layer stamp, PDMS with a mass ratio of A:B = 20:1 was mixed, vacuumed for 60 min, and spun onto the control layer master at 1100 rpm for 35s to a height of 108 μm thick. Both stamps were partially cured at 80 °C for 30 min. The fluidic layer was then peeled off from the fluidic layer master and aligned to the control layer. Two-layer PDMS were thermally bounded by baking at 80 °C for 60 min, and then peeled off from control layer master. Inlets and outlets of the device were punched by a 2 mm hole puncher. Finally, the two-layer PDMS and a pre-cleaned glass slide were treated with oxygen plasma cleaner (PDC-32G, Harrick Plasma) and brought together to form closed channels and chamber. The device was then baked at 80 °C for 60 min to strengthen the bonding between PDMS and glass. Glass slides were cleaned in a basic solution (H2O: 27% NH4OH: 30% H2O2 = 5:1:1, volumetric ratio) at 75 °C for 2 h and then rinsed with ultrapure water and thoroughly air blown to dry.

Setup of the microfluidic device

The microfluidic experiment was monitored by a charge-coupled device (CCD) camera (ORCA-285, Hamamatsu) attached to the port of an inverted microscope (IX 71, Olympus). The experiment started with prefilling the control channel with water to prevent air bubble defusing into fluidic channel. The reagents were flowed into the inlet via perfluoroalkoxy alkane (PFA) high-purity tubing (1622L, ID: 0.02 in. and OD: 0.0625 in., IDEX Health & Science) driven by a syringe pump (Fusion 400, Chemyx). The micromechanical valve was actuated by a solenoid valve (18801003-12V, ASCO Scientific), which was connected to a pressure source (a gas cylinder or a compressed air outlet) and controlled by a data acquisition card (NI SCB-68, National Instruments) and a LabVIEW (LabVIEW 2012, National Instruments) program for its switching function. The pressure (30 – 35 p.s.i) that was applied to control channel deformed PDMS membrane between fluidic channel and control channel and formed a partially closed valve to stop beads while allowing fluids to pass. The oscillatory washing was conducted by connecting inlet and outlet of microfluidic device to solenoid valves via PFA tubing. A digital pulse signal was first created in LabVIEW program, and it got converted to electric signal by data acquisition card, then it was sent out to solenoid valves to achieve automation.

Preparation of MNase digested/sonicated chromatin

To prepare from 4×10^6 GM12878 cells for Pol2 ChIP-seq

4×10^6- cell samples were centrifuged at 1,600g for 5 min at room temperature in a swing bucket centrifuge with soft deceleration. Cells were then washed twice with 1.0 ml 1× Phosphate-buffered saline (PBS) (14190-144, SigmaAldrich) at room temperature by centrifugation at

1,600g for 5 min and resuspension. Cells were resuspended in 1.0 ml 1× PBS and cross-linked by adding 67 µl of 16% freshly prepared formaldehyde (28906, Thermo Scientific) to 1 ml cell suspension (to a 1% final formaldehyde concentration) for 5 min. Cross-linking was terminated by adding 71 µl of 2 M freshly prepared glycine (to a 0.125 M final glycine concentration) and shaking for 5 min at room temperature. Two-step fixation was achieved by first adding 4 µl of 250X cross-link gold (C01019027, Diagenode) to 1 ml cell suspension in 1× PBS/$MgCl_2$ (to a 1X final cross-link gold concentration) for 30 min (Magnesium-containing PBS was used to ensure that the monolayer remained attached during the fixation), then followed by formaldehyde and glycine treatment as described above. Cross-linked cells were pelleted at 1,600g for 5 min and washed with 1 ml precooled PBS buffer and resuspended in 1 ml TM2+ buffer (10 mM Tris, pH 7.4, 2 mM $MgCl_2$, 0.5 mM PMSF). To release nuclei, 64 µl 10% NP-40 was added to the fixed cells (to a 0.6% final NP-40 concentration). Cells were kept on ice for 4 min with occasional mild vortexing. Nuclei were pelleted at 1,600g for 5 min, washed with 1 ml TM2+, and resuspended in 100 µl TM2+IC (TM2+ with 1× Roche EDTA-free protease inhibitor cocktail). MNase digestion was conducted by preheating each sample at 37 °C for 3 min. 25 U of Micrococcal Nuclease Solution (88216, Thermo Scientific) and 1 µl 0.1 M $CaCl_2$ (to a 1 mM final $CaCl_2$ concentration) were added into each sample. Samples were incubated at 37 °C for 6 min. Digestion was stopped by adding 4 µl 50 mM EGTA (to a 2 mM final EGTA concentration). Nuclei were pelleted by centrifugation at 15,700g for 1 min at 4 °C, and each sample was resuspended in 200 µl 80T+IC buffer (70 mM NaCl, 10 mM Tris, pH 7.4, 2 mM $MgCl_2$, 2 mM EGTA, 0.1% Triton X-100, 0.5 mM PMSF, 1× EDTA-free Roche Complete Protease Inhibitor Cocktail). To completely release the chromatin from nuclei, nuclear membrane was broken by sonication (Bioruptor300, Diagenode) with 3 sets of 30-sec pulses and

30-sec rest with power setting on high. Lysates were clarified by centrifugation at 16,100g for 10 min at 4 °C. Approximately 200 µl MNase digested/Sonicated chromatin in the supernatant was transferred to a new 1.5-ml LoBind Eppendorf tube (17014013, Denville) for MOWChIP-seq. From this stock chromatin preparation, samples equivalent to 10^6, 10^5 and 5×10^4 cells were divided into aliquots and diluted with IP buffer to give a final volume of 50 µl for MOWChIP-seq. Same amount of the sample was used as the input. After this procedure, we typically obtained ~3.3 pg DNA per cell from the pre-ChIP chromatin samples.

To prepare from $1x10^6$ MCF-7 cells for ERα ChIP-seq

Stock chromatin was prepared from 1 million cells . The cells were counted using a haemocytometer and centrifuged at 1600g for 5 min at room temperature. Media was aspirated and the cells were washed twice with PBS at room temperature by centrifugation and resuspension. Formaldehyde was freshly diluted to 1% using PBS and the cells were incubated at room temperature (15min) on a rocking platform. Formaldehyde was quenched by adding 66.7uL of 2M glycine and shaken for 5 - 10 min at room temperature on the rocker. Cells were collected by centrifugation (1600g for 5 min at 4 ℃), then washed twice with ice cold solution of PBS(+1X Protease Inhibitor Cocktail + 5 mM Sodium Butyrate). The cells were then centrifuged and resuspended in 130µL Sarkosyl Buffer(0.1% SDS, 1% Triton X-100,10 mM Tris HCl--pH7.4, 1mM EDTA--pH 8.0, 0.1% Na Deoxycholate, 0.25% Sarkosyl, 0.3M NaCl + 1X PIC+ 5 mM sodium butyrate) and were then put on ice for 10-15 min. Cross-linked cells were sonicated with a Covaris M220 with peak incident power 50, duty factor 20%,cycles per burst 200,time, temp 20°C for varying times (12-20 min). After sonication, the solution was centrifuged for 10

min at 4°C, 14,000 x g. The supernatant was carefully transferred to a new tube. This lysate contained 1 million cells. An appropriate fraction was used for each reaction. For example for 10,000 cells reaction, 1/100 of this solution was used in a pool-split manner.

Preparation of immunoprecipitation (IP) beads.

Pol2

Dynabeads Protein A (10001D, Invitrogen) were used for immunoprecipitation. They are 2.8 µm superparamagnetic beads with recombinant Protein A (~45 kDa) covalently bound to the surface. 5 µl of 30 mg/ml beads (equivalent to 150 µg) were washed twice with freshly prepared IP buffer (20 mM Tris-HCl, pH 8.0, 140 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% (w/v) sodium deoxycholate, 0.1% SDS, 1% (v/v) Triton X-100) and resuspended in 150 µl IP buffer containing antibody. Beads were gently mixed with the antibody at 4 °C on a rotator mixer at 24 r.p.m. for 2 h. Antibody-coated beads were washed twice with the IP buffer and then resuspended in 5 µl IP buffer. We optimized the antibody concentration for the bead coating step on the basis of our ChIP-qPCR results. The optimal antibody concentration for MOWChIP-seq with Anti-RNA polymerase II CTD repeat YSPTSPS (phospho S5) antibody (ab5131, Abcam) was 6.67 µg/ml for $5x10^4$cells, 10 µg/ml for 10^5 cells and 13.33µg/ml for 10^6 cells. These conditions were equivalent to using 1000, 1500 and 2000 ng of antibody in the preparation of 150 µg IP beads.

ERα

On ice, 2.5 µL of Protein A (Thermo Fisher Cat. No. 10001D) and 2.5 µL of Protein G (Thermo Fisher Cat. No.100003D) were mixed together and placed on a magnetic rack. The supernatant was removed and the beads were washed 3 times with 125 µL of ice-cold PBS+ 0.5% BSA. The beads were then resuspended in 125 µL of PBS-BSA. ERα Antibody (Santa Cruz Biotechnology Cat. No. sc-8002X) were added to the bead solution and the whole mixture was gently incubated with the antibody at 4°C on a rotator mixer at 24 rpm overnight. The antibody coated beads were then rinsed twice with 125 µL of PBS plus BSA (ice-cold). They were then resuspended in 5µL of Sarkosyl Buffer.

MOWChIP

Pol2

The MOWChIP procedure started with rising the fluidic chamber with the IP buffer at a flow rate of 20 µl/min for 30 s. After found no air bubble in the microfluidic chamber, the micromechanical valve was partially closed. The 5 µl (optimal condition) magnetic beads coated with antibody were flowed into fluidic chamber by the syringe pump at 20 µl/min, and aided with a cylindrical permanent magnet (NdFeB, D48-N52, 0.25 inch dia. and 0.5 inch thick, K&J Magnetics) to help beads travel through FPA tubing. The beads were packed against the partially closed valve to form a packed bed. The 0.5 µl of 100 mM PMSF (P7626-1G, Sigma-Aldrich) and 0.5 µl 100× protease inhibitor cocktail (P8340, Sigma-Aldrich) were freshly added to the 50 µl chromatin sample (to a 1 mM final PMSF concentration and 1× final protease inhibitor cocktail concentration). The chromatin sample was then flowed through packed bed with a flow rate of 1.5 µl/min. The immunoprecipitation step was finished around 60 min under this flow rate.

After ChIP, a low-salt washing buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% (v/v) Triton X-100) was introduced into the fluidic chamber at a flow rate of 2 µl/min for 2 min. Tubing that connected to inlet and outlet for oscillatory washing were prefilled with 20 µl low-salt washing buffer. Oscillatory washing was done by applying pressure pulses (each at 1 p.s.i., with a pulse width of 0.5 s and an interval of 0.5 s between two pulses) alternatingly at inlet and outlet of the fluidic chamber for 5 min while keeping the micromechanical valve open. The pulse signals were created in LabVIEW program, and got converted to electric signals by data acquisition card, then were sent out to solenoid valves to achieve automation of oscillatory washing. After oscillatory washing, beads were retained on one side of the fluidic chamber by a magnet. The unbound chromatin fragments and other debris/waste were flushed out of the microfluidic chamber by flowing a high-salt washing buffer (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% (v/v) Triton X-100) through the chamber at 2 µl/min for 2 min. Tubing that connected to inlet and outlet for oscillatory washing were prefilled with 20 µl high-salt washing buffer. Oscillatory washing was again conducted for 5 min. Beads were retained, and the unbound chromatin was flushed out by flowing IP buffer through the chamber at 2 µl/min for 2 min. Finally, IP beads were flowed out by IP buffer under a flow rate of 50 µl/min for about 5 min and collected into a 1.5-ml LoBind Eppendorf tube. The tube was then put on a magnetic stand, and the IP buffer was removed. Beads were kept in the tube and immediately proceeded for DNA extraction.

ERα

The microfluidic device was first rinsed with the Sarkosyl Buffer. The antibody-coated magnetic beads were then loaded into the microfluidic chamber using pressure-driven flow of the syringe pump. A magnet was also used to assist the flow of beads into the device. After the magnetic beads were loaded and packed into a bed, 50µl of sonicated chromatin fragments mixed with 1.05 µl of carrier [20:1 ratio of recombinant histone 2B (M2505S; New England Biolabs) and human mRNA (Invitrogen)] suspended in Sarkosyl Buffer were flowed in via syringe pump. This chromatin and carrier mixture was flowed through the packed bed of beads at a rate of 1.5 µl/min. After ChIP, the beads are washed sequentially with three ice-cold buffers : low salt RIPA 0 buffer(0.1%SDS, 1% Triton--X100, 10 mM Tris HCl pH7.6, 1 mM EDTA, 0.1% NaDOC), high salt RIPA 0.3M NaCl buffer (0.1%SDS, 1% Triton--X100, 10 mM Tris HCl pH 7.6, 1mM EDTA, 0.1% NaDOC, 0.3M NaCl) and a LiCl Buffer (250mM LiCl, 0.5% NP40, 0.5% Na DOC, 1mM EDTA, 10mM Tris HCl pH 8.1). After each oscillatory washing step, the beads were held in place by the magnet while the washed off debris was flushed out of the microfluidic chamber by a flow of fresh washing buffer of the subsequent step flow at a rate of 2 µl/min. In the end, the beads were flowed out of the microfluidic chamber with TE Buffer (pH 8) at flow rate of 50 µl/min and collected into an Eppendorf tube.

Extraction of ChIP DNA and input DNA

Pol2

Chromatin samples (either ChIPed chromatin on beads or input chromatin) were reverse cross-linked in reverse cross-linking buffer (200 mM NaCl, 50 mM Tris-HCl, 10 mM EDTA, 1% SDS, 0.1 M $NaHCO_3$) with 2 µl of 20 mg/ml proteinase K (26160, Thermo Scientific) added before

use (to a 200 µg/ml final proteinase K concentration) and had their volumes adjusted to 200 µl. Samples were then incubated at 65 °C for at least 4 hours or overnight. An equal volume (200 µl) of Phenol-chloroform-isoamylalcohol (25:24:1) was added to sample, mixed by vortexing, and centrifuged at 16,100g for 5 min at room temperature. DNA was extracted by collecting ~200 µl aqueous phase to a fresh 1.5 ml Eppendorf tube. The 50 µl of 10 M ammonium acetate was then added (to a 2 M final ammonium acetate concentration) resulting in a 250 µl solution. The 750 µl (3 times volume of 250 µl) of 100% ethanol with 2 µl of 20 µg/µl glycogen (10814010, Invitrogen) were finally added to carry out ethanol precipitation for DNA purification. Samples were placed at -20 °C for at least 2 hours or overnight for ethanol precipitation. Next, samples were centrifuged at 16,100g for 10 min at 4 °C, and supernatants were carefully removed. Samples were washed with 70% ice-cold ethanol without disturbing the pellets. The supernatants were removed again after another centrifugation at 16,100g for 5 min at 4 °C. The final pellets were air dried for 5 min and resuspend in 10 µl DNase-free water. This purified DNA can be used directly for ChIP-qPCR or for sequencing library construction. DNA concentrations were measured using a Qubit 2.0 fluorometer with dsDNA HS Assay kit (Q32851, Life Technologies).

ERα

The beads were placed on a magnetic rack and the buffer was replaced by 100 µl of fresh TE Buffer. The 1 µl of RNase (10 mg/mL, Roche) was added and the tube was incubated at 37 °C for 30 min to remove any residual RNA from the carrier mix. In the case of the input samples, because there were no beads in volved, the total volume of chromatin was made up to 100 µl with TE buffer and 1 µl of RNase was added). After the RNase digestion of immunoprecipitated

samples and input control samples was over, 5 µl of Proteinase K (Thermo Fisher Cat. no. EO0492) was added to each sample and the tube was placed at 65 °C for 8 hrs for reverse crosslinking. The reverse cross-linked DNA was then purified using standard phenol-chloroform extraction.

Construction of sequencing libraries

Sequencing libraries were prepared by Accel-NGS 2S Plus DNA Library Kit (21096, Swift Biosciences). This kit provides high complexity next-generation sequencing libraries and compatibility with ultra-low inputs (~10 pg). The library preparation process involved two steps of repairs and two steps of ligations to repair both 5' and 3' termini and sequentially attach Illumina adapter sequences to the ends of fragmented double-stranded DNA. Bead-based SPRI clean-ups were used to remove oligonucleotides and small fragments, and to change enzymatic buffer composition between steps. Different SPRIselect bead-to-sample ratios were used for different input quantities. PCR amplification (98 °C for 30 s, followed by 98 °C for 10 s, 60 °C for 30 s, 68 °C for 60 s for each cycle) was conducted to increase yield of indexed libraries. We used ~6 cycles for ChIP DNA from 10^6 cells, ~7 cycles for ChIP DNA from 10^5 cells, and ~9 cycles for ChIP DNA from $5x10^4$ or fewer cells. Library fragment sizes were determined using high-sensitivity DNA analysis kit (5067-4626, Agilent) on an Agilent 2200 TapeStation. The Kapa library quantification kit (KK4809, Kapa Biosystems) was used to determine effective library concentrations. The final concentrations of libraries submitted for sequencing were ~10 nM. The libraries were sequenced on an Illumina HiSeq 4000 with single-end 50-nt reads. Typically, 15–20 million reads were generated per library.

MNase digested followed by short sonication produced average DNA size around 200 bp (**Figure 2.4a**). After library preparation, the average peak size shifted to 350 as library construction added about 150 bp adaptor (75 bp on each side) to the ChIP DNA (**Figure 2.4b**).

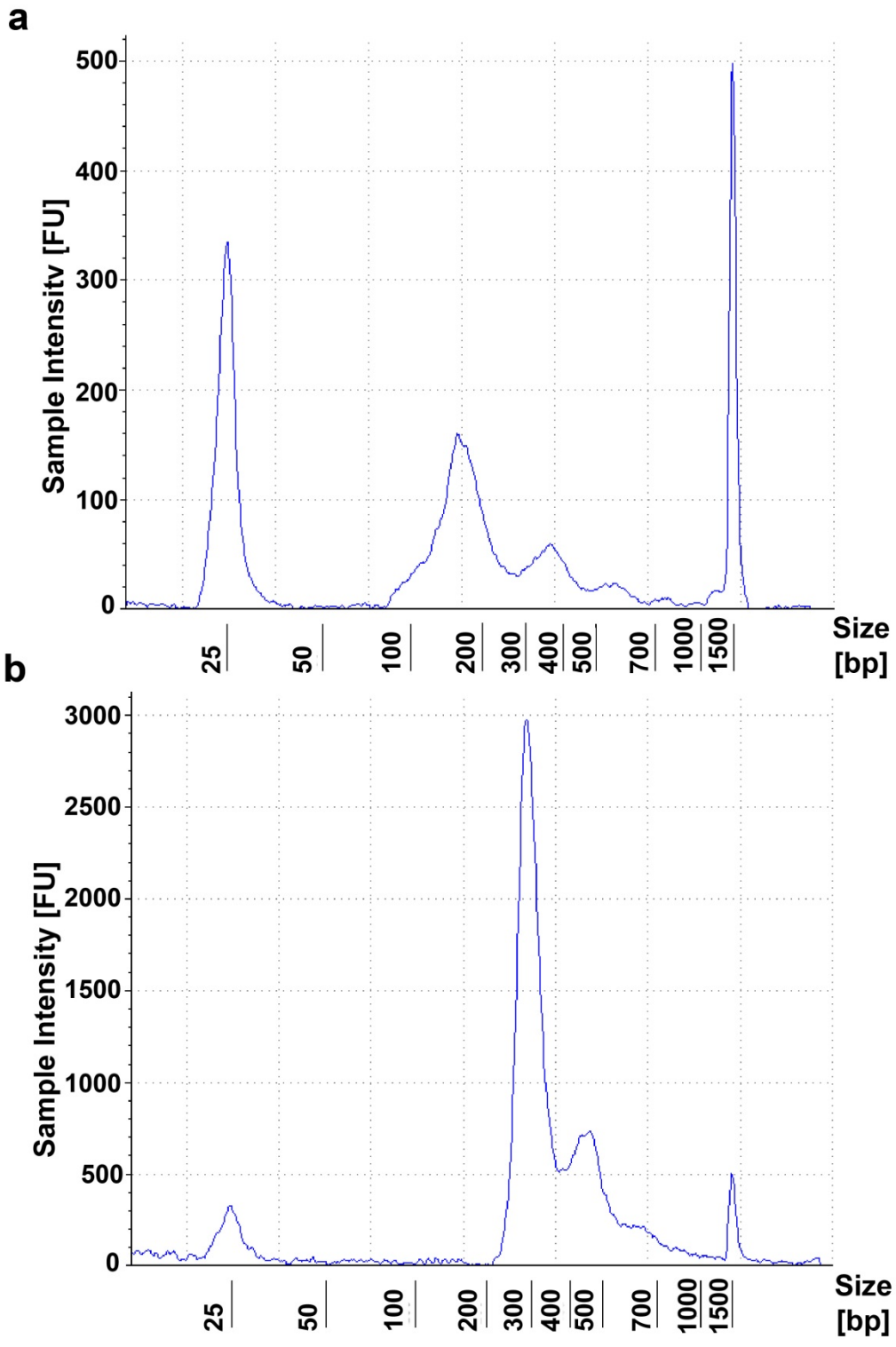

a
Sample Intensitv [FU]
500
400
300
200
100
0
25
50
100
200
300
400
500
700
1000
1500
Size
[bp]
b
Sample Intensity [FU]
3000
2500
2000
1500
1000
500
0
25
50
100
200
300
400
500
700
1000
1500
Size
[bp]

Figure 2.4 DNA fragment size profiles. **(a)** Sonicated DNA profile before ChIP ($4x10^6$ GM12878 with two-step fixation and 25U MNase digestion). **(b)** DNA profile after library preparation.

ChIP-qPCR data analysis

Real-time PCR was done using iQ SYBR Green Supermix (Bio-Rad, Hercules, CA, USA) on an CFX96 real-time PCR machine (Biorad) with C1000Tm thermal cycler base. All PCR assays were performed using the following thermal cycling condition: 95°C for 10 min followed by 40 cycles of (95°C for 15 s, 56°C for 30 s, 72°C for 30s). Primer concentrations were 400 nM. All primers were ordered from Integrated DNA Technologies (Coralville, IA, USA). The ChIP-qPCR results were represented as relative fold enrichment, which is the ratio of percent input between a positive locus and a negative locus. Percent input was calculated as the following equation: $2^{(C_qINPUT - C_qIP)} * 100\%$, where C_qINPUT is amplification cycle number run by real-time PCR for the chromatin sample without immunoprecipitation, C_qIP is the amplification cycle number for the chromatin sample with immunoprecipitation.

Cell culture

GM12878 cells were obtained from Coriell Institute for Medical Research. Species of origin of the cell line was confirmed by PCR targeting the gene encoding glucose-6-phosphate dehydrogenase. The donor subject has a single base-pair (G-to-A) transition at nucleotide 681 in exon 5 of the CYP2C19 gene (CYP2C19*2), which creates an aberrant splice site. Donor origin

of the cell line was confirmed using PCR against the point mutation. The cell line was tested for mycoplasma contamination using ABI MycoSEQ mycoplasma detection assay (Applied Biosystems). GM12878 cells were cultured in RPMI 1640 (Invitrogen, Carlsbad, CA, USA) with 15% fetal bovine serum, 100 U penicillin, 100mg streptomycin/ml (Invitrogen) at 37°C in a humidified incubator containing 5% CO_2 . Cells were sub-cultured every two to three days to maintain exponential growth.

MCF-7, an adherent estrogen receptor (ER) positive cell line, was cultured in DMEM (ATCC) with 10% FBS and 1% Penicillin-Streptomycin (PS) at 37°C and 5% CO2. Cells were harvested at 90% confluency.

ChIP-seq data analysis

ChIP sequencing reads were mapped to the human genome (hg19) using BWA (v0.7.17) with default parameter settings. Peaks of each ChIP sample were called against input by SPP (v1.14) with -npeak=300,000. 300,000 peaks were further filtered by IDR (v2.0.4) with -idr_thresh set at different thresholds.

To evaluate the quality of our ChIP-seq data, we selected ENCODE published data sets of RNA polymerase II using GM12878. ChIP-seq signals were normalized based on signal per million reads. Normalized signals in all promoter regions in the genome were extracted. Promoter regions were defined as upstream 2000 bp and downstream 2000 bp around transcription starting sites (TSSs). Promoter regions with zero signals in both data sets were excluded for computing Pearson correlation coefficient.

Samtools (-F 1804) was used to remove reads unmapped, not primary alignment, reads failing platform quality checks, and duplicates. Parameter (-q 30) was used to remove multi-mapped reads that has low mapping quality score. PCR duplicates were marked by Picard's MarkDuplicates.

Receiver operating characteristic (ROC) curve was generated as previously described. Briefly, we used peaks called from ENCODE as gold standard. We then defined positive regions as overlapping regions of peaks called from our experiment with gold standard at promoter regions, and negative regions as non-overlapping regions of peaks called from our experiment with gold standard at promoter regions. Both positive and negative regions are fed into ROCR (v1.0-5) for computing true positive rate and false positive rate.

2.3 Results and Discussion

Histone modifications can be easily detected by ChIP since histone is wrapped around by DNA in its original state. Other protein-DNA interactions (e.g. RNA polymerases, transcription factors) are difficult to be detected by ChIP owing to their low-occupancy and transient mechanisms[86, 151]. A two-step fixation method, which uses disuccinimidyl glutarate (DSG) to cross-link RNA polymerases /transcription factor to protein complexes before formaldehyde treatment to cross-link them to DNA has been proposed to improve the ChIP signal[50, 152](**Figure 2.5**). However, the number of cell used for ChIP was excessive in those studies ($6x10^6$)[50, 152, 153]. Fixation time also plays an important role. For high on-rate RNA polymerases and transcription factors, the longer the cross-linking time the more unbound chromatin sites will become occupied. However, if the

on-rate is low (e.g. Pol2), as fixation time increased, there would not have more occupied chromatin sites, suggesting a low quality ChIP-seq signal[151].

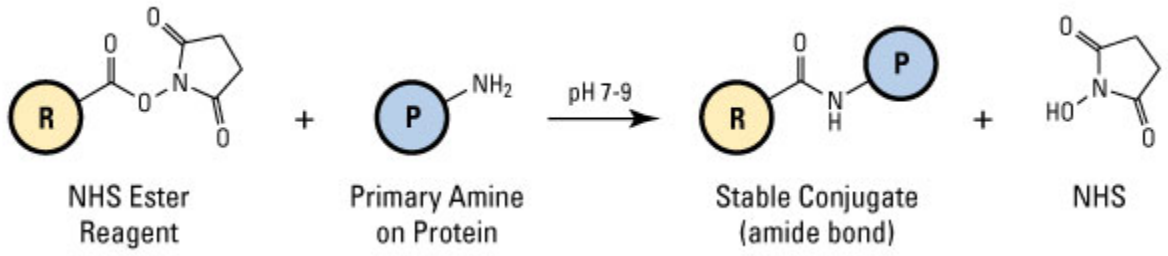


Figure 2.5 DSG chemistry (adapted from https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/amine-reactive-crosslinker-chemistry.html).

We started with 10^6 GM12878 cells, we first tried different fixation times and kept sonication time constant at 14 mins (**Figure 2.6**). We found that as fixation time increased, the amount of DNA recovered after fixation and sonication decreased from 70% DNA recovery for 5 mins down to 15% DNA recovery for 2 hrs. Therefore, we choose minimum fixation time as 5 mins to obtain largest DNA recovery since rapid binding dynamic saturate with ChIP-seq signal at around 1 minute[151].

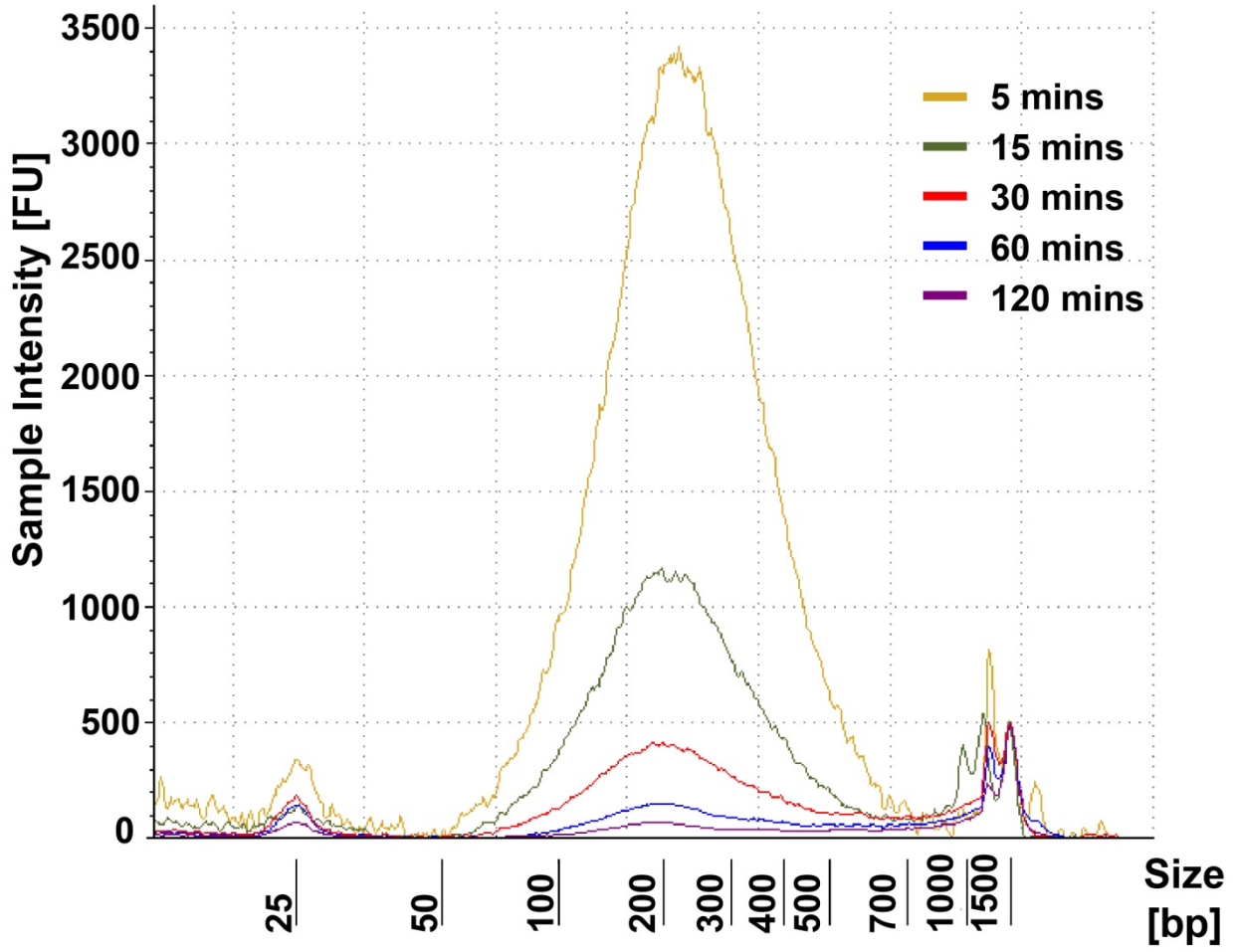


Figure 2.6 DNA fragment size profiles with fixed sonication time (14 min) but various fixation time.

We then kept fixation time constant as 5 mins, and tried different sonication times (**Figure 2.7**). As we gradually increased sonication time from 14 mins to 24 mins, the peak barely shifted to the left. Therefore, the selected sonication time range did not have much effect on average size. We chose minimal sonication time as14 mins for the experiment.

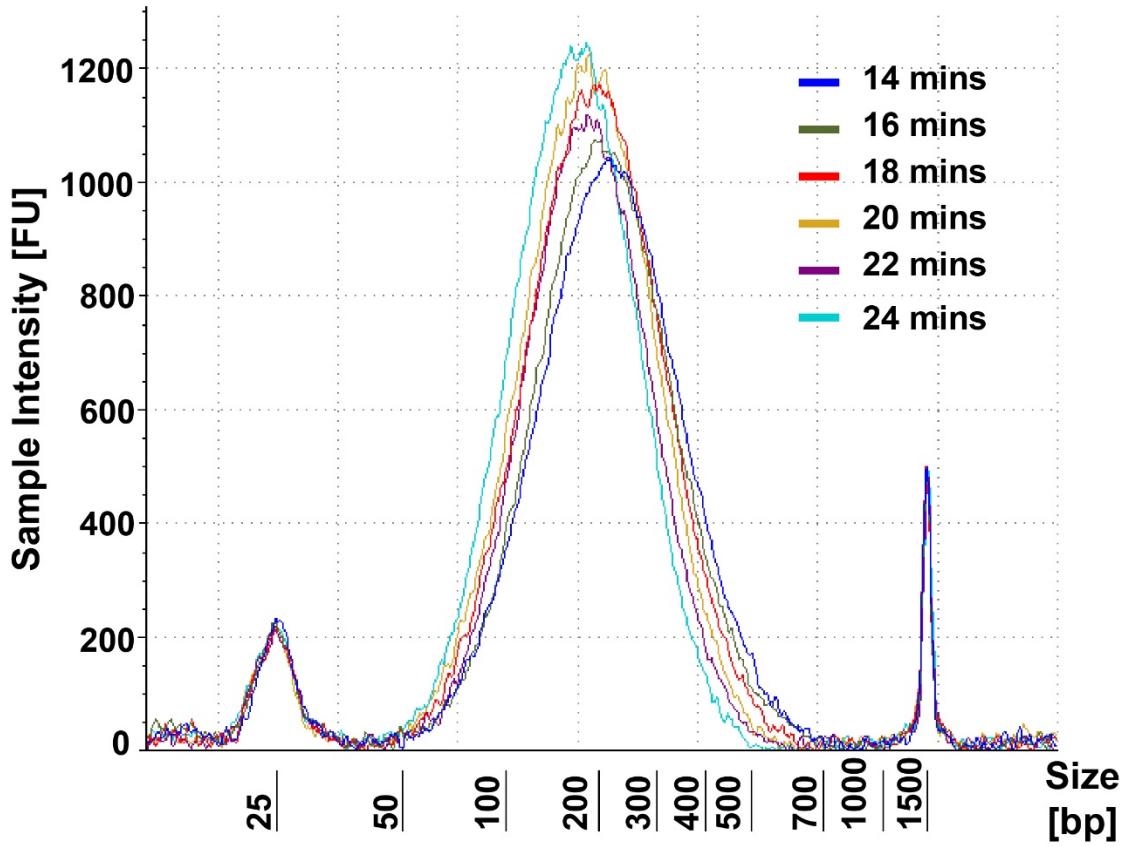


Figure 2.7 DNA fragment size profiles with fixed fixation time (5 min) but various sonication time.

Real time polymer chain reaction (qPCR) was used for quality control before ChIP-seq library preparation. We conducted ChIP-qPCR based on 10^6 cells that were extracted from cell line GM12878 with 13.3 μg/ml (previously optimized) anti-RNA polymerase II CTD repeat YSPTSPS (phospho S5) antibody. ACTIN-1 and GAPDH-2 are two known positive loci, MYOD and AFM are two known negative loci. The percent input was calculated, representing how much chromatin that had been pulled down by anti-pol2 antibody during the ChIP process when comparef to the input amount. Fold enrichment revealed how good the ChIP efficiency was from

positive loci to negative loci. The error bars were calculated as standard deviation from 3 ChIP-qPCR replicates. The results confirmed that percent inputs and fold enrichment at positive loci were substantially higher than percent inputs at negative loci, suggesting successful ChIP (**Figure 2.8**).

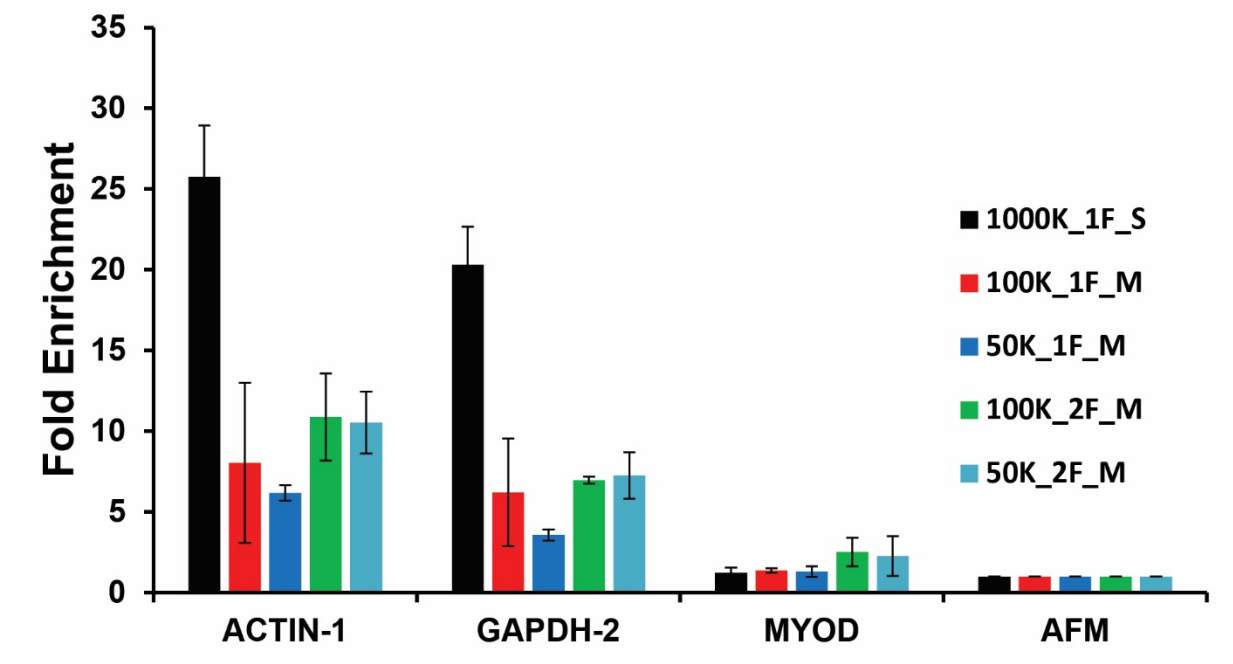


Figure 2.8 Fold enrichment (ratio between % input at positive loci and % input at negative loci) under different conditions. ACTIN-1 GAPDH-2 are two known positive loci, MYOD and AFM are two known negative loci. The error bars were calculated as standard deviation from 3 ChIP-qPCR replicates. One-step fixation as 1F, two-step fixation as 2F, sonication as S, and MNase digestion as M.

However, when we tried to reduce amount of cells for ChIP down to 10^5, the fold enrichment for two positive loci dropped down to around 1 to 2, and there was no ChIP-seq signal.

We suspected that mechanical shearing might be too harsh for the delicate protein like RNA polymerase II and cause damage to epitopes of interest. We then tried MNase digestion to fragment chromatin to avoid damaging the protein. However, chromatin recovery was only 15%. We added minimal amount of sonication (3 mins) after MNase digestion for the purpose of lysing the nucleus membrane to improve chromatin recovery up to 70%.

Using too much MNase will over digest the chromatin and too little will under digest the chromatin. The 4×10^6 GM12878 cells were used for MNase digestion. As the amount of MNase used in digestion was increased, not only were oligonucleotides other than mononucleotide digested (over digestion), but also the chromatin recovery was reduced (indicated as lower peak height in **Figure 2.9**). In the case of under digestion, there was no major peak and the majority of peaks were located in 700 bp -1500 bp (**Figure 2.10**). Using too much MNase will decrease the chromatin recovery, using too little MNase will end up under digest the chromatin. To balance this, we used 25 U MNase and followed by 3 sets of 30 sec on high power and 30 sec off with biorupter to obtain 70% recovered chromatin after MNase digestion.

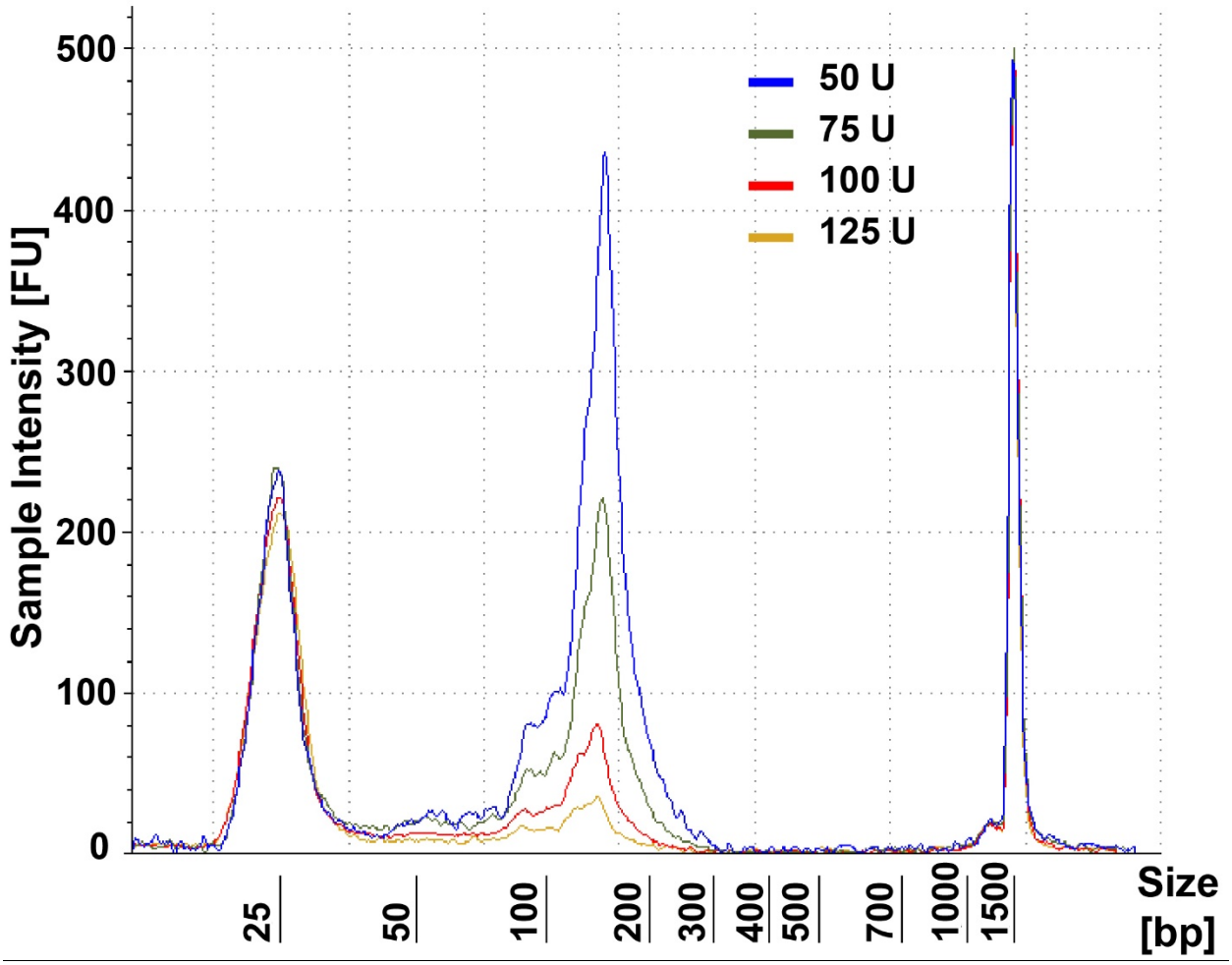


Figure 2.9 DNA fragment size over digestion profiles with two-step fixation and various amount of MNase used for digestion.

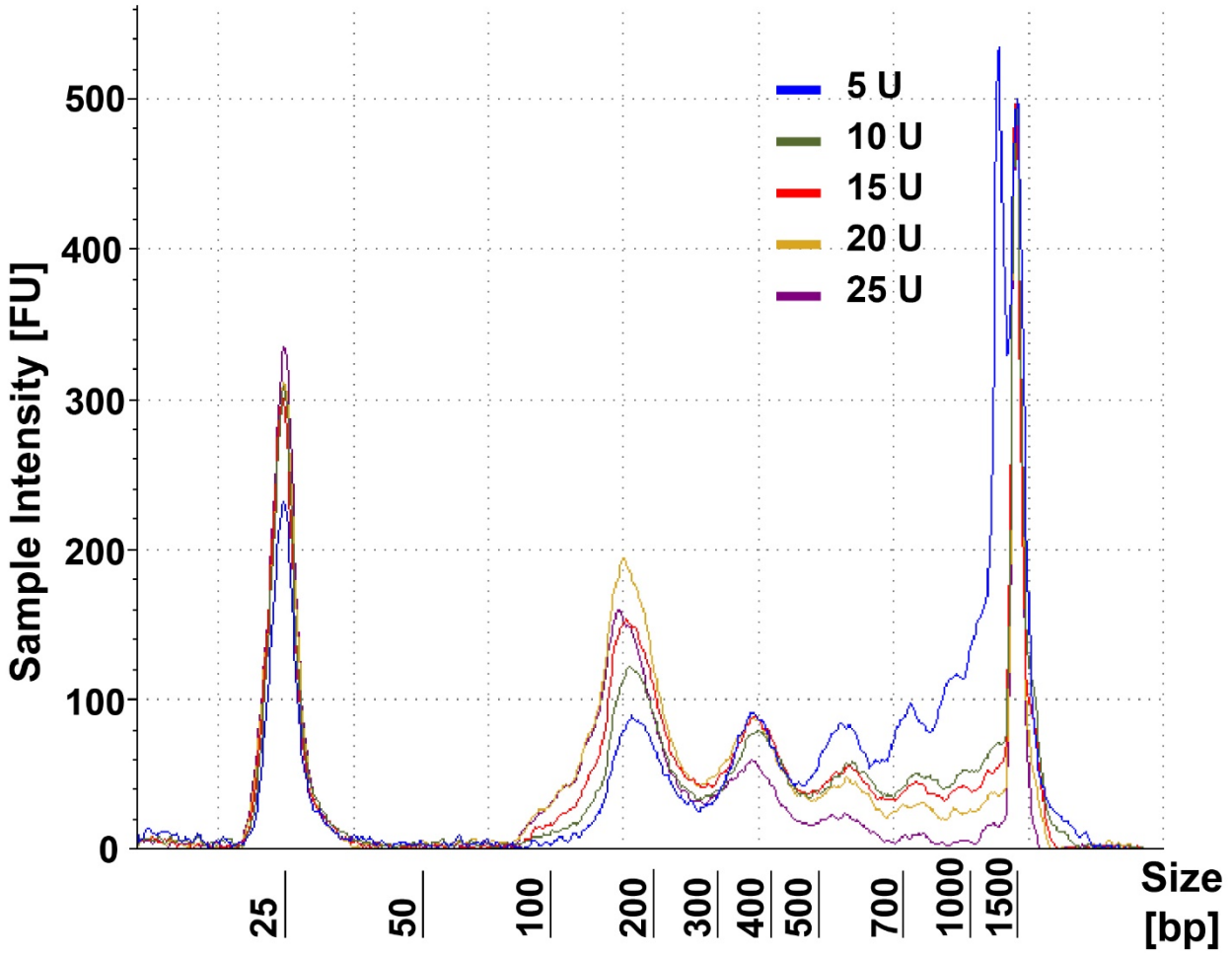


Figure 2.10 DNA fragment size under digestion profiles with two-step fixation and various amount of MNase used for digestion.

Using the optimized MNase digestion protocol, we conducted a one-step fixation ChIP-seq on RNA polymerase II with different input chromatin amounts: 100,000 and 50,000 GM12878 cells. As cell number decreased, the fold enrichment decreased as well. However, fold enrichment at positive loci were still higher than negative loci, suggesting ChIP was successful but efficiency decreased as cell number went down (**Figure 2.8**). We further conducted two-steps fixation

ChIP-seq on RNA polymerase II with different input chromatin amounts: 100,000, and 50,000 GM12878 cells. There was a slight improvement in fold enrichment at positive loci.

Our microfluidic ChIP process collected 8.25 ng, 0.75 ng and 0.3 ng DNA from 1000K, 100 K, and 50 K starting cells, respectively. ChIP-seq results were obtained using our microfluidic technology with various amounts of starting cell (1000 K-50 K) from human cell line GM12878. As expected, the quality of pol2 binding profiles slowly declined when the starting cell number decreased from 1000 K to 50 K (**Figure 2.11a**). One-step fixation for 100 K and 50 K had more noise comparing to corresponding two-step fixation.

Two replicates of one-step fixation followed by sonication 10^6 GM12878 cells RNA polymerase II ChIP-seq have averaged ~0.83 Pearson correlation to the ENCODE SRX 100530 data (600 million cells) and ~0.95 self-correlation (**Figure 2.11b**).

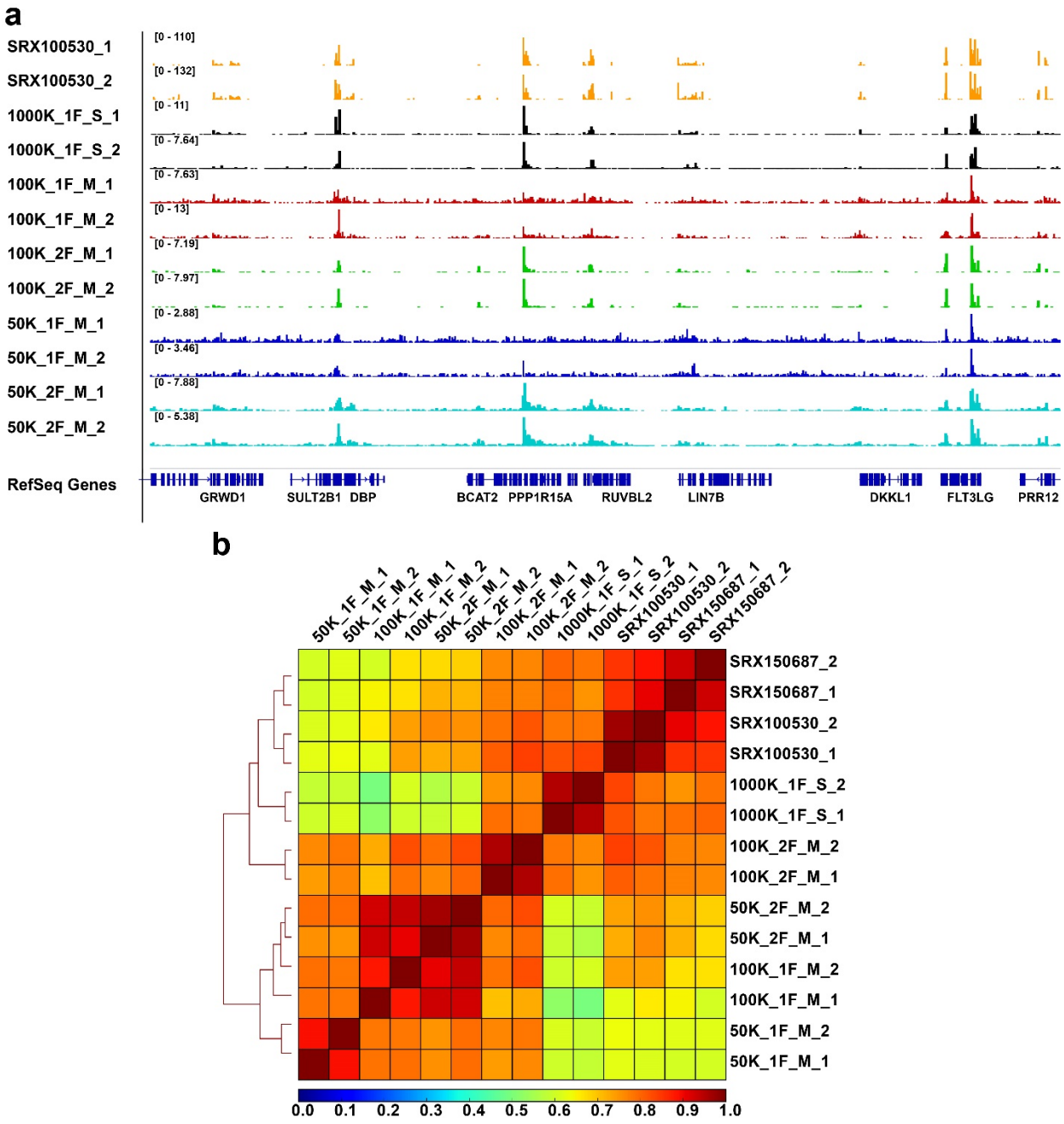


Figure 2.11 Microfluidic pol2 ChIP-seq data on GM12878 cell line. **(a)** Genome browser tracks for our microfluidic ChIP-seq using various amounts of cell number (1000 - 50 K). Two replicates were profiled for each sample. SRX100530 is GM12878 ENCODE pol2 data. One-step fixation as 1F, two-step fixation as 2F, sonication as S, and MNase digestion as M. **(b)** Genome-wide correlations among ChIP-seq data sets of various sample sizes. ChIP-seq genomic

coverage profiles at promoter regions (TSS +/- 2000 bp) were used for computing correlations. Colors represent Pearson correlation coefficients.

Two replicates of one-step fixation followed by MNase digestion 10^5 GM12878 cells RNA polymerase II ChIP-seq have averaged ~0.69 Pearson correlation to the ENCODE data and ~0.90 self-correlation. Two replicates of two-steps fixation followed by MNase digestion 10^5 GM12878 cells RNA polymerase II ChIP-seq have averaged ~0.83 Pearson correlation to the ENCODE data and ~0.96 self-correlation.

Two replicates of one-step fixation followed by MNase digestion 50,000 GM12878 cells RNA polymerase II ChIP-seq have averaged ~0.64 Pearson correlation to the ENCODE data and ~0.90 self-correlation. Two replicates of two-step fixation followed by MNase digestion 50,000 GM12878 cells RNA polymerase II ChIP-seq have averaged ~0.73 Pearson correlation to the ENCODE data and ~0.97 self-correlation.

Two replicates of no fixation followed by MNase digestion 50,000 GM12878 cells RNA polymerase II ChIP-seq have averaged ~0.2 Pearson correlation to the ENCODE data and ~0.96 self-correlation (data not shown).

Two replicates of two-step fixation followed by MNase digestion 10,000 GM12878 cells RNA polymerase II ChIP-seq have averaged ~0.3 Pearson correlation to the ENCODE data and ~0.96 self-correlation (data not shown).

As result, MNase digestion suggests better ChIP-seq signal than sonication. Two-steps fixation with MNase digestion resulted in the best ChIP-seq quality followed by one-step fixation with MNase digestion, and lastly, no fixation with MNase digestion.

With IDR set to 0.05, SPP identified 4162 peaks for 1000K_1F_S, 4840 and 6217 peaks for 100K_2F_M and 50K_2F_M, 446 and 290 for 100K_1F_M and 50K_1F_M. There were 12167 and 26039 peaks called from ENCODE SRX150687 and SRX100530. When relaxed the IDR threshold to 0.2, SPP called more peaks, specifically, 10896 peaks for 1000K_1F_S, 11338 and 23031 peaks for 100K_2F_M and 50K_2F_M, 3483 and 290 for 100K_1F_M and 50K_1F_M.

We then used the receiver operating characteristic (ROC) curve (IDR=0.05) to compare the data quality of our data to ENCODE data (SRX100530). The IDR statistic splits peaks presented in both replicates into either a reproducible group or an irreproducible group. The signals in the reproducible group have a better correlation coefficient, and are ranked higher than the irreproducible group. The area under the curve (AUC) values for pol2 are relative low compared to histone modification[149]. Even between two ENCODE sets (SRX100530 and SRX150687), the AUC value is only 0.69. The 100K_2F_M has AUC value of 0.68, which shows the best performance, followed by 1000K_1F_S 0.66, and 50K_2F_M 0.61 (**Figure 2.12**). The 100K_1F_M has AUC value of 0.53, suggesting lower data quality. Although 50K_1F_M had an AUC value of 0.73, the number is inconclusive since the peak number is very low.

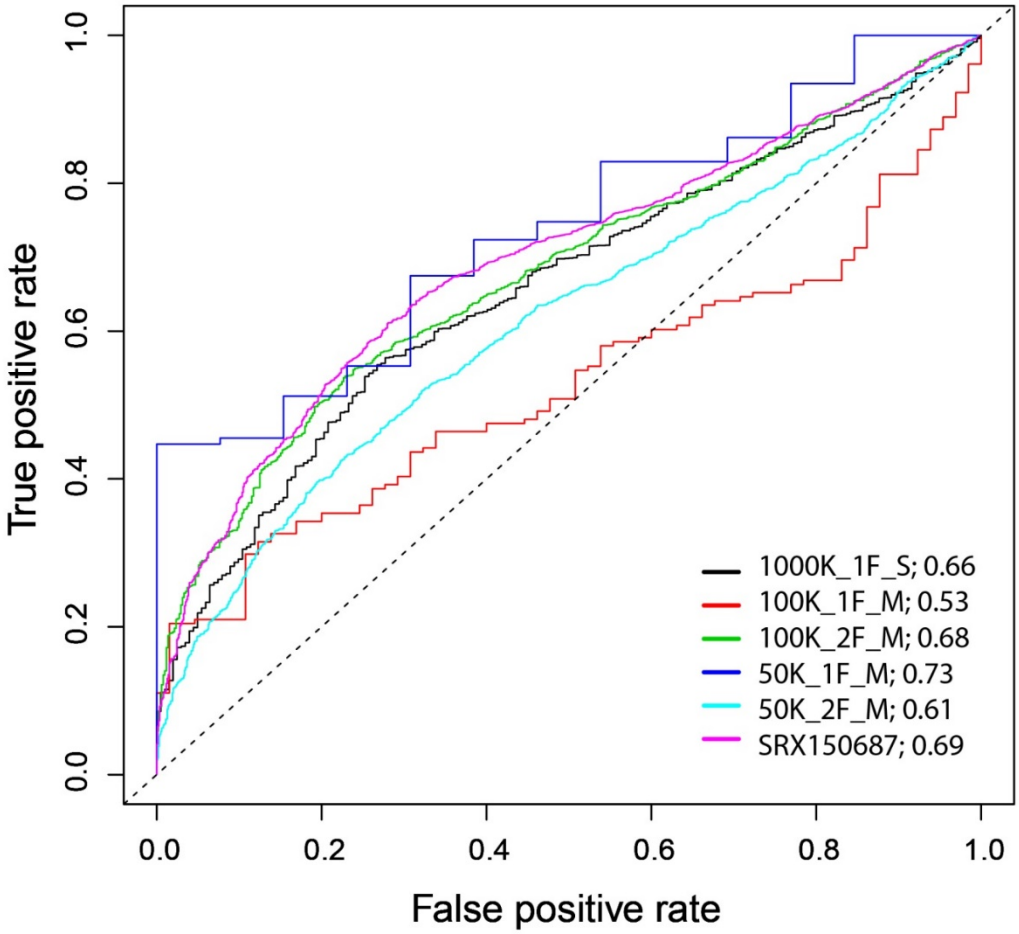


Figure 2.12 Receiver operating characteristic (ROC) curve of pol2 data. ROC curves were constructed by comparing the ChIP-seq data generated by various cell number and conditions to published ENCODE SRX100530 pol2 data generated using conventional protocols with millions of cells. SRX150687 is another set of GM12878 ENCODE pol2 data. Values shown are area under the curve (AUC) of two replicate experiments. One-step fixation as 1F, two-step fixation as 2F, sonication as S, and MNase digestion as M.

The performance (AUC) varies along with IDR threshold. To find the IDR threshold that provides the best performance. We plotted AUC of 100K_2F_M at different IDR thresholds (**Figure 2.13**). When IDR threshold was set to 0.05, the AUC value reached the maximum value of 0.68. Peak information lost from low-input data could be rescued by relaxing the IDR threshold since AUC only dropped 0.01 when IDR was changed from 0.05 to 0.2.

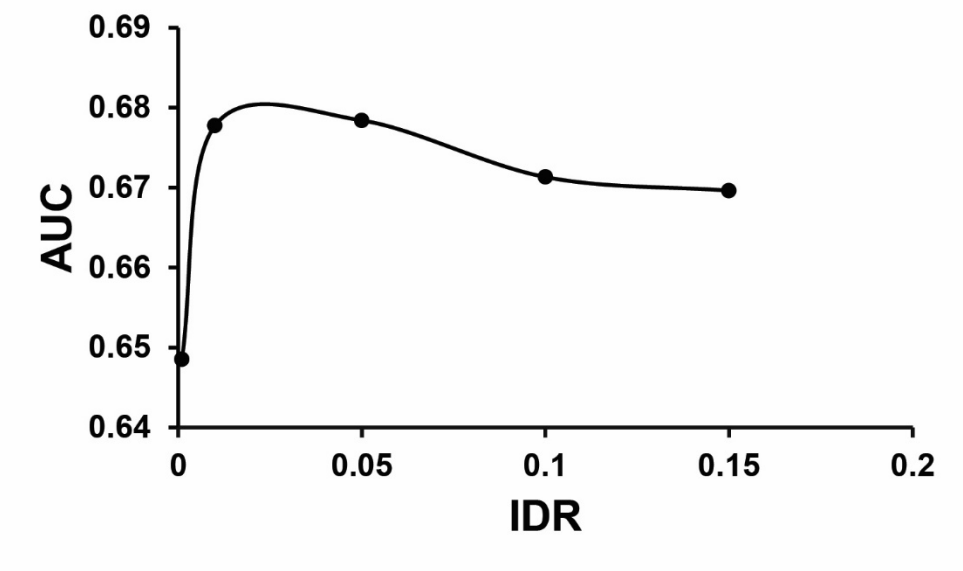


Figure 2.13 ChIP-seq performance at different IDR threshold. Area under the curve (AUC) was calculated from receiver operating characteristic (ROC) curve of 100 K cell with two-step fixation and MNase digestion.

Since ERα binds 200000 bp around transcription starting site[154], we compared microfluidic ERα untreated ChIP-seq data to ENCODE untreated data at TSS +/- 200000 bp regions (**Figure 2.14**). Two replicates of 10 K ERα untreated ChIP-seq have averaged ~0.86 Pearson correlation to the ENCODE data and ~0.98 self-correlation. Two replicates of 7.5 K ERα untreated ChIP-seq have averaged ~0.86 Pearson correlation to the ENCODE data and ~0.97 self-correlation. Two replicates of 5 K ERα untreated ChIP-seq have averaged ~0.78 Pearson correlation to the ENCODE data and ~0.88 self-correlation. Two replicates of 2.5 K ERα untreated ChIP-seq have averaged ~0.72 Pearson correlation to the ENCODE data and ~0.87 self-correlation.

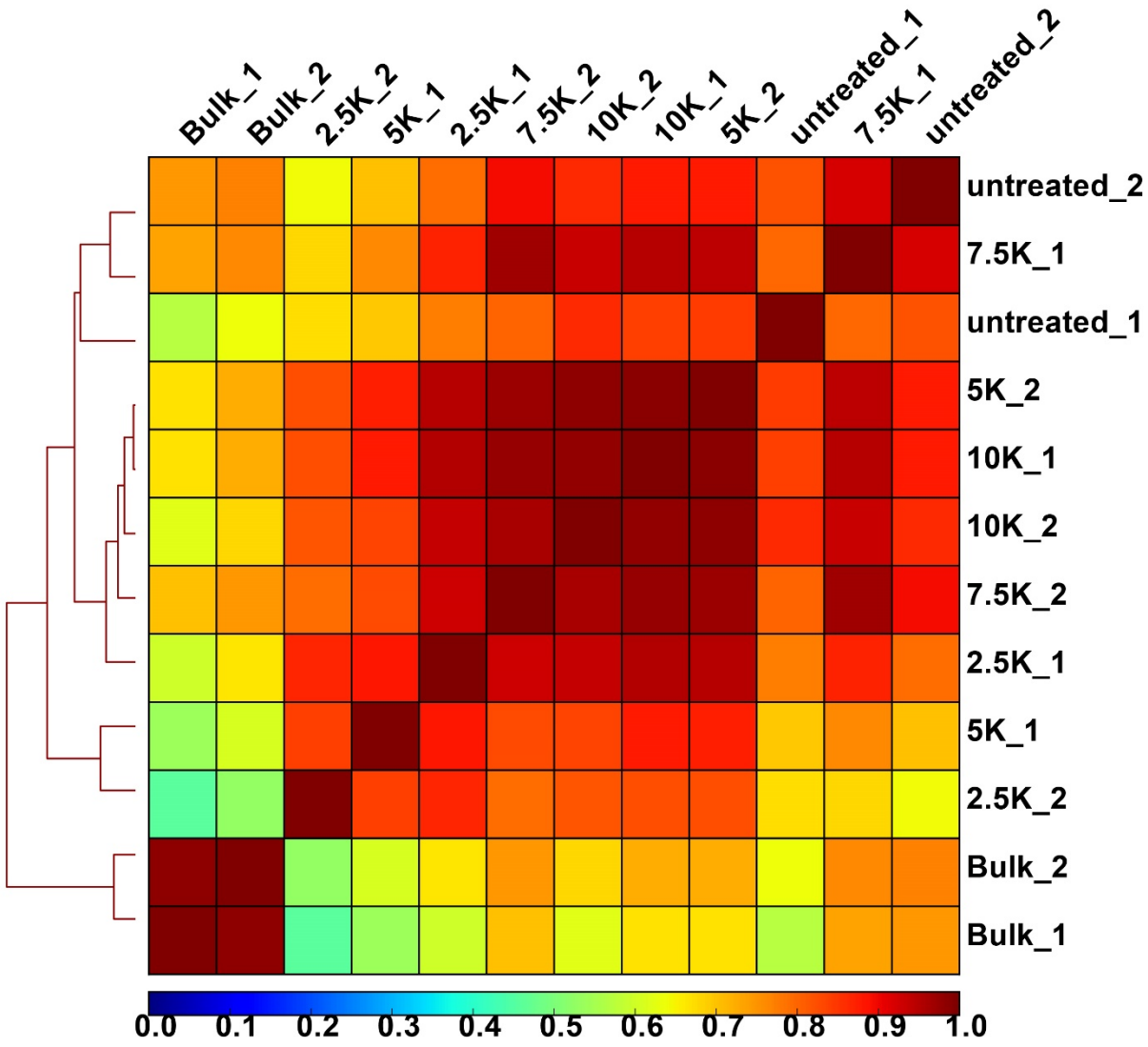


Figure 2.14 Microfluidic ERα ChIP-seq data on MCF-7 cell line. Genome-wide correlations among ChIP-seq data sets of various sample sizes. ChIP-seq genomic coverage profiles at TSS +/- 200000 bp regions were used for computing correlations. Colors represent Pearson correlation coefficients.

High mapping rate and low duplicates rate suggest successful pol2 libraries and ChIP-seq data (**Table 2.1**).

	Total	Mapped	%Mapped	% Dupes/100
1000K_1F_S_1	8471392	8177687	96.53%	0.05%
1000K_1F_S_2	10585890	10150025	95.88%	0.09%
100K_1F_M_1	22532970	22109700	98.12%	0.07%
100K_1F_M_2	17389782	16839120	96.83%	0.09%
100K_2F_M_1	19519594	12124186	62.11%	0.10%
100K_2F_M_2	43037945	32027532	74.42%	0.28%
50K_1F_M_1	30277555	29535828	97.55%	0.17%
50K_1F_M_2	32119009	31194046	97.12%	0.18%
50K_2F_M_1	32393450	31699424	97.86%	0.08%
50K_2F_M_2	31993937	31023117	96.97%	0.10%

Table 2.1 Mapping rate of Pol2 data.

By increasing IDR thresholds, we were able to rescue peak information from the low input pol2 data (**Table 2.2**).

	Peak Number IDR_0.05	Peak Number IDR_0.2
1000K1FS	4162	10896
100K1FM	446	3483
100K2FM	4840	11338
50K1FM	290	290
50K2FM	6217	23031
SRX150687	12167	-
SRX100530	26039	-

Table 2.2 Peak called with SPP at different IDR thresholds.

For peaks called from MACS, by relaxing P-value of 10^6 and 10^5 pol2 ChIP-seq used in peak calling to 0.005, we were able to obtain 15578 peaks for 10^6 pol2 ChIP-seq. Among those peaks, 14808 (95%) peaks were found in ENCODE peaks called using P-value of 0.00001 as shown in Venn diagram (**Figure 2.15**). And 10213 peaks for 10^5 pol2 ChIP-seq, 7279 (71%) peaks were found in ENCODE peaks. 6056 peaks for $5x10^4$ pol2 ChIP-seq, 4795 (79%) peaks were found in ENCODE peaks. BedTools intersectBed was used to find intersection regions of pol2 ChIP-seq data with ENCODE data.

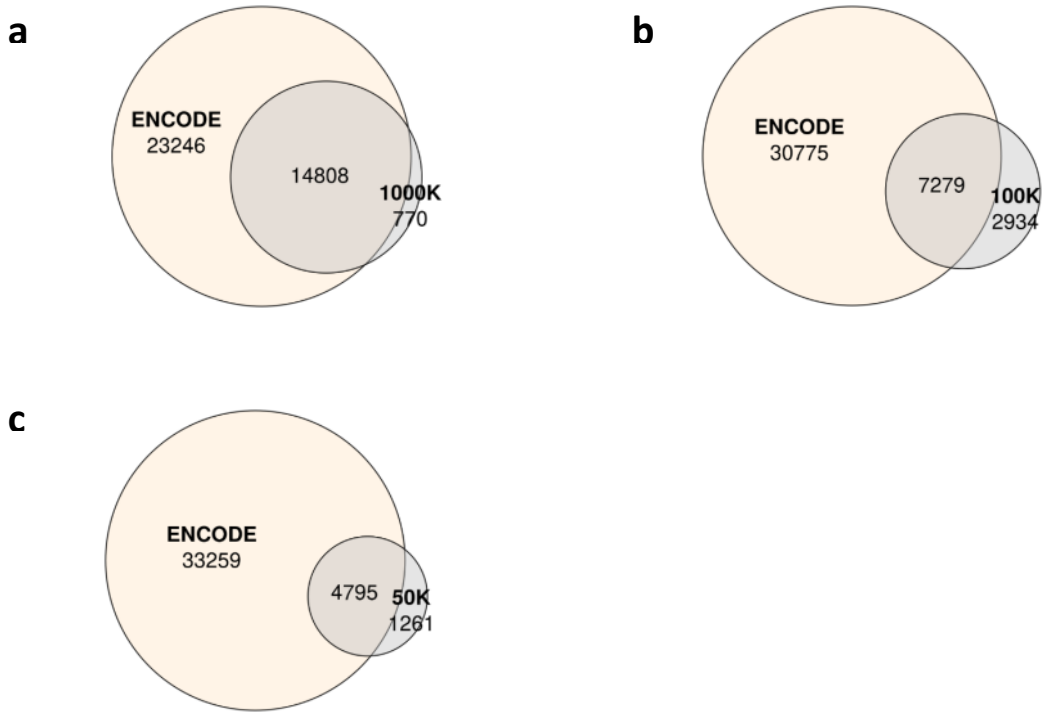


Figure 2.15 Venn diagrams of peaks intersected with ENCODE peaks: **(a)** 1000K_1F_S, **(b)** 100K_2F_M, **(c)** 50K_2F_M on the bottom left.

Fingerprints plots of 10^6 (1000K_1F_S) and 10^5 (100K_1F_M) pol2 ChIP-seq comparing to ENCODE are shown (**Figure 2.16**). The tool developed by Diaz et al.[155] determines how the signal in the ChIP sample can be differentiated from the background distribution of reads in the control sample. The result shows the strength of a ChIP experiment. The broader the peak (e.g., transcription factor), the less clear the plot.

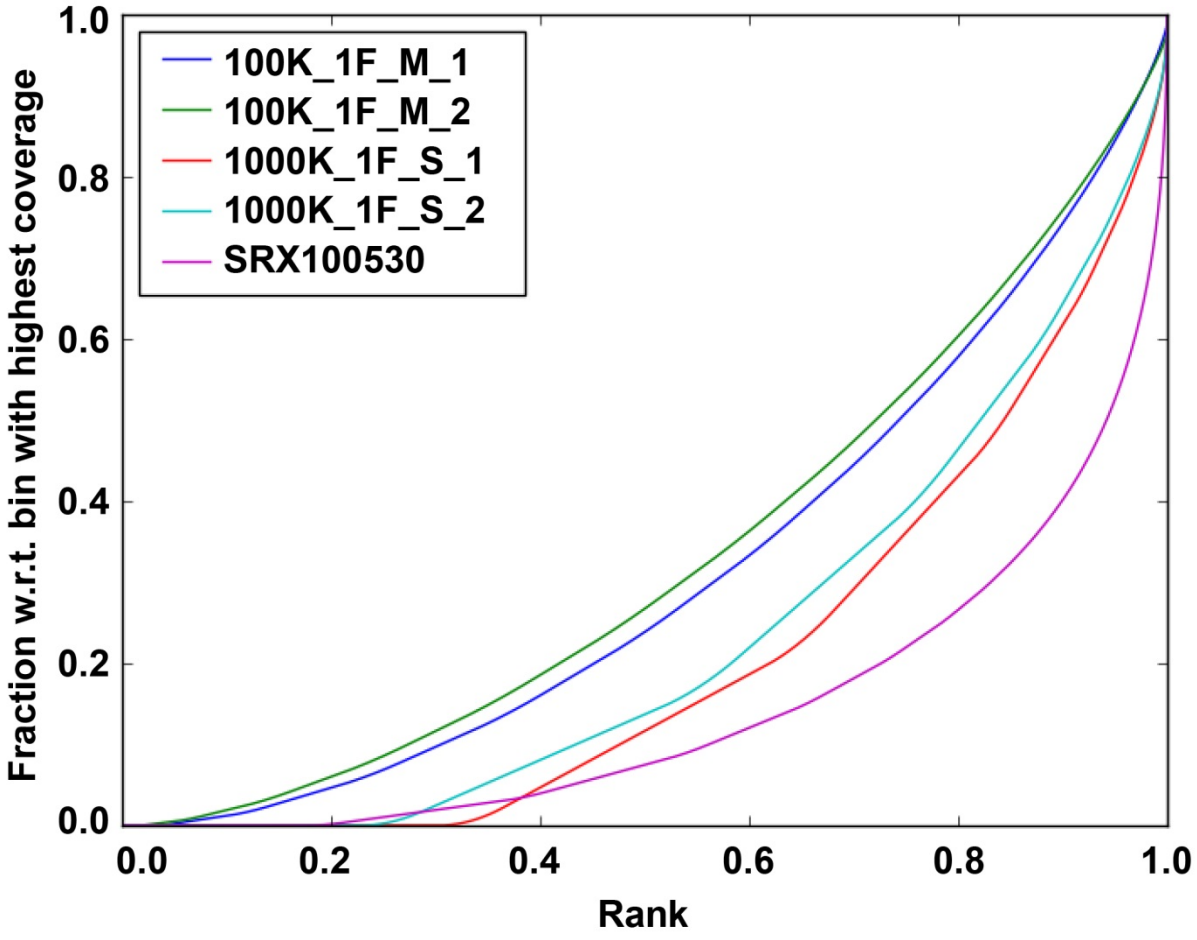


Figure 2.16 Fingerprints Plot of 10^6 and 10^5 pol2 ChIP-seq comparing to ENCODE.

Enrichment is calculated as the number of reads overlapping each hg19 genes exported from UCSC table browser. ENCODE data had the highest enrichment, and it decreased as cell number was decreased as shown (**Figure 2.17**).

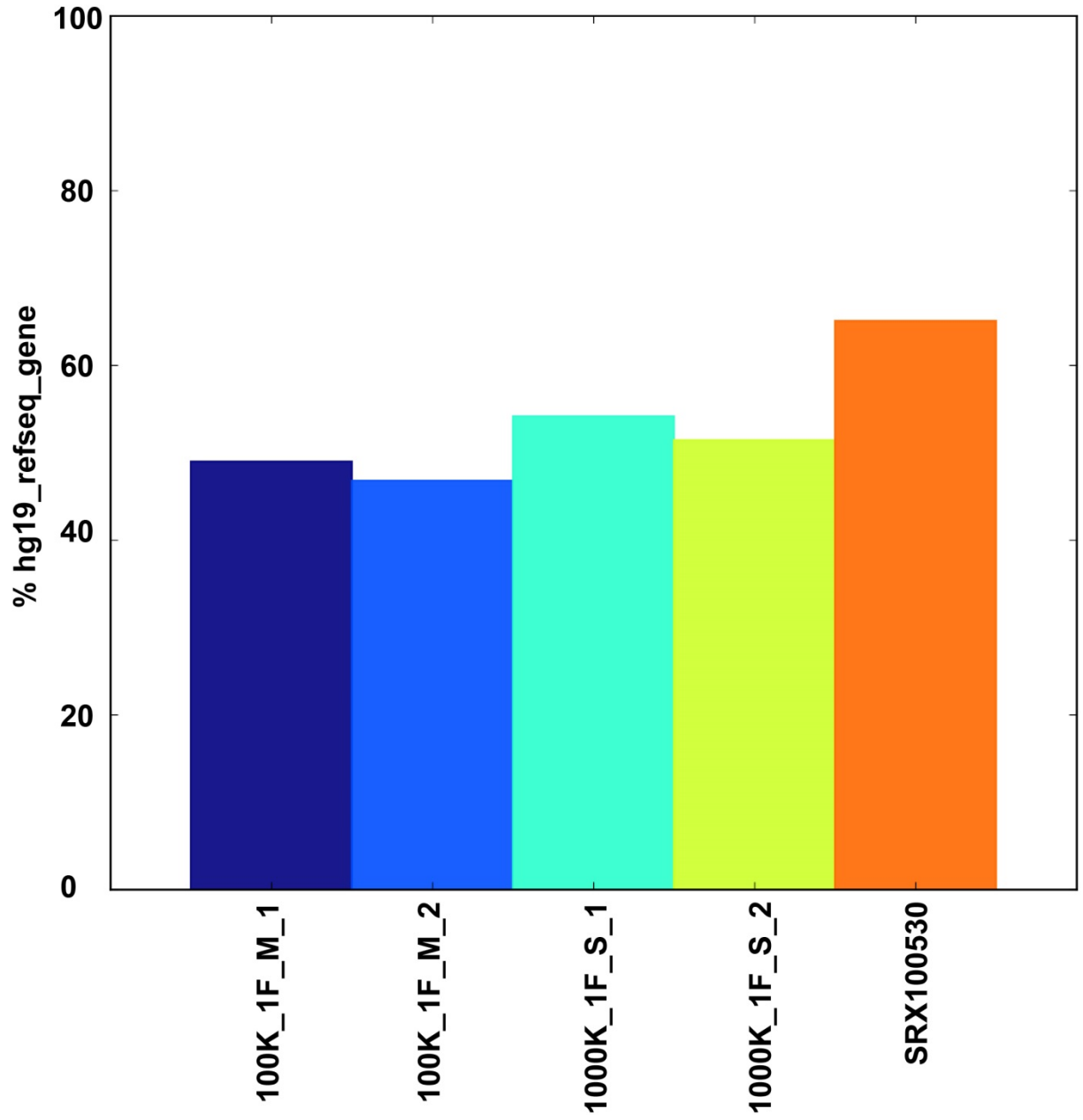


Figure 2.17 Enrichment plot of 10^6 and 10^5 pol2 ChIP-seq comparing to ENCODE.

Distributions of sequencing depths can be viewed as frequencies of the found read coverages. The 1 million bps were sampled and counted the number of overlapping reads. One can identify how many bases were covered by how many times. For example, 10% of the sampled 10^6 bp are covered once. The figure on the right shows the reverse cumulative sum of the left figure to show area under the curve, and it reveals what percentage of genome has certain sequencing depth. For example, 20% of the sampled 10^6 bp have up to 1 overlapping read (**Figure 2.18**).

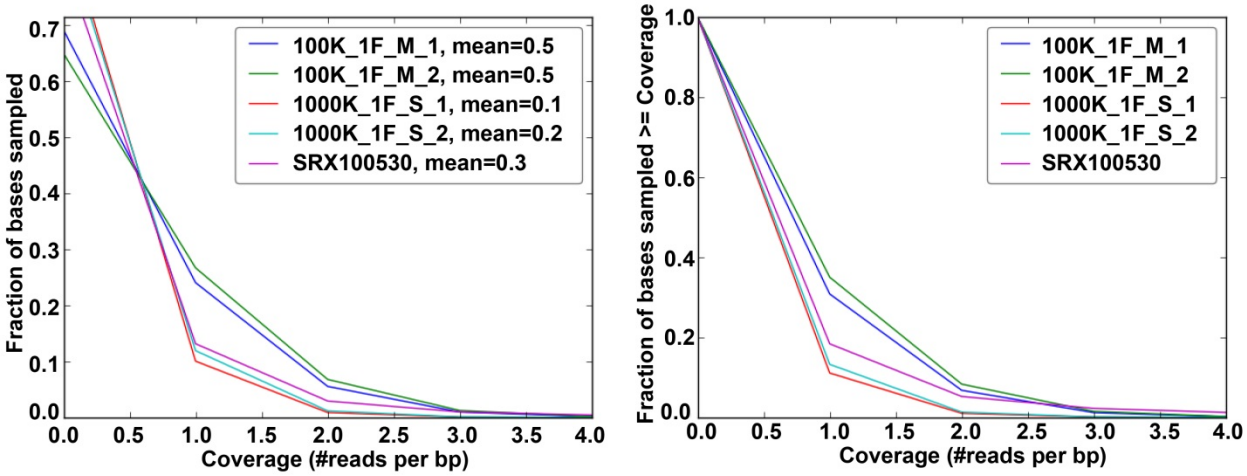


Figure 2.18 Sequencing coverage profile of 10^6 and 10^5 pol2 ChIP-seq comparing to ENCODE.

Since there were read coverage values from thousands of regions, the dimensions could be reduced by conducting principal component analysis (PCA). Here, principal components represent the directions along data which the variation in the data is maximal. It is useful to determine whether samples show greater variability between experimental conditions than between replicates of the same treatment. It is also useful to identify unexpected patterns. PC1, the first principal component is the axis that spans the most variation. PC2 is the axis that spans the second most variation. 10^6 cluster with ENCODE and they have similar PC1 values as 10^5

pol2 ChIP-seq. however, 10^5 have different PC2 values meaning they are not mainly varied from 10^6 and ENCODE PC1 direction, but are varied from PC2 direction. A Scree plot shows how much variation that each principal component can account for. Most of the variations is accounted for by the first two principal components. First principal component counts for more variation than the second principal component (**Figure 2.19**).

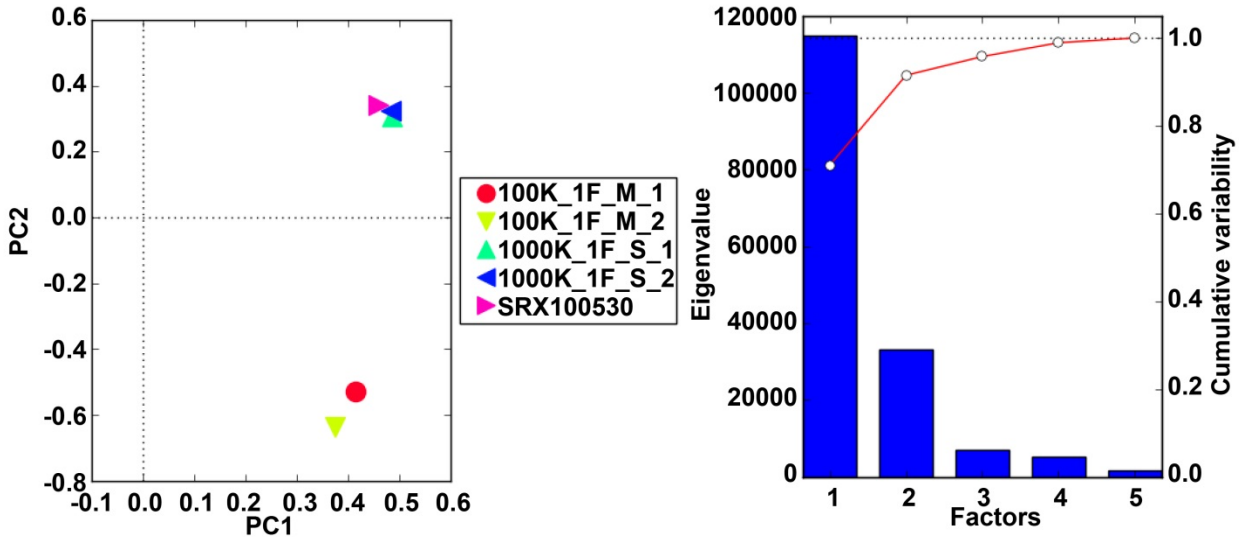


Figure 2.19 PCA of read counts and Scree plot.

Genome coverage profile was calculated. All hg19 genes (62332 genes) were exported from UCSC table browser. Heat map of ChIP-seq signal at transcription start site (TSS) of those genes was plotted based on coverage scores as shown (**Figure 2.20**). Parameter Kmeans =2 was used to distinguish active and inactive regions of pol2 signals. Cluster_1 represents active region, and Cluster_2 represents inactive region. Active regions are the regions have pol2 binding, and inactive regions are the regions do not have pol2 binding. We were able to identify pol2 binding events at TSS in both 10^6 and 10^5 ChIP-seq data, however, the signal was less localized as cell

number went down (as shown in the 10^5 data in **Fig 3.20**). On top of heat maps are average signal profiles represented by heat maps.

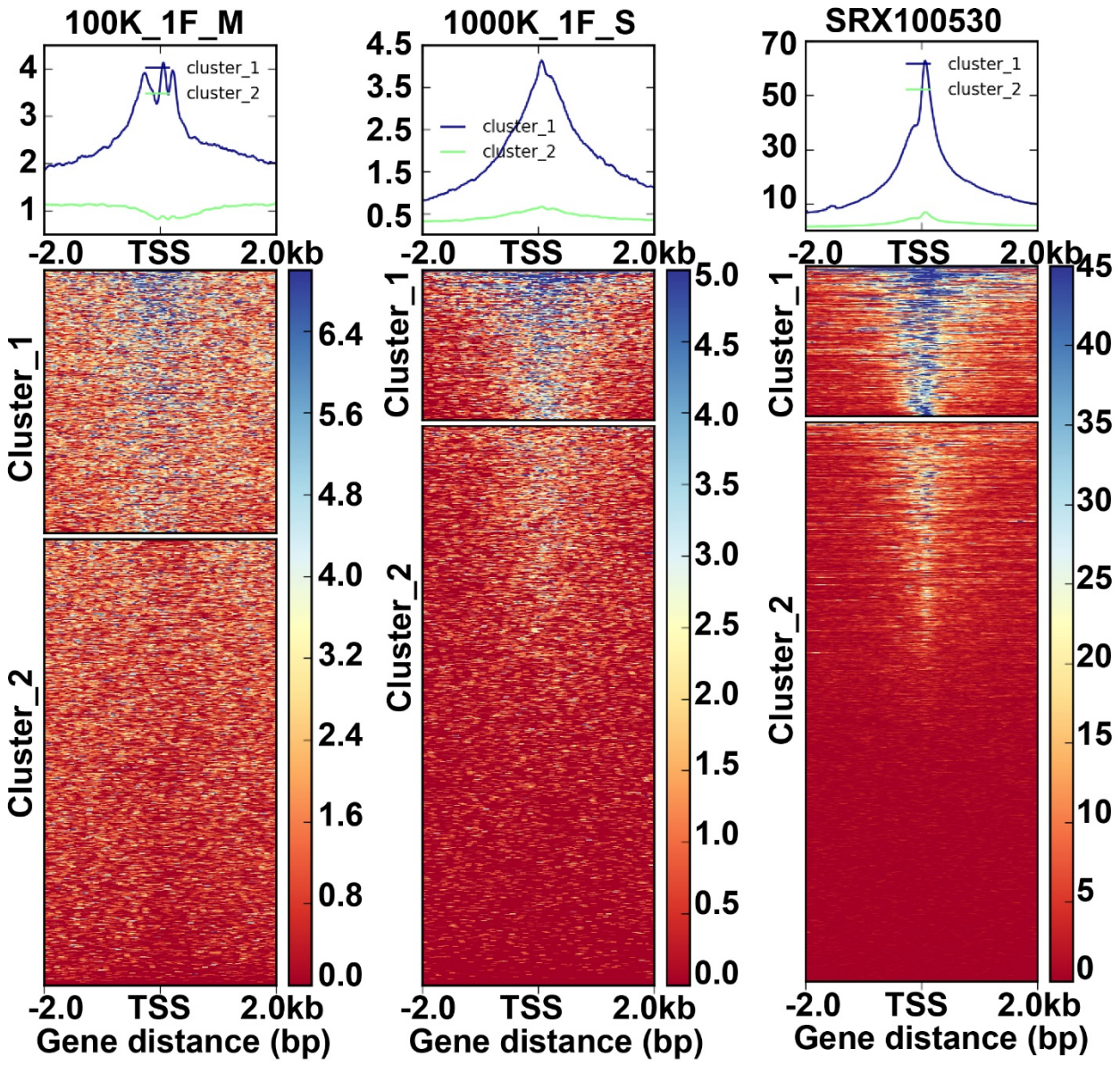


Figure 2.20 Heat map of active and inactive ChIP-seq signals of 10^6 and 10^5 pol2 ChIP-seq comparing to ENCODE around the transcription start site of genes.

Our microfluidic technology is superior for conventional ChIP for couple of reasons. First, high concentrations from trace amounts of molecules could be built up inside the tiny volumes that offered by microfluidic chamber. Adsorption kinetics and completeness was facilitated by such high concentration. Second, the IP beads take up a large fraction of the tiny volume so that the surface area/volume ratio (15-40% bead volume in 800 nl) is tremendously improved when compared to 5% bead volume in a conventional 1.5 ml ChIP assay[150]. The close proximity among beads greatly increased the efficiency and rate for chromatin adsorption on the bead surface due to the short diffusion lengths involved. The adsorption of a chromatin molecule among beads was rapid given that travel time $\tau_D \sim w^2/D$, where w is diffusion distance between two beads, and D is diffusivity. Third, by using a microfluidic technique that is uniquely suited for bead manipulation at the microscale, we effectively eliminated nonspecific adsorption by microfluidic oscillatory washing. This is critical for producing high quality ChIP DNA that preserves desired biological information. Finally, our microfluidic device integrates various steps and minimizes material loss among steps.

3. Study of the Temporal Dynamics in the DNA Methylome during Cancer Development Using a Transgenic Mouse Model with Microfluidic MeDIP-seq Assays

3.1 Introduction

Gene activation and expression are not only associated with alteration in the DNA sequence, but also affected by other changes to DNA and histones (i.e. epigenetics). One of well-studied epigenetic modification, DNA methylation, plays critical roles in gene expression and regulation, and are highly involved in biological processes such as embryonic development and tumorigenesis. DNA methylation refers to the addition of a methyl group at the carbon-5 position of cytosine residues within CpG dinucleotides, forming 5-methylcytosine (5mC). This process is catalyzed by enzymes called DNA methyltransferases (DNMT1, DNMT3a, and DNMT3b)[156, 157]. Clusters of CpG sites ("CpG islands" or CGIs) commonly span promoters of housekeeping genes. Promoter CGIs typically remain unmethylated in normal cells and are associated with active gene expression during differentiation, while methylated CGIs are associated with gene repression. The genome-wide DNA methylation profile (i.e. the DNA methylome) is established early in development for regulation and maintenance of gene expression during differentiation. The methylome is altered and disrupted in disease states. For example, the hypermethylation at various genes in tumors has been recognized as a common epigenetic feature for all types of human cancers. Genes that were critically involved in cancer biology, including the cell-cycle inhibitor p16-INK4a, the DNA-repair genes MLH1 and BRCA1, and the tumor-suppressor genes

Rb and p53 have been shown to undergo methylation-associated silencing in tumor cells[40, 41]. While acquiring hypermethylation at specific promoters, overall human tumors undergo a global genomic hypomethylation. Since DNA methylation protects the integrity of chromosomes, this global DNA hypomethylation potentially contributes to the large-scale genetic changes that are hallmarks of tumorigenesis. Therefore, a better understanding of the dynamics in the methylome during disease development will improve various aspects of biomedicine, including risk stratification, disease diagnostics, and drug/therapeutic discovery.

Bisulfite sequencing is generally considered the gold standard for DNA methylation analyses[158, 159]. Bisulfite treatment converts cytosine residues to uracil, but leaves 5mC residues unaffected. Such treatment introduces changes in the DNA sequence based on the methylation status of individual cytosine residues. Combined with next-generation sequencing (NGS), this approach generates methylomic profile with single-nucleotide resolution. In spite of the ultrahigh resolution, bisulfite sequencing requires a large amount of DNA (1-5 µg) and the cost associated with deep sequencing is high[160-162]. It normally requires at least 500 million reads per sample to cover about 95 % of the genome[163, 164]. This can be prohibitive when various samples with repeats are analyzed[165]. An alternative to bisulfite sequencing is reduced representation bisulfite sequencing (RRBS)[166, 167]. Although RRBS provides single-nucleotide resolution, it only covers 5-10 % of genome (mostly CGIs and promoter regions). Enrichment-based technologies have also been applied to study methylomes[168-172]. These approaches dramatically reduce the required sequencing depth by enriching methylated DNA fragments based on affinity purification. Methylated DNA immunoprecipitation followed by next-generation sequencing (MeDIP-seq) uses a monoclonal antibody specific for 5mC to target single-stranded methylated DNA

fragments (sheared by sonication) and identifies these fragments using sequencing[102, 170-173]. In comparison to bisulfite sequencing, MeDIP requires only 30-60 million reads per sample with a 100-300 bp resolution[95, 174]. MeDIP-seq provides an unbiased tool for probing DNA methylation events both within CGIs and non-CGI regions throughout the entire genome. There have been a number of microfluidic technologies developed for profiling genome-wide epigenetic changes[149, 175-177].

DNA methylomes, similar to other epigenetic information, are specific to cell and tissue types, the disease condition and its developmental stage. Profiling methylomes associated with various tissues and diseases is important for understanding the dynamics in methylomes during disease development and establishing epigenomic signatures for disease diagnosis and prognosis. However, one critical challenge in these efforts is that current methylomic profiling tools often do not offer sufficient sensitivity to examine tiny quantities of cell samples from scarce sources such as small lab animals and patients. For example, conventional whole-genome bisulfite sequencing requires 1-5 µg DNA and MeDIP-seq requires 5-20 µg DNA due to the low efficiency associated with immunoprecipitation. Given that a diploid mammalian cell typically contains 4-8 pg of DNA, bisulfite sequencing and MeDIP-seq require at least 10^6-10^7 cells. In contrast, mouse and patient samples do not yield large quantities of cells. For example, in murine splenocytes, there is only about ~ 10,000 per spleen and ~5000 per ml peripheral blood leukocyte naturally occurring T regulatory cells. Circulating tumor cells (CTCs) are found by the frequency of 1-10 per ml of whole blood in metastatic cancer patients. Furthermore, the isolation of a homogenous single cell type always generates further loss in the sample amount. Thus, highly sensitive technologies for methylome profiling are in high demand in order to facilitate

generation of data with direct biomedical relevance. For example, heterogeneity of DNA methylation profile bwtween two mice can be differentiated with low-input microfluidic technology (**Figure 3.1**).

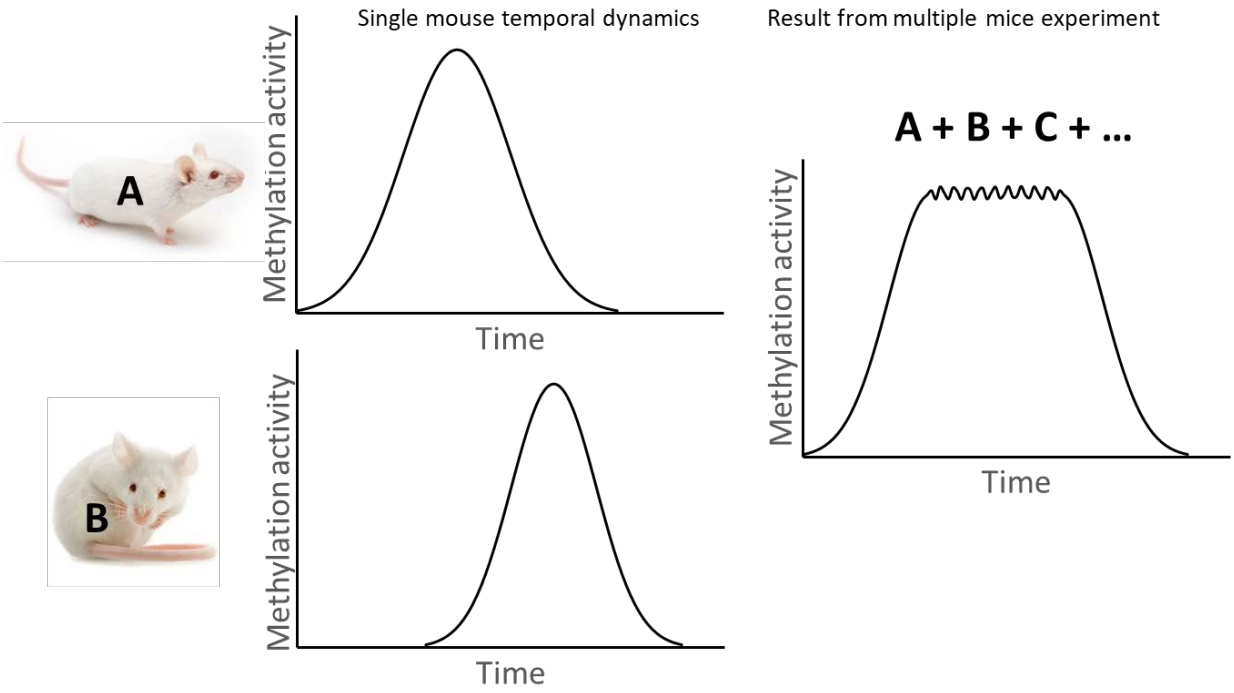


Figure 3.1 The advantages of single mouse experiments for studying temporal dynamics in DNA methylome.

In this project, we developed an ultrasensitive microfluidic MeDIP-seq technology for profiling methylomes with entire 1.5 hr on-chip process. The microfluidic MeDIP-seq with high immunoprecipitation efficiency produces a sensitivity of 0.5 ng DNA (or ~50-100 cells). This is roughly 4-5 orders of magnitude higher than the prevailing protocol and 2-3 orders of magnitude higher than the-state-of-the-art (~50 ng)[174, 178]. We optimized the technology to reach

ultrasensitive with a lymphoblastic cell line GM12878. We found that the combined use of a packed bed of beads for MeDIP and effective oscillatory washing for removing nonspecific adsorption/trapping is key to extremely high yield of highly-enriched MeDIP DNA. We then obtained methylomic profiles in a transgenic mouse model during mammary cancer development.

Our microfluidic technology is superior for conventional MeDIP for several reasons. First, high concentrations from trace amounts of molecules could be built up inside the tiny volumes that offered by microfluidic chamber. Adsorption kinetics is facilitated by such high concentration. Second, the IP beads occupy a large fraction of the tiny volume so that the surface area/volume ratio (15-40%) is tremendously improved when compared to 5% in the conventional MeDIP[150]. The close proximity among beads greatly increases the efficiency and rate for chromatin adsorption on the bead surface due to the short diffusion lengths involved. The adsorption of a chromatin molecule among beads is rapid given that travel time $\tau_D \sim w^2/D$, where w is diffusion distance between two beads, and D is diffusivity. Third, by using microfluidic technique uniquely suited for bead manipulation at the microscale, we effectively remove nonspecific adsorption after high efficiency adsorption using microfluidic oscillatory washing. This is critical for producing high quality MeDIP DNA that preserves desired biological information. Finally, the microfluidic device integrates various steps and minimizes material loss among steps.

The new capability of our technology will allow establishing genome-wide methylation profiles with less than 100 cells. Our technology dramatically widens the sample range for methylomic profiling to include primary cell samples from scarce sources. Dynamic epigenomic information during disease development from tiny cell samples of a patient that used to be not accessible to

researchers and clinicians due to the technological limitation now could become attainable. With such information, one can build epigenomic signatures for disease diagnosis and prognosis in the context of personalized medicine.

3.2 Methods and Materials

Fabrication of the microfluidic MeDIP device

The microfluidic device was composed of a microfluidic chamber (~800 nl), connecting channels, and a micromechanical valve that could be partially closed to stop magnetic beads while allowing liquid to pass. The main chamber was in elliptic shape with a major axis of 6 mm, a minor axis of 3 mm, and a depth of 40 μm. 27 micro-pillars were placed inside the main chamber to prevent collapsing of PDMS.

Multilayer soft lithography was used to fabricate the microfluidic device[111, 149]. Briefly, two photomasks (one for fluidic layer, and one for control layer) that had desired microscale patterns were designed with computer aided design software FreeHand MX (Macromedia) and printed on high-resolution (5,080 d.p.i.) transparencies. To make fluidic layer master, 40 μm thick photoresist (SU-8 2025, Microchem) was spun on a 3-inch silicon wafer (978, University Wafer) at 500 rpm for 10s and 2500 rpm for 30s followed by soft bake at 65°C for 1 min and 95°C for 7 min. To make control layer master, 50 μm thick SU-8 was spun at 500 rpm for 10s and 1500 rpm for 30s and followed by the same bake condition as the fluidic master. The coated wafers were exposed to UV light for 17s at 580 mW exposure intensity and a post exposure bake at 65°C for

1 min and 95°C for 7 min. Masters were then developed in SU-8 developer for 2-3 min, rinsed with isopropyl alcohol and blown dry by pressured air. To make fluidic layer stamp, PDMS (General Electric silicone RTV 615, MG chemicals) with a mass ratio of A:B = 5:1 was thoroughly mixed and vacuumed for 1 h. It was then poured onto the fluidic layer master in a Petri dish to a depth of ~5 mm thick. To make control layer stamp, PDMS with a mass ratio of A:B = 20:1 was mixed, vacuumed for 1 h, and spun onto the control layer master at 1100 rpm for 35s to achieve a depth of 108 μm thick. Both PDMS stamps were partially cured at 80 °C for 30 min. The fluidic layer was then peeled off from the fluidic layer master and aligned to the control layer PDMS stamp. Two-layered PDMS was bonded by baking at 80 °C for 1 h, and then peeled off from the control layer master. Inlets and outlets of the device were punched by a 2 mm hole puncher. Finally, the two-layered PDMS structure and a pre-cleaned glass slide were treated in an oxygen plasma cleaner (PDC-32G, Harrick Plasma) and brought into contact to form closed channels and chamber. The device was then baked at 80 °C for 1 h to strengthen the bonding between PDMS and glass. Glass slides were cleaned in a basic solution (H_2O: 27% NH_4OH: 30% H_2O_2 = 5:1:1, volumetric ratio) at 75 °C for 2 h and then rinsed with ultrapure water and thoroughly air blown to dry.

Setup of the microfluidic device

The microfluidic experiment was monitored by a charge-coupled device (CCD) camera (ORCA-285, Hamamatsu) attached to the port of an inverted microscope (IX 71, Olympus). The reagents were flowed into the inlet via a tubing driven by a syringe pump (Fusion 400, Chemyx). The micromechanical valve was actuated by a solenoid valve (18801003-12V, ASCO Scientific),

which was connected to a pressure source and controlled by a data acquisition card (NI SCB-68, National Instruments) and a LabVIEW (LabVIEW 2012, National Instruments) program. Pressure (30- 35 p.s.i.) that was applied to the control channel deformed the PDMS membrane between fluidic channel and control channel and formed a partially closed valve to stop beads while allowing fluid to pass. The oscillatory washing was conducted by connecting the inlet and outlet of the microfluidic chamber to solenoid valves[149]. Alternating pressure pulses programmed by a LabVIEW program were exerted on the two ends of the chamber to move the beads back and forth.

Cell culture

GM12878 cells were obtained from Coriell Institute for Medical Research. GM12878 cells were cultured in RPMI 1640 (Invitrogen, Carlsbad, CA, USA) with 15% fetal bovine serum, 100 U penicillin, 100mg streptomycin/ml (Invitrogen) at 37°C in a humidified incubator containing 5% CO_2. Cells were sub-cultured every two to three days to maintain exponential growth.

Mouse

Hemizygous transgenic female mice (FVB-Tg(C3-1-TAg)cJeg/JegJ, Jackson Laboratory) were used in this study. Mice were euthanized at age of 6, 16, 23 weeks by exposing mice to 100% carbon dioxide at 1 – 2 l/min for 5 min followed by cervical dislocation. Mice were sprayed with 70% ethanol on the ventral side. Skin on the ventral side was cut through, pulled away from

mouse body, and pinned down to expose mammary glands. Mammary tumors were identified and cut off. The size and weight of a tumor were recorded.

Preparation of ssDNA

Genomic DNA from GM12878 cells was extracted from 10^6 cells using Blood & Cell Culture DNA Mini Kit (Qiagen). Mouse genomic DNA was extracted from mouse mammary tissues using DNeasy Blood & Tissue Kit (Qiagen). Extracted genomic DNA was sonicated with a Covaris E220 sonicator for 180 s with 10 % duty cycle, 50 peak incident power and 200 cycles per burst. The concentration of sonicated DNA was quantified using a Qubit 2.0 fluorometer with dsDNA HS Assay kit (Q32851, Life Technologies). Different sample sizes (0.5~100 ng) were aliquoted and diluted with MeDIP buffer to generate a final volume of 50 µl for microfluidic MeDIP. 0.5 ng, 5 ng, 10 ng, or 100 ng of the sonicated sample was used as the input. Sonicated DNA samples were freshly denatured into single strand DNA (ssDNA) under 95℃ for 15 min and put on ice for 5 min before microfluidic MeDIP.

Preparation of immunoprecipitation (IP) beads

Dynabeads Protein A (10001D, Invitrogen) were used for MeDIP. They are 2.8 µm superparamagnetic beads with recombinant Protein A (~45 kDa) covalently bound to the surface. 5 µl of 30 mg/ml beads (equivalent to 150 µg) were washed twice with freshly prepared MeDIP buffer (10 mM monobasic sodium phosphate dihydrate, 10 mM dibasic sodium phosphate, 140 mM NaCl, 0.05%(v/v) Triton-100X) and resuspended in 150 µl MeDIP buffer containing 5-

methylcytosine (5mC) antibody (pAb) (61255, Active Motif). The antibody concentration for coating was 3.33 µg/ml for 0.5 ng DNA, 5 µg/ml for 1 ng DNA, 6.67 µg/ml for 10 ng DNA, and 40 µg/ml for 100 ng DNA. Beads were gently mixed with the antibody at 4 °C on a rotator mixer at 24 rpm for 2 h. Antibody-coated beads were washed twice with the MeDIP buffer and then resuspended in 5 µl MeDIP buffer.

Microfluidic MeDIP

The MeDIP procedure started with rinsing the fluidic chamber with the MeDIP buffer at a flow rate of 20 µl/min for 30 s. The micromechanical valve was then partially closed. The 5 µl antibody-coated immunomagnetic beads were flowed into the microfluidic chamber driven by the syringe pump at 20 µl/min and aided with a cylindrical permanent magnet (NdFeB, D48-N52, 0.25 inch dia. and 0.5 inch thick, K&J Magnetics). The beads were packed against the partially closed valve to form a packed bed. The MeDIP buffer containing denatured ssDNA fragments with a total volume of 50 µl was flowed through the packed bed of MeDIP beads at a flow rate of 1.5 µl/min. After MeDIP, the MeDIP buffer was flowed into the fluidic chamber at a flow rate of 2 µl/min for 2 min. Two tubings each prefilled with 20 µl MeDIP buffer were plugged into the inlet and outlet of the chamber for oscillatory washing. Oscillatory washing[149] was done by applying pressure pulses (each at 0.5 p.s.i., with a pulse duration of 0.5 s and an interval of 0.5 s between two pulses) alternatingly at inlet and outlet of the fluidic chamber for 5 min while keeping the micromechanical valve open. The pulsing was controlled by the data acquisition card and a LabVIEW program. After oscillatory washing, beads were retained on one side of the chamber by the magnet. The unbound ssDNA was flushed out of the chamber by flowing MeDIP

buffer through the chamber at 2 µl/min for 2 min. Finally, MeDIP beads were flushed out by MeDIP buffer under a flow rate of 50 µl/min for about 5 min and collected into a 1.5-ml LoBind Eppendorf tube.

Extraction of MeDIP DNA and input DNA

MeDIPed ssDNA on beads were obtained by pipetting out MeDIP buff on a magnet and were mixed with 198 µl digestion buffer (200 mM NaCl, 50 mM Tris-HCl, 10 mM EDTA, 1% SDS, 0.1 M $NaHCO_3$) and 2 µl of 20 mg/ml proteinase K (26160, Thermo Scientific). Input ssDNA (50 µl) was mixed with 148 µl digestion buffer and 2 µl 20 mg/ml proteinase K. Samples were then incubated at 50 °C for 4 h. Equal volume (200 µl) of Phenol-chloroform-isoamylalcohol (25:24:1) was added to the sample, mixed by vortexing, and centrifuged at 16,100g for 5 min at room temperature. ssDNA was extracted by collecting ~200 µl aqueous phase to a fresh 1.5 ml Eppendorf tube. The 50 µl of 10 M ammonium acetate was then added to the sample. The 750 µl of 100% ethanol and 2 µl of 20 µg/µl glycogen (10814010, Invitrogen) were added for ethanol precipitation and ssDNA purification. Samples were placed at -20 °C overnight for ethanol precipitation. Next, samples were centrifuged at 16,100g for 10 min at 4 °C before the supernatant was carefully removed. The ssDNA pellet was washed with 500 µl of 70% ice-cold ethanol without breaking the pellet. The supernatant was then removed after centrifugation at 16,100g for 5 min at 4 °C. The pellet was air dried for 5 min and resuspended in 10 µl DNase-free water. This purified ssDNA was used directly for MeDIP-qPCR or for sequencing library construction. DNA concentrations were measured using a Qubit 2.0 fluorometer with ssDNA Assay kit (Q10212, Invitrogen).

Construction of sequencing libraries

Sequencing libraries were prepared by DNA SMART™ ChIP-Seq kit (634865, Clontech). This kit generates sequencing libraries from low-input ssDNA MeDIP samples (100 pg - 10 ng) without using ligation. Sequencing libraries were amplified by PCR using primers containing Illumina adapters. The libraries were purified using Ampure XP beads (A63880, Beckman Coulter). The library fragment size was determined using high sensitivity DNA analysis kit (5067-4626, Agilent) on an Agilent 2200 TapeStation. Sonicated DNA for MeDIP had average fragment size around 200 bp (**Figure 3.2a**). After library preparation, the average peak size shifted to 350 bp as library construction added about 150 bp adaptor (75 bp on each end) to the MeDIP DNA (**Figure 3.2b**). KAPA library quantification kit (KK4809, Kapa Biosystems) was used to determine effective library concentrations. The final concentrations of libraries submitted for sequencing were ~8 nM. The libraries were sequenced on an Illumina HiSeq 4000 with single-end 50 nt read. Typically, 15-20 million reads were generated per library.

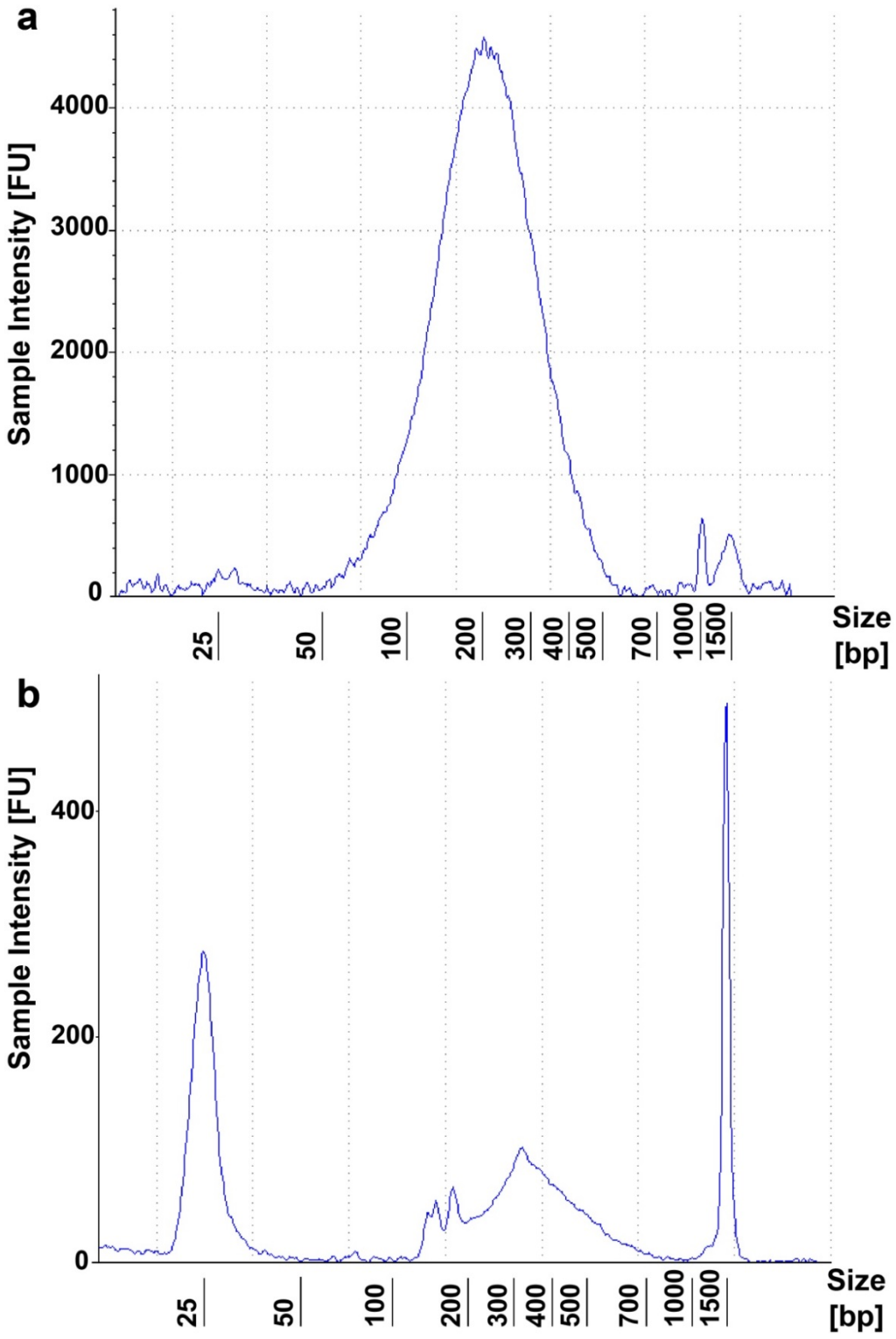


Figure 3.2 DNA fragment size profiles. **(a)** Sonicated DNA profile before MeDIP. **(b)** DNA profile after library preparation.

MeDIP-qPCR data analysis

qPCR was done using iQ SYBR Green Supermix (Bio-Rad, Hercules, CA, USA) on an CFX96 real-time PCR machine (Biorad) with C1000Tm thermal cycler base. All PCR assays were performed using the following thermal cycling condition: 95°C for 10 min followed by 40 cycles of (95°C for 15 s, 56°C for 30 s, 72°C for 30s). GEMIN4 and ZC3H13 are two known positive loci for human while GRICK3 and ZC3H13 are two known positive loci for mouse. N1 and N2 are two known negative loci for human and mouse. Primer concentration was 400 nM. All primers were ordered from Integrated DNA Technologies (Coralville). The MeDIP-qPCR results were represented as relative fold enrichment (**Figure 3.3**), which is the ratio of percent input (**Figure 3.4**) between a positive locus and a negative locus. Percent input was calculated using the following equation: $2^{(C_q INPUT - C_q IP)} \times 100\%$, where C_q is amplification cycle number run by qPCR, $INPUT$ is the DNA sample without immunoprecipitation, IP is DNA sample after MeDIP.

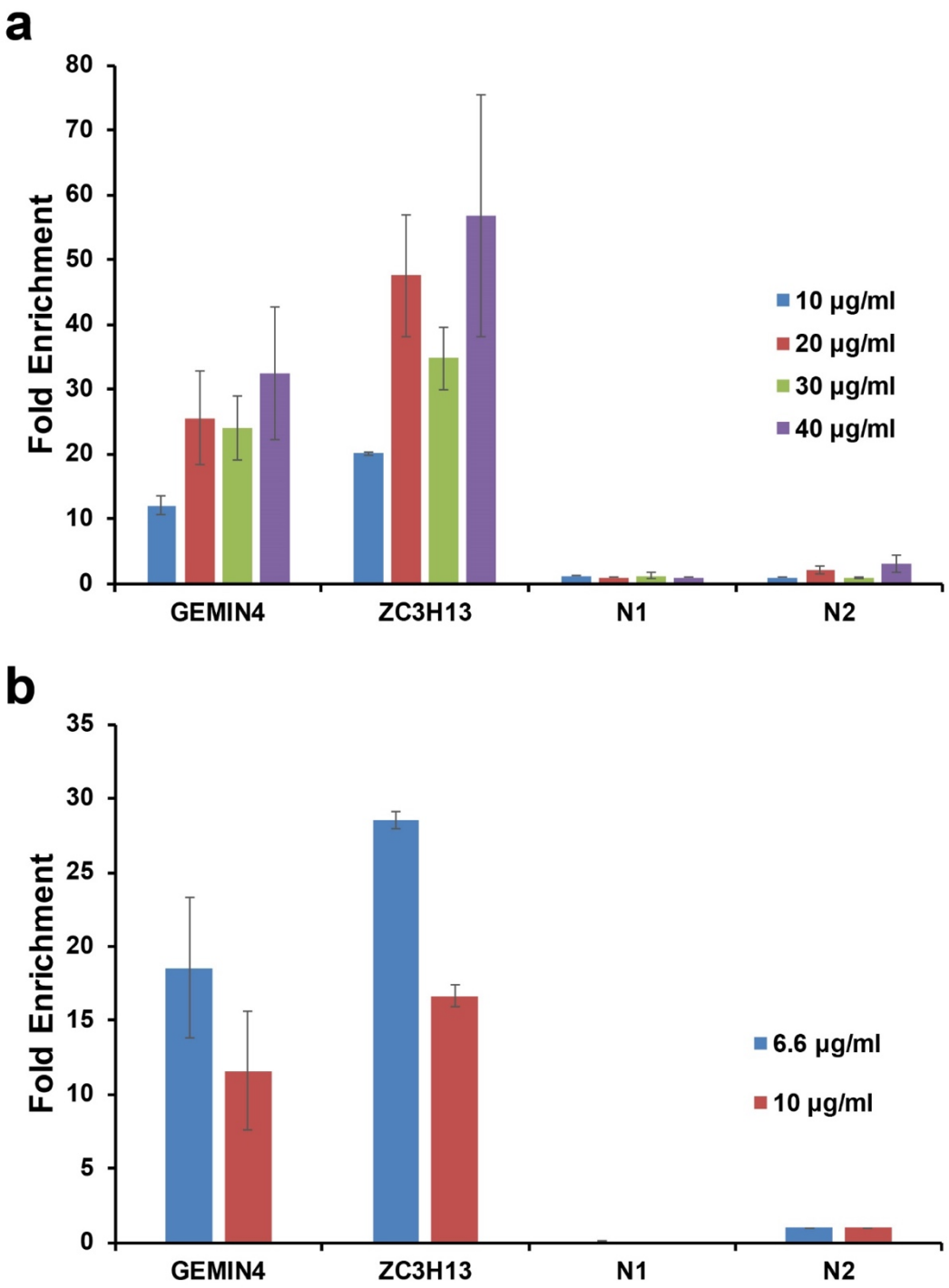


Figure 3.3 Fold enrichment (ratio between % input at positive loci and % input at negative loci) under different antibody coating concentrations. **(a)** 10 µg/ml, 20 µg/ml, 30 µg/ml, 40 µg/ml with

100 ng DNA from GM12878 cells. **(b)** 6.6 µg/ml, 10 µg/ml with 10 ng DNA from GM12878 cells.

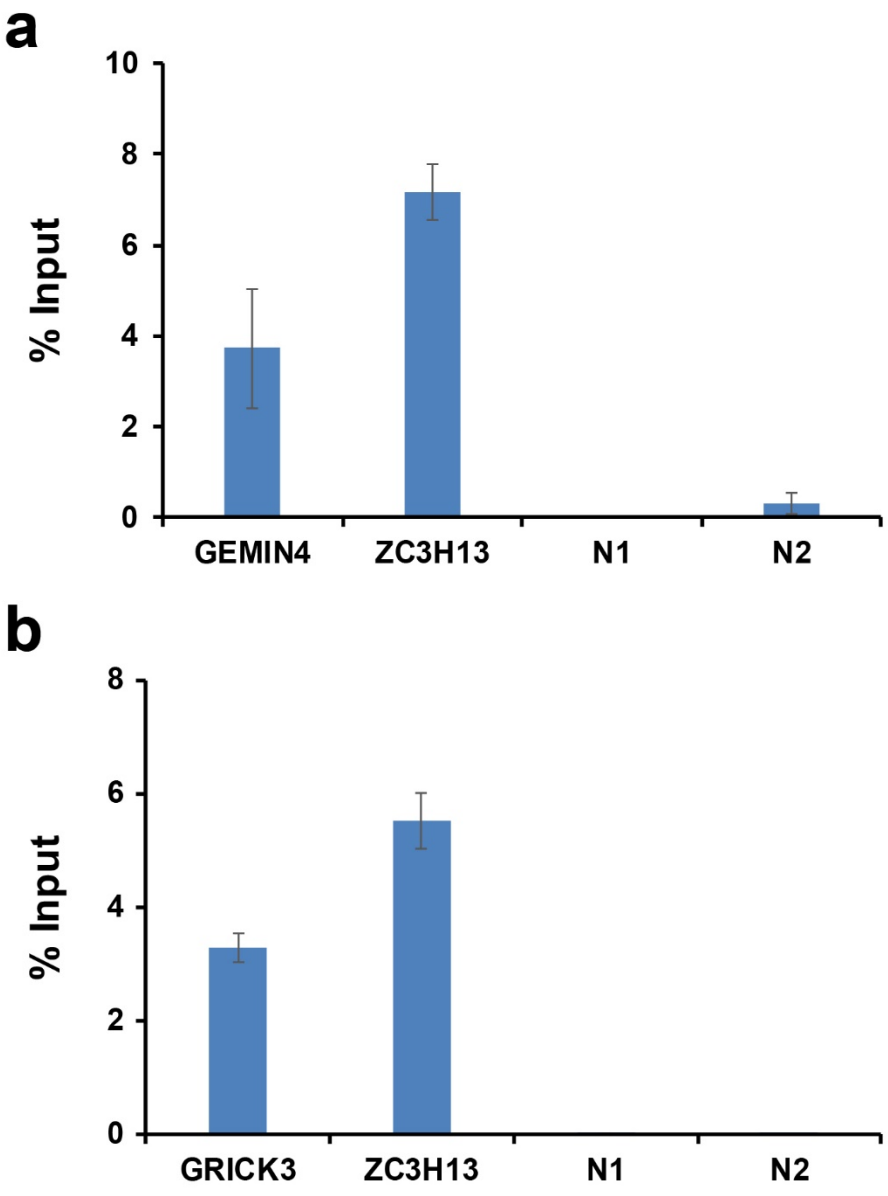


Figure 3.4 Percent input data by MeDIP-qPCR. GEMIN4 and ZC3H13 are two known positive loci for human, GRICK3 and ZC3H13 are two known positive loci for mouse. N1 and N2 are two known negative loci for human and mouse. The error bars were calculated as standard deviation from 3 MeDIP-qPCR replicates**. (a)** Percent input data for human GM12878 cell line.

The 10 ng DNA was used as starting material. **(b)** Percent input data on 6 weeks mouse mammary gland tissue. The 10 ng DNA was used as starting material.

MeDIP-seq data analysis

MeDIP sequencing reads were mapped to the human genome (hg19) or mouse genome (mm10) using Bowtie2 (v2.2.5)[84] with default parameter settings. Peaks of each MeDIP sample were called against input by MACS (v2.1.0)[85] with p-value < 10^{-5} and other parameters set at default values. Pearson correlations of genome-wide coverage profiles between samples were calculated using Bioconductor R package "MeDIPS" with MeDIPS.correlation function[179].

To identify DMRs between tumor and normal tissues, MeDIPS.meth function from MeDIP package[179] with parameters set as p.adj = "fdr", diff.method = "edgeR", MeDIP = T, CNV = F, minRowSum = 10 and MeDIPS. selectSig function with parameters set as p.value = 0.01, adj = F, ratio = NULL, bg.counts = NULL, CNV = F were used. We first obtained the 250 bp windows that were differentially methylated (only uniquely mapped reads were used), and then we merged neighboring significant windows into a larger continuous region with MeDIPS.mergeFrames function to generate differentially methylated regions. GO analysis was performed by GREAT[180] with default parameter settings.

The edgeR method was applied to the counts of the genome wide windows for testing differential coverage. The weighted trimmed mean of M-values (TMM) method was used to

calculate scale factors between libraries. The edgeR method adds four vectors to the results table which are edgeR's exactTest standard output: edgeR.logFC, edgeR.logCPM, edgeR.p.value, and edgeR.adj.p.value. A CpG density dependent relative methylation scores (rms) was calculated for the two conditions.

3.3 Results and Discussion

The chip device contained 7 reaction chambers that allowed us to run 7 MeDIP assays at the same time (**Figure 3.5a**). The device and operation were similar to what we demonstrated previously for MOWChIP-seq[149]. MeDIP started with flowing a suspension of antibody-coated magnetic beads (IP beads) into the microfluidic chamber (~800 nl) and the IP beads were packed against a partially closed microvalve to form a packed bed (**Figure 3.5b**). Sonicated single-stranded DNA (ssDNA) (100-500 bp, in 50 µl, **Figure 3.2**) were then flowed through the bead-packed bed, allowing targeted methylated DNA to adsorb onto the bead surface (**Figure 3.5c**). After MeDIP, we applied alternating pressure pulses at the two ends of the microfluidic chamber (at 0.5 p.s.i. and with a duration of 0.5 s for each pulse) for 5 min to create oscillatory washing to effectively remove nonspecific binding (**Figure 3.5d**). Such washing was conducted in MeDIP buffer. After the oscillatory washing, the beads were retained by a magnet on one side of the chamber while the unbound ssDNA fragments and other debris/waste were flushed out of the microfluidic chamber by MeDIP buffer (**Figure 3.5e**). Finally, the MeDIP beads with adsorbed ssDNA fragments were flashed out of the chamber and collected for off-chip processing.

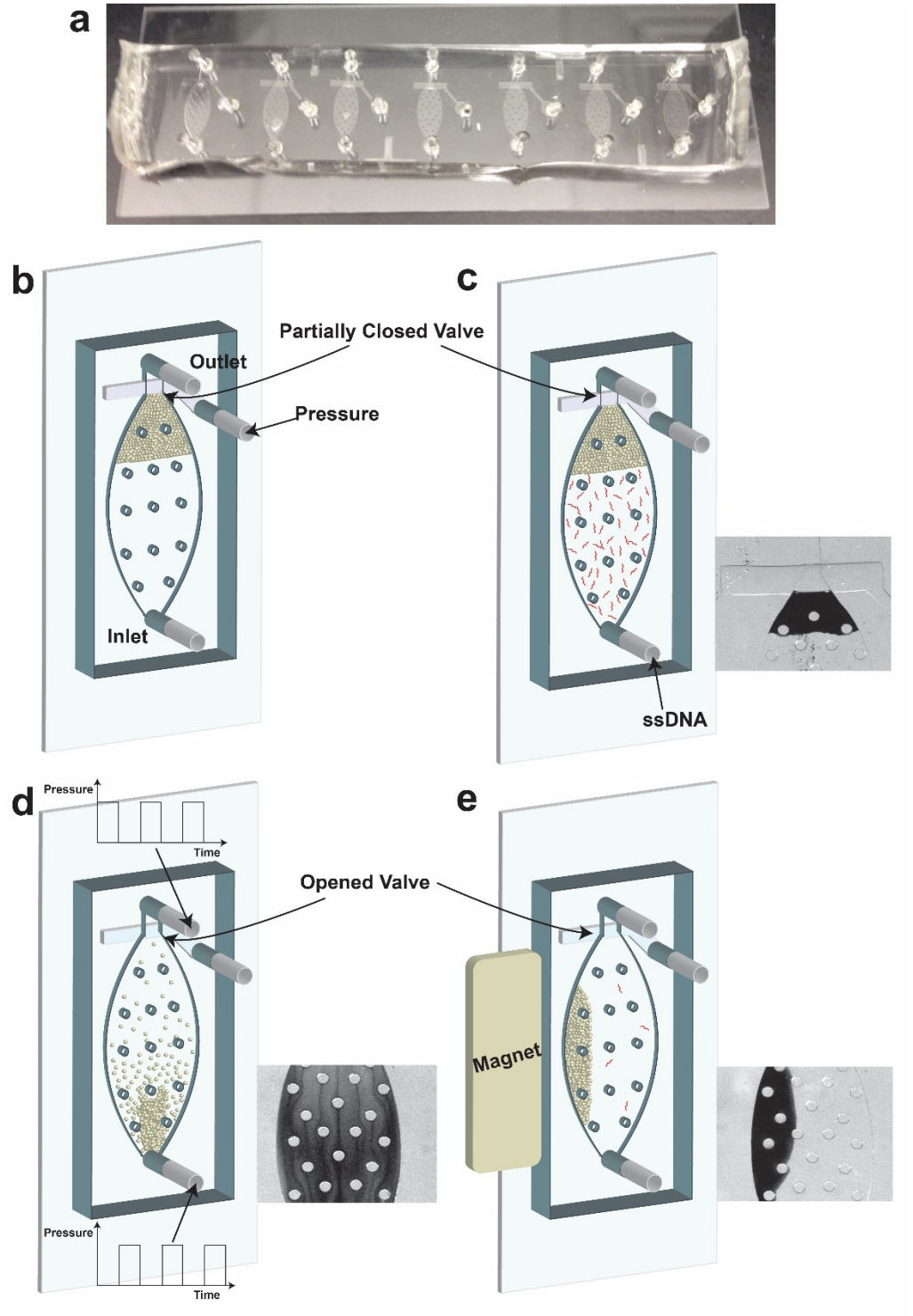


Figure 3.5 Microfluidic MeDIP-seq device and its operation. **(a)** The microfluidic device (containing 7 independent units). **(b)** Formation of packed bed of IP beads while micro-valve is closed. **(c)** MeDIP by flowing ssDNA fragments through the packed bed. **(d)** Oscillatory

washing. **(e)** Retaining IP beads on one side of chamber to remove unbound chromatin fragments and debris.

We first tested the technology using a lymphoblastic cell line GM12878. Our microfluidic MeDIP process collected 0.2 ng and 1.2 ng DNA out of the inputs of 10 and 100 ng DNA, respectively. MeDIP-seq results were obtained using our microfluidic technology with various amounts of starting DNA (100-0.5 ng) from the human cell line GM12878 (**Table 3.1**). As expected, the quality of methylomic profiles slowly declined when the DNA sample amount decreased from 100 to 0.5 ng (**Figure 3.6a**). The correlation coefficients r between the two replicates were excellent for samples in the range of 100-5 ng (0.93, 0.91, 0.90 for 100-, 10-, 5-ng samples, respectively) (**Figure 3.6b**). Data obtained using 10- and 5-ng samples were also highly correlated with those of 100-ng samples (with average r of 0.91 and 0.90, respectively). For 0.5 ng samples, the correlation between the two replicates was 0.80 and the correlations to 100 ng data were 0.69 and 0.60 for the two replicates. Nevertheless, most of the major MeDIP peaks can be seen in the 0.5-ng data (**Figure 3.6a**).

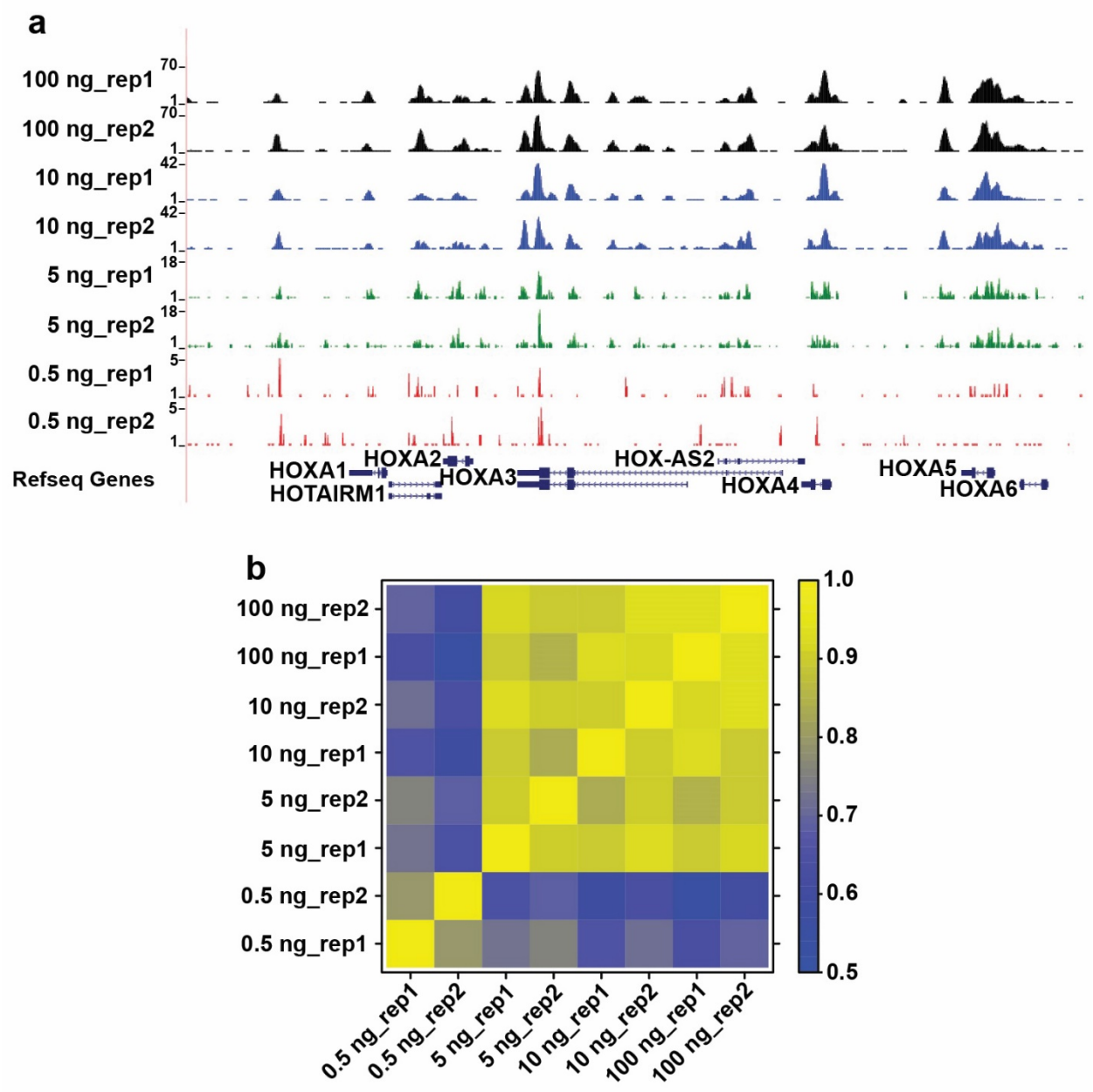


Figure 3.6 Microfluidic MeDIP-seq data on GM12878 cell line. **(a)** Genome browser tracks for our microfluidic MeDIP-seq using various amounts of DNA (100-0.5 ng). Two replicates were profiled for each sample. **(b)** Genome-wide correlations among MeDIP-seq data sets of various sample sizes. MeDIP-seq genomic coverage profiles were used for computing correlations. Colors represent Pearson correlation coefficients.

Sample	# Total Reads (M)	% Unique Mapped	% Total Mapped	Redundancy Rate	# Unique Reads(M)	# Peaks
100ng_rep1	23.8	47.41	76.7	0.43	11.3	104141
100ng_rep2	22	52.35	79.34	0.25	11.5	106679
10ng_rep1	20.9	48.15	74.14	0.58	10.1	71960
10ng_rep2	26.6	49.54	73.16	0.44	13.2	130783
5ng_rep1	20.7	54.5	79.07	0.45	11.3	165800
5ng_rep2	24.1	58.73	81.86	0.33	14.2	198914
0.5ng_rep1	27.8	59.61	78.82	0.76	16.6	91840
0.5ng_rep2	21.7	59.48	76.56	0.61	12.9	53024

Table 3.1 Summary of MeDIP-seq data on GM12878 from various starting amounts (0.5 ng-100 ng).

To quality-control MeDIP-seq data, saturation analysis was conducted by determining the number of reads needed to generate a genome-wide profile that could be reproduced by another independent set of reads of similar number. We defined saturation sequencing depth as the minimum number of reads needed for each of two subsets of sequencing reads that yielded a >0.95 Pearson correlation coefficient between them. 0.5 ng DNA samples required an average 3.17 million reads to saturate (**Figure 3.7**). With the input amount DNA going up, 100 ng samples only required an average of 0.89 million reads to saturate (**Figure 3.7**).

The saturation analysis splits the total set of regions into two different random sets of equal size A and B. Both sets A and B are further divided into 10 random subsets of equal size. For each set, A and B, the saturation analysis iteratively picks an increasing number of subsets and computes short read coverage at genome wide windows of 250 bp. The Pearson correlation of genome wide coverages for subsets of A and B then are calculated after each iteration step. It is assumed

that the resulting genome wide coverages become more similar as the number of reads in each subsets increases. However, such analysis only covers half of available short reads for a given sample. To examine the reproducibility for the total set of available short reads of the given sample, the full set of given regions is artificially doubled by considering each given region twice to create an estimated saturation. Subsequently, the described saturation analysis is performed on the artificially doubled set of regions.

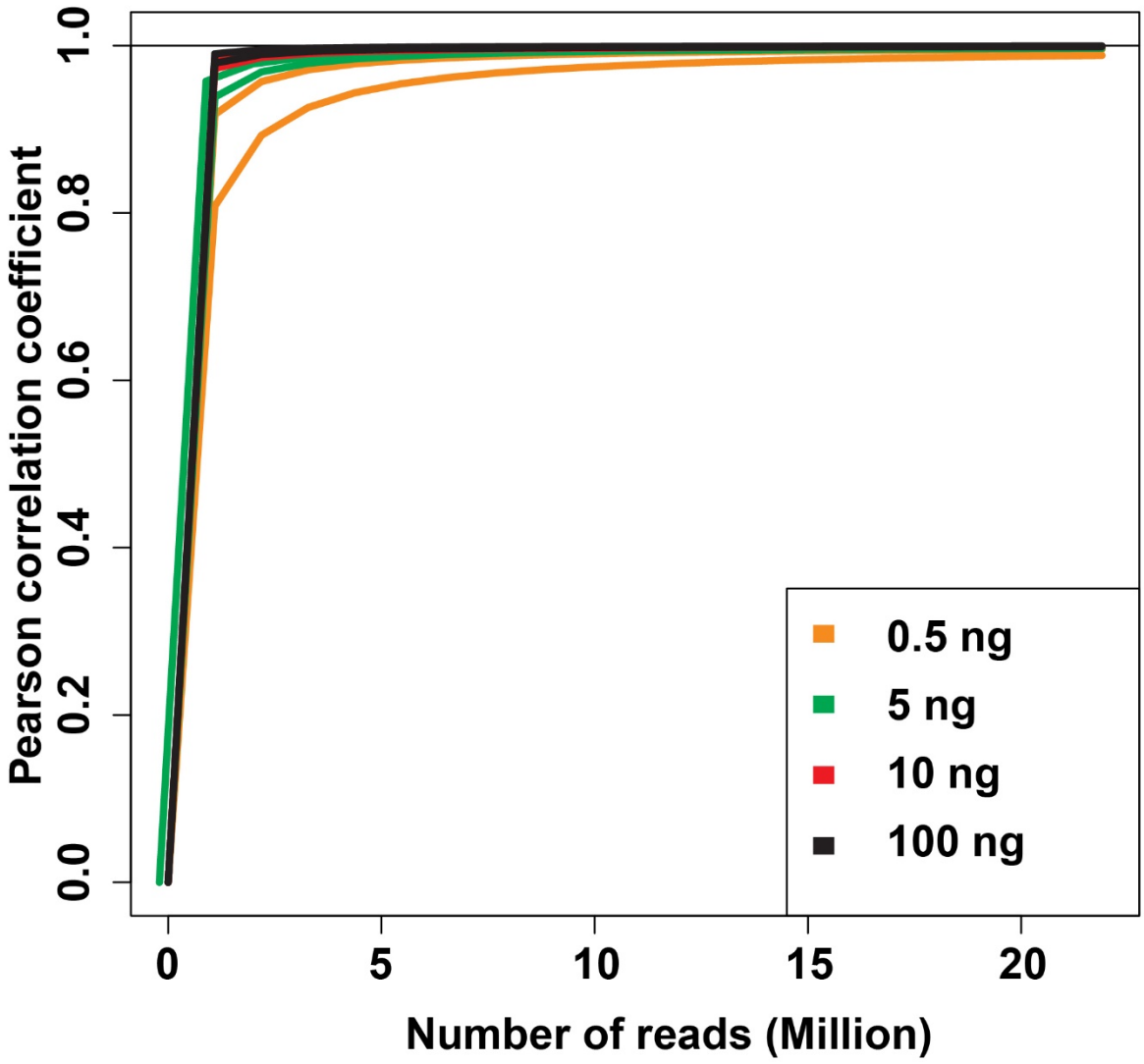
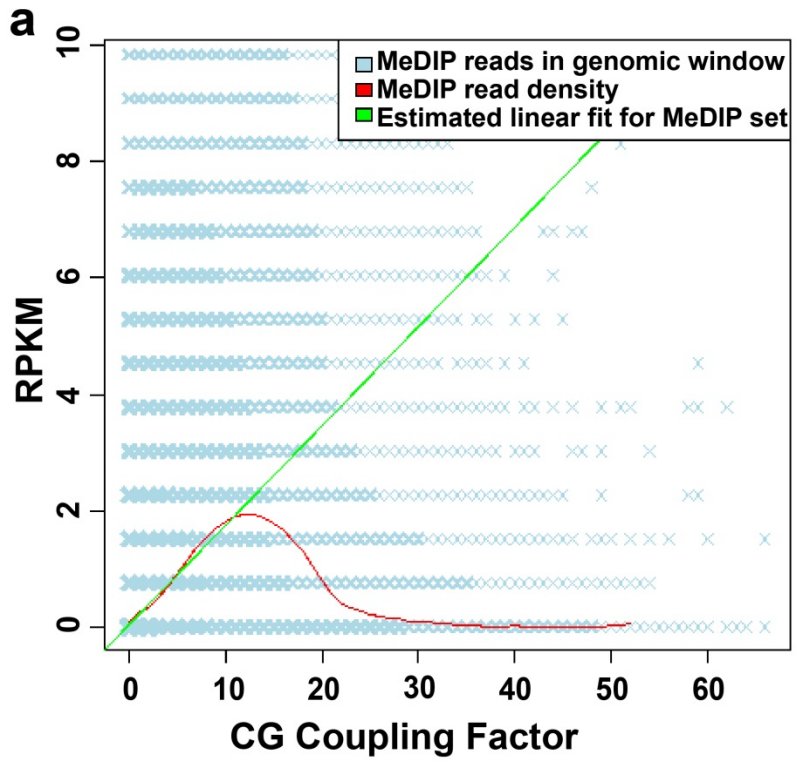


Figure 3.7 Saturation analysis on GM12878 MeDIP-seq data taken with different input amounts (0.5-100 ng) indicate the reproducibility of the genome wide coverage at regular genomic intervals given an increasing sequencing depth.

MeDIP-seq data were normalized based on the coverage and CpG density in genome-wide windows (of 250 bp each)[172]. This can be visualized as calibration plots. The 0.5 ng (**Figure 3.8a**) and 100 ng (**Figure 3.8b**) plots showed that most CpG-poor regions were methylated, whereas CpG-rich regions were generally unmethylated, revealing the linear relationships between the MeDIP-seq reads and the density of methylated CpGs.


a
MeDIP reads in genomic window
MeDIP read density
Estimated linear fit for MeDIP set
RPKM
0
2
4
6
8
10
0
10
20
30
40
50
60
CG Coupling Factor

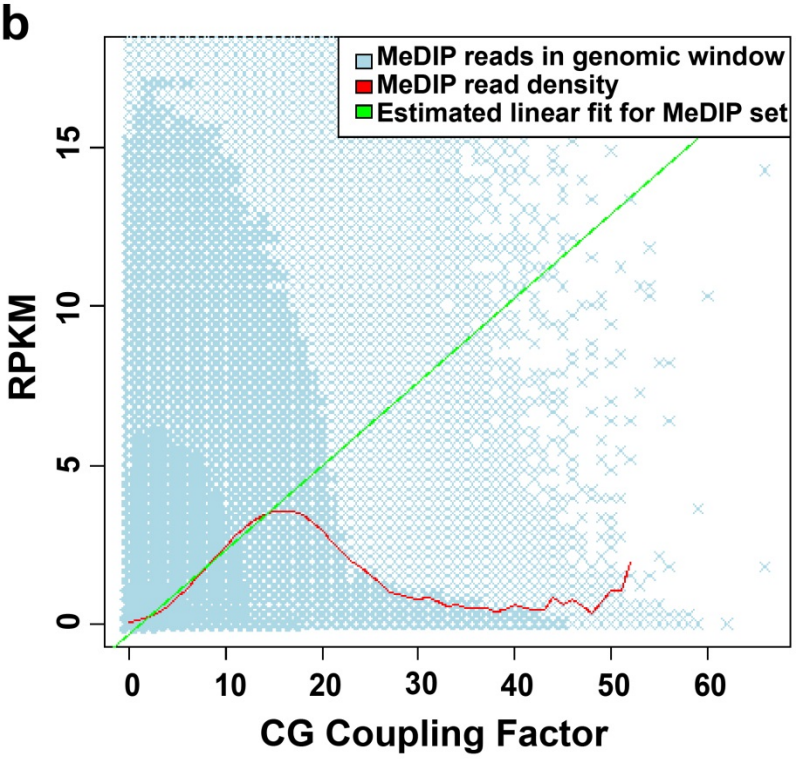

b
MeDIP reads in genomic window
MeDIP read density
Estimated linear fit for MeDIP set
RPKM
0
5
10
15
0
10
20
30
40
50
60
CG Coupling Factor

Figure 3.8 Calibration plots. RPKM is MeDIP Reads Per Kilobase Million. CG Coupling Factor is CpG density within given genomic widow. MeDIP-seq data taken with **(a)** 0.5 ng DNA from GM12878 cell line. **(b)** 100 ng DNA from GM12878 cell line.

We obtained methylomic profiles on mammary tumors of different stages that harvested from C3(1)/SV40 T-antigen transgenic mice (**Figure 3.9a & Table 3.2**). This transgenic mouse model contains a recombinant gene expressing simian virus 40 early-region transforming sequences under the regulatory control of the rat prostatic steroid binding protein C3(1) gene[123]. The expression of TAg in the mammary epithelium results in progressive lesions and tumor development in all female mice. Atypia of the mammary ductal epithelium develops at about 8 weeks of age, representing low-grade mammary intraepithelial neoplasia (MIN). The atypical lesions progress to high-grade MIN, resembling human DCIS (ductal carcinoma in situ) at about 12 weeks of age. Invasive carcinomas are typically observed at about 16 weeks of age[122].

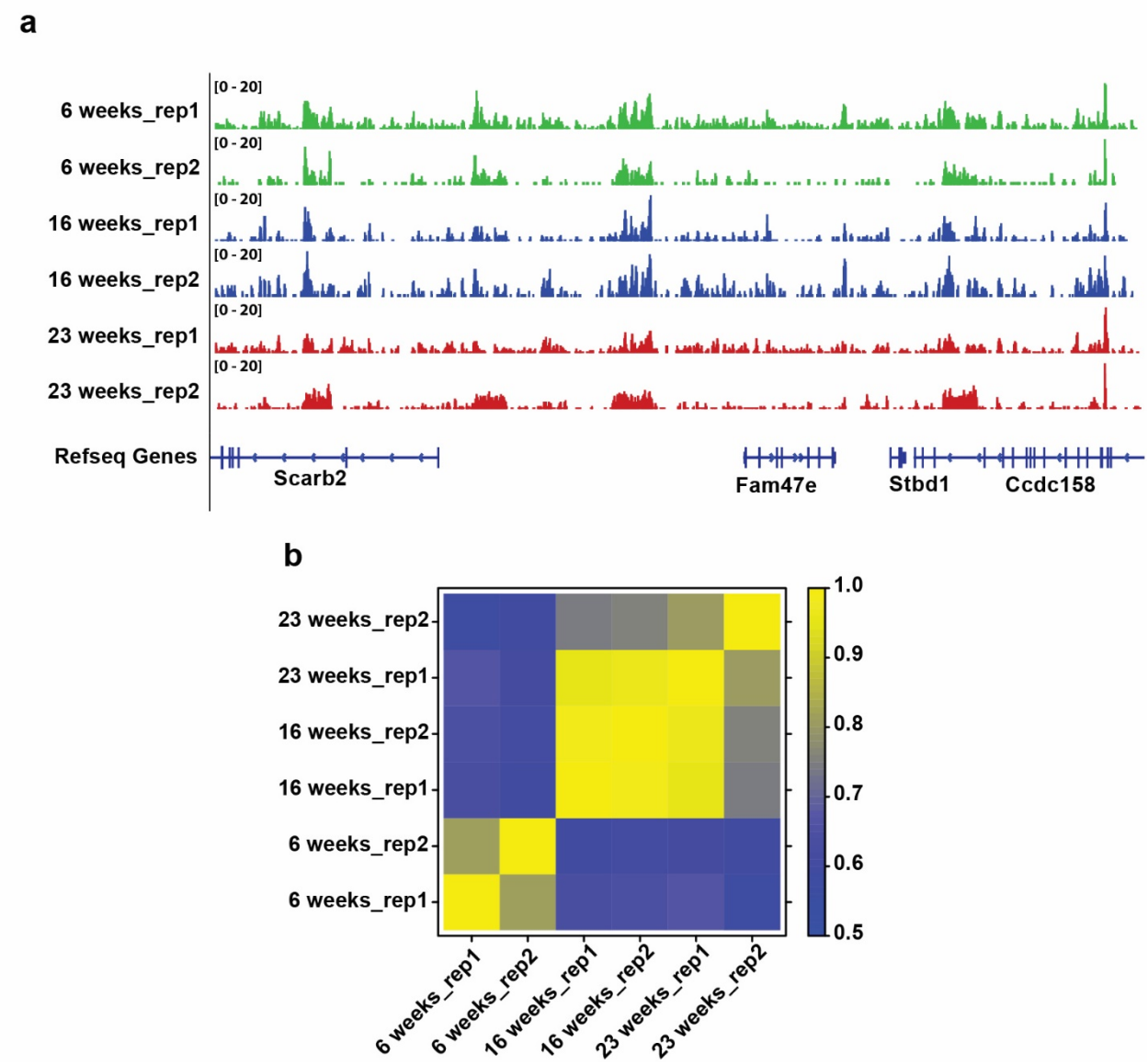


Figure 3.9 Microfluidic MeDIP-seq data on transgenic mouse samples. **(a)** Genome browser tracks for microfluidic MeDIP-seq using 10 ng DNA from different stages of mammary tumor harvested from C3(1)/Tag transgenic mice. Two replicates were profiled for each sample. **(b)** Genome-wide correlations among MeDIP-seq data sets of different stages of mammary cancer. MeDIP-seq genomic coverage profiles were used for computing correlations. Colors represent Pearson correlation coefficients.

Sample	# Total Reads (M)	% Unique Mapped	% Total Mapped	Redundancy Rate	# Unique Reads(M)	# Peaks
6wk_rep1	43.5	66.69	89	0.43	29.0	54506
6wk_rep2	54.9	11.77	42.52	0.84	6.5	12594
16wk_rep1	22.2	41.7	73.71	0.63	9.3	49562
16wk_rep2	22.4	53	84.98	0.4	11.9	57306
23wk_rep1	31.9	52.82	85.45	0.38	16.8	28772
23wk_rep2	84.9	44.67	60.66	0.88	37.9	10627

Table 3.2 Summary of MeDIP-seq data on transgenic mice mammary gland tissue from different stages of tumorigenesis (6 weeks – 23 weeks).

Histologic progression of formalin-fixed, paraffin-embedded mammary normal tissue and mammary tumor tissue were stained by Hematoxilin and Eosin then imaged. Mammary tissues were first harvested from 6 week old C3(1)/Tag transgenic mice (**Figure 3.10a**). Early stage of mammary intraepithelial neoplasia progressed into invasive carcinoma beginning at about 16 weeks, with cribriform patterns formed in mammary tumor tissue (**Figure 3.10b**). As tumor progressed, more cribriform patterns were found in 23 week mammary tumor tissue (**Figure 3.10c**).

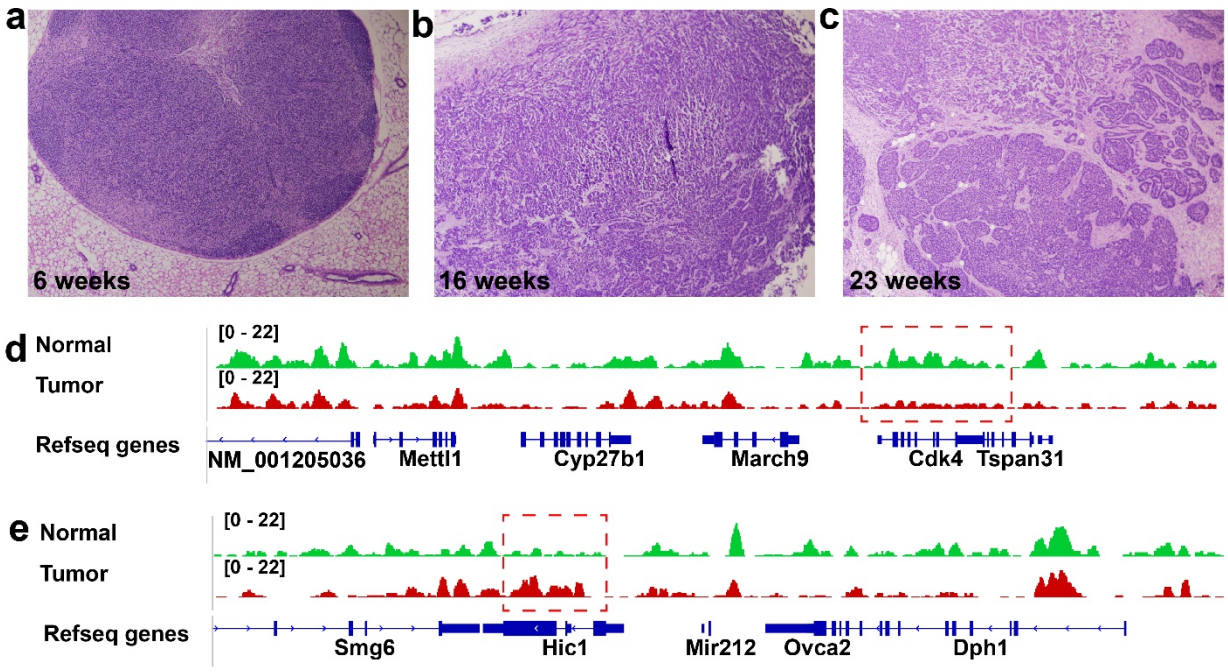


Figure 3.10 Histologic progression of mammary tumors in C3(1)/Tag transgenic mice and MeDIP-seq profile comparing 23 weeks tumor to 6 weeks normal tissue. **(a)** Normal mammary tissue in a 6-week-old mouse. **(b)** Invasive carcinoma with cribriform pattern in a 16 week-old mouse. **(c)** Invasive carcinoma with extensive cribriform patterns in a 23 week-old mouse. **(d)** Hypomethylation at oncogene Cdk4 with tumor development. **(e)** Hypermethylation at TSG Hic1 with tumor development.

The correlation coefficient r between two replicates was 0.82, 0.99, and 0.81 for 6 week 16 week, 23 week samples, respectively. The 23 week samples were more consistent with 16 week samples (average r = 0.86) than with 6 week samples (average r = 0.63) (**Figure 3.9b**). There were substantial changes in the MeDIP-seq peaks at critical genes with tumor development. For example, when comparing MeDIP-seq profiles of 23 week samples (invasive carcinomas tumor tissues) and 6 week samples (healthy mammary tissue), we observed peak decrease (decreased

level of methylation or hypomethylation) at oncogene CDK4 with development of tumor (**Figure 3.10 d**). In contrast, at tumor suppress gene HIC1, higher peaks (hypermethylation) were observed in mammary tumor tissues of 23 weeks (**Figure 3.10e**).

Using EdgeR test ($p < 0.01$), we were able to identify 29984 differentially methylated regions (DMRs) between 6 and 16 week samples, 16483 DMRs between 16 and 23 week samples. The 3938 DMRs were included in both sets. This means that 12545 DMRs were newly developed during the period of 16-23 weeks (**Figure 3.11 and Table 3.3**). In addition, we tracked hypermethylation and hypomethylation, which were increase or decrease in the methylation level over the course of mammary tumor development, respectively. When the 16 week samples were compared to 6 week ones, out of the total of 29984 DMRs, 12532 (42%) were hypermethylated and 17452 (58%) were hypomethylated. The 9747 (59%) DMRs were hypermethylated and 6736 (41%) DMRs were hypomethylated when comparing 23 week data with 16 week data. Interestingly, 1772 out of the 12532 hypermethylated DMRs found at 16 weeks were also identified as hypermethylated during the period of 16 to 23 weeks (i.e. there were 7975 newly developed hypermethylated DMRs from 16 to 23 weeks). In contrast, none of the hypomethylated DMRs found during 6-16 weeks were identified as hypomethylated during 16 to 23 weeks (i.e. all the 6736 hypomethylated DMRs developed during 16 to 23 weeks without any prior decrease in the methylation level during 6-16 weeks).

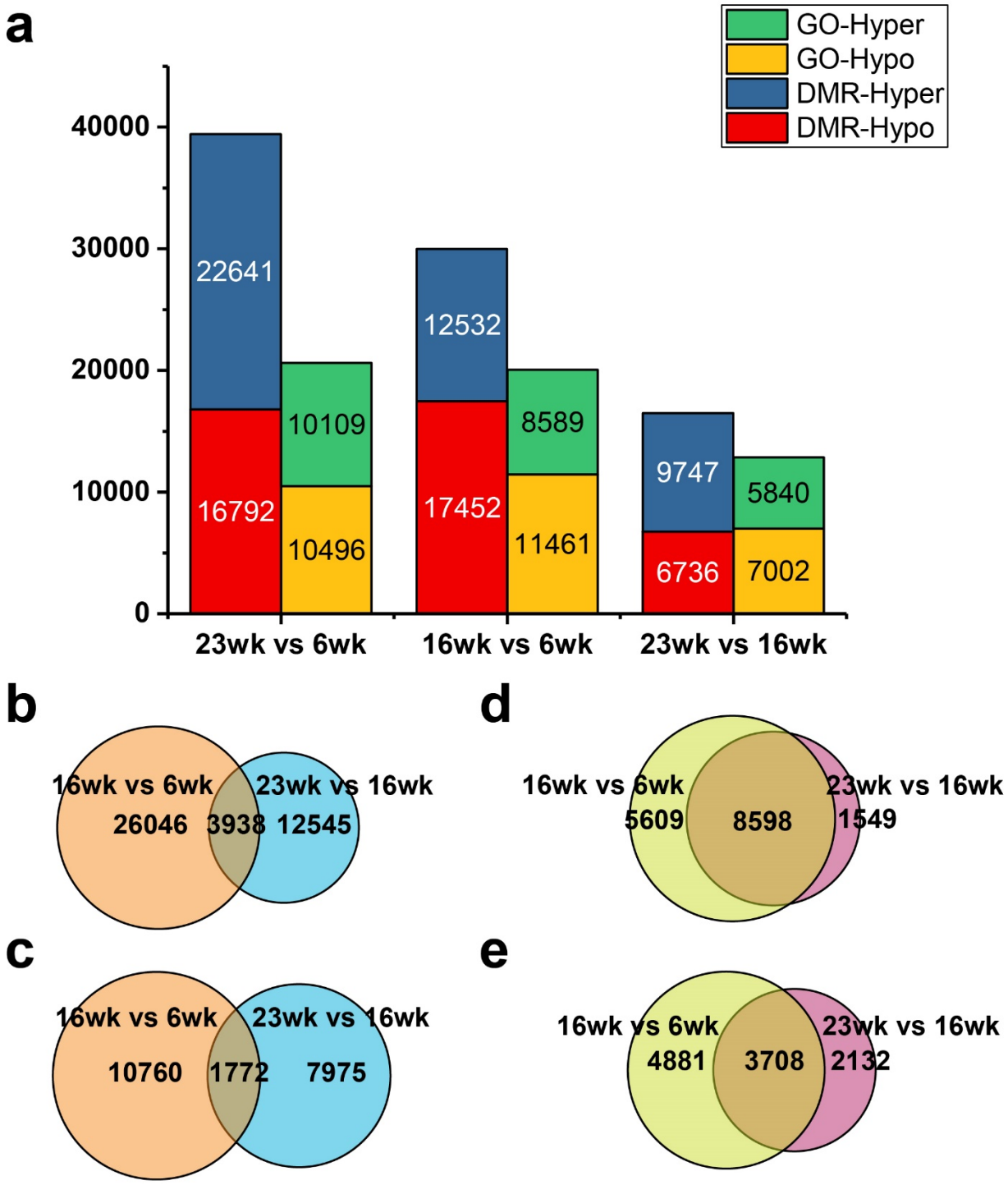


Figure 3.11 Summary of DMRs and GO genes. **(a)** Hyper and Hypo methylation occurred at DMRs and GO genes identified from comparing 16wk vs 6wk group, 23wk vs 16wk group, and

23wk vs 6wk group. **(b)** Comparison of all DMRs identified from 16wk vs 6wk group and 23wk vs 16wk group. **(c)** Comparison of Hyper methylation DMRs identified from 16wk vs 6wk group and 23wk vs 16wk group. **(d)** Comparison of all GO genes identified from 16wk vs 6wk group and 23wk vs 16wk group. **(e)** Comparison of Hyper methylated GO genes identified from 16wk vs 6wk group and 23wk vs 16wk group.

	DMR-Total	DMR-Hyper	DMR-Hypo	GO-Total	GO-Hyper	GO-Hypo
23wk vs 6wk	39433	22641	16792	14856	10109	10496
16wk vs 6wk	29984	12532	17452	14207	8589	11461
23wk vs 16wk	16483	9747	6736	10147	5840	7002
16wk vs 6wk Unique	26046	10706	17452	5609	4881	11461
16wk vs 6wk Intersect w/ 23wk vs 16wk	3938	1772	0	8598	3708	0
23wk vs 16wk Unique	12545	7975	6736	1549	2132	7002

Table 3.3 Summary of DMRs and GO genes identified by comparing different stages of tumorigenesis.

We also divided DMRs to genomic regions including introns, exons, promoters, CpG islands, CpG islands in promoter regions (**Figure 3.12**). There were generally more hypermethylation DMRs than hypomethylation ones in the overall level and all these specific genomic regions for the two analyses (6-16 and 16-23 week). During 6-16 weeks, we found particularly large differences between hyper- and hypo-methylation in CpG islands and CpG islands in promoter regions (**Figure 3.12a**). In comparison, in the 16-23 week analysis, hypermethylation was substantially more prevalent than hypomethylation in all categories of genomic regions analyzed

(**Figure 3.12b**). This also led to a much larger margin between hyper- and hypo-methylation in the entire genome in 16-23 weeks than in 6-16 weeks.

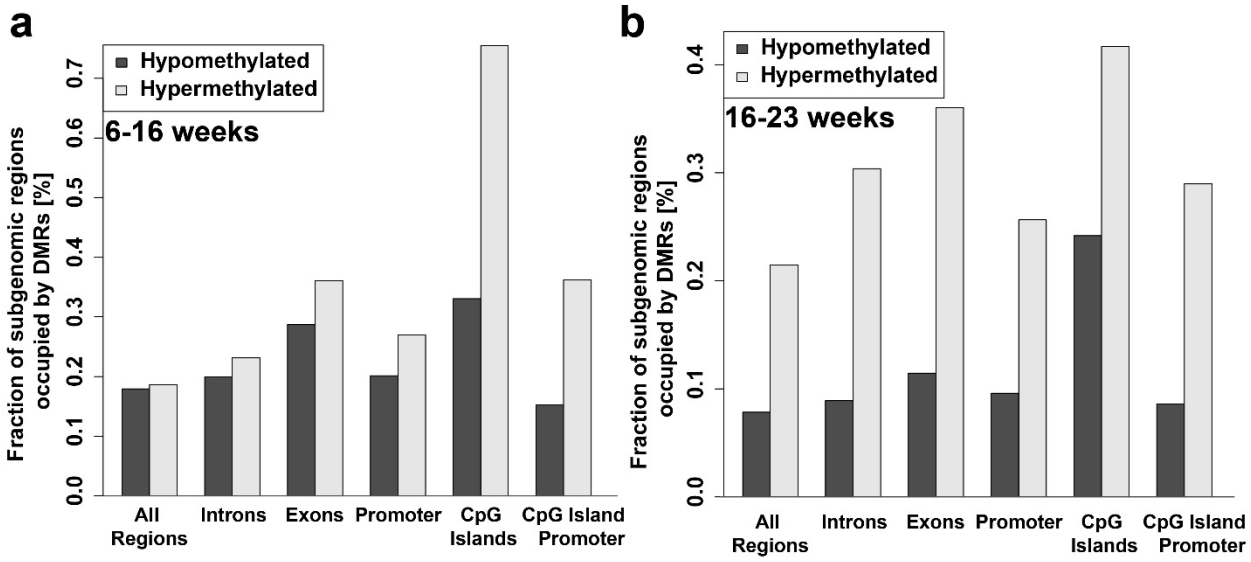


Figure 3.12 Fraction of subgenomic regions occupied by DMRs. Fraction is calculated as sum of all the 250 bp windows that were either hypo- or hypermethylated ($p<0.01$) in given subgenomic region divided by the overall size of that subgenomic region. **(a)** Fraction of subgenomic regions occupied by DMRs for hypo- and hypermethylation (16 weeks mammary tumor compared to 6 weeks mammary normal tissue). **(b)** Fraction of subgenomic regions occupied by DMRs for hypo- and hypermethylation (23 weeks mammary tumor compared to 16 weeks mammary tumor).

Genes associated with DMRs were identified by Genomic Regions Enrichment of Annotations Tool (GREAT)[180]. We were able to identify 14207 genes associated with 29984 DMRs when tumor progressed from 6 weeks to 16 weeks, and 10147 genes associated with 16483 DMRs when tumor progressed from 16 weeks to 23 weeks. A large percentage of the DMR-associated

genes in 16-23 weeks (84.7%, or 8598 out 10147) were also identified as DMR-associated during 6-16 weeks. GREAT analysis specifically identified enrichment of DMRs near genes involved in biological processes, mouse phenotype, and disease ontology in mammary tumorigenesis (**Figure 3.13**). GO terms are defined significant by both the binomial test over genomic regions (top 100 binomial P-value) and the hypergeometric test over genes (top 1500

hypergeometric FDR Q-value).

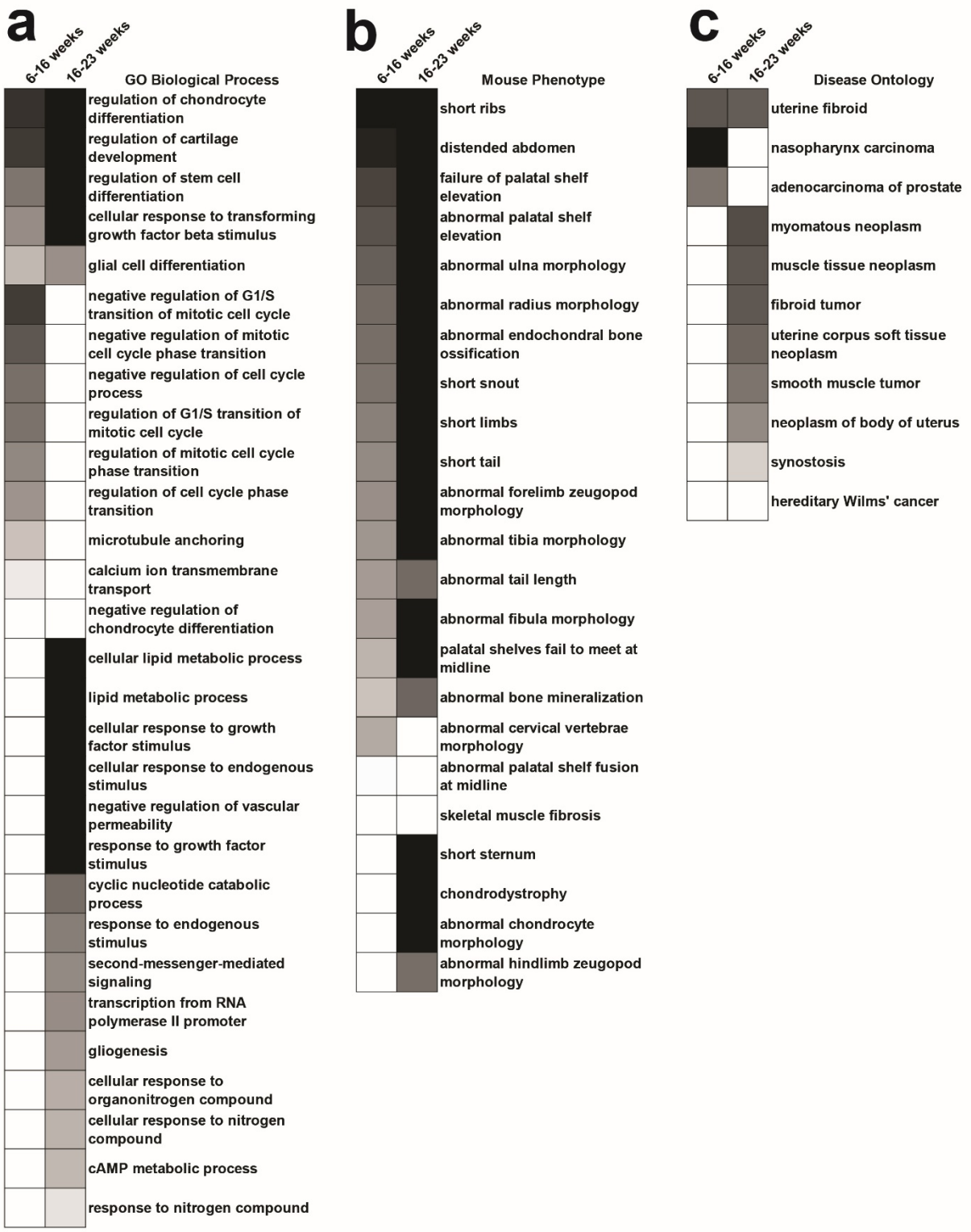


Figure 3.13 Enriched GO terms on **(a)** Biological Process, **(b)** Mouse Phenotype, (**c)** Disease Ontology in the two periods of 6-16 weeks and 16-23 weeks for mouse mammary tumor development. Annotation terms were in different shades that represent their statistical enrichments (black for the highest statistical significance, and white for the lowest statistical significance).

GREAT analysis specifically identified enrichment of hypermethylated DMRs near genes involved in biological processes, mouse phenotype, and disease ontology in mammary tumorigenesis from 6 weeks to 16 weeks (**Figure 3.14**). Regulation of cartilage development, regulation of chondrocyte differentiation, and mammary gland duct morphogenesis were identified in the top biologic processes. GREAT detected genes whose DNA methylation generated mouse phenotype such as short ribs, abnormal sternum ossification, and abnormal chondrocyte morphology. Adenosquamous carcinoma, muscle tissue neoplasm, and smooth muscle tumor were also identified as top category in GREAT disease ontology analysis.

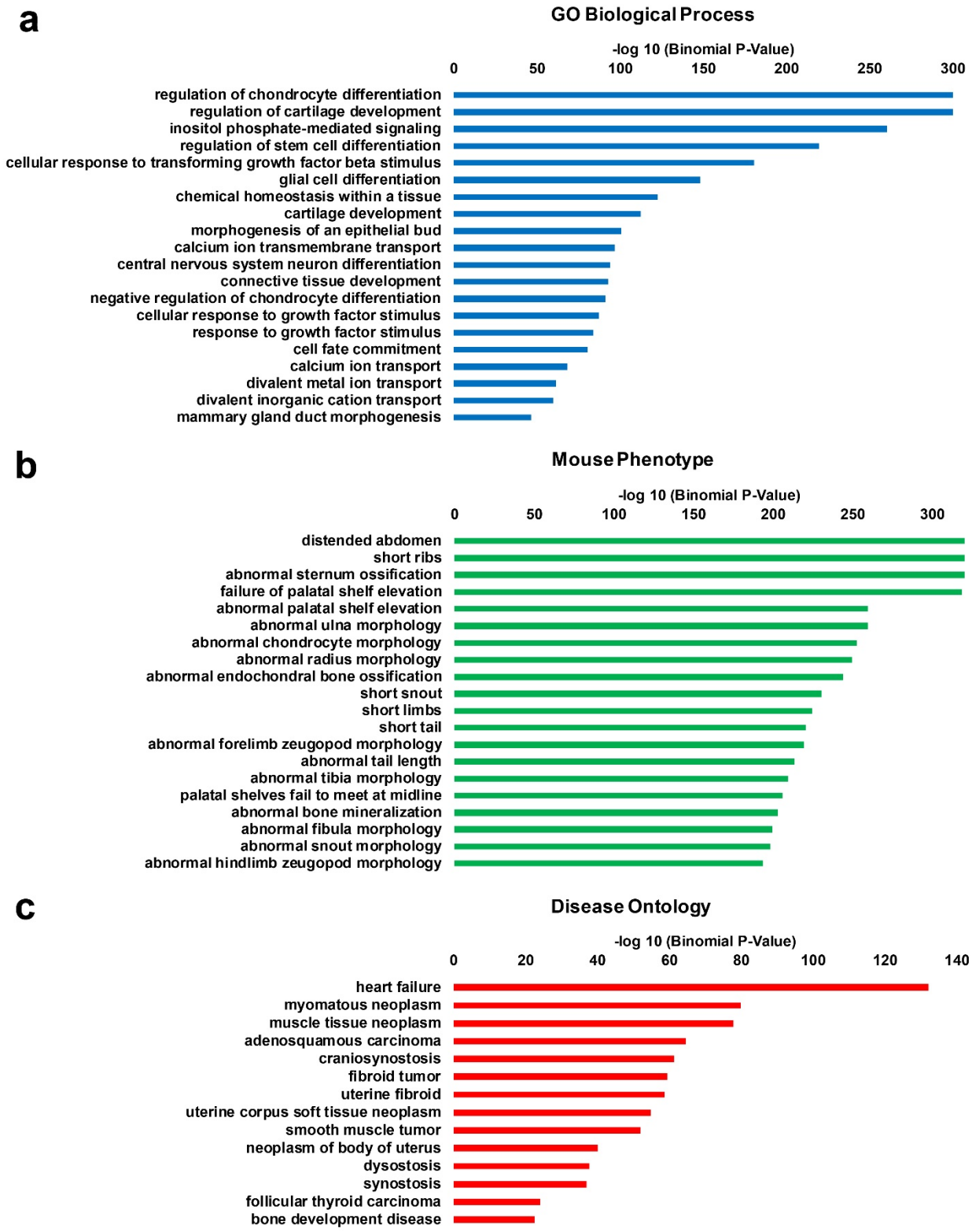
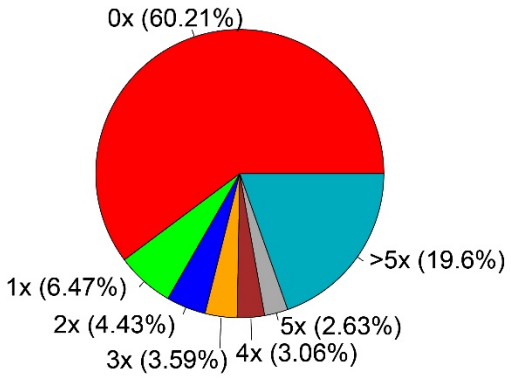

a
GO Biological Process
-log 10 (Binomial P-Value)
0
50
100
150
200
250
300
regulation of chondrocyte differentiation
regulation of cartilage development
inositol phosphate-mediated signaling
regulation of stem cell differentiation
cellular response to transforming growth factor beta stimulus
glial cell differentiation
chemical homeostasis within a tissue
cartilage development
morphogenesis of an epithelial bud
calcium ion transmembrane transport
central nervous system neuron differentiation
connective tissue development
negative regulation of chondrocyte differentiation
cellular response to growth factor stimulus
response to growth factor stimulus
cell fate commitment
calcium ion transport
divalent metal ion transport
divalent inorganic cation transport
mammary gland duct morphogenesis
b
Mouse Phenotype
-log 10 (Binomial P-Value)
0
50
100
150
200
250
300
distended abdomen
short ribs
abnormal sternum ossification
failure of palatal shelf elevation
abnormal palatal shelf elevation
abnormal ulna morphology
abnormal chondrocyte morphology
abnormal radius morphology
abnormal endochondral bone ossification
short snout
short limbs
short tail
abnormal forelimb zeugopod morphology
abnormal tail length
abnormal tibia morphology
palatal shelves fail to meet at midline
abnormal bone mineralization
abnormal fibula morphology
abnormal snout morphology
abnormal hindlimb zeugopod morphology
c
Disease Ontology
-log 10 (Binomial P-Value)
0
20
40
60
80
100
120
140
heart failure
myomatous neoplasm
muscle tissue neoplasm
adenosquamous carcinoma
craniosynostosis
fibroid tumor
uterine fibroid
uterine corpus soft tissue neoplasm
smooth muscle tumor
neoplasm of body of uterus
dysostosis
synostosis
follicular thyroid carcinoma
bone development disease

Figure 3.14 Functional enrichment analyses of DMRs related hypermethylation genes identified from 16wk vs 6wk group for **(a)** GO Biological Process, **(b)** GO Mouse Phenotype, **(c)** GO Disease Ontology.

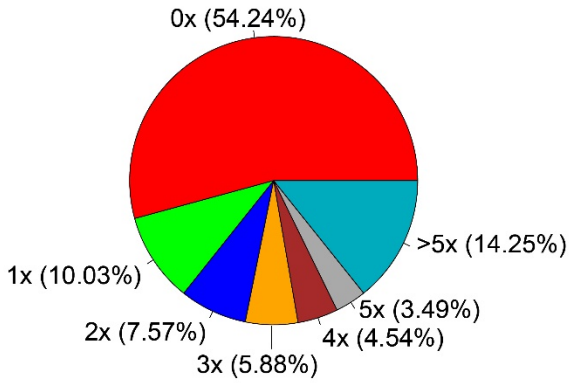
DNA methylation level is known to have strong association with chondrocyte differentiation activity and cartilage development [181-185]. Some important biological processes such as regulation of chondrocyte differentiation, regulation of cartilage development, and regulation of stem cell differentiation were found by GREAT in tumor progression in both the periods of 6 to 16 weeks and 16 to 23 weeks (**Figure 3.13a**). Mouse phenotype of abnormal endochondral bone ossification was another process related to DNA methylation and breast cancer development[186-188] that was found in tumor progression during both 6-16 weeks and 16-23 weeks (**Figure 3.13b**). Some unique stage-specific biological processes such as negative regulation of G1/S transition of mitotic cell cycle, negative regulation of mitotic cell cycle phase transition were only found when tumor progressed from 6 to 16 weeks. Other biological processes such as cellular lipid metabolic process, cellular response to growth factor stimulus were only found when tumor progressed from 16 to 23 weeks. Mouse phenotypes such as abnormal cervical vertebrae morphology, abnormal palatal shelf fusion at midline were only found when tumor progressed from 6 to 16 weeks while short sternum, chondrodystrophy were only found when tumor progressed from 16 to 23 weeks. Disease ontologies such as nasopharynx carcinoma, adenocarcinoma of prostate were only found when tumor progressed from 6 to 16 weeks and myomatous neoplasm, muscle tissue neoplasm were only found when tumor progressed from 16 to 23 weeks (**Figure 3.13c**).

We compared DMR-associated genes involved in our early stage mouse mammary tumors (6 to 16 weeks) to those of early stage mouse intestinal tumor (normal to adenoma) from previous publication[170]. 15 to 19 weeks isogenic B6-$APC^{Min/+}$ (APC^{Min}) and B6-$APC^{+/+}$ (B6) wild type littermates were used for comparison in their study of mouse intestinal tumors. We identified 37175 DMRs ($p<0.01$) between normal intestinal tissue and adenoma tissue, which were associated with 15761 genes. We found that 11814 of these genes (75.0 %) overlapped with DMR-associated genes involved in 6-16 week mouse mammary tumor development. This suggests that a largely common set of genes experience methylation variation during development of different types of cancers.

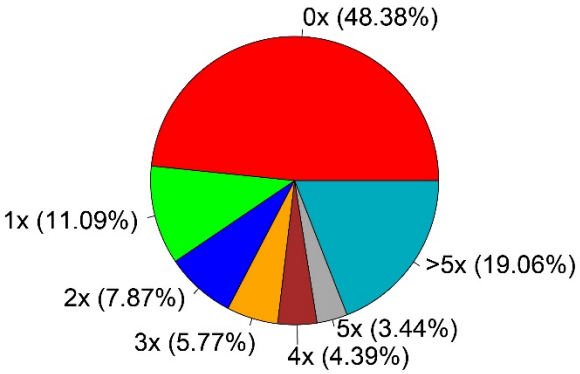
The sequence pattern coverage analysis (The coverage analysis shows the fraction of CpGs covered by the given reads according to their coverage level) was performed to test the number of CpGs covered by the given short reads, indicating an effective MeDIP enrichment[189, 190] (**Figure 3.15**).



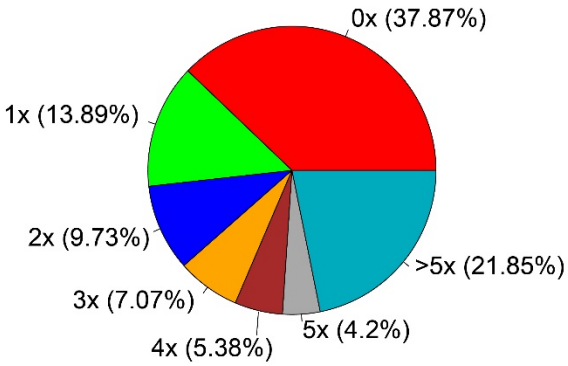
0.5 ng_rep1



0.5 ng_rep2



5 ng_rep1



5 ng_rep2

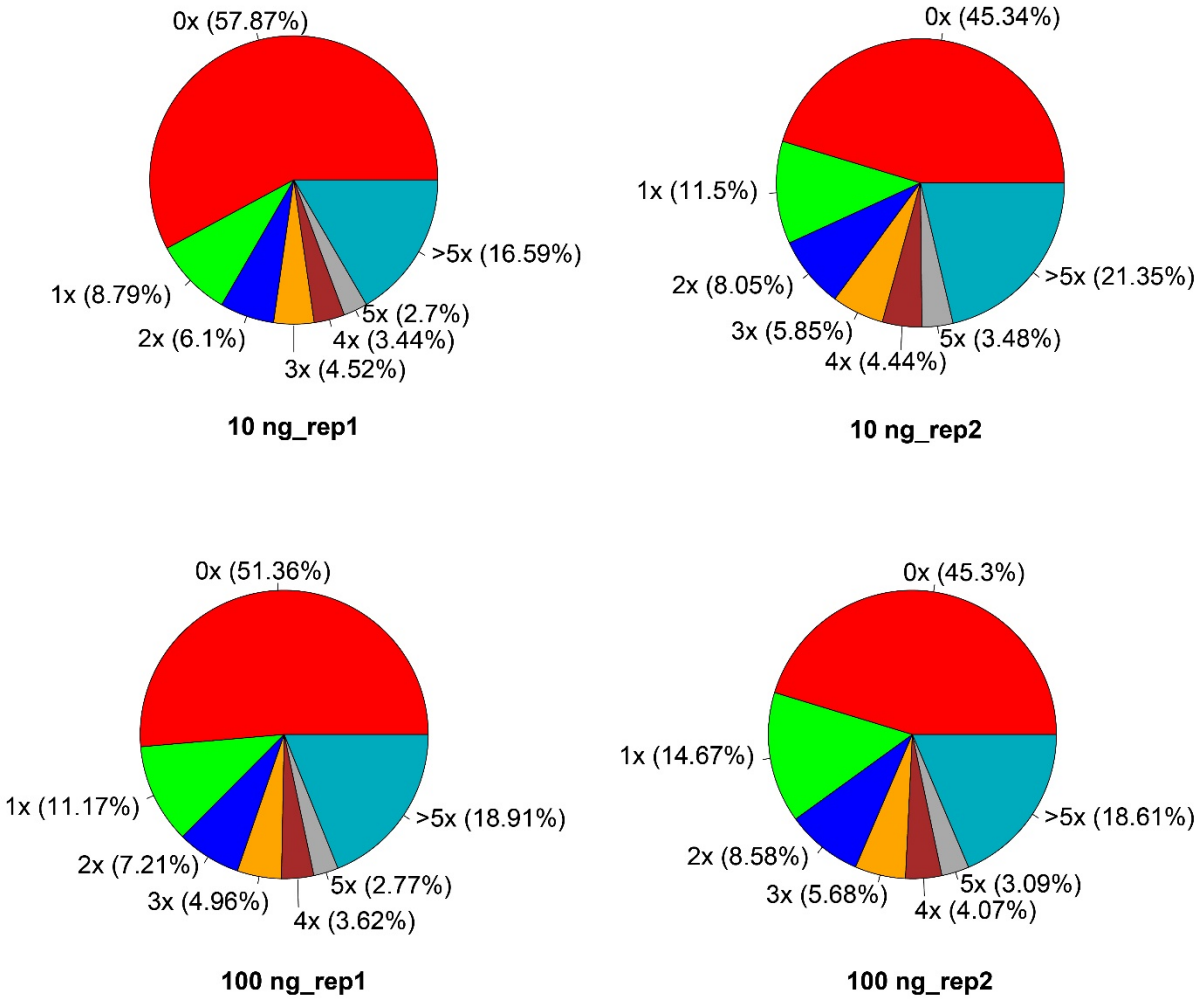


Figure 3.15 Coverage analysis - pie chart of **different input amount** GM12878 MeDIP-seq. The coverage analysis shows the fraction of CpGs covered by the given reads according to their coverage level.

The CpGs coverages can also be viewed as how frequent CpGs are covered by reads as shown (**Figure 3.16**). For example, 1 ng GM12878 MeDIP-seq has higher frequency for uncovered CpGs when compared to 10ng. The 10 ng has higher frequency for CpGs covered more than 5 times, also suggesting better data quality.

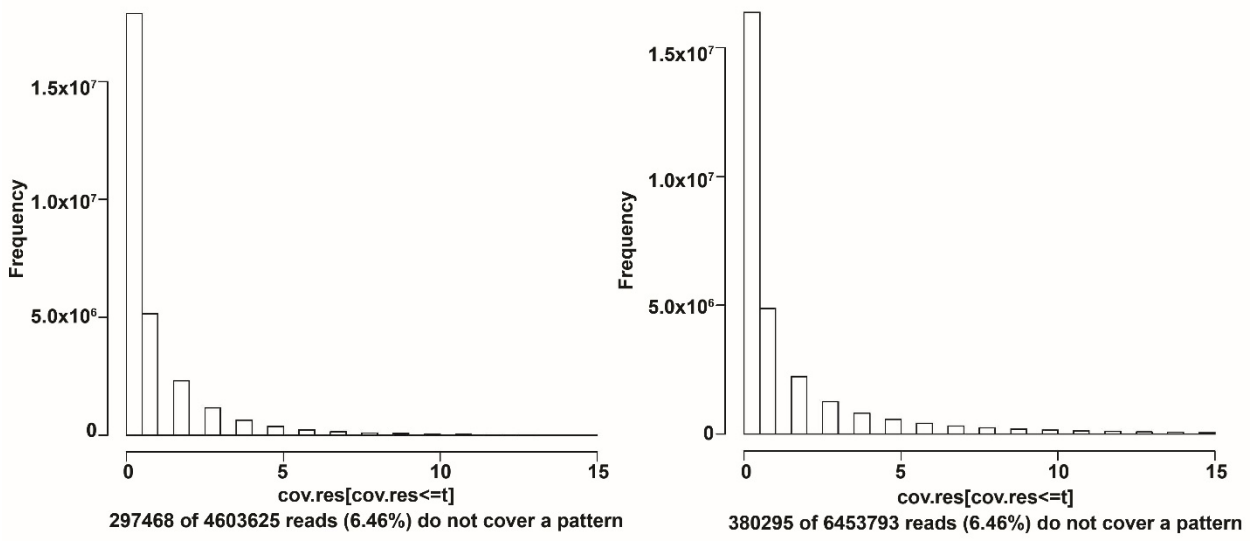


Figure 3.16 Coverage analysis - histogram of **(a)** 1 ng vs **(b)** 10 ng GM12878 MeDIP-seq. The histogram shows the distribution of CpG coverages.

To further check the enrichment of CpG rich ssDNA fragments obtained by the MeDIP step, the enrichment of CpGs within the genomic regions covered by the given set of short reads was compared to the full reference genome to generate CpG enrichment values. The number of Cs, the number of Gs, the number CpGs, and the total number of bases within the specified reference genome were counted. The relative frequency of CpGs and the observed/expected ratio of CpGs in the reference genome and immunoprecipitated regions were then calculated. The CpG enrichment is calculated by dividing the relative frequency of CpGs (or observed/expected ratio of CpGs) in the immunoprecipitated regions by the relative frequency of CpGs (or observed/expected ratio of CpGs) in the reference genome.

100 ng GM12878 MeDIP-seq has both higher enrichment of CpGs and enrichment obs/exp of 2.98 and 1.71 respectively comparing to 1.64 and 1.23 in 0.5ng data, suggesting better data quality (**Table 3.4**).

Sample	# Unique Reads(M)	# Cs (M)	# Gs (M)	# CpGs (M)	Enrichment of CpGs	Enrichment obs/exp
100ng_rep1	11.3	789	786.5	85.9	2.98	1.71
100ng_rep2	11.5	961.5	958.7	92.4	2.55	1.55
10ng_rep1	10.1	474.8	472.8	49.1	2.77	1.66
10ng_rep2	13.2	790.7	788.4	74.5	2.49	1.53
5ng_rep1	11.3	650.8	648.6	60.1	2.43	1.5
5ng_rep2	14.2	944.7	942.6	79	2.15	1.39
0.5ng_rep1	16.6	360.5	359.6	28.4	1.95	1.37
0.5ng_rep2	12.9	424.9	423.4	29.1	1.64	1.23

Table 3.4 CpG enrichment of MeDIP-seq data on GM12878 from various starting amounts (0.5 ng-100 ng).

The number of short reads in 250 bp window was counted across genome. Pearson correlation based on $\log_2$(counts) was calculated for different input DNA amount. The graph shows a more dispersed genomic coverage counts profile, there is relative low 0.68 Pearson correlation between 0.5 ng and 5 ng GM12878 MeDIP-seq (**Figure 3.17a**). The graph shows a condensed linear genomic coverage counts profile, pearson correlation between 10 ng and 5 ng GM12878 MeDIP-seq is 0.9, suggesting 5 ng is better correlated with 10 ng data rather than 0.5 ng data (**Figure 3.17b)**.

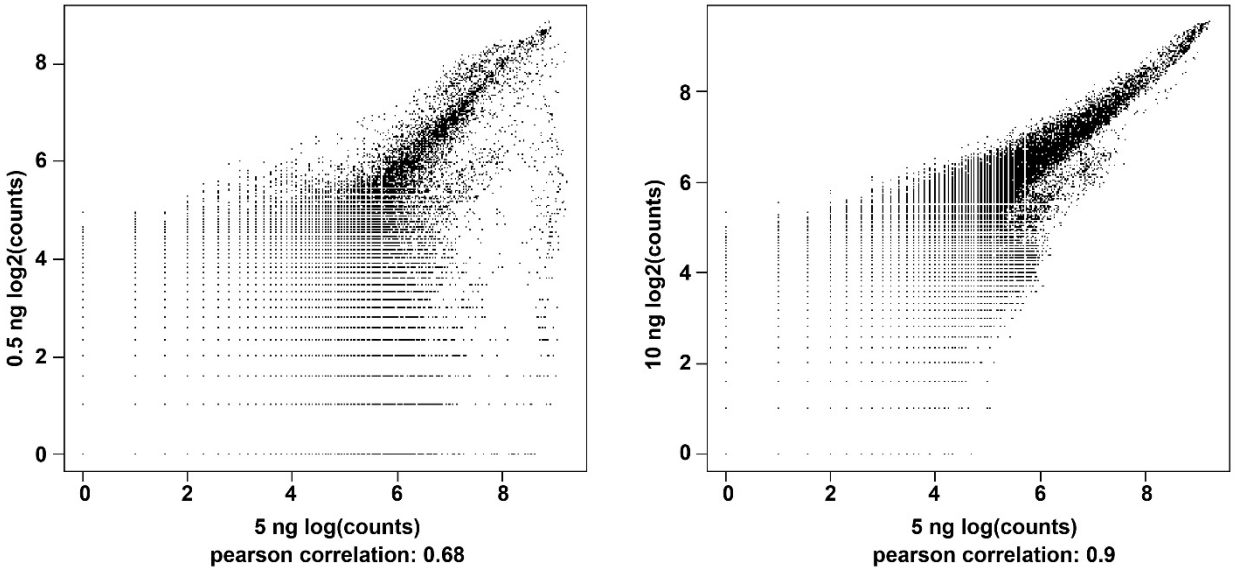


Figure 3.17 Pearson correlation based on genome coverage profile calculated as log2(counts). **(a)** Pearson correlation of 0.5 ng GM12878 MeDIP-seq genome coverage profile compared to 5 ng. **(b)** Pearson correlation of 10 ng GM12878 MeDIP-seq genome coverage profile compared to 5 ng.

To correct for multiple testing, regions at 10% false discovery rate (FDR) were identified. DMRs of 16 weeks tumor to 6 weeks normal were shown in MA plot (**Figure 3.18a**). The log ratios of the two conditions are called M values (from "minus" in the log scale). The mean values of the two conditions are called A values (from "average" in the log scale). Ones above x-axis were hypermethylation regions, and ones below x-axis were hypomethylation regions. The 29984 DMRs were found at P<0.01 depicted as orange points and 3215 DMRs were found at FDR<0.1 depicted as red crosses. The 39433 DMRs were found at P<0.01 and 11825 DMRs were found at FDR<0.1 when compared 23 week tumor to 6 week normal (**Figure 3.19a**) 16483 DMRs were

found at P<0.01 and 5764 DMRs were found at FDR<0.1 when compared 23 week tumor to 16 week tumor (**Figure 3.20a**)

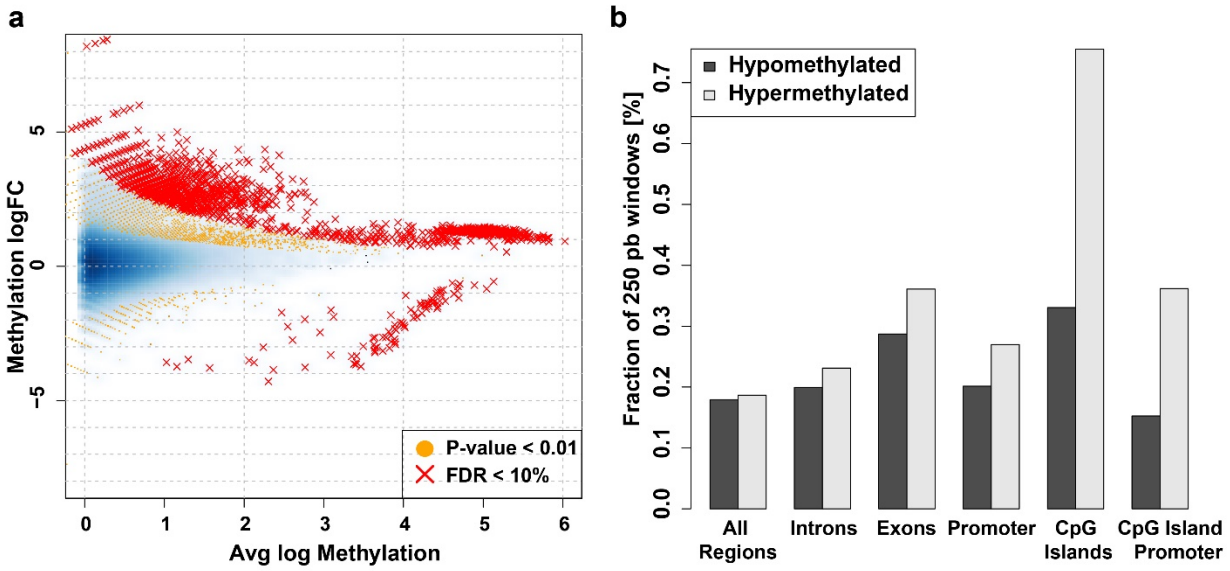


Figure 3.18 DMRs and their distribution in sub-genomic regions. **(a)** MA plot of 16 week tumor to 6 week normal. P-value was adjusted for multiple testing. **(b)** Fraction of differentially methylated regions calculated at P-value<0.01. Hypo- and hypermethylation of 16 week mammary tumor to 6 week mammary normal tissue at different genomic locations.

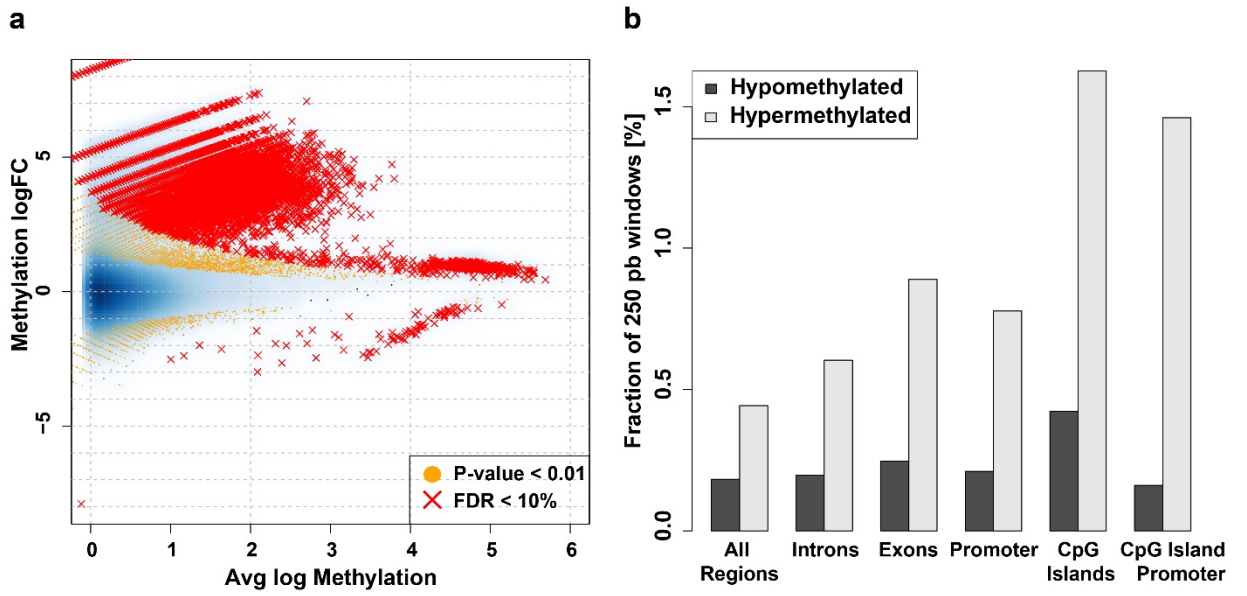


Figure 3.19 DMRs and their distribution in sub-genomic regions. **(a)** MA plot of 23 week tumor to 6 week normal. P-value was adjusted for multiple testing. **(b)** Fraction of differentially methylated regions calculated at P-value<0.01. Hypo- and hypermethylation of 23 week mammary tumor to 6 week mammary normal tissue at different genomic locations.

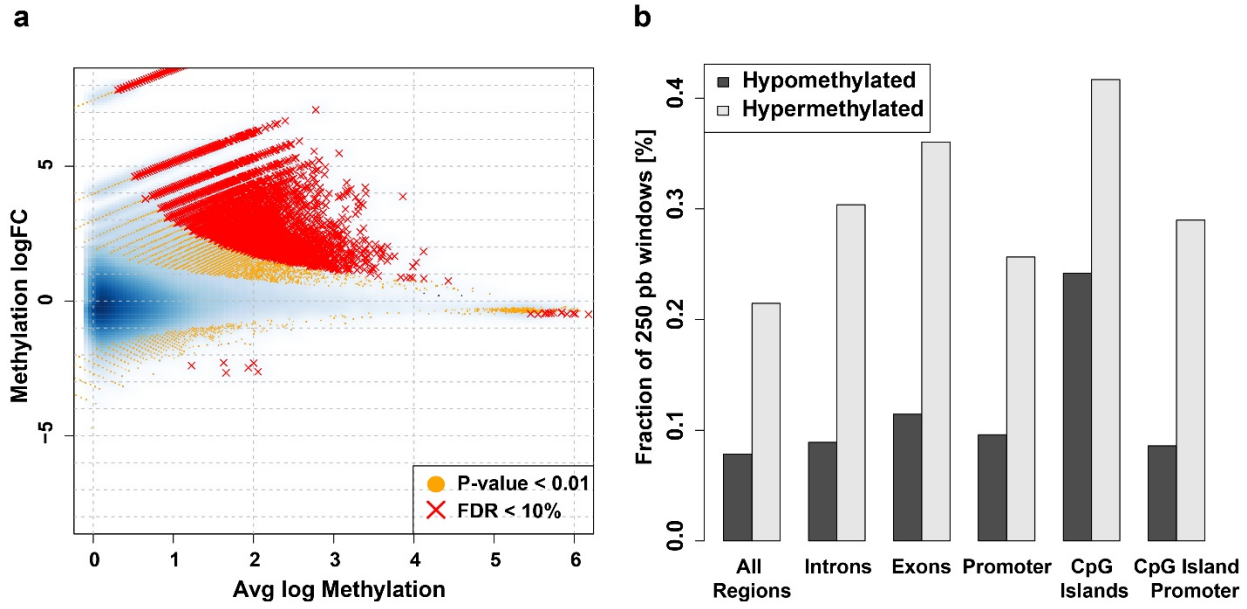


Figure 3.20 DMRs and their distribution in sub-genomic regions. **(a)** MA plot of 23 week tumor to 16 week tumor. P-value was adjusted for multiple testing. **(b)** Fraction of differentially methylated regions calculated at P-value<0.01. Hypo- and hypermethylation of 23 week mammary tumor to 16 week mammary tumor tissue at different genomic locations.

Our microfluidic technology is superior for conventional MeDIP for several reasons. First, high concentrations from trace amounts of molecules can be built up inside the tiny volumes that offered by microfluidic chamber. Adsorption kinetics and completeness was facilitated by such high concentration. Second, the IP beads take up a large fraction of the tiny volume so that the surface area/volume ratio (15-40% bead volume in 800 nl) is tremendously improved when compared to 5% bead volume in 1.5 ml conventional MeDIP[150]. The close proximity among beads greatly increased the efficiency and rate for chromatin adsorption on the bead surface due to the short diffusion lengths involved. The adsorption of a chromatin molecule among beads was rapid given that travel time $\tau_D \sim w^2/D$, where w is diffusion distance between two beads, and D is diffusivity. Third, by using microfluidic technique uniquely suited for bead manipulation at the microscale, we effectively remove nonspecific adsorption after high efficiency adsorption using microfluidic oscillatory washing. This is critical for producing high quality MeDIP DNA that preserves desired biological information. Finally, microfluidic device integrates various steps and minimizes material loss among steps.

4. Study of Epigenomic Regulations in Circulating Tumour Cells with Microfluidic Assays

4.1 Introduction

Cancer, a terrifying term, owes its ability to spread in the body. Tumors metastasize when they break away from original site and travel through the bloodstream to form new tumors at distant sites, such as brain, liver, and lungs (**Figure 4.1**). Those tumor cells detected in the bloodstream are known as circulating tumor cells (CTCs). Although CTCs are rarely found in healthy people, it has been documented that they can be severed as prognostic markers for breast, prostate, or colorectal cancer[191-195]. It is notoriously known that CTCs occur at very low concentrations in the peripheral blood. There are about 1–10 cells per 10 mL in most cancer patients[196], making it a challenge for CTCs detection. Although recent effort has been made in the developing microfluidic devices to enrich CTCs, but the discovery and validation of new CTC markers is still in its infancy.

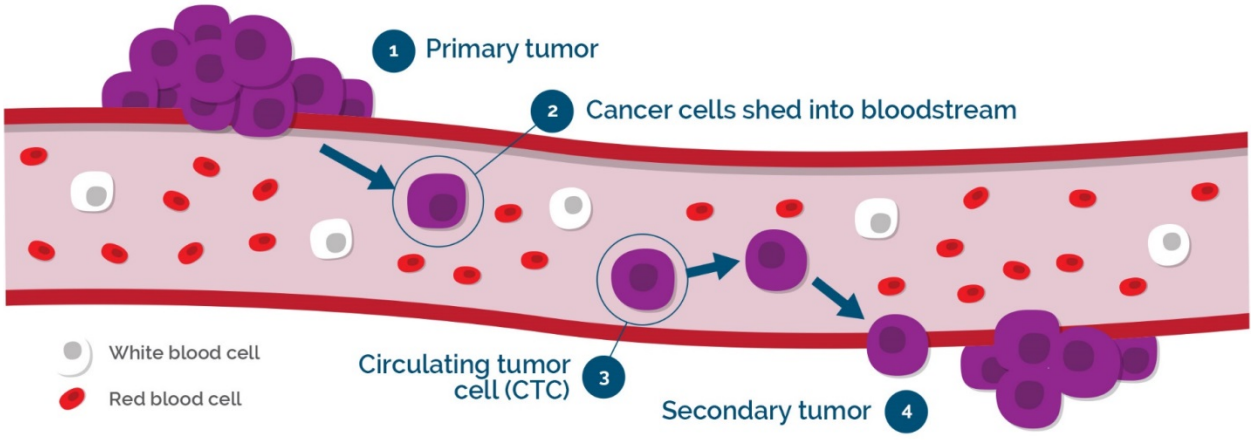


Figure 4.1 Generation of CTCs (adapted from https://vortexbiosciences.com/technology/).

In general, there are two ways to enrich CTCs (**Figure 4.2**). First, it is through physical properties of CTCs. There are several devices featured using physical properties to isolate CTCs. For example, a membrane and filtration-based system on the basis of the CTC size has been developed[197]. A microposts chip was built on basis of CTC size and its deformability.[198] Or centrifugation on a Ficoll density gradient on the basis of the CTC density has been adapted.[199] and a spiral CTC chip on the basis of the CTC size under the influence of Dean drag forces has also been created[200]. However, some CTCs can undergo epithelial-to-mesenchymal transition (EMT), and result in more deformable than regular CTCs[201]. More recently, CTCs of various sizes have been identified as well[202]. Isolating CTCs by just their physical properties is not sufficient.

The second way is through protein expression-based technologies. Epithelial cell adhesion molecule (EPCAM) and cytokeratins (cytoskeletal proteins that are specific for epithelial cells) are two makers used for positive enrichment of epithelial CTCs[203-205]. As mentioned above, some CTCs undergo EMT, resulting in decreased expression of epithelia markers. By only using epithelial markers may lead to false negative findings. Therefore N-cadherin and vimentin, which are both expressed in mesenchymal cells can be used for positive enrichment of mesenchymal CTCs.[206] Other positive enrichment markers include tumor-specific markers (e.g., HER2, EGFR) that are specific to certain tumor types[207-209] and tissue-specific markers (e.g., prostate-specific antigen (PSA) for prostate, mammaglobin for breast cancer) that have high specificity to certain tissue types[210, 211]. However, binding of those positive enrichment markers to antibodies may induce cytotoxicity[212], thus affecting the original state of CTCs. Magnetic beads that bind to CD45+ leukocytes were used for negative enrichment to remove normal

CD45+ haematopoietic cells in order to provide tumor antigen–independent enrichment performance[213, 214]. There have been efforts made on developing density-gradient centrifugation and subsequent negative enrichment for CTCs[215-217]. These methods could achieve a 1.4-3.1 log depletion of white blood cells with a 40–90% cancer cell yield. However, owing to manual sample pre-processing steps such as red blood cell lysis (loss quantified as ~11%) and density-gradient centrifugation (loss quantified as ~27%), a robust and automated platform that exhibits high recovery while performing negative depletion is still in high demand.

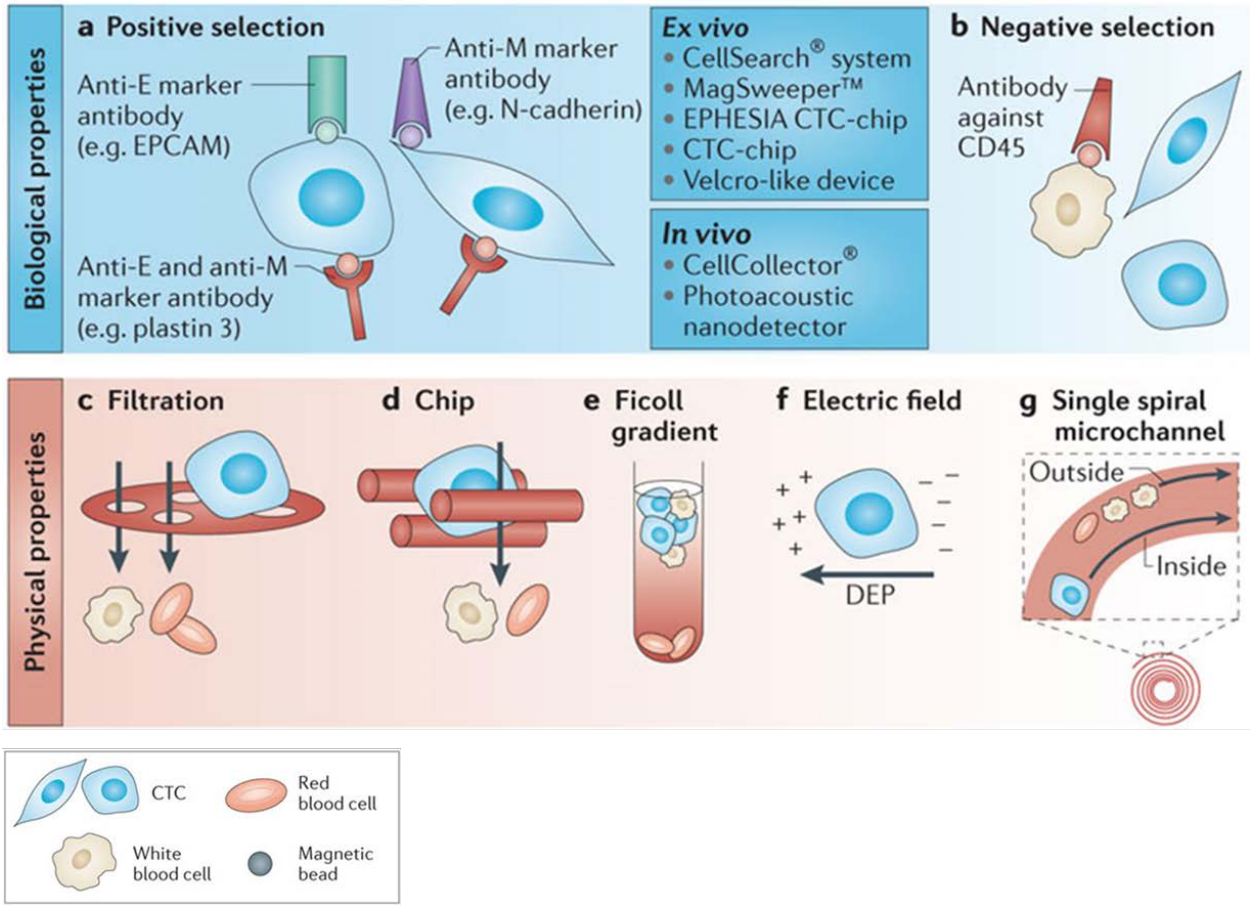


Figure 4.2 CTC Enrichment Technologies (adapted from Alix-Panabières et al. Nat Rev Cancer. 14 (2014) 624)[218].

CTCs have been studied for the targets of molecular therapies. For example, three out of four breast cancer patients with HER2-positive CTCs benefited from HER2-targeted therapy, even though the primary tumors were HER2-negative[219]. Another study shows that metastatic castration-resistant prostate cancer (mCRPC) patients with androgen receptor splice variant 7 (AR-V7) mRNA in CTCs resist to enzalutamide and abiraterone targeting the AR-response[220]. Recent studies of CTCs have focused on their genomes, transcriptomes and proteomes[221-228]. However, the epigenetics of CTCs remains an unexplored field with great potential. It has long been recognized that the phenotype of a cell is determined by its gene expression profile and its response to environmental cures. Gene activation and repression are not only associated with alteration in the DNA sequence, but also by mechanisms that affect gene expression and cellular phenotypes without altering DNA sequence (i.e. epigenetics). Cellular processes are regulated by a complex interplay among different layers of epigenetic information, including DNA methylation, and histone modifications. Epigenetic regulations play critical roles in gene expression and regulation, and are highly involved in biological processes such as embryonic development and tumorigenesis.

The number of studies on epigentics analysis of CTCs are so far small and they do not provide genome-wide information. This is most likely due to combined technical challenges of CTC isolation and difficulties with tools for profiling epigenetics on tiny cell samples. Some pioneering studies on methylation analysis of CTCs have been revealing exciting insights on temporal dynamic DNA methylation during cancer metastasis. For example, Chimonidou et al showed that the frequency of methylation at the CST6, BRMS1, and SOX17 gene promoters was significantly higher in EpCAM-enriched CTCs from patients with metastatic tumors than in those from patients with operable tumors, and least frequent from healthy donors.[229-231] Another

study by Friedlander et al showed that methylation in the CTCs occurred more frequently in genes associated with apoptosis, angiogenesis, and VEGF signaling, which suggests that some of the epigenetic programs associated with tumor progression in primary tumors are maintained in CTCs[232]. A better understanding of genome-wide epigenetics regulations of CTCs during cancer metastasis will improve various aspects of biomedicine, including risk stratification, disease diagnostics, and drug/therapeutic discovery.

Profiling epigenomics of CTCs contributes to cancer intervention in the context of personalized medicine[233, 234]. Epigenomic profiles of CTCs can served as highly-sensitive cancer cell markers for cancer diagnosis, prognosis, and eventually provide guidance for therapeutic medicine. Epigenomic profiles can provide valuable information for prognosis because certain patterns of epigenomic-based silencing are associated with tumor aggressivity and angiogenesis[96]. Moreover, epigenomic profiles have also been used to predict responses to a therapeutic procedure or reagent[97]. Finally, epigenomic marks have been used as therapeutic targets. DNA-demethylating agents have been explored as promising drugs for reversing promoter DNA hypermethylation and associated gene silencing[98].

There have been a number of molecular techniques for profiling epigenomics and identifying the interaction between regulatory proteins and DNA in general. However, in vitro methods such as electrophoresis mobility shift assays (EMSA) and DNase 1 protection assay cannot properly validate in vivo relevance[138]. In comparison, the chromatin immunoprecipitation (ChIP) assay has become the technique of choice for examining in vivo DNA-protein interactions over the years[80, 139-141]. Common ChIP assays confirm potential interactions between protein of interest and promoters of interest by conducting locus-specific qPCR analysis. On the other hand, unbiased and genome-wide mapping of protein binding is also possible with techniques such as

ChIP-ChIP and ChIP-seq[140]. Such examination provides a snapshot of the dynamic processes of epigenomic regulation, without overexpressing any component.

In a typical ChIP experiment, DNA is first cross-linked with protein by cross-linking agents (e.g., formaldehyde), in order to freeze protein–DNA interactions. Subsequently, cells are lysed in order to release chromatin. Then, chromatin is sonicated/MNase digested to yield fragments of protein-bound DNA that are typically 200–600 base pair long. These fragments are then immunoprecipitated onto the surface of antibody-coated beads. The antibody here specifically targets a transcription factor, a specific modified form of histone, or a cytosine methyl group. Finally, the immunoprecipitated fraction is isolated. The cross-linking is reversed and the released DNA is assayed to determine the sequences[142]. The identification of the DNA sequences can be done by qPCR if there are known candidates. Alternatively, these binding sites can be mapped at the genome scale by sequencing (ChIP-seq) using high-throughput sequencing technology (e.g., Illumina HiSeq 2500 System)[81, 89].

Epigenetic information are specific to cell and tissue types, the disease condition and its developmental stage. Profiling epigenomics of CTCs is important for understanding the dynamics and molecular mechanisms during cancer metastasis and establishing epigenomic signatures for cancer diagnosis and prognosis. However, one critical challenge in these efforts is that current epigenomic profiling tools often do not offer sufficient sensitivity to examine tiny quantities of cell samples from scarce sources such as small lab animals and patients. Although ChIP-qPCR/seq has been tremendously useful for epigenetic/ epigenomic studies, the technology suffers from serious limitations. First, the key limitation is that it requires a large number of cells

(>10^6 cells per IP for ChIP-qPCR and 10^7-10^8 cells for ChIP-seq)[141, 143, 144]. This is feasible when using cell lines. However, such requirement has become a major challenge when primary cells are used because very limited amounts of samples can be generated from lab animals or patients. For example, there are around 10,000 naturally occurring T regulatory cells per murine spleen, and ~5000 per ml peripheral blood. For metastatic cancer patients, there are only about 1–10 circulating tumor cells per ml of whole blood. Furthermore, primary samples are typically a mixture of different cell types. Isolating single cell type from a mixture brings further loss in the sample amount. Second, the outcome of a ChIP assay can be affected when a large cell number is used. Population heterogeneity may contribute to the large standard deviations among trials. Finally, most ChIP assays involve lengthy manual handling and they normally take 3–4 days or longer to finish. These cumbersome procedures not only create loss of materials but also introduce human/ technical errors that lead to large variations between replicates[145, 146]. In general, ChIP assays with ultrahigh sensitivity and high degree of automation and integration are highly desirable.

There have been several attempts to reduce the quantity of starting material required for ChIP-seq. A carrier molecule (mRNA and Histone H2B mix) was used to enable ChIP-Seq from 10,000 cells[134]. How the carrier molecule exactly works to enhance ChIP-Seq signal is not clearly known but it is hypothesized that the bulky material of the carrier helps to retain the small amounts of relevant chromatin throughout the procedure and the carrier also reduces nonspecific binding by acting as a competitor for nonspecific binding sites on the magnetic beads and elsewhere. Single-tube linear DNA amplification (LinDA) was developed to study ERα binding for as low as 5000 cells[147]. In this method, DNA was ligated with poly T and and in vitro

transcripted to RNA. The RNA was then reverse transcribed and amplified using the T7 promoter-BpmI-oligo(dA)15 primer. Moreover, other groups have successfully performed ChIP-Seq of histone modifications on microfluidics devices to further reduce input requirement. A microfluidic device with sensitivity of 1000 cells was used to study histone modifications of mouse early embryonic cells[148]. Microfluidic oscillatory washing based ChIP-seq (MOWChIP-seq) with sensitivity of a 100 cells was developed to study histone modification in GM12878 cells and fetal liver cells[149]. However, not much progress has been made in the field of low-input CTC ChIP-seq with microfluidic device.

In this project, we developed a negative enrichment of CTCs followed by ultrasensitive microfluidic ChIP-seq technology for profiling histone modification (H3K4Me3) of CTCs to resolve the technical challenges associated with CTC isolation and difficulties related with tools for profiling whole genome histone modification on tiny cell samples. The microfluidic ChIP-seq with high immunoprecipitation efficiency produced a sensitivity of 0.5 ng DNA (or ~50-100 cells). This is roughly 4-5 orders of magnitude higher than the prevailing protocol. With indexing first technology, input requirement could be further reduced by an order of magnitude, which enabled study of temporal dynamics in histone modification of CTC during cancer metastasis using a mouse model. The technology is “first-in-class” with unprecedented sensitivity. The new capability of our technology will allow establishing genome-wide histone modification profiles with less than 10 cells. Our technology dramatically widens the sample range for epigenomic profiling to include not only CTCs, but also other primary cell samples from scarce sources. Dynamic epigenomic information during disease development from tiny cell samples of a patient that used to be not accessible to the researchers and clinicians due to the technological limitation

now could become attainable. With such information, one can build epigenomic signatures for disease diagnosis and prognosis in the context of personalized medicine.

Our microfluidic technology is innovative for several reasons. First, one-chip negative enrichment of CTCs not only effectively depletes white blood cells, but also avoids false negative findings of CTCs associated with positive enrichment methods and preserves original state of CTCs at same time. Second, indexing first enables multiple DNA-barcoded samples to be pooled for ChIP, further reducing initial input requirements and increasing cross-sample reproducibility. Third, high concentrations from trace amounts of molecules could be built up inside the tiny volumes that offered by microfluidic chamber. Adsorption kinetics and completeness was facilitated by such high concentration. Moreover, the IP beads take up a large fraction of the tiny volume so that the surface area/volume ratio (15-40% bead volume in 800 nl) is tremendously improved when compared to 5% bead volume in 1.5 ml conventional MeDIP[150]. The close proximity among beads greatly increased the efficiency and rate for chromatin adsorption on the bead surface due to the short diffusion lengths involved. The adsorption of a chromatin molecule among beads was rapid given that travel time $\tau_D \sim w^2/D$, where w is diffusion distance between two beads, and D is diffusivity. Furthermore, by using microfluidic technique uniquely suited for bead manipulation at the microscale, we effectively remove nonspecific adsorption after high efficiency adsorption using microfluidic oscillatory washing. This is critical for producing high quality ChIP DNA that preserves desired biological information. Finally, the microfluidic device integrates various steps and minimizes material loss among steps.

4.2 Methods and materials

Fabrication of the microfluidic ChIP device

The microfluidic ChIP device is composed of a microfluidic chamber (~800 nl), connecting channels, and a micromechanical vale that can be partially closed to stop magnetic beads while allowing liquid to pass. The main chamber is in elliptic shape with a major axis of 6 mm, a minor axis of 3 mm, and a depth of 40 μm. 27 micro-pillars are spotted inside the main chamber to prevent collapsing of PDMS.

Multilayer soft lithography was adopted to fabricate microfluidic ChIP device. Briefly, two photomasks (one for fluidic layer, and one for control layer) that had desired microscale patterns were designed with computer aided design software FreeHand MX (Macromedia) and printed on high-resolution (5,080 d.p.i.) transparencies. To make fluidic layer master (~40 μm thick), photoresist (SU-8 2025, Microchem) was spun on a 3-inch silicon wafer (978, University Wafer) at 500 rpm for 10s and 2500 rpm for 30s followed by soft bake at 65°C for 1 min and 95°C for 7 min. To make control layer master (~50 μm thick), SU-8 was spun at 500 rpm for 10s and 1500 rpm for 30s and followed by same soft bake condition. Each master covered with its photomask was UV exposed for 17s at 580 mW exposure intensity and followed by a post exposure bake at 65°C for 1 min and 95°C for 7 min. masters were then developed in SU-8 developer for 2-3 min, rinsed with IPA and air blown to dry. To make fluidic layer stamp, PDMS (General Electric silicone RTV 615, MG chemicals) with a mass ratio of A:B = 5:1 was thoroughly mixed and vacuumed for 1 hr. it is then poured onto the fluidic layer master in a Petri dish to a height ~5 mm thick. To make control layer stamp, PDMS with a mass ratio of A:B = 20:1 was mixed,

vacuumed for 60 min, and spun onto the control layer master at 1100 rpm for 35s to a height of 108 μm thick. Both stamps were partially cured at 80 °C for 30 min. The fluidic layer was then peeled off from the fluidic layer master and aligned to the control layer. Two-layer PDMS were thermally bounded by baking at 80 °C for 60 min, and then peeled off from control layer master. Inlets and outlets of the device were punched by a 2 mm hole puncher. Finally, the two-layer PDMS and a pre-cleaned glass slide were treated with oxygen plasma cleaner (PDC-32G, Harrick Plasma) and brought together to form closed channels and chamber. Device was then baked at 80 °C for 60 min to strengthen the bonding between PDMS and glass. Glass slides were cleaned in a basic solution (H2O: 27% NH4OH: 30% H2O2 = 5:1:1, volumetric ratio) at 75 °C for 2 h and then rinsed with ultrapure water and thoroughly air blown to dry.

Setup of the microfluidic device

The microfluidic experiment was monitored by a charge-coupled device (CCD) camera (ORCA-285, Hamamatsu) attached to the port of an inverted microscope (IX 71, Olympus). The experiment started with prefilling the control channel with water to prevent air bubble defusing into fluidic channel. The reagents were flowed into the inlet via perfluoroalkoxy alkane (PFA) high-purity tubing (1622L, ID: 0.02 in. and OD: 0.0625 in., IDEX Health & Science) driven by a syringe pump (Fusion 400, Chemyx). The micromechanical valve was actuated by a solenoid valve (18801003-12V, ASCO Scientific), which was connected to a pressure source (a gas cylinder or a compressed air outlet) and controlled by a data acquisition card (NI SCB-68, National Instruments) and a LabVIEW (LabVIEW 2012, National Instruments) program for its switching function. The pressure (30 – 35 p.s.i) that was applied to control channel deformed

PDMS membrane between fluidic channel and control channel and formed a partially closed valve to stop beads while allowing fluids to pass. The oscillatory washing was conducted by connecting inlet and outlet of microfluidic device to solenoid valves via PFA tubing. A digital pulse signal was created in LabVIEW program, converted to electric signal by a data acquisition card, and sent to solenoid valves.

Preparation of sonicated chromatin

10 K cell

10 K cell samples were centrifuged at 1,600g for 5 min at room temperature. Cells were then washed with 1.0 ml 1× PBS (14190-144, Sigma-Aldrich) at room temperature by centrifugation at 1,600g for 5 min and resuspension. Cells were resuspended in 1.0 ml 1× PBS and cross-linked by adding 67 µl of 16% freshly prepared formaldehyde (28906, Thermo Scientific) to 1 ml cell suspension (to a 1% final formaldehyde concentration) for 5 min. Cross-linking was terminated by adding 71 µl of 2 M freshly prepared glycine (to a 0.125 M final glycine concentration) and shaking for 5 min at room temperature. Cross-linked cells were pelleted at 1,600g for 5 min and washed with precooled (4°C) PBS buffer and resuspended in 130 µl of the sonication buffer (10 mM Tris-HCl, pH 8.1, 1 mM EDTA, 0.1% SDS and 1× protease inhibitor cocktail). Cross-linked cells were sonicated with a Covaris E220 sonicator for 14 min with 5% duty cycle, 105 peak incident power and 200 cycles per burst. The sonicated lysate was centrifuged at 16,100g for 10 min at 4 °C. Sonicated chromatin in the supernatant was transferred to a new 1.5-ml LoBind Eppendorf tube (17014013, Denville) for MOWChIP-seq. From this stock chromatin preparation, samples equivalent to 4500, 450, 500 and 50 cells were divided into aliquots and diluted to give

a final volume of 50 µl for MOWChIP-seq. Same amount of the sample was used as the input. After this procedure, we typically obtained ~3.3 pg DNA per cell from the pre-ChIP chromatin samples.

Indexing Chromatin

After sonication, chromatin was diluted 1 to 5 with Sonication Equilibration Buffer (10 mM TrisCl, 140 mM NaCl, 0.1 % Sodium Deoxycholate, 1% Tx-100, 1 mM EDTA, 1X Protease Inhibitors (Roche) to achieve an SDS concentration of 0.1%. Chromatin were then incubated with 15 ul of Dynabeads Protein A (10001D, Invitrogen) magnetic beads and 1.3 ug of anti-H3 antibody (ab1791) overnight at 4°C.

Magnetic beads were used to efficiently add, wash and remove the different master mixes used in the indexing process. All the reactions were done while chromatin was bound to the H3 coated magnetic beads. The immunocomplexes were magnetized and washed 3 times with 150 ul of 10 mM Tris Cl + 1X Protease Inhibitors EDTA free (Roche). After the washes bead bound chromatin was resuspended in 20 ul of the same buffer. Chromatin End Repair was performed by NEBNext® End Repair Module (E6050S, NEB) following manufacture's protocol. After end repair, bead bound chromatin was washed once with 150 ul of 10mM TrisCl + Protease Inhibitors and re-suspended in 40 ul of the same buffer. Chromatin was A-tailed by NEBNext® dA-Tailing Module (E6053S, NEB) following manufacture's protocol. After A-Tailing bead bound chromatin was washed once with 150 ul of 10mM TrisCl + Protease Inhibitors and re-suspended in 20 ul of the same buffer. Chromatin was indexed by adding 5 ul of 0.75 uM Yshapped Indexed Adaptors (KK8711, KAPA Single-Indexed AdapterSet A) and ligated by

NEBNext® Quick Ligation Module (E6056S, NEB) following manufacture's protocol. Bead bound indexed chromatin was washed once with 150 ul of 10mM TrisCl + Protease Inhibitors to remove the non ligated adaptors. The 4500, 450, 500 and 50 cells were indexed differently so they could be pooled together later.

To release the indexed chromatin from the antibody coated magnetic beads, denaturing conditions (DTT, high salt and detergent) and heat were used. After the post-Indexing wash, beads were re-suspended in 12.5 ul of 100 mM DTT and incubated for 5 min at room temperature. Then, samples were incubated with 12.5 ul of Chromatin Release Buffer (500mM NaCl, 2% SDS, 2% Sodium Deoxycholate, 2X protease Inhibitors) at 37C for 30 min. After the release incubation, magnetic beads were again thoroughly re-suspended (25 ul) and pooled together with another sample (25 ul) resulting in a pool volume of 50 ul. 4500 cells pooled with 500 cells, and 450 cells pooled with 50 cells. The pooled indexed chromatin samples were concentrated using a 50Kda cutoff Centricon (Amicon).

Y-shaped indexed adaptors could also be obtained by annealing single stranded universal adaptor and single stranded indexed adaptors. A phosphate group were added to 5' end of indexed adaptors to allow ligation of the adapter to DNA fragment of interest. A phosphorothioate bond were also added in between C and T at 3' end of universal adaptors. This is to provide a strong bond between the C and T overhang and prevent T nucleotide to be removed by exonuclease activity, which is critical for annealing to the 3' end A overhang created from A-tailing. Denaturing reagents and heat will then be applied to digest proteins such as nucleases, histones and release the indexed DNA from the antibody coated magnetic beads. Multiple indexed DNA can be pooled together to reduce initial input requirements for ChIP-seq and increase cross-sample reproducibility.

Normally, two unique adapters, A and B, are wanted to be on either end of unknown insert sequences. Cohesive-end ligation is difficult since the insert sequences are unknown. Blunt-end adapter ligation is used in a reaction that contains Unknown Inserts, adapter A, and adapter B. There will be three different possible insert-ligated products, A-insert-A, B-insert-B and A-insert-B. A-insert-B is the only desired product. Ligating Y-shaped adapter to an A-tailed inserts will provide 100% A-insert-B. Moreover, if directionality is assigned to the insert, Y-adapter ligation will provide A-Insert (1-->200)-B and A-Insert (200-->1)-B two different kinds of insert-ligated products per insert. Each will create a separate clonal cluster and sequence information starting at both base 1 and base 200 of the insert can be obtained since single-read sequencing always sequences from Adapter A. For 1-->200 case, Read 1 + will have TATGCA, Read 1 - will have TGCATA. For 200-->1case, Read 1 + will have TGCATA, Read 1 – will have TATGCA. They all can be either aligned back to + or – strand of the original genome.

Preparation of immunoprecipitation (IP) beads.

Dynabeads Protein A (10001D, Invitrogen) were used for immunoprecipitation. They are 2.8 µm superparamagnetic beads with recombinant Protein A (~45 kDa) covalently bound to the surface. The 5 µl of 30 mg/ml beads (equivalent to 150 µg) were washed twice with freshly prepared IP buffer (20 mM Tris-HCl, pH 8.0, 140 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% (w/v) sodium deoxycholate, 0.1% SDS, 1% (v/v) Triton X-100) and resuspended in 150 µl IP buffer containing antibody. Beads were gently mixed with the antibody at 4 °C on a rotator mixer at 24 r.p.m. for 2 h. Antibody-coated beads were washed twice with the IP buffer and then resuspended in 5 µl IP buffer. We optimized the antibody concentration for the bead coating step

on the basis of our ChIP-qPCR results. The optimal antibody concentration for MOWChIP-seq with anti-H3K4me3 antibody (07-473, Millipore) was 3.3 μg/ml for 500 cells, 5 μg/ml for 5,000 cells. These conditions were equivalent to using 500 and 750 ng of antibody in the preparation of 150 μg IP beads.

MOWChIP

The MOWChIP procedure started with rising the fluidic chamber with the IP buffer at a flow rate of 20 μl/min for 30 s. After found no air bubble in the microfluidic chamber, micromechanical valve was partially closed. The 5 μl (optimal condition) magnetic beads coated with antibody were flown into fluidic chamber by the syringe pump at 20 μl/min, and aided with a cylindrical permanent magnet (NdFeB, D48-N52, 0.25 inch dia. and 0.5 inch thick, K&J Magnetics) to help beads travel through FPA tubing. The beads were packed against the partially closed valve to form a packed bed. 0.5 μl of 100 mM PMSF (P7626-1G, Sigma-Aldrich) and 0.5 μl 100× protease inhibitor cocktail (P8340, Sigma-Aldrich) were freshly added to the 50 μl chromatin sample (to a 1 mM final PMSF concentration and 1× final protease inhibitor cocktail concentration). The chromatin sample was then flowed through packed bed with a flow rate of 1.5 μl/min. The immunoprecipitation step was finished around 60 min under this flow rate. After ChIP, a low-salt washing buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% (v/v) Triton X-100) was introduced into the fluidic chamber at a flow rate of 2 μl/min for 2 min. Tubing that connected to inlet and outlet for oscillatory washing were prefilled with 20 μl low-salt washing buffer. Oscillatory washing was done by applying pressure pulses (each at 1 p.s.i., with a pulse width of 0.5 s and an interval of 0.5 s between two pulses) alternatingly at

inlet and outlet of the fluidic chamber for 5 min while keeping the micromechanical valve open. The pulse signals were created in LabVIEW program, converted to electric signals by a data acquisition card, and sent to solenoid valves to achieve automation of oscillatory washing. After oscillatory washing, beads were retained on one side of the fluidic chamber by magnet. The unbound chromatin fragments and other debris/waste were flushed out of the microfluidic chamber by flowing a high-salt washing buffer (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% (v/v) Triton X-100) through the chamber at 2 µl/min for 2 min. Tubing that connected to inlet and outlet for oscillatory washing now were prefilled with 20 µl high-salt washing buffer. Oscillatory washing was again conducted for 5 min. Beads were retained, and the unbound chromatin were flushed out by flowing IP buffer through the chamber at 2 µl/min for 2 min. Finally, IP beads were flown out by IP buffer under a flow rate of 50 µl/min for about 5 min and collected into a 1.5-ml LoBind Eppendorf tube. The tube was then put on a magnetic stand, IP buffer were removed. Beads were kept in the tube and immediately proceeded for DNA extraction.

Extraction of ChIP DNA and input DNA

Chromatin samples (either ChIPed chromatin on beads or input chromatin) were reverse cross-linked in reverse cross-linking buffer (200 mM NaCl, 50 mM Tris-HCl, 10 mM EDTA, 1% SDS, 0.1 M $NaHCO_3$) with 2 µl 20 mg/ml proteinase K (26160, Thermo Scientific) added before use (to a 200 µg/ml final proteinase K concentration) and had their volumes adjusted to 200 µl. Samples were then incubated at 65 °C for at least 4 hours or overnight. Equal volume (200 µl.) of Phenol-chloroform-isoamylalcohol (25:24:1) was added to sample, mixed by vortexing, and

centrifuged at 16,100g for 5 min at room temperature. DNA was extracted by collecting ~200 µl aqueous phase to a fresh 1.5 ml Eppendorf tube. The 50 µl of 10 M ammonium acetate was then added (to a 2 M final ammonium acetate concentration) resulting in 250 µl solution. The 750 µl (3 times volume of 250 µl) 100% ethanol with 2 µl 20 µg/µl glycogen (10814010, Invitrogen) were finally added to carry ethanol precipitation for DNA purification. Samples were placed at -20 °C for at least 2 hours or overnight for ethanol precipitation. Next, samples were centrifuged at 16,100g for 10 min at 4 °C, supernatant were carefully removed. Samples were washed with 70% ice-cold ethanol without disturbing the pellet. The supernatants were removed again after another centrifugation at 16,100g for 5 min at 4 °C. The final pellet was air dried for 5 min and resuspend in 10 µl DNase-free water. This purified DNA can be used directly for ChIP-qPCR or for sequencing library construction. DNA concentrations were measured using a Qubit 2.0 fluorometer with dsDNA HS Assay kit (Q32851, Life Technologies).

Construction of sequencing libraries

Sequencing libraries were prepared by Accel-NGS 2S Plus DNA Library Kit (21096, Swift Biosciences). This kit provides high complexity next-generation sequencing libraries and compatibility with ultra-low inputs (~10 pg). The library preparation process involved two steps of repairs and two steps of ligations to repair both 5’ and 3’ termini and sequentially attach Illumina adapter sequences to the ends of fragmented double-stranded DNA. Bead-based SPRI clean-ups were used to remove oligonucleotides and small fragments, and to change enzymatic buffer composition between steps. Different SPRIselect bead-to-sample ratios were used for different input quantities. PCR amplification (98 °C for 30 s, followed by 98 °C for 10 s, 60 °C

for 30 s, 68 °C for 60 s for each cycle) was conducted to increase yield of indexed libraries. We used ~18 cycles for ChIP DNA from 500 cells, ~16 cycles for ChIP DNA from 5000 cells. Library fragment size was determined using high-sensitivity DNA analysis kit (5067-4626, Agilent) on an Agilent 2200 TapeStation. The Kapa library quantification kit (KK4809, Kapa Biosystems) was used to determine effective library concentrations. The final concentrations of libraries submitted for sequencing were ~10 nM. The libraries were sequenced on an Illumina HiSeq 4000 with single-end 50-nt reads. Typically, 15–20 million reads were generated per library.

ChIP-qPCR data analysis

Real-time PCR was done using iQ SYBR Green Supermix (Bio-Rad, Hercules, CA, USA) on an CFX96 real-time PCR machine (Biorad) with C1000Tm thermal cycler base. All PCR assays were performed using the following thermal cycling condition: 95°C for 10 min followed by 40 cycles of (95°C for 15 s, 56°C for 30 s, 72°C for 30s). UNKL and C9orf3 are two known positive loci, N1 and N2 are two known negative loci. Primer concentrations were 400 nM. All primers were ordered from Integrated DNA Technologies (Coralville, IA, USA). The ChIP-qPCR results were represented as relative fold enrichment (**Figure 4.3**). which is the ratio of percent input between a positive locus and a negative locus. Percent input was calculated as the following equation: $2^{(C_q INPUT - C_q IP)} * 100\%$, where C_q is amplification cycle number run by real-time PCR, $INPUT$ is the chromatin sample without immunoprecipitation, IP is chromatin sample with immunoprecipitation.

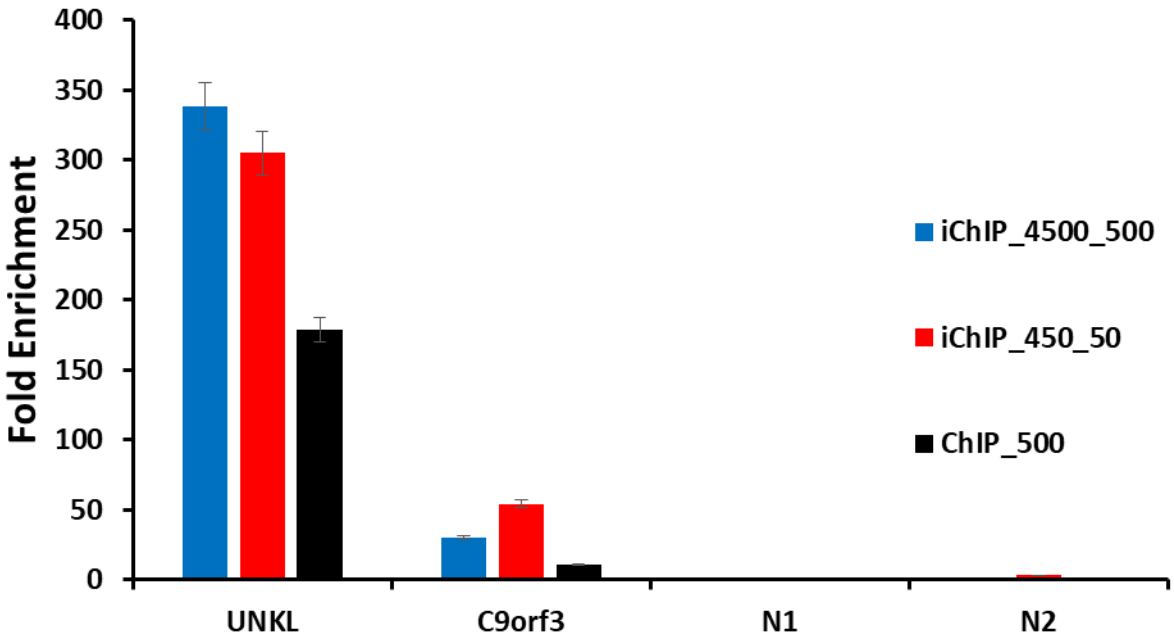


Figure 4.3 Fold enrichment (ratio between % input at positive loci and % input at negative loci) under different conditions. UNKL and C9orf3 are two known positive loci, N1 and N2 are two known negative loci. The error bars were calculated as standard deviation from 3 ChIP-qPCR replicates.

Cell culture

MDA-MB-231 cells were cultured in L15 medium (ATCC) with 15% fetal bovine serum, 100 U penicillin, 100mg streptomycin/ml (Invitrogen) at 37°C in a humidified incubator containing 5% CO_2 . Cells were sub-cultured every two to three days to maintain exponential growth.

Generating primary tumor single cell suspension

Primary mammary solid tumors were converted into single cell suspension with Gentle Collagenase/Hyaluronidase (07919, Stem Cell Technologies) following the manufacturer's protocol.

Isolation of CTCs

CTCs were isolated from blood with EasySep™ Mouse/Human Chimera Isolation Kit (19849, Stem Cell Technologies) following the manufacturer's protocol.

Mice

6 weeks old immunodeficient mice (NOD SCID gamma, Jackson) were used in this study.

Figure 4.4 Mice under anesthesia.

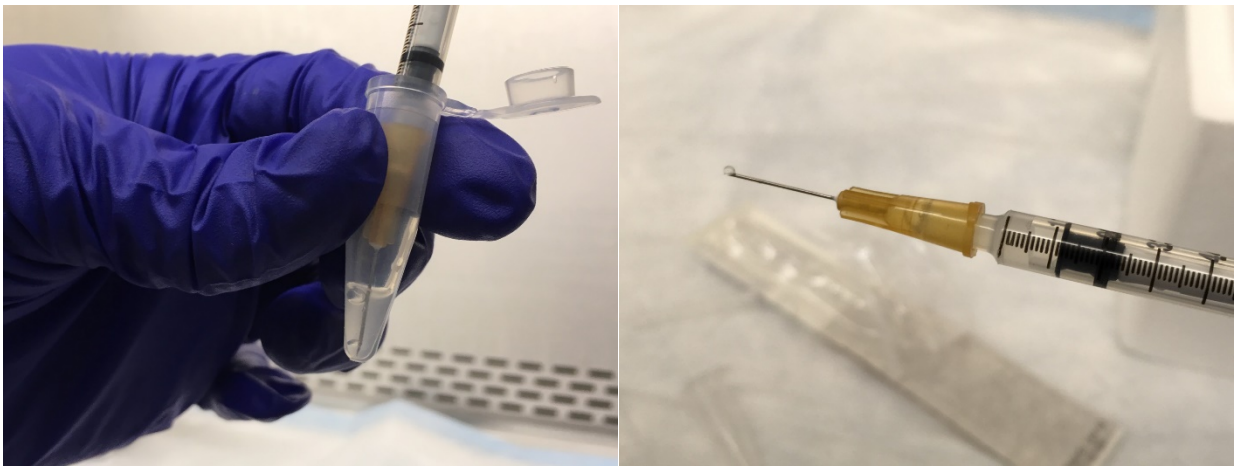


Figure 4.5 1000 K MDA-MB-231 cell suspension before injection.

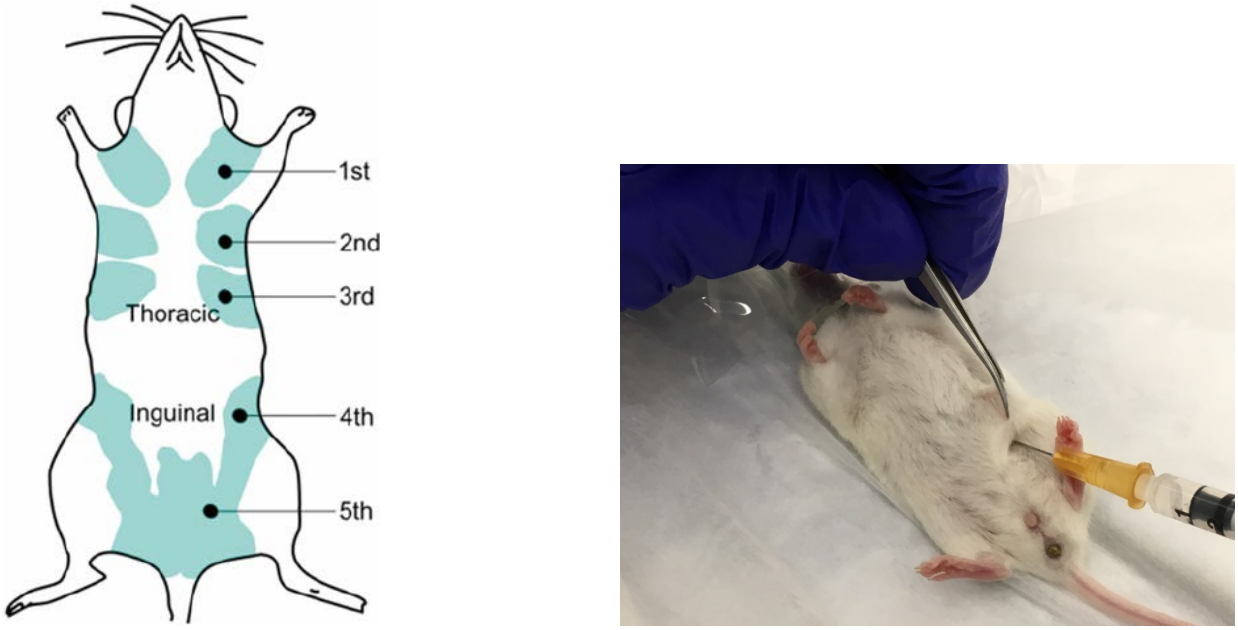


Figure 4.6 1-2 x 10^6 MDA-MB-231 cell suspension was injected into fourth inguinal mammary fat pad of 6 weeks old mouse under anesthesia.

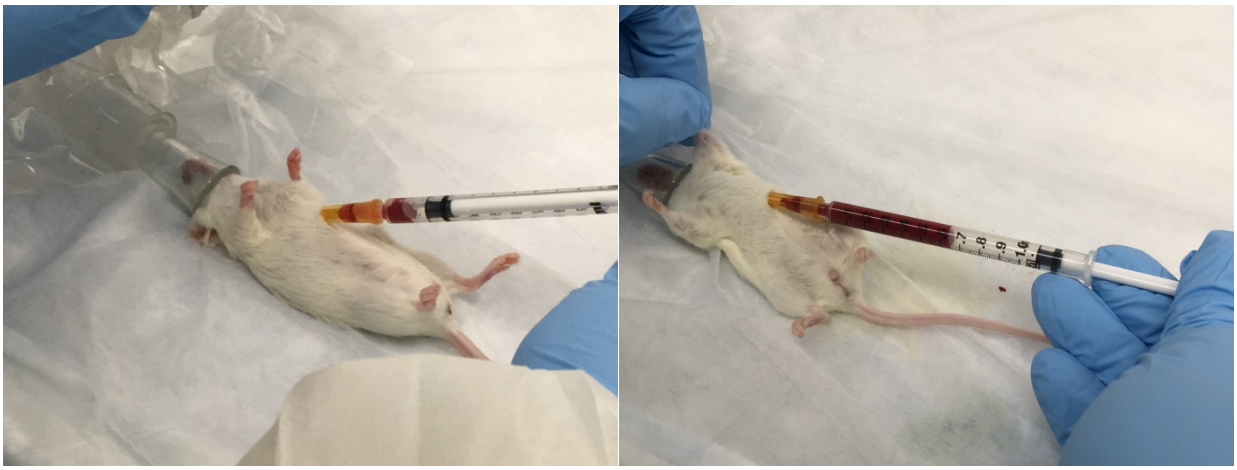


Figure 4.7 Cardiac puncture under anesthesia – blood draw about 1 ml per mouse after 6 week of injection (mouse was 12 weeks old).

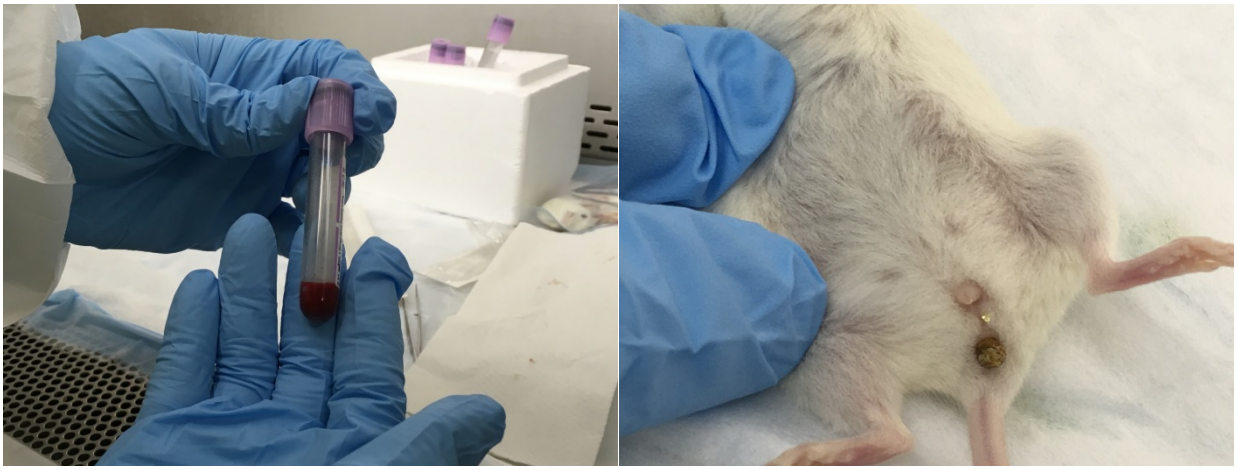


Figure 4.8 Collected blood and appearance of solid mammary tumor.

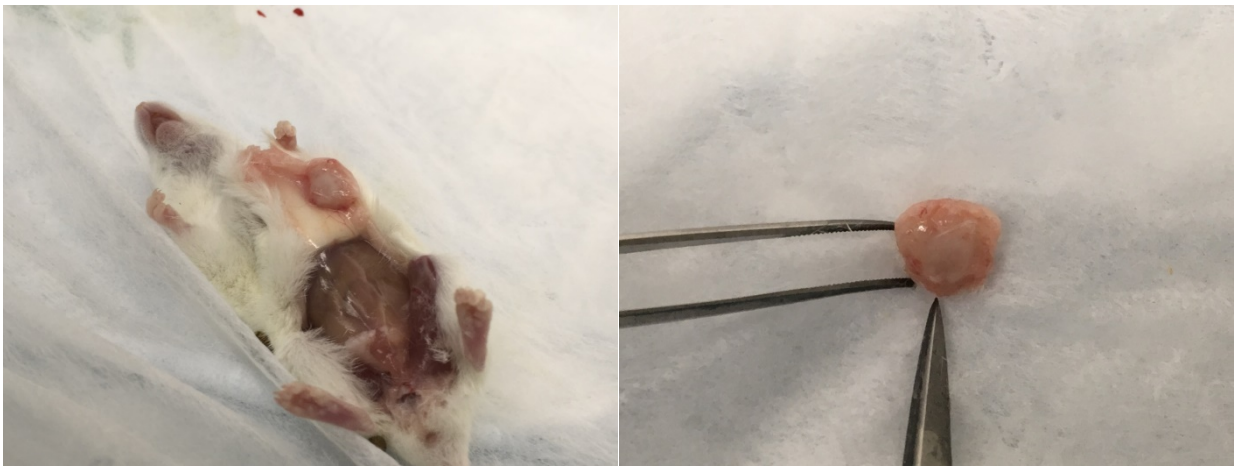


Figure 4.9 Mouse was euthanized, solid mammary tumor was harvested after 6 week of injection (mouse was 12 weeks old). Mice were sprayed with 70% ethanol on the ventral side. Skin on the ventral side was cut through, pulled away from mouse body, and pinned down to expose mammary glands. Mammary tumors were identified and cut off. The size and weight of a tumor were recorded.

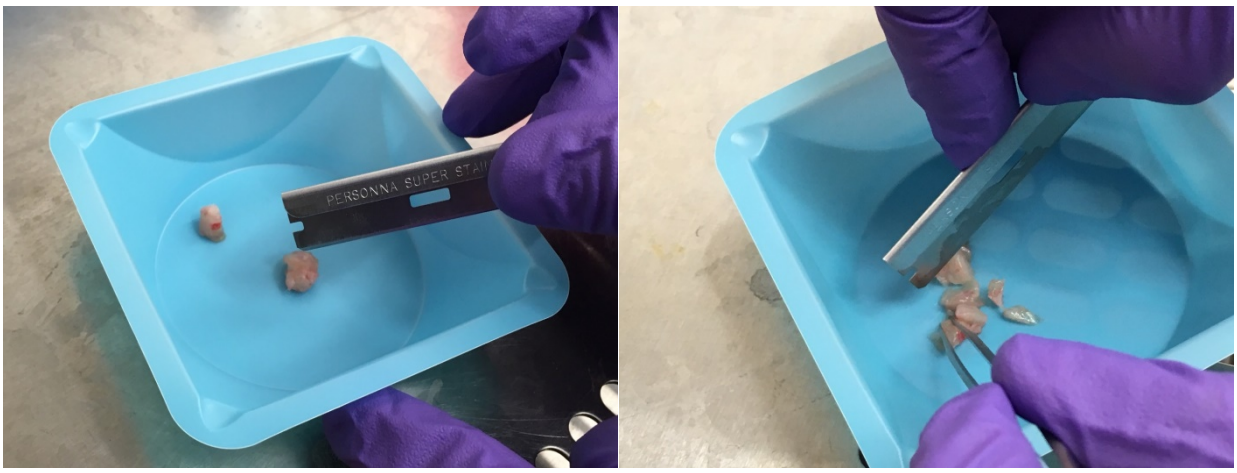


Figure 4.10 Dissect solid mammary tumor.

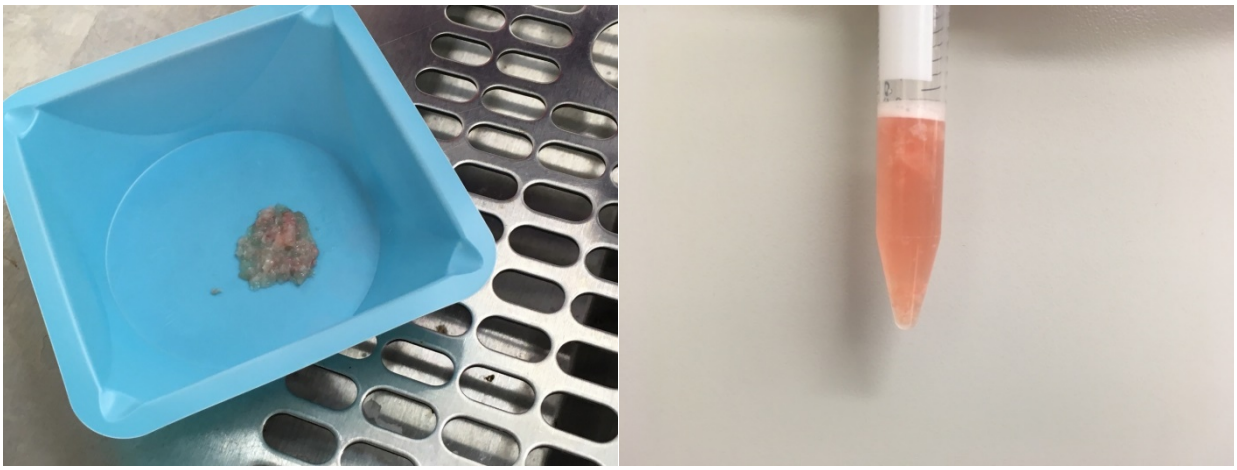


Figure 4.11 Dissected solid mammary tumor and single cell suspension from dissected solid mammary tumor.

ChIP-seq data analysis

ChIP sequencing reads were mapped to the human genome (hg19) using BWA (v0.7.17) with default parameter settings. Peaks of each ChIP sample were called against input by MACS (v2.4.2) with (p-value < 10^{-5}) and other parameters set at default values.

4.3 Results and discussion

We first removed red blood cells through a centrifugation on a Ficoll density gradient on the basis of the CTC density or ammonium chloride lysis. The supernatant that contains CTCs and white blood cells were collected and flown into a microfluidic chamber for negative enrichment. Immunomagnetic enrichment targeting solely CD45 results in a large granulocyte population consistently found in the product. CD66b, which is highly expressed in granulocytes and not observed in tumor cells were supplemented with CD45 for negative enrichment[235]. Here we used a cocktail of antibodies for negative selection. A large block of magnet was put under the microfluidic device to manipulate the beads inside the chamber. After on-chip negative enrichment, CTCs are flowed into next microfluidic chamber for cell lysis to release chromatin. Optimized amount of MNase will be then flowed into chamber to digest chromatin into 200-600 bp. We just did sonication off-chip instead. Next, Dynabeads Protein A coated with anti-H3 antibody were added to the chromatin for immobilization and to efficiently add, wash and remove the different master mixes that would be used in the next indexing process (**Figure 4.12**).

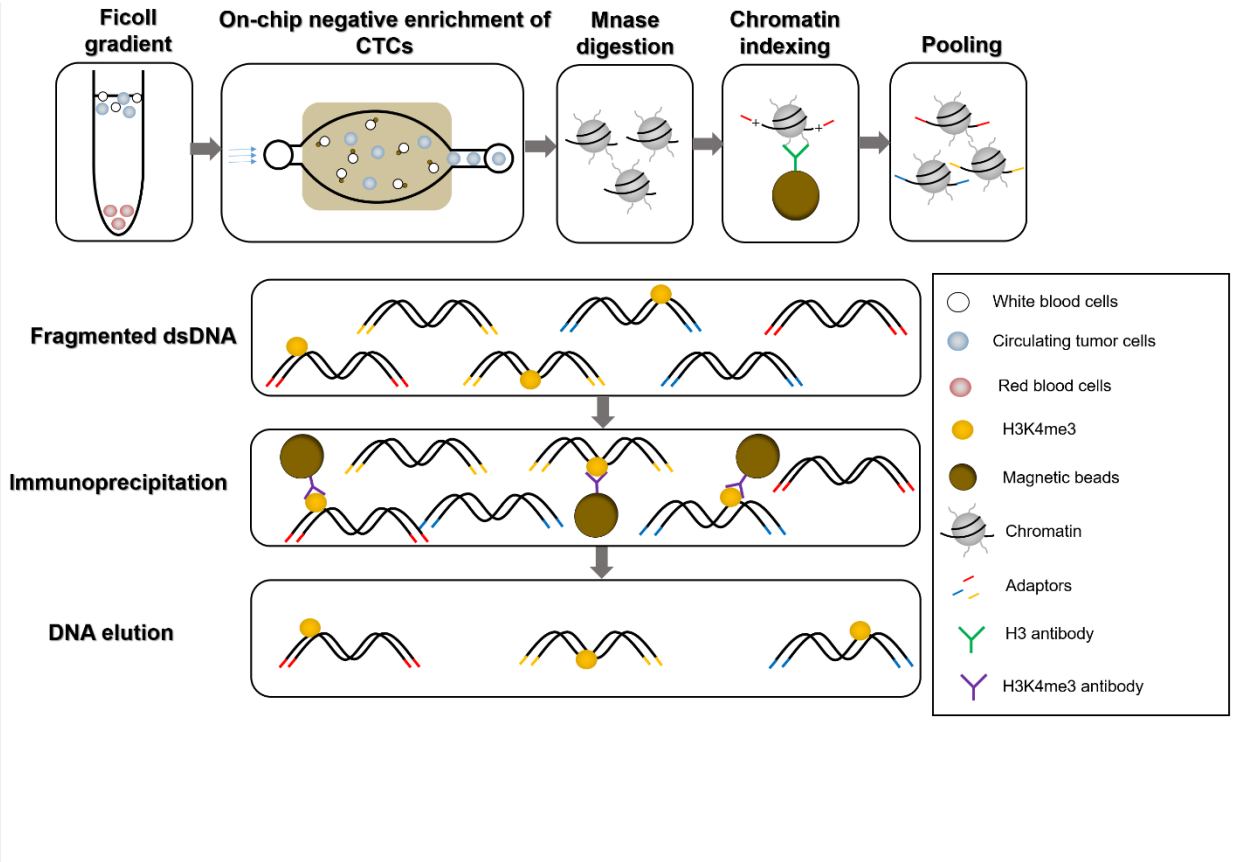


Figure 4.12 Schematic diagram of the on-chip negative enrichment of CTCs approach followed by ChIP involving an initial chromatin barcoding step.

For indexing, chromatin end was first repaired, then A-tailed, and indexed by ligating Y-shaped indexed adaptors containing Illumina P5 and P7 sequences to the chromatin's DNA ends (**Figure 4.13**).

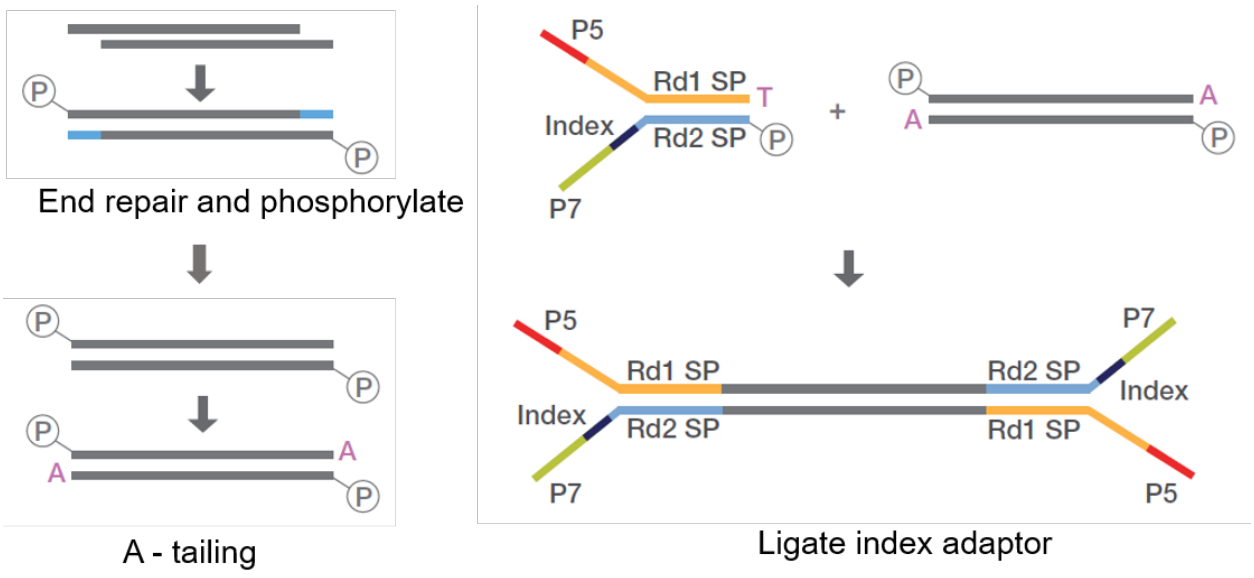


Figure 4.13 End repair, A-tailing, and ligation.

The indexing-first method will allow us to pool indexed CTCs from different mice together to resolve the low input issue for ChIP-seq. For example, we can collect 10 CTCs from 1 mL blood draw, which is not enough for our microfluidic ChIP-seq assay. With indexing first, we could pool 10 sets of 10 CTCs with each set barcoded with different adapters together to reach 100 cells minimum requirement and demultiplex after sequencing (**Figure 4.14**).

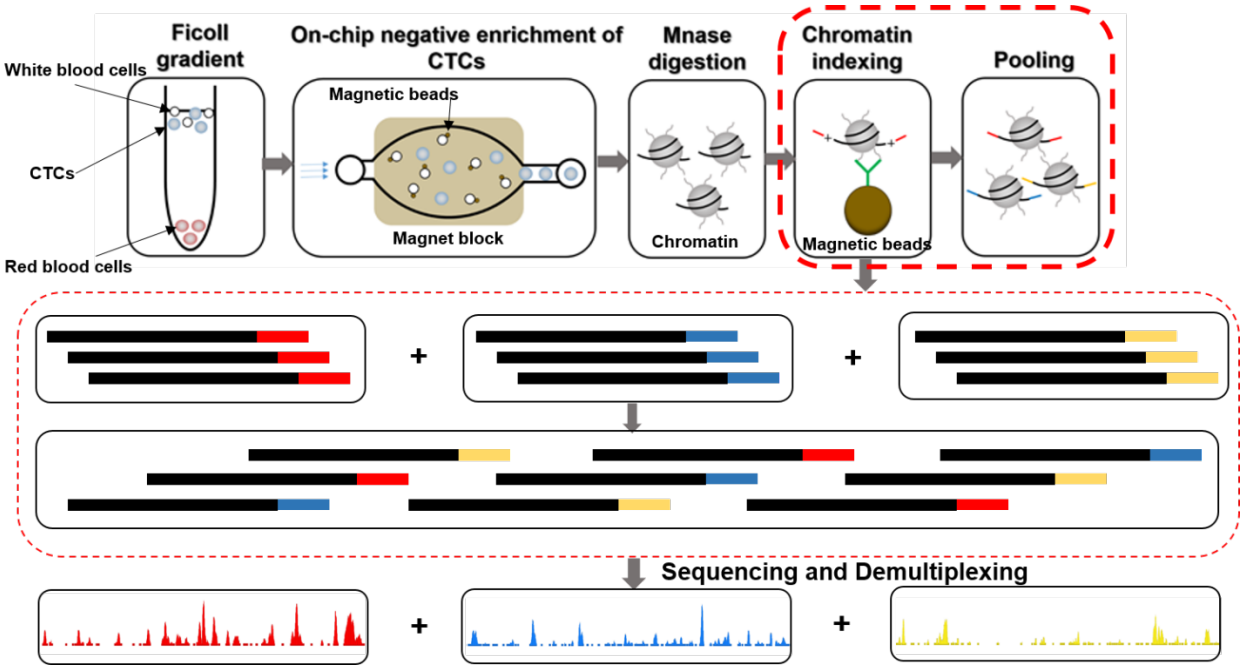


Figure 4.14 Indexing first to pool sample together and demultiplex after sequencing.

The chip device contained 7 reaction chambers that allow us to run 7 parallel MeDIP assays at the same time (**Figure 4.15a**). The device and operation were similar to what we demonstrated previously for MOWChIP-seq[149]. ChIP started with flowing a suspension of antibody-coated magnetic beads (IP beads) into the microfluidic chamber (~800 nl) and the IP beads were packed against a partially closed microvalve to form a packed bed (**Figure 4.15b**). Indexed chromatin were then flowed through the bead-packed bed, allowing targeted chromatin to adsorb onto the bead surface coated with H3K4me3 antibody (**Figure 4.15c**). After ChIP, we applied alternating pressure pulses at two ends of the microfluidic chamber (at 0.5 p.s.i. and with a duration of 0.5 s for each pulse) for 5 min to create oscillatory washing to effectively remove nonspecific binding (**Figure 4.15d**). Such washing was conducted in ChIP buffer. After the oscillatory washing, the beads were retained by a magnet on one side of the chamber while the unbound chromatin

fragments and other debris/waste were flushed out of the microfluidic chamber by ChIP buffer (**Figure 4.15e**). Finally, the ChIP beads with adsorbed chromatin fragments were flashed out of the chamber and collected for off-chip processing.

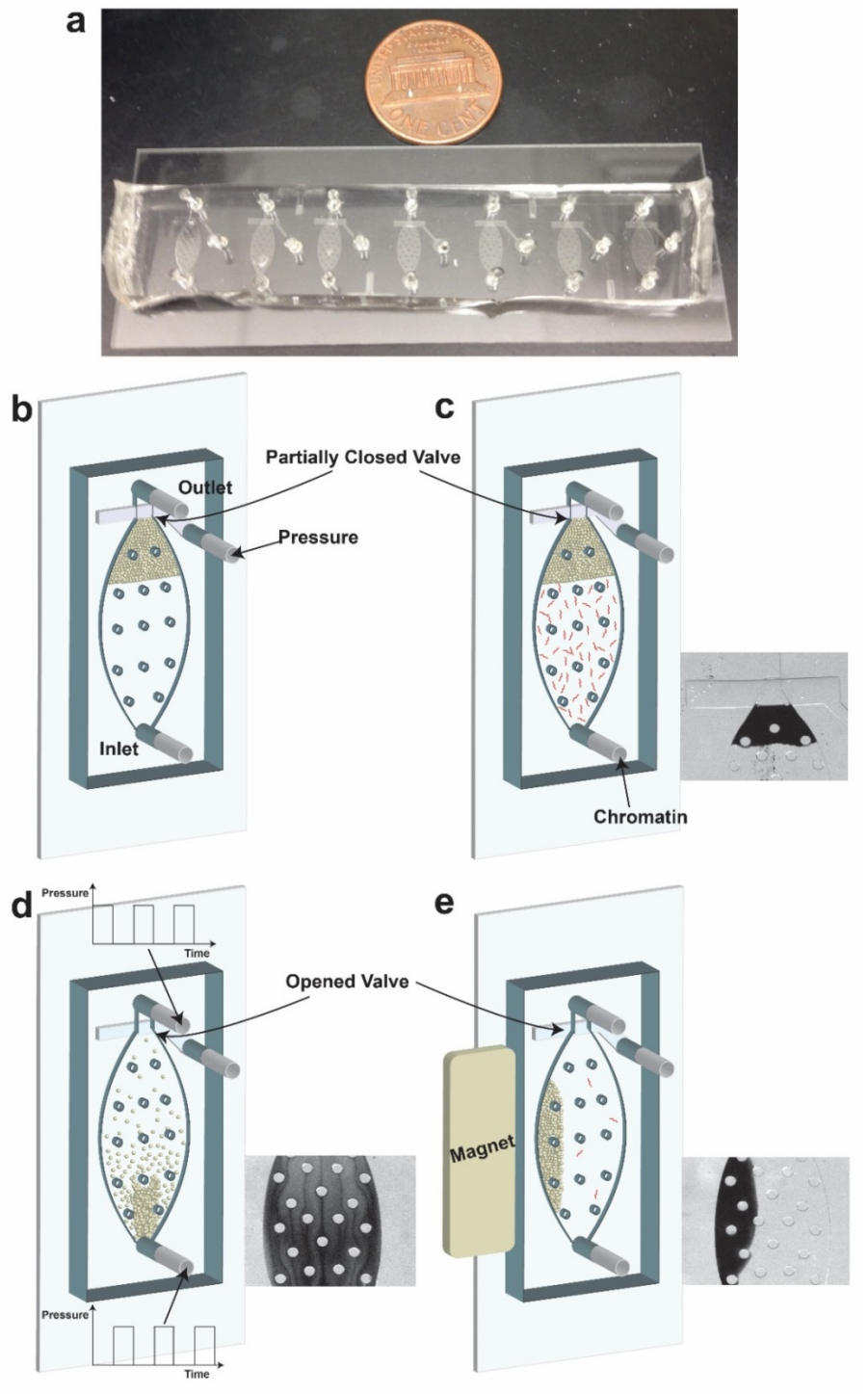


Figure 4.15 Microfluidic ChIP-seq device and its operation. **(a)** The microfluidic device (containing 7 independent units). **(b)** Formation of packed bed of IP beads while micro-valve is closed. **(c)** ChIP by flowing chromatin fragments through the packed bed. **(d)** Oscillatory

washing. **(e)** Retaining IP beads on one side of chamber to remove unbound chromatin fragments and debris.

To test indexing first technology, we started 5000 GM12878 cells (28.4ng), 4500 cells were ligated with barcode1, and 500 cells were ligated with barcode2. After ichip chromatin elution and concentrator, we were able to collect 2.35 ng DNA, which was about 8.28% recovery rate. The added barcode size can be visualized in (**Figure 4.16**).

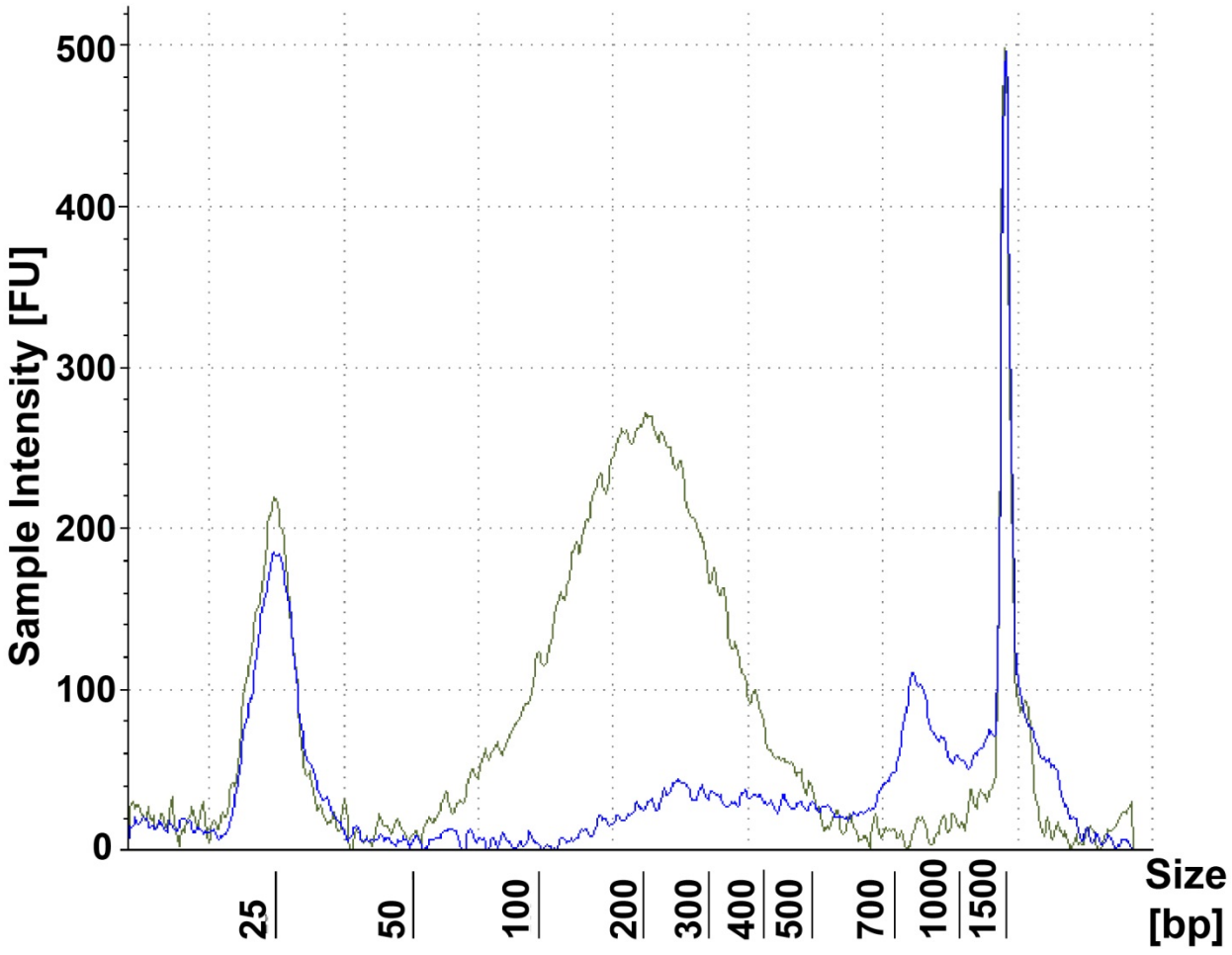


Figure 4.16 DNA size shifted after adapter ligation.

After entire ichip procedure, we were able to collect about 0.2 ng DNA, which further reduced recovery rate to 0.7%. Without ichip, we were able to obtain 3.69 ng DNA after MOWChIP procedure, which was about 13% recovery rate from starting 5000 cells. ichip elution and using the concentrator caused serious sample lost. However, we suspect that such elution may also cause damage to the protein of the interest, making it less likely to bind with its antibody, which was observed as result from low signal to noise ratio of sequencing results. As comparison, we started with about 500 cells (2.5 ng) for MOWChIP, the signal to noise ratio was high in this case (**Figure 4.17**).

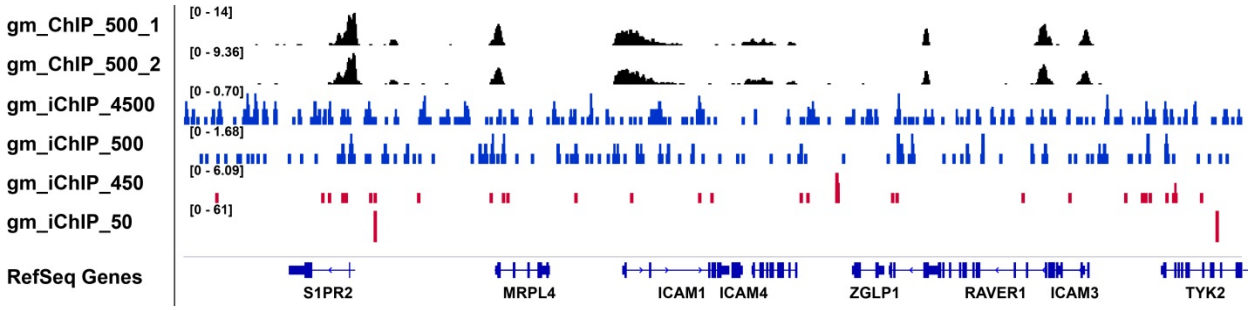


Figure 4.17 Demultiplexing after sequencing. 500 gm12878 cell ChIP-seq. 50 cells were pooled with 450 cells for iChIP-seq. 500 cells were pooled with 4500 cells for iChIP-seq.

After we optimized the technology with several parameters as reported in previous paper (e.g. antibody concentrations, washing conditions), we conducted in vivo experiment of xenotransplantation of CTCs into 6 weeks old immunodeficient mice (NOD SCID gamma, Jackson) to study temporal dynamics of CTCs during cancer metastasis[236].

First, we generated tumor xenografts from cell lines. 1-2 x 10^6 MDA-MB-231 cells were injected orthotopically into bilateral inguinal mammary fat pads of three 6 weeks old mice[237, 238]. About 1 mL blood samples were collected from each mouse by cardiac puncture of the left ventricle when animals were euthanized for tumor burden at 12 weeks. We collected about 5000 K nucleated cells from 1 mL blood per mouse after red blood lysis. After negative selection, we collected 1000 - 10000 CTCs/mL blood per mouse.

We took 300 cells from 1000-10000 cells that were isolated from negative selection for H3K4me3 ChIP-seq. 300 CTCs had only an average 10.18% mapping rate to hg19 and an average 5210 called peaks (**Figure 4.18**), but an average 84.47% mapping rate to mm10 and an average 865 called peaks. This suggests that CTCs isolated from negative selection were not pure, they were mainly mouse white blood cells (**Figure 4.19**). We then took 300 cells that genereated from primary mammary tumor for H3K4me3 ChIP-seq. 300 primary tumor cells had an average 60.42% mapping rate to hg19 and an average 879 called peaks (**Figure 4.18**), and an average 14.28% mapping rate to mm10.

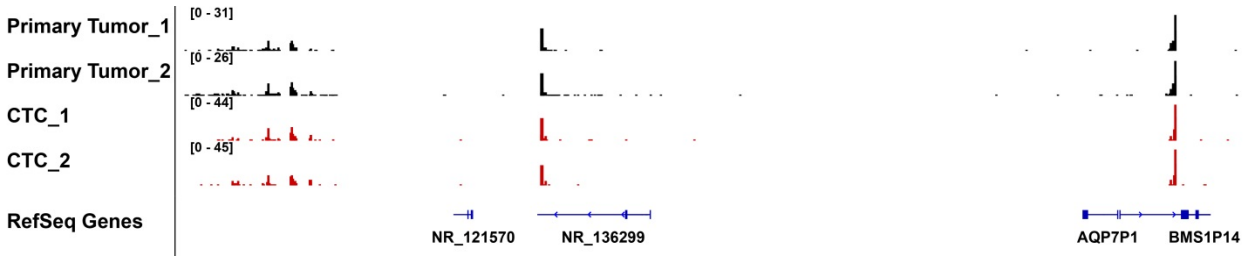


Figure 4.18 300 CTCs and Primary Tumor ChIP-seq. CTC reads shown are 10.18% mapped to human genome. Primary reads shown are 60.42% mapped to human genome.

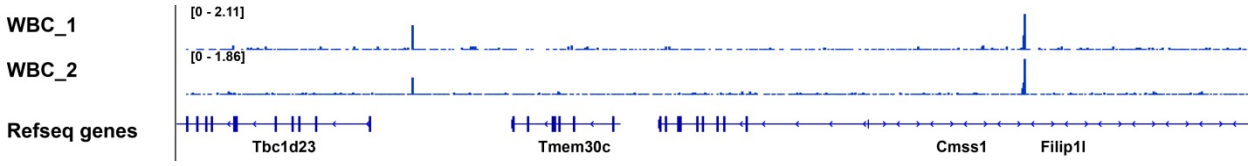


Figure 4.19 300 WBCs ChIP-seq. Reads shown are 84.47% mapped to mouse genome.

5. Summary and Future Work

Dynamic epigenomic changes during cell differentiation or disease development often require profiling epigenomes using a tiny quantity of tissue samples. Conventional epigenomic tests do not support such analysis due to the large amount of materials required by these assays. I developed a low input and high throughput microfluidics-based immunoprecipitation followed by next-generation sequencing technology for profiling epigenomes using as little as 0.5 ng DNA (or ~100 cells) with 1.5 h on-chip process for immunoprecipitation. This technology enabled me to examine epigenomic regulations (e.g., DNA methylation, histone modifications, and transcriptional regulations) during cancer development.

Although our microfluidic can process as little as 0.5 ng DNA as starting material. The study using single live mouse is challenging especially in the case of analyzing early-staged PDX (which do not produce enough CTCs cells) with ChIP-seq (which requires 10 fold more cells than ChIP-qPCR). Therefore, in the future, we will try to make three improvements to our process. First, we will improve CTC isolation by using spiral channel microfluidics developed by University of Michigan[239]. This is important because we can only obtain small volumes of blood from mice (75-100 μL).

Second, we will try to improve the purity of primary tumor by isolating the "host" cells[240]. We will generate tumor xenografts from mouse tumor cell line 4T1[226] by injecting suspensions of tumor cells into the fourth inguinal mammary fat pad of mice.

Third, while our goal is to study temporal dynamic histone modification during the course of cancer metastasis with our optimized ChIP-seq technology, low levels of CTCs (0–7 cells at days 8-23 and 26-55 cells at days 8–23) are expected after negative enrichment[237]. Because we would

not have 100 cells after 4 weeks, we plan to barcode the chromatin with before immunoprecipitation by tagmentation (transposase Tn5)[241]. This will allow us to pool CTC samples from each week together along with other samples (e.g., primary tumor samples) to meet 0.5 ng minimum input requirement for ChIP-seq. We can then compare Epigenetics with bioinformatics tool, such as, DiffBind[242] to reveal different levels of histione modification happed in CTCs and primary tumor during mammary cancer metastasis.

With these proposed technologies, we hope to detect histone modification heterogeneity among different live mice that share similar metastatic characteristics. Even for earlier stage PDX, the tagmentation method will allow us to pool indexed CTCs from different mice together to resolve the low input issue for ChIP-seq. In addition, we expect that further modifications to the microfluidic architecture as well as the on-chip chemistries will be implemented to increase overall sensitivity and fidelity.

References

1. Sarma, K. & Reinberg, D. Histone variants meet their match. *Nat Rev Mol Cell Bio* **6**, 139-149 (2005).
2. Richmond, T.J. Hot papers - Crystal structure - Crystal structure of the nucleosome core particle at 2.8 angstrom resolution by K. Luger, A.W. Mader, R.K. Richmond, D.F. Sargent, T.J. Richmond - Comments. *Scientist* **13**, 15-15 (1999).
3. Routh, A., Sandin, S. & Rhodes, D. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *P Natl Acad Sci USA* **105**, 8872-8877 (2008).
4. Groth, A., Rocha, W., Verreault, A. & Almouzni, G. Chromatin challenges during DNA replication and repair. *Cell* **128**, 721-733 (2007).
5. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).
6. Ruthenburg, A.J., Li, H., Patel, D.J. & Allis, C.D. Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Bio* **8**, 983-994 (2007).
7. Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **442**, 772-778 (2006).
8. Venkatesh, S. & Workman, J.L. Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Bio* **16**, 178-189 (2015).
9. Jenuwein, T. & Allis, C.D. Translating the histone code. *Science* **293**, 1074-1080 (2001).
10. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
11. Koch, C.M. et al. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* **17**, 691-707 (2007).
12. Talbert, P.B. & Henikoff, S. Spreading of silent chromatin: inaction at a distance. *Nat Rev Genet* **7**, 793-803 (2006).

13. Huang, J.Q. et al. Lsh, an epigenetic guardian of repetitive elements. *Nucleic Acids Res* **32**, 5019-5028 (2004).

14. Ruthenburg, A.J., Li, H., Patel, D.J. & Allis, C.D. Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol* **8**, 983-994 (2007).

15. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465-476 (2008).

16. Egger, G., Liang, G.N., Aparicio, A. & Jones, P.A. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429**, 457-463 (2004).

17. Kashyap, V. et al. Regulation of stem cell pluripotency and differentiation involves a mutual regulatory circuit of the NANOG, OCT4, and SOX2 pluripotency transcription factors with polycomb repressive complexes and stem cell microRNAs. *Stem Cells Dev* **18**, 1093-1108 (2009).

18. Boland, M.J., Nazor, K.L. & Loring, J.F. Epigenetic regulation of pluripotency and differentiation. *Circ Res* **115**, 311-324 (2014).

19. Zhu, J. et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642-654 (2013).

20. Capell, B.C. & Berger, S.L. Genome-wide epigenetics. *J Invest Dermatol* **133**, e9 (2013).

21. Romanoski, C.E., Glass, C.K., Stunnenberg, H.G., Wilson, L. & Almouzni, G. Epigenomics: Roadmap for regulation. *Nature* **518**, 314-316 (2015).

22. Bernstein, B.E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* **28**, 1045-1048 (2010).

23. Rajender, S., Avery, K. & Agarwal, A. Epigenetics, spermatogenesis and male infertility. *Mutat Res* **727**, 62-71 (2011).

24. Krivtsov, A.V. et al. H3K79 Methylation Profiles Define Murine and Human MLL-AF4 Leukemias. *Cancer Cell* **14**, 355-368 (2008).

25. Feinberg, A.P. & Tycko, B. Timeline - The history of cancer epigenetics. *Nat Rev Cancer* **4**, 143-153 (2004).

26. Bannister, A.J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res* **21**, 381-395 (2011).

27. Bernstein, B.E., Meissner, A. & Lander, E.S. The mammalian epigenome. *Cell* **128**, 669-681 (2007).

28. Edmunds, J.W., Mahadevan, L.C. & Clayton, A.L. Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. *Embo J* **27**, 406-420 (2008).

29. Steger, D.J. et al. DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. *Mol Cell Biol* **28**, 2825-2839 (2008).

30. Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R. & Young, R.A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77-88 (2007).

31. Jones, P.A. et al. Moving AHEAD with an international human epigenome project. *Nature* **454**, 711-715 (2008).

32. Ehrlich, M. et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* **10**, 2709-2721 (1982).

33. Blackledge, N.P. & Klose, R. CpG island chromatin: a platform for gene regulation. *Epigenetics* **6**, 147-152 (2011).

34. Deaton, A.M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-1022 (2011).

35. Robertson, K.D. DNA methylation and human disease. *Nat Rev Genet* **6**, 597-610 (2005).

36. Esteller, M. Epigenetics in cancer. *N Engl J Med* **358**, 1148-1159 (2008).

37. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature reviews. Genetics* **8**, 286-298 (2007).

38. Jones, P.A. & Baylin, S.B. The epigenomics of cancer. *Cell* **128**, 683-692 (2007).

39. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33 Suppl**, 245-254 (2003).

40. Herman, J.G. & Baylin, S.B. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* **349**, 2042-2054 (2003).

41. Feinberg, A.P. & Tycko, B. The history of cancer epigenetics. *Nature reviews. Cancer* **4**, 143-153 (2004).

42. Gonzalo, S. Epigenetic alterations in aging. *J Appl Physiol (1985)* **109**, 586-597 (2010).

43. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

44. Cosma, M.P. Ordered recruitment: Gene-specific mechanism of transcription activation. *Mol Cell* **10**, 227-236 (2002).

45. Bentley, D. The mRNA assembly line: transcription and processing machines in the same factory. *Curr Opin Cell Biol* **14**, 336-342 (2002).

46. Jimenez-Sanchez, G., Childs, B. & Valle, D. Human disease genes. *Nature* **409**, 853-855 (2001).

47. Darnell, J.E. Transcription factors as targets for cancer therapy. *Nat Rev Cancer* **2**, 740-749 (2002).

48. Baldwin, A.S., Jr. Series introduction: the transcription factor NF-kappaB and human disease. *J Clin Invest* **107**, 3-6 (2001).

49. Beg, A.A., Finco, T.S., Nantermet, P.V. & Baldwin, A.S., Jr. Tumor necrosis factor and interleukin-1 lead to phosphorylation and loss of I kappa B alpha: a mechanism for NF-kappa B activation. *Mol Cell Biol* **13**, 3301-3310 (1993).

50. Nowak, D.E., Tian, B. & Brasier, A.R. Two-step cross-linking method for identification of NF-kappaB gene network by chromatin immunoprecipitation. *Biotechniques* **39**, 715-725 (2005).

51. Shendure, J. & Ji, H.L. Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135-1145 (2008).

52. Schuster, S.C. Next-generation sequencing transforms today's biology. *Nat Methods* **5**, 16-18 (2008).

53. Metzker, M.L. APPLICATIONS OF NEXT-GENERATION SEQUENCING Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46 (2010).

54. Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *P Natl Acad Sci USA* **100**, 8817-8822 (2003).

55. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732 (2005).

56. Kim, J.B. et al. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481-1484 (2007).

57. Leamon, J.H. et al. A massively parallel PicoTiterplate (TM) based platform for discrete picoliter-scale polymerase chain reactions (vol 24, pg 3769, 2003). *Electrophoresis* **25**, 1176-1176 (2004).

58. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* **34** (2006).

59. Harris, T.D. et al. Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106-109 (2008).

60. Eid, J. et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133-138 (2009).

61. Metzker, M.L. et al. Termination of DNA-Synthesis by Novel 3'-Modified-Deoxyribonucleoside 5'-Triphosphates. *Nucleic Acids Res* **22**, 4259-4267 (1994).

62. Canard, B. & Sarfati, R.S. DNA-Polymerase Fluorescent Substrates with Reversible 3'-Tags. *Gene* **148**, 1-6 (1994).

63. Ju, J.Y. et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *P Natl Acad Sci USA* **103**, 19635-19640 (2006).

64. Guo, J. et al. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *P Natl Acad Sci USA* **105**, 9145-9150 (2008).

65. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).

66. Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363-+ (1998).

67. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242**, 84-89 (1996).

68. Levene, M.J. et al. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682-686 (2003).

69. Tomkinson, A.E., Vijayakumar, S., Pascal, J.M. & Ellenberger, T. DNA ligases: Structure, reaction mechanism, and function. *Chem Rev* **106**, 687-699 (2006).

70. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A Ligase-Mediated Gene Detection Technique. *Science* **241**, 1077-1080 (1988).

71. Valouev, A. et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**, 1051-1063 (2008).

72. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133-141 (2008).

73. Valouev, A. et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5**, 829-834 (2008).

74. Orlando, V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* **25**, 99-104 (2000).

75. O'Neill, L.P. & Turner, B.M. Immunoprecipitation of native chromatin: NChIP. *Methods* **31**, 76-82 (2003).

76. Tolstorukov, M.Y., Kharchenko, P.V., Goldman, J.A., Kingston, R.E. & Park, P.J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome Res* **19**, 967-977 (2009).

77. Henikoff, S., Henikoff, J.G., Sakai, A., Loeb, G.B. & Ahmad, K. Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome Res* **19**, 460-469 (2009).

78. Kidder, B.L., Hu, G.Q. & Zhao, K. ChIP-Seq: technical considerations for obtaining high-quality data. *Nat Immunol* **12**, 918-922 (2011).

79. Dahl, J.A. & Collas, P. mu ChIP - a rapid micro chromatin immunoprecipitation assay for small cell samples and biopsies. *Nucleic Acids Res* **36** (2008).

80. Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-680 (2009).

81. Next Generation Microarray Bioinformatics: Methods and Protocols. *Methods Mol Biol* **802**, 1-401 (2012).

82. Dahl, J.A. & Collas, P. A rapid micro chromatin immunoprecipitation assay (microChIP). *Nat Protoc* **3**, 1032-1045 (2008).

83. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

84. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

85. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

86. Landt, S.G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**, 1813-1831 (2012).

87. Kharchenko, P.V., Tolstorukov, M.Y. & Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**, 1351-1359 (2008).

88. Boyle, A.P., Guinney, J., Crawford, G.E. & Furey, T.S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537-2538 (2008).

89. Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-680 (2009).

90. Mikkelsen, T.S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-U552 (2007).

91. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9** (2008).

92. Landt, S.G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**, 1813-1831 (2012).

93. Chen, Y.W. et al. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat Methods* **9**, 609-+ (2012).

94. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods* **58**, 189-191 (2012).

95. Sims, D., Sudbery, I., Ilott, N.E., Heger, A. & Ponting, C.P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**, 121-132 (2014).

96. Esteller, M. Aberrant DNA methylation as a cancer-inducing mechanism. *Annu Rev Pharmacol* **45**, 629-656 (2005).

97. Glasspool, R.M., Teodoridis, J.M. & Brown, R. Epigenetics as a mechanism driving polygenic clinical drug resistance. *Brit J Cancer* **94**, 1087-1092 (2006).

98. Tsai, H.C. et al. Transient low doses of DNA-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells. *Cancer Cell* **21**, 430-446 (2012).

99. Tost, J. DNA Methylation: An Introduction to the Biology and the Disease-Associated Changes of a Promising Biomarker. *Mol Biotechnol* **44**, 71-81 (2010).

100. Down, T.A. et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology* **26**, 779-785 (2008).

101. Pelizzola, M. et al. MEDME: An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome Res* **18**, 1652-1659 (2008).

102. Chavez, L. et al. Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res* **20**, 1441-1450 (2010).

103. Hohne, D.N., Younger, J.G. & Solomon, M.J. Flexible Microfluidic Device for Mechanical Property Characterization of Soft Viscoelastic Solids Such as Bacterial Biofilms. *Langmuir* **25**, 7743-7751 (2009).

104. Whitesides, G.M., Ostuni, E., Takayama, S., Jiang, X.Y. & Ingber, D.E. Soft lithography in biology and biochemistry. *Annu Rev Biomed Eng* **3**, 335-373 (2001).

105. Huh, D., Torisawa, Y.S., Hamilton, G.A., Kim, H.J. & Ingber, D.E. Microengineered physiological biomimicry: Organs-on-Chips. *Lab Chip* **12**, 2156-2164 (2012).

106. Folch, A. & Toner, M. Microengineering of cellular interactions. *Annu Rev Biomed Eng* **2**, 227-+ (2000).

107. Duffy, D.C., McDonald, J.C., Schueller, O.J.A. & Whitesides, G.M. Rapid prototyping of microfluidic systems in poly(dimethylsiloxane). *Anal Chem* **70**, 4974-4984 (1998).

108. Anderson, J.R. et al. Fabrication of topologically complex three-dimensional microfluidic systems in PDMS by rapid prototyping. *Anal Chem* **72**, 3158-3164 (2000).

109. Greener, J. et al. Rapid, cost-efficient fabrication of microfluidic reactors in thermoplastic polymers by combining photolithography and hot embossing. *Lab Chip* **10**, 522-524 (2010).

110. Studer, V. et al. Scaling properties of a low-actuation pressure microfluidic valve. *J Appl Phys* **95**, 393-398 (2004).

111. Unger, M.A., Chou, H.P., Thorsen, T., Scherer, A. & Quake, S.R. Monolithic microfabricated valves and pumps by multilayer soft lithography. *Science* **288**, 113-116 (2000).

112. Thorsen, T., Maerkl, S.J. & Quake, S.R. Microfluidic large scale integration. *Science* **298**, 580-584 (2002).

113. Burns, M.A. et al. An integrated nanoliter DNA analysis device. *Science* **282**, 484-487 (1998).

114. Hong, J.W., Studer, V., Hang, G., Anderson, W.F. & Quake, S.R. A nanoliter-scale nucleic acid processor with parallel architecture. *Nature Biotechnology* **22**, 435-439 (2004).

115. Pal, R. et al. An integrated microfluidic device for influenza and other genetic analyses. *Lab Chip* **5**, 1024-1032 (2005).

116. Witek, M.A., Llopis, S.D., Wheatley, A., McCarley, R.L. & Soper, S.A. Purification and preconcentration of genomic DNA from whole cell lysates using photoactivated polycarbonate (PPC) microfluidic chips. *Nucleic Acids Res* **34** (2006).

117. Yeung, S.W., Lee, T.M.H., Cai, H. & Hsing, I.M. A DNA biochip for on-the-spot multiplexed pathogen identification. *Nucleic Acids Res* **34** (2006).

118. Cipriany, B.R. et al. Single molecule epigenetic analysis in a nanofluidic channel. *Anal Chem* **82**, 2480-2487 (2010).

119. Ittner, L.M. & Goetz, J. Pronuclear injection for the production of transgenic mice. *Nat Protoc* **2**, 1206-1215 (2007).

120. Gordon, J.W., Scangos, G.A., Plotkin, D.J., Barbosa, J.A. & Ruddle, F.H. Genetic transformation of mouse embryos by microinjection of purified DNA. *Proc Natl Acad Sci U S A* **77**, 7380-7384 (1980).

121. Thomas, K.R. & Capecchi, M.R. Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell* **51**, 503-512 (1987).

122. Green, J.E. et al. The C3(1)/SV40 T-antigen transgenic mouse model of mammary cancer: ductal epithelial cell targeting with multistage progression to carcinoma. *Oncogene* **19**, 1020-1027 (2000).

123. Maroulakou, I.G., Anver, M., Garrett, L. & Green, J.E. Prostate and Mammary Adenocarcinoma in Transgenic Mice Carrying a Rat C3(1) Simian-Virus-40 Large Tumor-Antigen Fusion Gene. *P Natl Acad Sci USA* **91**, 11236-11240 (1994).

124. Weiss, A.T.A., Delcour, N.M., Meyer, A. & Klopfleisch, R. Efficient and Cost-Effective Extraction of Genomic DNA From Formalin-Fixed and Paraffin-Embedded Tissues. *Vet Pathol* **48**, 834-838 (2011).

125. Fischer, A.H., Jacobson, K.A., Rose, J. & Zeller, R. Hematoxylin and eosin staining of tissue and cell sections. *CSH Protoc* **2008**, pdb prot4986 (2008).

126. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-731 (2012).

127. Sun, H. et al. Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Res* **39**, 190-201 (2011).

128. Mokry, M. et al. Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes. *Nucleic Acids Res* **40**, 148-158 (2012).

129. Raha, D. et al. Close association of RNA polymerase II and many transcription factors with Pol III genes. *P Natl Acad Sci USA* **107**, 3639-3644 (2010).

130. Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* **11**, 394-403 (2004).

131. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-731 (2012).

132. Mokry, M. et al. Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes. *Nucleic Acids Res* **40**, 148-158 (2012).

133. Brivanlou, A.H. & Darnell, J.E., Jr. Signal transduction and the control of gene expression. *Science* **295**, 813-818 (2002).

134. Zwart, W. et al. A carrier-assisted ChIP-seq method for estrogen receptor-chromatin interactions from breast cancer core needle biopsy samples. *Bmc Genomics* **14**, 232 (2013).

135. Stender, J.D. et al. Genome-wide analysis of estrogen receptor alpha DNA binding and tethering mechanisms identifies Runx1 as a novel tethering factor in receptor-mediated transcriptional activation. *Mol Cell Biol* **30**, 3943-3955 (2010).

136. Yaghmaie, F. et al. Caloric restriction reduces cell loss and maintains estrogen receptor-alpha immunoreactivity in the pre-optic hypothalamus of female B6D2F1 mice. *Neuro Endocrinol Lett* **26**, 197-203 (2005).

137. Lumachi, F., Santeufemia, D.A. & Basso, S.M. Current medical treatment of estrogen receptor-positive breast cancer. *World J Biol Chem* **6**, 231-239 (2015).

138. Alberts, B. et al. Molecular biology of the cell, Edn. 4th. (Garland Science, New York; 2002).

139. Massie, C.E. & Mills, I.G. ChIPping away at gene regulation. *Embo Rep* **9**, 337-343 (2008).

140. Farnham, P.J. Insights from genomic profiling of transcription factors. *Nat Rev Genet* **10**, 605-616 (2009).

141. Collas, P. The current state of chromatin immunoprecipitation. *Mol Biotechnol* **45**, 87-100 (2010).

142. Dahl, J.A. & Collas, P. mu ChIP - a rapid micro chromatin immunoprecipitation assay for small cell samples and biopsies. *Nucleic Acids Res* **36** (2008).

143. O'Neill, L.P., VerMilyea, M.D. & Turner, B.M. Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. *Nat Genet* **38**, 835-841 (2006).

144. Nelson, J.D., Denisenko, O. & Bomsztyk, K. Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nat Protoc* **1**, 179-185 (2006).

145. O'Neill, L.P. & Turner, B.M. Immunoprecipitation of chromatin. *Methods Enzymol* **274**, 189-197 (1996).

146. Spencer, V.A., Sun, J.M., Li, L. & Davie, J.R. Chromatin immunoprecipitation: a tool for studying histone acetylation and transcription factor binding. *Methods* **31**, 67-75 (2003).

147. Shankaranarayanan, P. et al. Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. *Nat Methods* **8**, 565-567 (2011).

148. Shen, J. et al. H3K4me3 epigenomic landscape derived from ChIP-Seq of 1,000 mouse early embryonic cells. *Cell Res* **25**, 143-147 (2015).

149. Cao, Z., Chen, C., He, B., Tan, K. & Lu, C. A microfluidic device for epigenomic profiling using 100 cells. *Nat Methods* **12**, 959-962 (2015).

150. Chellappan, S.P. Chromatin Protocols Third Edition Preface. *Chromatin Protocols, 3rd Edition* **1288**, V-Vi (2015).

151. Poorey, K. et al. Measuring chromatin interaction dynamics on the second time scale at single-copy genes. *Science* **342**, 369-372 (2013).

152. Tian, B., Yang, J. & Brasier, A.R. Two-step cross-linking for analysis of protein-chromatin interactions. *Methods Mol Biol* **809**, 105-120 (2012).

153. Bosisio, D. et al. A hyper-dynamic equilibrium between promoter-bound and nucleoplasmic dimers controls NF-kappa B-dependent gene activity. *Embo J* **25**, 798-810 (2006).

154. Carroll, J.S. et al. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38**, 1289-1297 (2006).

155. Diaz, A., Park, K., Lim, D.A. & Song, J.S. Normalization, bias correction, and peak calling for ChIP-seq. *Stat Appl Genet Mol* **11** (2012).

156. Okano, M., Bell, D.W., Haber, D.A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257 (1999).

157. Jeltsch, A. Molecular enzymology of mammalian DNA methyltransferases. *Curr Top Microbiol* **301**, 203-225 (2006).

158. Frommer, M. et al. A Genomic Sequencing Protocol That Yields a Positive Display of 5-Methylcytosine Residues in Individual DNA Strands. *P Natl Acad Sci USA* **89**, 1827-1831 (1992).

159. Clark, S.J., Harrison, J., Paul, C.L. & Frommer, M. High-Sensitivity Mapping of Methylated Cytosines. *Nucleic Acids Res* **22**, 2990-2997 (1994).

160. Eckhardt, F. et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38**, 1378-1385 (2006).

161. Rakyan, V.K. et al. DNA methylation profiling of the human major histocompatibility complex: A pilot study for the Human Epigenome Project. *Plos Biology* **2**, 2170-2182 (2004).

162. Taylor, K.H. et al. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res* **67**, 8511-8518 (2007).

163. Cokus, S.J. et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215-219 (2008).

164. Lister, R. et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-536 (2008).

165. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465-476 (2008).

166. Meissner, A. et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868-5877 (2005).

167. Meissner, A. et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-U791 (2008).

168. Bock, C. et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol* **28**, 1106-1114 (2010).

169. Weber, M. et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**, 853-862 (2005).

170. Grimm, C. et al. DNA-methylome analysis of mouse intestinal adenoma identifies a tumour-specific signature that is partly conserved in human colon cancer. *PLoS Genet* **9**, e1003250 (2013).

171. Harris, R.A. et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* **28**, 1097-1105 (2010).

172. Down, T.A. et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26**, 779-785 (2008).

173. Weber, M. et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**, 853-862 (2005).

174. Taiwo, O. et al. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* **7**, 617-636 (2012).

175. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **33**, 1165-1172 (2015).

176. Stark, A., Shin, D.J., Pisanic, T., 2nd, Hsieh, K. & Wang, T.H. A parallelized microfluidic DNA bisulfite conversion module for streamlined methylation analysis. *Biomed Microdevices* **18**, 5 (2016).

177. Zhu, Y. & Lu, C. in Microfluidic Methods for Molecular Biology. (eds. C. Lu & S.S. Verbridge) 349-363 (Springer International Publishing, Cham; 2016).

178. Zhao, M.T., Whyte, J.J., Hopkins, G.M., Kirk, M.D. & Prather, R.S. Methylated DNA Immunoprecipitation and High-Throughput Sequencing (MeDIP-seq) Using Low Amounts of Genomic DNA. *Cell Reprogram* **16**, 175-184 (2014).

179. Lienhard, M., Grimm, C., Morkel, M., Herwig, R. & Chavez, L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* **30**, 284-286 (2014).

180. McLean, C.Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).

181. Zimmermann, P. et al. Correlation of COL10A1 induction during chondrogenesis of mesenchymal stem cells with demethylation of two CpG sites in the COL10A1 promoter. *Arthritis Rheum* **58**, 2743-2753 (2008).

182. Hiramatsu, K. et al. Generation of hyaline cartilaginous tissue from mouse adult dermal fibroblast culture by defined factors. *J Clin Invest* **121**, 640-657 (2011).

183. Barter, M.J., Bui, C. & Young, D.A. Epigenetic mechanisms in cartilage and osteoarthritis: DNA methylation, histone modifications and microRNAs. *Osteoarthr Cartilage* **20**, 339-349 (2012).

184. Goldring, M.B. & Marcu, K.B. Epigenomic and microRNA-mediated regulation in cartilage development, homeostasis, and osteoarthritis. *Trends Mol Med* **18**, 109-118 (2012).

185. Im, G.I. & Choi, Y.J. Epigenetics in osteoarthritis and its implication for future therapeutics. *Expert Opin Biol Th* **13**, 713-721 (2013).

186. Bradley, E.W., McGee-Lawrence, M.E. & Westendorf, J.J. Hdac-Mediated Control of Endochondral and Intramembranous Ossification. *Crit Rev Eukar Gene* **21**, 101-113 (2011).

187. den Hollander, W. & Meulenbelt, I. DNA Methylation in Osteoarthritis. *Curr Genomics* **16**, 419-426 (2015).

188. Patel, S. & Maso, C.J. Ossification in metastases from carcinoma of the breast. *JAMA* **198**, 1309-1311 (1966).

189. Harris, R.A. et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* **28**, 1097-U1194 (2010).

190. Bock, C. et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol* **28**, 1106-U1196 (2010).

191. Scher, H.I. et al. Circulating tumour cells as prognostic markers in progressive, castration-resistant prostate cancer: a reanalysis of IMMC38 trial data. *Lancet Oncol* **10**, 233-239 (2009).

192. Hou, J.M. et al. Clinical Significance and Molecular Characteristics of Circulating Tumor Cells and Circulating Tumor Microemboli in Patients With Small-Cell Lung Cancer. *J Clin Oncol* **30**, 525-532 (2012).

193. Krebs, M.G. et al. Evaluation and Prognostic Significance of Circulating Tumor Cells in Patients With Non-Small-Cell Lung Cancer. *J Clin Oncol* **29**, 1556-1563 (2011).

194. Aggarwal, C. et al. Relationship among circulating tumor cells, CEA and overall survival in patients with metastatic colorectal cancer. *Ann Oncol* **24**, 420-428 (2013).

195. Deneve, E. et al. Capture of Viable Circulating Tumor Cells in the Liver of Colorectal Cancer Patients. *Clin Chem* **59**, 1384-1392 (2013).

196. Alix-Panabieres, C. & Pantel, K. Challenges in circulating tumour cell research. *Nat Rev Cancer* **14**, 623-631 (2014).

197. Lin, H.K. et al. Portable Filter-Based Microdevice for Detection and Characterization of Circulating Tumor Cells. *Clin Cancer Res* **16**, 5011-5018 (2010).

198. Nagrath, S. et al. Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* **450**, 1235-1239 (2007).

199. Hou, H.W. et al. Isolation and retrieval of circulating tumor cells using centrifugal forces. *Sci Rep-Uk* **3** (2013).

200. Sollier, E. et al. Size-selective collection of circulating tumor cells using Vortex technology. *Lab Chip* **14**, 63-77 (2014).

201. Rao, C.G. et al. Expression of epithelial cell adhesion molecule in carcinoma cells present in blood and primary and metastatic tumors. *Int J Oncol* **27**, 49-57 (2005).

202. Yu, M. et al. Circulating Breast Tumor Cells Exhibit Dynamic Changes in Epithelial and Mesenchymal Composition. *Science* **339**, 580-584 (2013).

203. Pantel, K. & Alix-Panabieres, C. The clinical significance of circulating tumor cells. *Nat Clin Pract Oncol* **4**, 62-63 (2007).

204. Pantel, K., Alix-Panabieres, C. & Riethdorf, S. Cancer micrometastases. *Nat Rev Clin Oncol* **6**, 339-351 (2009).

205. Pantel, K. & Alix-Panabieres, C. Circulating tumour cells in cancer patients: challenges and perspectives. *Trends Mol Med* **16**, 398-406 (2010).

206. Bednarz, N. et al. BRCA1 Loss Preexisting in Small Subpopulations of Prostate Cancer Is Associated with Advanced Disease and Metastatic Spread to Lymph Nodes and Peripheral Blood. *Clin Cancer Res* **16**, 3340-3348 (2010).

207. Bidard, F.C. et al. Clinical application of circulating tumor cells in breast cancer: overview of the current interventional trials. *Cancer Metast Rev* **32**, 179-188 (2013).

208. Riethdorf, S. et al. Detection and HER2 Expression of Circulating Tumor Cells: Prospective Monitoring in Breast Cancer Patients Treated in the Neoadjuvant GeparQuattro Trial. *Clin Cancer Res* **16**, 2634-2645 (2010).

209. Ignatiadis, M. et al. HER2-Positive Circulating Tumor Cells in Breast Cancer. *Plos One* **6** (2011).

210. Miyamoto, D.T. et al. Androgen Receptor Signaling in Circulating Tumor Cells as a Marker of Hormonally Responsive Prostate Cancer. *Cancer Discov* **2**, 995-1003 (2012).

211. Markou, A., Strati, A., Malamos, N., Georgoulias, V. & Lianidou, E.S. Molecular Characterization of Circulating Tumor Cells in Breast Cancer by a Liquid Bead Array Hybridization Assay. *Clin Chem* **57**, 421-430 (2011).

212. Munz, M. et al. Side-by-side analysis of five clinically tested anti-EpCAM monoclonal antibodies. *Cancer Cell Int* **10** (2010).

213. Hsieh, H.B. et al. High speed detection of circulating tumor cells. *Biosens Bioelectron* **21**, 1893-1899 (2006).

214. Vona, G. et al. Isolation by size of epithelial tumor cells - A new method for the immunomorphological and molecular characterization of circulating tumor cells. *Am J Pathol* **156**, 57-63 (2000).

215. Karabacak, N.M. et al. Microfluidic, marker-free isolation of circulating tumor cells from blood samples. *Nat Protoc* **9**, 694-710 (2014).

216. Lara, O., Tong, X.D., Zborowski, M. & Chalmers, J.J. Enrichment of rare cancer cells through depletion of normal cells using density and flow-through, immunomagnetic cell separation. *Exp Hematol* **32**, 891-904 (2004).

217. Yang, L.Y. et al. Optimization of an Enrichment Process for Circulating Tumor Cells From the Blood of Head and Neck Cancer Patients Through Depletion of Normal Cells. *Biotechnol Bioeng* **102**, 521-534 (2009).

218. Alix-Panabieres, C. & Pantel, K. Challenges in circulating tumour cell research. *Nat Rev Cancer* **14**, 623-631 (2014).

219. Meng, S.D. et al. HER-2 gene amplification can be acquired as breast cancer progresses. *P Natl Acad Sci USA* **101**, 9393-9398 (2004).

220. Antonarakis, E.S. et al. AR-V7 and Resistance to Enzalutamide and Abiraterone in Prostate Cancer. *New Engl J Med* **371**, 1028-1038 (2014).

221. Pixberg, C.F., Schulz, W.A., Stoecklein, N.H. & Neves, R.P. Characterization of DNA Methylation in Circulating Tumor Cells. *Genes (Basel)* **6**, 1053-1075 (2015).

222. Pestrin, M. et al. Correlation of HER2 status between primary tumors and corresponding circulating tumor cells in advanced breast cancer patients. *Breast Cancer Res Tr* **118**, 523-530 (2009).

223. Fehm, T. et al. HER2 status of circulating tumor cells in patients with metastatic breast cancer: a prospective, multicenter trial. *Breast Cancer Res Tr* **124**, 403-412 (2010).

224. Powell, A.A. et al. Single Cell Profiling of Circulating Tumor Cells: Transcriptional Heterogeneity and Diversity from Breast Cancer Cell Lines. *Plos One* **7** (2012).

225. Strati, A. et al. Gene expression profile of circulating tumor cells in breast cancer by RT-qPCR. *Bmc Cancer* **11** (2011).

226. Baccelli, I. et al. Identification of a population of blood circulating tumor cells from breast cancer patients that initiates metastasis in a xenograft assay. *Nat Biotechnol* **31**, 539-U143 (2013).

227. Zhang, L. et al. The Identification and Characterization of Breast Cancer CTCs Competent for Brain Metastasis (vol 5, 189er5 2013). *Sci Transl Med* **5** (2013).

228. Hodgkinson, C.L. et al. Tumorigenicity and genetic profiling of circulating tumor cells in small-cell lung cancer. *Nat Med* **20**, 897-903 (2014).

229. Chimonidou, M. et al. DNA Methylation of Tumor Suppressor and Metastasis Suppressor Genes in Circulating Tumor Cells. *Clin Chem* **57**, 1169-1177 (2011).

230. Chimonidou, M., Strati, A., Malamos, N., Georgoulias, V. & Lianidou, E.S. SOX17 Promoter Methylation in Circulating Tumor Cells and Matched Cell-Free DNA Isolated from Plasma of Patients with Breast Cancer. *Clin Chem* **59**, 270-279 (2013).

231. Balkouranidou, I. et al. Breast cancer metastasis suppressor-1 promoter methylation in cell free DNA provides prognostic information in non-small cell lung cancer. *Cancer Res* **74** (2014).

232. Friedlander, T.W. et al. Detection and Characterization of Invasive Circulating Tumor Cells Derived from Men with Metastatic Castration- Resistant Prostate Cancer. *Int J Cancer* **134**, 2284-2293 (2014).

233. Toss, A., Mu, Z., Fernandez, S. & Cristofanilli, M. CTC enumeration and characterization: moving toward personalized medicine. *Ann Transl Med* **2**, 108 (2014).

234. Ignatiadis, M. & Dawson, S.J. Circulating tumor cells and circulating tumor DNA for precision medicine: dream or reality? *Ann Oncol* **25**, 2304-2313 (2014).

235. Karabacak, N.M. et al. Microfluidic, marker-free isolation of circulating tumor cells from blood samples. *Nat Protoc* **9**, 694-710 (2014).

236. Kim, M.Y. et al. Tumor Self-Seeding by Circulating Cancer Cells. *Cell* **139**, 1315-1326 (2009).

237. Eliane, J.P. et al. Monitoring serial changes in circulating human breast cancer cells in murine xenograft models. *Cancer Res* **68**, 5529-5532 (2008).

238. Smith, M.C.P. et al. CXCR4 regulates growth of both primary and metastatic breast cancer. *Cancer Res* **64**, 8604-8612 (2004).

239. Lin, E. et al. High-Throughput Microfluidic Labyrinth for the Label-free Isolation of Circulating Tumor Cells. *Cell Syst* **5**, 295-304 e294 (2017).

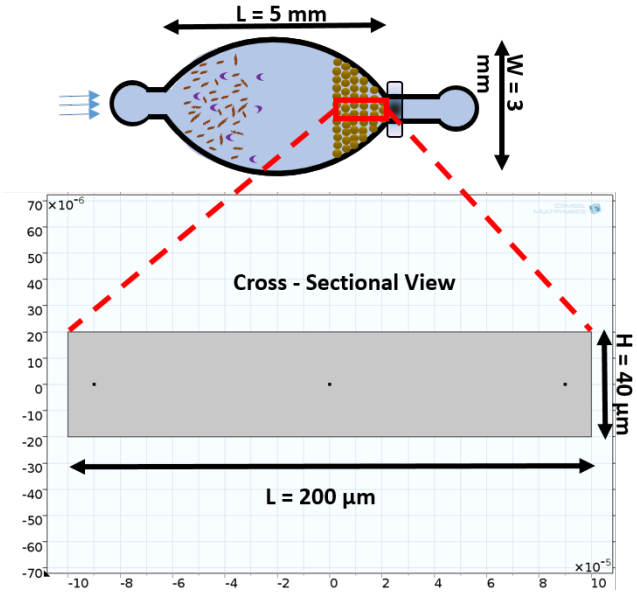
240. Ullal, A.V. et al. Cancer cell profiling by barcoding allows multiplexed protein analysis in fine-needle aspirates. *Sci Transl Med* **6**, 219ra219 (2014).

241. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).

242. Ross-Innes, C.S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389-393 (2012).

Appendix

COMSOL Simulation



Adsorption Diffusion Convection

$$\left(\varepsilon + \rho_b k_{P,c}\right)\frac{\partial c_i}{\partial t} - D_{e,i}\nabla^2 c_i + u\nabla c_i = 0$$

Initial Conditions:

$u_x = 40\ \mu m/s$

$c_{0,c} = 1\ mol/m^3$

$c(t = 0) = 0\ mol/m^3$

Boundary Conditions:

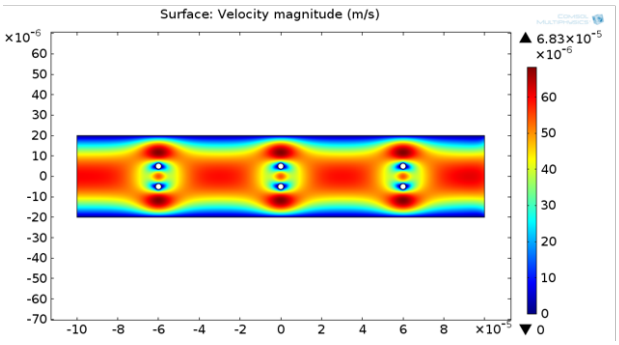
$-D_{e,i}\nabla c_i + uc_i = 0 \mid (y = \pm 20)$

$-D_{e,i}\nabla c_i = 0 \mid (x = 10)$

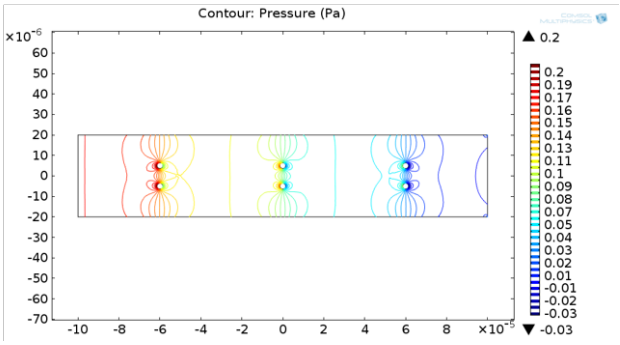
Viscous Resistance Inertial Resistance

$$\Delta P = -\left(\frac{\mu}{\alpha}u_i + C\frac{1}{2}\rho u u_i\right)\Delta n$$

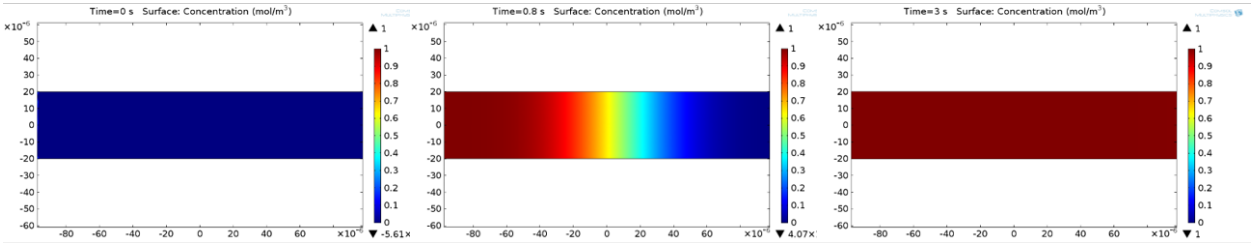
Velocity Profile



Pressure Profile



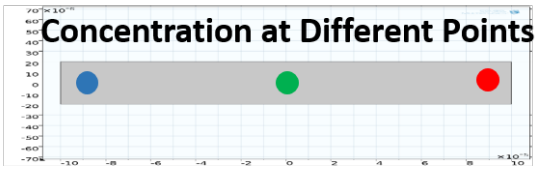
Concentration Profile



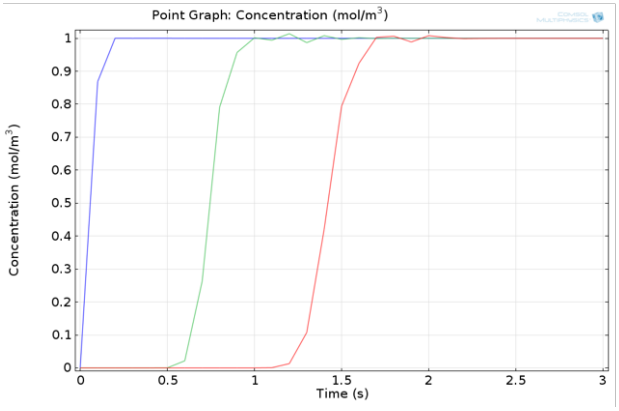
t = 0 s

t = 0.8 s

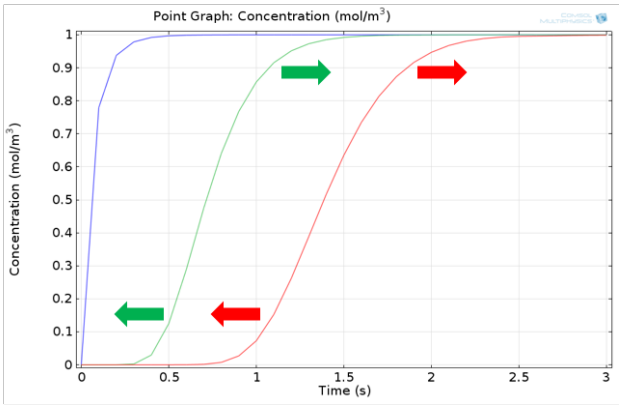
t = 3 s

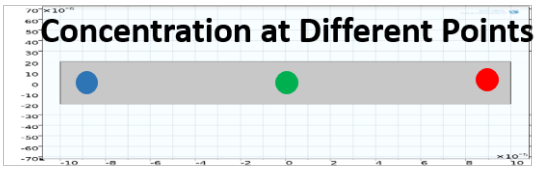


$D = 10^{-10}$, $k_D = 10^{-6}$

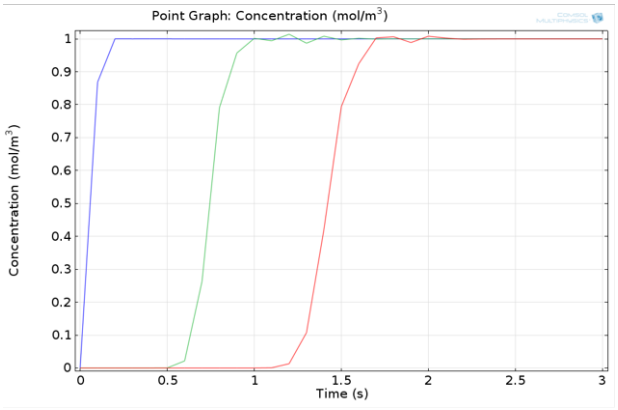


$D = 10^{-9}$, $k_D = 10^{-6}$

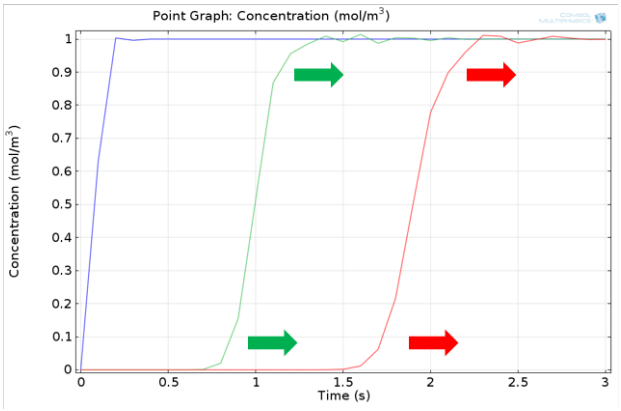




$D = 10^{-10}$, $k_D = 10^{-6}$



$D = 10^{-10}$, $k_D = 10^{-4}$



D=diffusivity

D=10^{-10} small protein

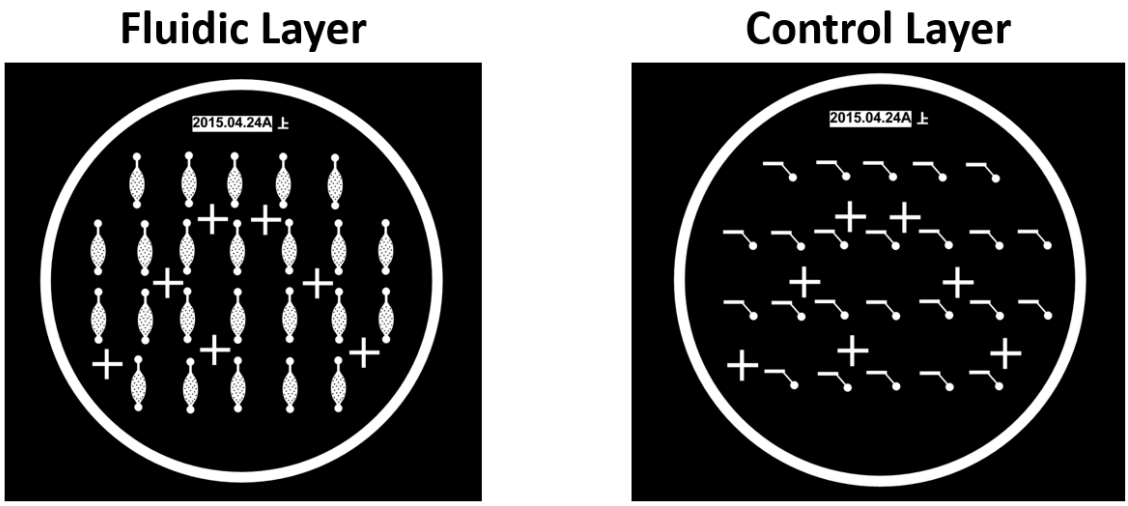
D=10^{-9} water

KD=adsorption rate constant (porous medium)

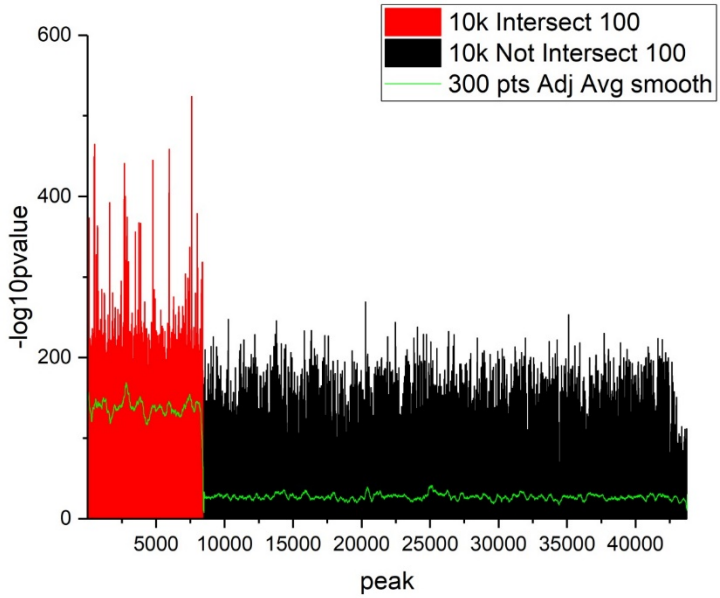
KD=10^{-6} moderate protein affinity

KD=10^{-4} high protein affinity

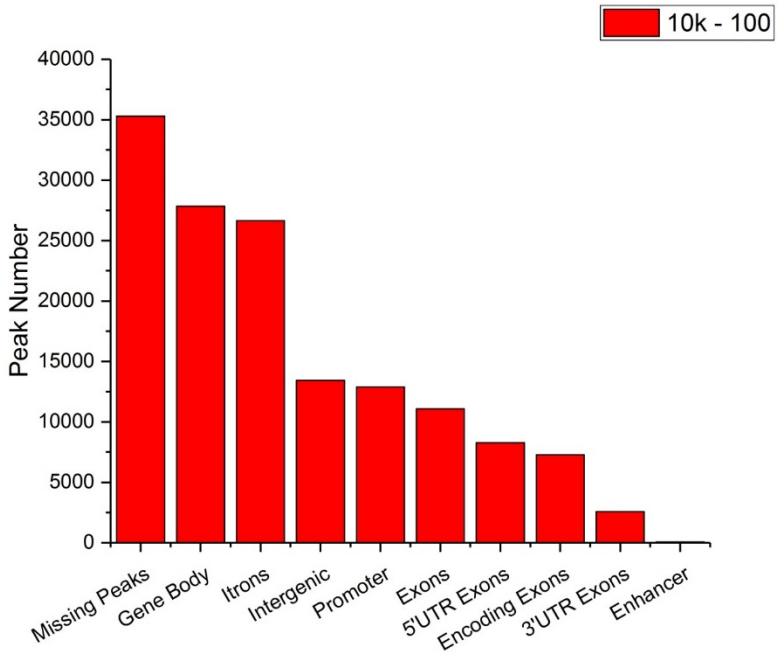
Better diffusion, concentration build up faster



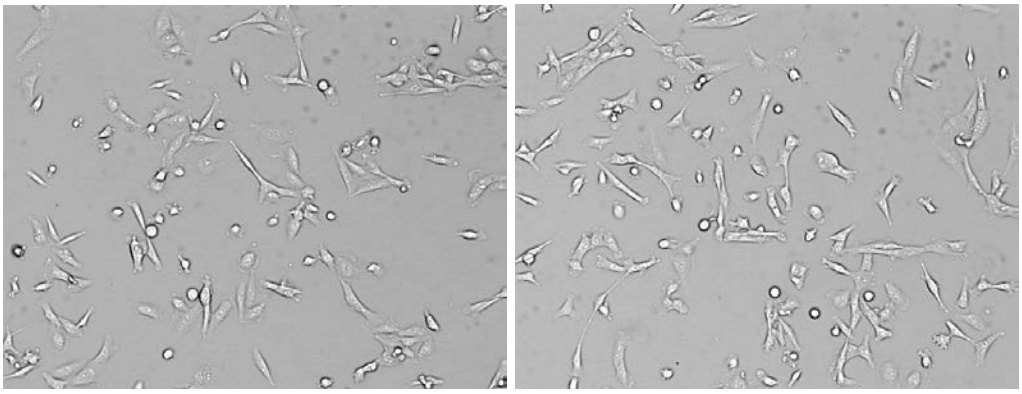
27 pillars in each chamber → Aspect Ratio $\frac{H}{W} > \frac{1}{10}$



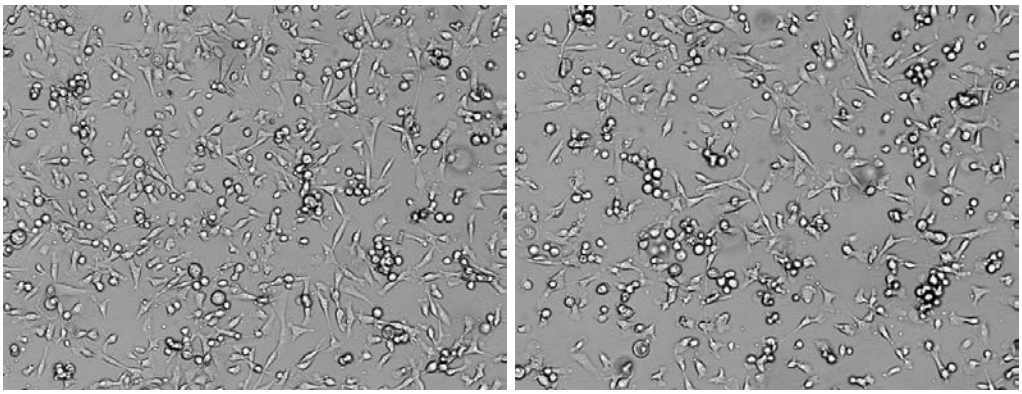
GM12878 H3K4ME3 10k – 100 cells missing peaks are those peaks with low intensity



GM12878 H3K4ME3 10k – 100 cells missing peaks in different regions



MDA-MB-231



Isolated CTC from mice blood draw after 6 wk of tumor injection (12 wk old mice)

Bioinformatics

General

Download data from Chicago
wget -r ftp://clu:FXu7CNs2@fgfftp.uchicago.edu/FOLDER
Browser based login: https://fgfftp.uchicago.edu/
Your login name: clu
Your password: FXu7CNs2

Find file and move to a directory
find ./ -name '*.sra' | xargs mv -t
/lustre/work/blueridge/yanzhu/yan/medip/2016/06032016/Grimm
Concatenate fastq files
cat *.fastq >> 1.fastq

use GSM to find SRR#
grep ^SRR SRA_Accessions.tab | grep GSM
fastq-dump SRR...

SRA to fastq (sratoolkit)
fastq-dump xxx.sra
To see if SRA file is paired-ended
sra-paired xxx.sra
Extract paired-end reads from SRA file
fastq-dump --split-3 xxx.sra

Sickle trimming (Illumina quality using CASAVA >= 1.8 is Sanger encoded)
sickle se -f 1.fastq -t sanger -o trimmed_1.fastq

Align to genome(Bowtie, paired-ended use -1 -2)
bowtie -p 16 -m 1 -S -q /work/blueridge/yanzhu/BowtieIndexes/hg19 xxx.fastq xxx.sam

Align to genome(Bowtie2, paired-ended use -1 -2)
bowtie2 -p 16 -x /work/blueridge/yanzhu/BowtieIndexes/2/hg19 xxx.fastq > xxx.sam

Align to genome(BWA)
bwa aln -t
16 /work/blueridge/yanzhu/BwaIndexes/Homo_sapiens.GRCh38.dna.primary_assembly.fa
xxx.fastq > xxx.sai
bwa
samse /work/blueridge/yanzhu/BwaIndexes/Homo_sapiens.GRCh38.dna.primary_assembly.fa
xxx.sai xxx.fastq > xxx.sam
paired-ended

bwa sampe
/work/blueridge/yanzhu/BwaIndexes/Homo_sapiens.GRCh38.dna.primary_assembly.fa xxx_1.sai xxx_2.sai xxx_1.fastq xxx_2.fastq > xxx.sam
samtools flagstat xxx.sam > bwa.result

Check the number of lines containing the 'XS:' tag corresponds to the number of reads showing > 1 alignments
grep 'XS:' aligned.bam | wc -l
Remove the lines containing the ':XS' tag
grep -v "XS:" aligned.bam > aligned_unique.bam

Skip alignments with MAPQ smaller than 10
samtools view -@ 16 -b -q 10 aligned.bam > aligne.filtered.bam

SAM to BAM(samtools)
samtools view -@ 16 -S -b xxx.sam > xxx.bam

Sorting and Indexing(samtools)
samtools sort -@ 16 xxx.bam xxx.sorted
sort -o xxx.sorted -O sam -T tmpprefix xxx.sam
grep -v '^[[:space:]]*@' xxx.sam | sort -k3,3 -k4,4n > xxx.sorted.sam (remove sam header and sort to prepare sam file for methylkit)
samtools index xxx.sorted.bam

Genome Coverage
bedtools genomecov -ibam b2_1.sorted.bam -g /home/yanzhu/blueridge/bin/chromInfo.txt > b2_1_coverage.bed

Fragment densities counting(igvtools)
igvtools count -w 100 -e 200 xxx.sorted.bam xxx.sorted.wig hg19.chrom.sizes

wigToBigWig
./wigToBigWig xxx.sorted.wig hg19.chrom.sizes xxx.sorted.bw

Call peaks(MACS)
macs14 -t treat.bam -c control.bam -g hs -n filename -w -S
macs2 callpeak -t treat.bam -c control.bam -g hs -n filename -B

SPP call peaks
Rscript /home/yanzhu/blueridge/bin/run_spp.R -c=b2_1.bam -savp -out=spp_gm_pol2_ref

Call peaks (SPP)
#1804=read unmapped, mate unmapped, not primary alignment, read fails platform/vendor quality checks, read is PCR or optical duplicate
module load gcc
samtools view -@ 16 -b -F 1804 -q 30 b2_10.bam | bamToBed -i | awk 'BEGIN{FS="\t";OFS="\t"}{$4="N"; print $0}' | gzip -c > b2_10.tagAlign.gz
samtools view -@ 16 -b -F 1804 -q 30 b2_11.bam | bamToBed -i | awk 'BEGIN{FS="\t";OFS="\t"}{$4="N"; print $0}' | gzip -c > b2_11.tagAlign.gz

module purge
module load gcc/5.1.0 openblas R/3.2.0
Rscript /home/yanzhu/blueridge/bin/run_spp.R -c=b2_10.tagAlign.gz -i=b2_11.tagAlign.gz -npeak=300000 -savr -savp -rf -out=phantomPeakStatsReps.tab -x=-500:85 -p=16

IDR Analysis
Rscript batch-consistency-analysis.r b2_9.tagAlign_VS_b2_11.tagAlign.regionPeak b2_10.tagAlign_VS_b2_11.tagAlign.regionPeak -1 b2_9_11_VS_b2_10_11 0 F signal.value

IDR Plot
Rscript batch-consistency-plot.r 1 9_VS_10_IDR_plot b2_9_11_VS_b2_10_11

Threshold
numPeaks_9_10=$(awk '$11 <= 0.01 {print $0}' b2_9_11_VS_b2_10_11-overlapped-peaks.txt | wc -l)
#numPeaks_9_10 = 60
zcat b2_9.tagAlign_VS_b2_11.tagAlign.regionPeak | sort -k7nr,7nr | head -n 60 | gzip -c > spp.conservative.9_VS_11.regionPeak.gz

module purge
module load gcc/5.1.0 gsl/1.15 python/3.5.0 atlas mvapich2/2.1

export PYTHONPATH=/home/yanzhu/.local/lib/python3.5/site-packages/:/opt/apps/gcc5_1/python/3.5.0/modules/lib/python:/opt/apps/gcc5_1/python/3.5.0/lib

idr

AQUAS TF and histone ChIP-seq pipeline
module purge
unset PYTHONPATH
unset R_LIBS_USER
(source activate aquas_chipseq_py3)

python3 /home/yanzhu/blueridge/TF_chipseq_pipeline/chipseq.py -species hg19 (-type TF -final-stage idr) -nth 16 -fastq1 /DATA/REP1.fastq.gz -fastq2 /DATA/REP2.fastq.gz -ctl_fastq1 /DATA/CTL.fastq.gz

python3 /home/yanzhu/blueridge/TF_chipseq_pipeline/chipseq.py -species hg19 -nth 16 -fastq1 1.fastq -fastq2 4.fastq -fastq3 6.fastq -ctl_fastq1 2.fastq

QC metrics spreadsheet (TSV) generation
python /home/yanzhu/blueridge/TF_chipseq_pipeline/utils/parse_summary_qc_recursively.py --out-file ENCODE_summary.tsv

---

-F 1804 Remove read unmapped, mate unmapped, not primary alignment, read fails platform/vendor quality checks, and read is PCR or optical duplicate
-q 30 Remove low mapping quality reads
samtools view -@ 16 -b -F 1804 -q 30 b2_10.sorted.bam > b2_10_FILT.bam

Mark duplicates
java -Xmx728M -jar /home/yanzhu/blueridge/bin/picard.jar MarkDuplicates INPUT=b2_10_FILT.bam OUTPUT=b2_10_FILT_DUPMARK.bam METRICS_FILE=b2_10.dup.qc VALIDATION_STRINGENCY=LENIENT ASSUME_SORTED=true REMOVE_DUPLICATES=false

Remove duplicates
samtools view -@ 16 -F 1804 -b b2_10_FILT_DUPMARK.bam > b2_10_FINAL.bam

Index Final BAM file
samtools index b2_10_FINAL.bam
samtools flagstat b2_10_FINAL.bam > b2_10_FINAL.flagstat.qc

Compute library complexity (PCR bottleneck coefficient (PBC))
bedtools bamtobed -i b2_10_FILT_DUPMARK.bam | awk 'BEGIN{OFS="\t"}{print $1,$2,$3,$6}' | grep -v 'chrM' | sort | uniq -c | awk 'BEGIN{mt=0;m0=0;m1=0;m2=0} ($1==1){m1=m1+1} ($1==2){m2=m2+1} {m0=m0+1} {mt=mt+$1} END{printf "“%d\t%d\t%d\t%d\t%f\t%f\t%f\n",mt,m0,m1,m2,m0/mt,m1/m0,m1/m2}' > b2_10_FINAL.pbc.qc

Create SE tagAlign file
bedtools bamtobed -i b2_10_FINAL.bam | awk 'BEGIN{OFS="\t"}{$4="N";$5="1000";print $0}' | gzip -nc > b2_10_FINAL.tagAlign.gz

Subsampled tagAlign file for CC analysis
zcat b2_10_FINAL.tagAlign.gz | grep -v “chrM” | shuf -n 15000000 | gzip -nc > b2_10_FINAL_15M.tagAlign.gz

Calculate Cross-correlation QC scores
module purge
module load gcc/5.1.0 openblas R/3.2.0

Rscript /home/yanzhu/blueridge/bin/run_spp_nodups.R -c=b2_10_FINAL_15M.tagAlign.gz -p=16 -filtchr=chrM -savp=b2_10_FINAL_15M.cc.plot.pdf -out=b2_10_FINAL_15M.cc.qc

take the top value of Estimated fragment length col3
sed -r 's/,[^\t]+//g' b2_10_FINAL_15M.cc.qc > temp
mv temp b2_10_FINAL_15M.cc.qc

Call peak with SPP
Rscript /home/yanzhu/blueridge/bin/run_spp_nodups.R -c=b2_10_FINAL.tagAlign.gz -i=b2_11_FINAL.tagAlign.gz -p=16 -npeak=300000 -speak=360 -savr -savp -rf -out=b2_10_11.ccscores

zcat b2_10_FINAL.tagAlign_VS_b2_11_FINAL.tagAlign.regionPeak.gz | awk 'BEGIN{OFS="\t"}{ if ($2<0) $2=0; print $1,int($2),int($3),$4,$5,$6,$7,$8,$9,$10;}' | gzip -f -nc > b2_10_FINAL.tagAlign_x_b2_11_FINAL.tagAlign.regionPeak.gz

bedtools intersect -v -a <(zcat -f b2_10_FINAL.tagAlign_x_b2_11_FINAL.tagAlign.regionPeak.gz) -b /home/yanzhu/blueridge/refGene_Prom/hg19.blacklist.bed | awk 'BEGIN{OFS="\t"} {if ($5>1000) $5=1000; print $0}' | grep -P 'chr[\dXY]+[\t]' | gzip -nc > b2_10_FINAL.tagAlign_x_b2_11_FINAL.tagAlign.filt.regionPeak.gz

IDR analysis
module purge
module load gcc/5.1.0 gsl/1.15 atlas mvapich2/2.1

idr --samples b2_9_FINAL.tagAlign_x_b2_11_FINAL.tagAlign.filt.regionPeak.gz b2_10_FINAL.tagAlign_x_b2_11_FINAL.tagAlign.filt.regionPeak.gz --input-file-type narrowPeak --output-file b2_9_10 --rank signal.value --soft-idr-threshold 0.05 --plot --use-best-multisummit-IDR

Get peaks passing IDR threshold of 5%

awk 'BEGIN{OFS="\t"} $12>='"awk -v p=0.05 'BEGIN{print -log(p)/log(10)}'" {print $1,$2,$3,$4,$5,$6,$7,$8,$9,$10}' b2_9_10 | sort | uniq | sort -k7n,7n | gzip -nc > 9_VS_10.IDR0.05.narrowPeak.gz

bedtools intersect -v -a <(zcat -f 9_VS_10.IDR0.05.narrowPeak.gz) -b /home/yanzhu/blueridge/refGene_Prom/hg19.blacklist.bed | grep -P 'chr[\dXY]+[\t]' | awk

'BEGIN{OFS="\t"} {if ($5>1000) $5=1000; print $0} | gzip -nc >
9_VS_10.IDR0.05.filt.narrowPeak.gz

Normalize bdg generated from MACS based on SPMR and background noise reduction
macs2 callpeak -t b2_2.bam -c b2_10.bam -B --nomodel --extsize 300 --SPMR -g hs -n
macs2_b2_2_10

macs2 bdgcmp -t macs2_b2_2_10_treat_pileup.bdg -c macs2_b2_2_10_control_lambda.bdg -o
macs2_b2_2_10_FE.bdg -m FE
macs2 bdgcmp -t macs2_b2_2_10_treat_pileup.bdg -c macs2_b2_2_10_control_lambda.bdg -o
macs2_b2_2_10_logLR.bdg -m logLR -p 0.00001
macs2 bdgcmp -t macs2_b2_2_10_treat_pileup.bdg -c macs2_b2_2_10_control_lambda.bdg -o
macs2_b2_2_10_subtract.bdg -m subtract

bdg2bw macs2_b2_2_10_treat_pileup.bdg /home/yanzhu/blueridge/bin/chromInfo.txt
bdg2bw macs2_b2_2_10_FE.bdg /home/yanzhu/blueridge/bin/chromInfo.txt bdg2bw
macs2_b2_2_10_logLR.bdg /home/yanzhu/blueridge/bin/chromInfo.txt bdg2bw
macs2_b2_2_10_subtract.bdg /home/yanzhu/blueridge/bin/chromInfo.txt

Remove anything<0 from macs subtracted bw
bwtool remove less 0 macs2_b2_10_11_subtract.bw macs2_b2_10_11_subtract_removed.bw

bedGraphToBigWig
find . -name *treat*.gz| xargs cp -t bdg2bw
or
bdg2bw xxx.bdg /home/yanzhu/bin/chromInfo.txt (module swap intel gcc/4.7.2)

Normalize ChIP-seq signal based on RKPM
bamCoverage -b treatment.bam -o treatment.bw -p max
bamCoverage -b control.bam -o control.bw -p max

Calculate the log2 ratio of treatment/input (or treatment-input using --ratio subtract)
bigwigCompare -b1 treatment.bw -b2 control.bw -o log2ratio.bw

Normalize ChIP-seq signal based on RKPM and subtract from input
bamCompare -b1 treatment.bam -b2 control.bam –-ratio subtract –-normalizeUsingRPKM -o
RKPM_subtract.bw
bamCompare -b1 b2_10.sorted.bam -b2 b2_11.sorted.bam --ratio subtract --
normalizeUsingRPKM --blackListFileName
/home/yanzhu/blueridge/refGene_Prom/hg19.blacklist.bed -p max --skipNonCoveredRegions --
ignoreDuplicates --minMappingQuality 30 --samFlagExclude 1804 -
o bamcompare_10_11_rpkm_subtract_filt_nodup.bw

bamCompare -b1 b2_10.sorted.bam -b2 b2_11.sorted.bam --ratio log2 --normalizeUsingRPKM --blackListFileName /home/yanzhu/blueridge/refGene_Prom/hg19.blacklist.bed -p max --skipNonCoveredRegions --ignoreDuplicates --minMappingQuality 30 --samFlagExclude 1804 -o bamcompare_10_11_rpkm_log2_filt_nodup.bw

bamCoverage -b b2_10.sorted.bam --normalizeUsingRPKM --blackListFileName /home/yanzhu/blueridge/refGene_Prom/hg19.blacklist.bed -p max --skipNonCoveredRegions --ignoreDuplicates --minMappingQuality 30 --samFlagExclude 1804 -o bamcoverage_10_rpkm_filt_nodup.bw

Calculate Pearson correlation coefficient at whole genome
multiBigwigSummary(bigwigCorrelate) bins -b file1.bw file2.bw -o heatmap_wg.png --outFileCorMatrix CorMatrix_wg --corMethod pearson

multiBamSummary bins --bamfiles *.sorted.bam --labels xxx -out results.npz -p max

multiBigwigSummary bins -b *.bw -out results.npz -p max

plotCorrelation --corData results.npz --plotFile heatmap_pearson.pdf --corMethod pearson --whatToPlot heatmap --skipZeros --removeOutliers --outFileCorMatrix pearson_cor_heatmap --plotTitle "Pearson Correlation of Read Counts"

plotCorrelation --corData results.npz --plotFile test.pdf --corMethod pearson --whatToPlot heatmap --skipZeros --removeOutliers --outFileCorMatrix test --plotTitle "Pearson Correlation of Read Counts" --colorMap gnuplot2

plotCorrelation --corData results.npz --plotFile heatmap_spearman.pdf --corMethod spearman --whatToPlot heatmap --skipZeros --removeOutliers --outFileCorMatrix spearman_cor_heatmap --plotTitle "Spearman Correlation of Read Counts"

plotCorrelation --corData results.npz --plotFile scatterplot.pdf --corMethod pearson --whatToPlot scatterplot --skipZeros --removeOutliers --outFileCorMatrix pearson_cor_scatterplot

Calculate Pearson correlation coefficient at promoter region

multiBigwigSummary(bigwigCorrelate) BED-file -b file1.bw file2.bw -o heatmap_promoter.png --outFileCorMatrix CorMatrix_promoter --corMethod pearson --BED /home/yanzhu/refGene_Prom/refGene_Prom_2000_2000_hg19.bed

multiBamSummary BED-file --BED /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed --bamfiles *.sorted.bam --labels xxx -out results_promoter.npz -p max

multiBigwigSummary bins -b *.bw -bl /home/yanzhu/blueridge/refGene_Prom/hg19.blacklist.bed -out results_bin.npz -p max --outRawCounts scores_per_bin.tab

plotCorrelation --corData results_bin.npz --plotFile heatmap_bin.pdf --corMethod pearson --whatToPlot heatmap --skipZeros --removeOutliers --outFileCorMatrix pearson_cor_heatmap_bin --plotTitle "Pearson Correlation of Read Counts" --plotNumbers

multiBigwigSummary BED-file --BED /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed -b *.bw --labels xxx -bl /home/yanzhu/blueridge/refGene_Prom/hg19.blacklist.bed -out results_promoter.npz -p max --outRawCounts scores_per_promoter.tab

plotCorrelation --corData results_promoter.npz --plotFile heatmap_promoter.pdf --corMethod pearson --whatToPlot heatmap --skipZeros --removeOutliers --outFileCorMatrix pearson_cor_heatmap_promoter --plotTitle "Pearson Correlation of Read Counts" --plotNumbers

plotFingerprint
plotFingerprint -b gm_pol2_1e5_1.sorted.bam gm_pol2_1e5_2.sorted.bam gm_pol2_1e6_1.sorted.bam gm_pol2_1e6_2.sorted.bam gm_pol2_ref.sorted.bam gm_pol2_1e5_input.sorted.bam --labels gm_pol2_1e5_1 gm_pol2_1e5_2 gm_pol2_1e6_1 gm_pol2_1e6_2 gm_pol2_ref gm_pol2_1e5_input --skipZeros -T "Fingerprints of different samples" --plotFile fingerprints.pdf --outRawCounts fingerprints.tab -p max

plotCoverage -b gm_pol2_1e5_1.sorted.bam gm_pol2_1e5_2.sorted.bam gm_pol2_1e6_1.sorted.bam gm_pol2_1e6_2.sorted.bam gm_pol2_ref.sorted.bam --labels gm_pol2_1e5_1 gm_pol2_1e5_2 gm_pol2_1e6_1 gm_pol2_1e6_2 gm_pol2_ref -o coverage.pdf --plotTitle "Coverage" --outRawCounts coverage.tab --ignoreDuplicates -p max

plotPCA
plotPCA -in results_promoter.npz -o PCA_readCounts.pdf -T "PCA of read counts"

computeMatrix
computeMatrix reference-point -S gm_pol2_1e6_1.bw -R hg19_refseq_genes.bed --referencePoint TSS -a 2000 -b 2000 -out matrix_refseq_pol2.tab.gz -p max -q

plotHeatmap
plotHeatmap -m matrix_refseq_pol2.tab.gz -out heatmap_refseq_pol2.pdf --plotTitle 'gm_pol2_1e6' --regionsLabel 'Refseq Genes' --heatmapHeight 15

To find active and inactive region
plotHeatmap -m matrix_refseq_pol2_ref.tab.gz -out heatmap_refseq_pol2_ref_cluster.pdf --heatmapHeight 15 --kmeans 2

plotEnrichment
plotEnrichment -b gm_pol2_1e5_1.sorted.bam gm_pol2_1e5_2.sorted.bam gm_pol2_1e6_1.sorted.bam gm_pol2_1e6_2.sorted.bam gm_pol2_ref.sorted.bam gm_pol2_1e5_input.sorted.bam --labels gm_pol2_1e5_1 gm_pol2_1e5_2 gm_pol2_1e6_1 gm_pol2_1e6_2 gm_pol2_ref gm_pol2_1e5_input --BED hg19_refseq_genes.bed -o enrichment.pdf -p max

Getting intergenic regions
complementBed -i hg19_refseq_genes.bed -g chromInfo_hg19.txt > refGene_Intergenic_hg19.bed

Enhancer
intersectBed -wa -wb -a Vista_Enhancers_hg19.bed -b 2_not_intersect_8.bed | awk 'BEGIN{FS=OFS="\t"}{print $6,$7,$8,$9,$10,$11,$12,$13,$14,$15}' | sort -k1,1 -k2,2n -k3,3n -u > missingpeak_enhancer_intersect.bed

Promoter
intersectBed -wa -wb -a refGene_Prom_2000_2000_hg19.bed -b 2_not_intersect_8.bed | awk 'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$10,$11,$12,$13,$14,$15,$16}' | sort -k1,1 -k2,2n -k3,3n -u > missingpeak_prom_intersect.bed

Exon
intersectBed -wa -wb -a refGene_Exons_hg19.bed -b 2_not_intersect_8.bed | awk 'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$10,$11,$12,$13,$14,$15,$16}' | sort -k1,1 -k2,2n -k3,3n -u > missingpeak_exon_intersect.bed

Encoding Exon
intersectBed -wa -wb -a refGene_EncodingExons_hg19.bed -b 2_not_intersect_8.bed | awk 'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$10,$11,$12,$13,$14,$15,$16}' | sort -k1,1 -k2,2n -k3,3n -u > missingpeak_encodingexon_intersect.bed

5'UTR Exons

intersectBed -wa -wb -a refGene_5UTRExons_hg19.bed -b 2_not_intersect_8.bed | awk 'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$10,$11,$12,$13,$14,$15,$16}' | sort -k1,1 -k2,2n -k3,3n -u > missingpeak_5utrexon_intersect.bed

3' UTR Exons
intersectBed -wa -wb -a refGene_3UTRExons_hg19.bed -b 2_not_intersect_8.bed | awk 'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$10,$11,$12,$13,$14,$15,$16}' | sort -k1,1 -k2,2n -k3,3n -u > missingpeak_3utrexon_intersect.bed

Intron
intersectBed -wa -wb -a refGene_Introns_hg19.bed -b 2_not_intersect_8.bed | awk 'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$10,$11,$12,$13,$14,$15,$16}' | sort -k1,1 -k2,2n -k3,3n -u > missingpeak_intron_intersect.bed

Intergenic
intersectBed -wa -wb -a refGene_Intergenic_hg19.bed -b 2_not_intersect_8.bed | awk 'BEGIN{FS=OFS="\t"}{print $4,$5,$6,$7,$8,$9,$10,$11,$12,$13}' | sort -k1,1 -k2,2n -k3,3n -u > missingpeak_intergenic_intersect.bed

Gene Body
intersectBed -wa -wb -a hg19_refseq_genes_uniq.bed -b 2_not_intersect_8.bed | awk 'BEGIN{FS=OFS="\t"}{print $13,$14,$15,$16,$17,$18,$19,$20,$21,$22}' | sort -k1,1 -k2,2n -k3,3n -u > missingpeak_gene_intersect.bed

ROC

Getting promoter regions
UCSC Table Browser -> Group: Genes and Gene Prediction ->Track: RefSeq Genes -> Table: refGene -> Output Format: all fields from selected table
1. Transforming refGene(with TSS Info) file into bed format
old
awk 'BEGIN{FS=OFS="\t"}{print $3,$5,$6,$2"|"$13,0,$4}' refGene.txt> refGene.bed
new
awk 'BEGIN{FS=OFS="\t"}{print $1,$2,$3,$4,0,$6}' refGene_hg19.txt> refGene_hg19.bed

2. Getting promoter regions of each transcript

```
awk 'BEGIN{FS=OFS="\t"}($6=="+"){print $1,$2,$2,$4,$5,$6}' refGene.bed >
refGene.TSS.bed
awk 'BEGIN{FS=OFS="\t"}($6=="-"){print $1,$3,$3,$4,$5,$6}' refGene.bed >>
refGene.TSS.bed

awk 'BEGIN{FS=OFS="\t"}($6=="+"){print $1,$2,$2,$4,$5,$6}' refGene_hg19.bed >
refGene.TSS_hg19.bed
awk 'BEGIN{FS=OFS="\t"}($6=="-"){print $1,$3,$3,$4,$5,$6}' refGene_hg19.bed >>
refGene.TSS_hg19.bed
```

3. Use slopBed to increase the size of each feature (TSS) to define a promoter region around the TSS ([-2000,+500])

```
slopBed -i refGene.TSS.bed -l 2000 -r 500 -s  -g chromInfo.txt  | sort -k1,1 -k2,2n -k3,3n -u |
sortBed > refGene_Prom_2000_500.bed

slopBed -i refGene.TSS_hg19.bed -l 2000 -r 2000 -s -g
/home/yanzhu/blueridge/refGene_Prom/chromInfo_hg19.txt | sort -k1,1 -k2,2n -k3,3n -u |
sortBed > refGene_Prom_2000_2000_hg19.bed

slopBed -i refGene.TSS_hg19.bed -l 800 -r 200 -s -g
/home/yanzhu/blueridge/refGene_Prom/chromInfo_hg19.txt | sort -k1,1 -k2,2n -k3,3n -u |
sortBed > refGene_Prom_800_200_hg19.bed
```

Intersecting promoter regions with exp/gs

To obtain only MACS peaks (get rid of first 6 columns info from refGene_Prom_2000_2000.bed)

```
intersectBed -wa -wb -
a  /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed -b xxx.bed |
awk 'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$11}' > xxx_prom_intersect.bed
```

Intersecting promoter regions with exp/gs (macs2, p-value)

To obtain only MACS peaks (get rid of first 6 columns info from refGene_Prom_2000_2000.bed)

```
intersectBed -wa -wb -
a  /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed -b
exp/gs.narrowPeak | awk 'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$14}' >
exp/gs_prom_intersect.bed
```

Intersecting promoter regions with exp/gs (spp, signal value)

To obtain only MACS peaks (get rid of first 6 columns info from refGene_Prom_2000_2000.bed)

```
intersectBed -wa -wb -
a  /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed -b
```

exp/gs.narrowPeak | awk 'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$13}' >
exp/gs_prom_intersect.bed

Intersecting gs with exp for positive set
To obtain only exp peaks(get rid of first 4 columns info from gs_prom_intersect_sorted.bed)
Output only unique lines based on chrom/start/end (columns 1/2/3)
Sort 4th column (p-value, or idr)
intersectBed -wa -wb -a gs_prom_intersect.bed -b exp_prom_intersect.bed | awk
'BEGIN{FS=OFS="\t"}{print $5,$6,$7,$8}' | sort -k1,1 -k2,2n -k3,3n -u | sort -k4n >
exp_x_gs_positive.sorted.bed

Intersecting gs with exp for negative set
Output only unique lines based on chrom/start/end (columns 1/2/3)
Sort 4th column (p-value, or idr)
intersectBed -v -a exp_prom_intersect.bed -b gs_prom_intersect.bed | sort -k1,1 -k2,2n -k3,3n -u
| sort -k4n > exp_x_gs_negative.sorted.bed

Preparing for ROCR input
awk 'BEGIN{FS=OFS="\t"}{print $4, $5=1}' exp_x_gs_positive.sorted.bed >
exp_x_gs_positive.pred.label.bed
awk 'BEGIN{FS=OFS="\t"}{print $4, $5=0}' exp_x_gs_negative.sorted.bed >
exp_x_gs_negative.pred.label.bed
cat exp_x_gs_positive.pred.label.bed exp_x_gs_negative.pred.label.bed | awk
'BEGIN{FS=OFS="\t"}{print $1}' > exp_x_gs_pred.txt
cat exp_x_gs_positive.pred.label.bed exp_x_gs_negative.pred.label.bed | awk
'BEGIN{FS=OFS="\t"}{print $2}' > exp_x_gs_label.txt

sort only uniq idr peak
sort -k1,1 -k2,2n -k3,3n -u xxx.bed>xxx.uniq.bed

intersect peak file
intersectBed -wa -a gs_prom_intersect.bed -b exp_prom_intersect.bed | sort -k1,1 -k2,2n -k3,3n -
u > exp_x_gs_positive.sorted.bed

-f 0.5
B intersect at least 50% of A

ROCR
# it allows two different plots in the same frame
par(mfrow = c(1,2))
# plot a ROC curve for a single prediction run
# and color the curve according to cutoff.
library(ROCR)
data(pred,label)

```
pred=read.table("exp_x_gs_pred.txt")
label=read.table("exp_x_gs_label.txt")

pred <- prediction(pred,label)
perf <- performance(pred,"tpr", "fpr")

pdf("exp_x_gs_ROC.pdf")
plot(perf,colorize = TRUE)
# plot a ROC curve for a single prediction run
# with CI by bootstrapping and fitted curve
library(verification)
roc.plot(label,pred, xlab = "False
positive rate",
ylab = "True positive rate", main = NULL, CI = T, n.boot = 100, plot =
"both", binormal = TRUE)


# calculating AUC
auc <- performance(pred,"auc")
# now converting S4 class to vector
auc <- unlist(slot(auc, "y.values"))
# adding min and max ROC AUC to the center of the plot
minauc<-min(round(auc, digits = 2))
#maxauc<-max(round(auc, digits = 2))
minauct <- paste(c("AUC = "),minauc,sep="")
#maxauct <- paste(c("max(AUC) = "),maxauc,sep="")
legend(0.3,0.6,c(minauct,"\n"),border="white",cex=1.7,box.col = "white")


#Performance vs. cutoff

perf <- performance(pred, "acc")
plot(perf, avg= "vertical",
    spread.estimate="boxplot",
    show.spread.at= seq(0.1, 0.9, by=0.1))


#multiple ROC

pred1 <- prediction(pred_1e6_1,label_1e6_1)
perf1 <- performance(pred1,"tpr", "fpr")
plot(perf1,col=1)

par(new=TRUE)

pred2 <- prediction(pred_1e6_2,label_1e6_2)
```

```
perf2<- performance(pred2,"tpr", "fpr")
plot(perf2,col=2)

par(new=TRUE)

pred3 <- prediction(pred_1e6_3,label_1e6_3)
perf3<- performance(pred3,"tpr", "fpr")
plot(perf3,col=3)

par(new=TRUE)

pred4 <- prediction(pred_1e5_1,label_1e5_1)
perf4<- performance(pred4,"tpr", "fpr")
plot(perf4,col=4)

par(new=TRUE)

pred5 <- prediction(pred_1e5_2,label_1e5_2)
perf5<- performance(pred5,"tpr", "fpr")
plot(perf5,col=5)


abline(a=0, b= 1)
```

Compare first column of two files print overlap

```
awk -F, 'FNR==NR {a[$1]; next}; $1 in a' small#row.file big#row.file
```

```
./ROC GS_file EXP_file
ROC
#GS="SRX100530_ppr.IDR0.05.filt.12-col.bed"
#EXP="gm_pol2_50K_2F_M_ppr.IDR0.05.filt.12-col.bed"

GS="$1"
EXP="$2"

GS_PREFIX=$(echo ${GS}|awk -F'[.]' '{print $1}')
EXP_PREFIX=$(echo ${EXP}|awk -F'[.]' '{print $1}')

module load gcc
```

```
intersectBed -wa -wb -a
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed -b ${GS} | awk
'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$13}' > ${GS_PREFIX}_prom_intersect.bed

intersectBed -wa -wb -a
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed -b ${EXP} | awk
'BEGIN{FS=OFS="\t"}{print $7,$8,$9,$13}' > ${EXP_PREFIX}_prom_intersect.bed

intersectBed -wa -wb -a ${GS_PREFIX}_prom_intersect.bed -b
${EXP_PREFIX}_prom_intersect.bed | awk 'BEGIN{FS=OFS="\t"}{print $5,$6,$7,$8}' | sort -
k1,1 -k2,2n -k3,3n -u | sort -k4n > ${EXP_PREFIX}_x_${GS_PREFIX}_positive.sorted.bed

intersectBed -v -a ${EXP_PREFIX}_prom_intersect.bed -b
${GS_PREFIX}_prom_intersect.bed | sort -k1,1 -k2,2n -k3,3n -u | sort -k4n >
${EXP_PREFIX}_x_${GS_PREFIX}_negative.sorted.bed

awk 'BEGIN{FS=OFS="\t"}{print $4, $5=1}'
${EXP_PREFIX}_x_${GS_PREFIX}_positive.sorted.bed >
${EXP_PREFIX}_x_${GS_PREFIX}_positive.pred.label.bed
awk 'BEGIN{FS=OFS="\t"}{print $4, $5=0}'
${EXP_PREFIX}_x_${GS_PREFIX}_negative.sorted.bed >
${EXP_PREFIX}_x_${GS_PREFIX}_negative.pred.label.bed
cat ${EXP_PREFIX}_x_${GS_PREFIX}_positive.pred.label.bed
${EXP_PREFIX}_x_${GS_PREFIX}_negative.pred.label.bed | awk
'BEGIN{FS=OFS="\t"}{print $1}' > ${EXP_PREFIX}_x_${GS_PREFIX}_pred.txt
cat ${EXP_PREFIX}_x_${GS_PREFIX}_positive.pred.label.bed
${EXP_PREFIX}_x_${GS_PREFIX}_negative.pred.label.bed | awk
'BEGIN{FS=OFS="\t"}{print $2}' > ${EXP_PREFIX}_x_${GS_PREFIX}_label.txt

module load intel

Rscript ROC.R ${EXP_PREFIX}_x_${GS_PREFIX}_pred.txt
${EXP_PREFIX}_x_${GS_PREFIX}_label.txt
${EXP_PREFIX}_x_${GS_PREFIX}_ROC.pdf
```

```
ROC.R
args <- commandArgs(trailingOnly = TRUE)
print(args)
library(ROCR)
pred=read.table(args[1])
label=read.table(args[2])
pred <- prediction(pred,label)
```

```
perf <- performance(pred,"tpr", "fpr")
pdf(args[3])
plot(perf,colorize = TRUE)
auc <- performance(pred,"auc")
auc <- unlist(slot(auc, "y.values"))
minauc<-min(round(auc, digits = 2))
minauct <- paste(c("AUC = "),minauc,sep="")
legend(0.3,0.6,c(minauct,"\n"),border="white",cex=1.7,box.col = "white")
dev.off()
```

1 black
2 red
3 green
4 blue
5 cyan
6 magenta
7 yellow
8 grey

RRBS

Bismark to get %methylation
bismark -p 4 --sam /home/yanzhu/Genomes/Human/GRCh38/ xxx.fastq

BedGraph and Coverage
bismark_methylation_extractor --multicore 16 -s --bedGraph --counts --buffer_size 20G xxx.bam

CpG Report, BedGraph and Coverage (will get xxx.CpG_report.txt)
bismark_methylation_extractor --multicore 16 -s --bedGraph --counts --buffer_size 20G --cytosine_report --genome_folder /home/yanzhu/Genomes/Human/GRCh38/ xxx.bam

Generate CpG Report using Coverage
coverage2cytosine --genome_folder /home/yanzhu/Genomes/Human/GRCh38/ -o CytosineReport.txt xxx.cov

To prepare Input for MethylKit form xxx.CpG_report.txt or CytosineReport.txt (MethylKit required format: Chr, Start, Strand, MethylationRatio, Coverage, and add chr_ to the beginning of first column)
awk '{OFS="\t";if($4+0 > 0 || $5+0 >0) print "chr_"$1,$2,$3,$4/($4+$5),$4+$5;}' xxx.CpG_report.txt/CytosineReport.txt > InputForMethylKit.txt

Remove sam header and sort to prepare sam file for methylkit
grep -v '^[[:space:]]*@' xxx.sam | sort -k3,3 -k4,4n > xxx.sorted.sam
Add chr to the beginning of the 3rd column of sorted sam file to prepare for read.bismark
awk 'BEGIN{OFS="\t"}$3="chr"$3' xxx.sorted.sam > InputForMethylKit.sam

library(methylKit)
Read in sorted sam file
my.methRaw=read.bismark("xxx.sam",sample.id="xxx",assembly="hg38",read.context="CpG", save.folder=getwd())
Read in xxx_CpG.txt
myobj=read("xxx_CpG.txt",sample.id="test1",assembly="hg38",context="CpG")

Read in promoter regions
promoters = read.bed("/home/yanzhu/refGene_Prom/refGene_Prom_2000_500_hg38.bed")
Summarizes the methylation information over a given set of promoter regions
reads_in_promoters = regionCounts(my.methRaw, promoters)

Output methylation information at promoter regions to a tab-delimitated txt file
write.table(reads_in_promoters, file="meth_rrbs_ref_prom.txt", sep = "\t", quote = F, row.names = F)
Covert coverage to percent methylation and only print chr, start, end, %methylation
awk '{OFS="\t";if($6+0 > 0 || $7+0 >0) print $1,$2,$3,$6/($6+$7)}' meth_rrbs_ref_prom.txt > percmeth_rrbs_ref_prom.txt
Appending row number to first column
awk '$1=(FNR FS $1)' percmeth_rrbs_ref_prom.txt > percmeth_rrbs_ref_prom_ID.txt
Adding hearder
sed -i '1i ID\tChr\tStart\tStop\tBS' percmeth_rrbs_ref_prom_ID.txt

Just perform methylation call
perl /home/yanzhu/R/x86_64-unknown-linux-gnu-library/3.2/methylKit/exec/methCall.pl --read1 xxx.sam --type paired_sam --nolap --CpG xxx.CpG.txt

Intersecting promoter regions with RRBS (get rid of 1st line of xx.bedGraph)
sed '1d' xx.bedGraph > xxx.bedGraph
Getting promoter regions
UCSC Table Browser -> Group: Genes and Gene Prediction ->Track: RefSeq Genes -> Table: refGene -> Output Format: all fields from selected table
1. Transforming refGene(with TSS Info) file into bed format
awk 'BEGIN{FS=OFS="\t"}{print $3,$5,$6,$2"|"$13,0,$4}' refGene.txt> refGene.bed
2. Getting promoter regions of each transcript
awk 'BEGIN{FS=OFS="\t"}($6=="+"){print $1,$2,$2,$4,$5,$6}' refGene.bed > refGene.TSS.bed
awk 'BEGIN{FS=OFS="\t"}($6=="-"){print $1,$3,$3,$4,$5,$6}' refGene.bed >> refGene.TSS.bed
Get rid of [0, 0] start end positions
awk '(($2 != $3) || (($2 == $3) && ($2 != 0)))' refGene.TSS.bed
3. Use slopBed to increase the size of each feature (TSS) to define a promoter region around the TSS ([-2000,+500])
slopBed -i refGene.TSS.bed -l 2000 -r 500 -s -g chromInfo.txt | sortBed > refGene_Prom_2000_500.bed
Output only unique lines based on chrom/start/end (columns 1/2/3)
sort -k1,1 -k2,2n -k3,3n -u refGene_Prom_2000_500.bed > refGene_Prom_2000_500_Unique.bed

MeDIP

QC
library(MeDIPS)
library(BSgenome.Hsapiens.UCSC.hg19)
library("MeDIPSData")
bam.file.gm.7.MeDIP = system.file("zhenning", "7.bam", package = "MeDIPSData")
BSgenome="BSgenome.Hsapiens.UCSC.hg19"
uniq=TRUE
extend=300
shift=0
ws=100
sr = MeDIPS.saturation(file = bam.file.gm.7.MeDIP, BSgenome = BSgenome, uniq = uniq, extend = extend, shift = shift, window_size =

```
ws, nit = 10, nrit = 1, empty_bins = TRUE, rank = FALSE)
pdf('Saturation.pdf')
MeDIPS.plotSaturation(sr)
cr = MeDIPS.seqCoverage(file = bam.file.gm.7.MeDIP, pattern = "CG",
BSgenome = BSgenome, extend = extend, shift = shift, uniq = uniq)
pdf('SeqCoverage.pdf')
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="pie", cov.level =
c(0,1, 2, 3, 4, 5))
pdf('SeqCoverageHistogram.pdf')
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="hist", t = 15,
main="Sequence pattern coverage, histogram")
er = MeDIPS.CpGenrich(file = bam.file.gm.7.MeDIP, BSgenome = BSgenome,
extend = extend, shift = shift, uniq = uniq)
er
dev.off()
```

---

uniq=1 (only uniq reads)
uniq=0 (all reads)

```
library(MeDIPS)
library(BSgenome.Hsapiens.UCSC.hg19)
BSgenome="BSgenome.Hsapiens.UCSC.hg19"
uniq=1
extend=300
shift=0
ws=250

bamfile_gm_0.5ng=c("0.5ng_1.bam","0.5ng_2.bam")
gm_0.5ng=lapply(X=bamfile_gm_0.5ng, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)
save(gm_0.5ng,file="gm_0.5ng.RData")

bamfile_gm_5ng=c("5ng_1.bam","5ng_2.bam")
gm_5ng=lapply(X=bamfile_gm_5ng, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)
save(gm_5ng,file="gm_5ng.RData")

bamfile_gm_10ng=c("10ng_1.bam","10ng_2.bam")
gm_10ng=lapply(X=bamfile_gm_10ng, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)
```

```
save(gm_10ng,file="gm_10ng.RData")

bamfile_gm_100ng=c("100ng_1.bam","100ng_2.bam")
gm_100ng=lapply(X=bamfile_gm_100ng, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)
save(gm_100ng,file="gm_100ng.RData")

CS = MeDIPS.couplingVector(pattern = "CG", refObj = gm_100ng[[1]])
save(CS,file="CS.RData")

bamfile_gm_all=c("0.5ng_1.bam","0.5ng_2.bam",
"5ng_1.bam","5ng_2.bam", "10ng_1.bam","10ng_2.bam", "100ng_1.bam","100ng_2.bam" )

gm_all=lapply(X=bamfile_gm_all, FUN=MeDIPS.createSet, BSgenome = BSgenome, extend =
extend, shift = shift, uniq = uniq,window_size = ws)
save(gm_all,file="gm_all.RData")




promoter=read.table("/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.
bed")
cgi_UCSC = read.table("/home/yanzhu/blueridge/refGene_Prom/UCSC_hg19_cpg.bed")
cgi_model = read.table("/home/yanzhu/blueridge/refGene_Prom/model-based-cpg-islands-
hg19.txt",
header=F)
gm_all_promoter=lapply(X=bamfile_gm_all, FUN=MeDIPS.createROIset , ROI=promoter,
BSgenome = BSgenome, extend = extend, shift = shift, uniq = uniq)
save(gm_all_promoter,file="gm_all_promoter.RData")
gm_all_cgi=lapply(X=bamfile_gm_all, FUN=MeDIPS.createROIset , ROI=cgi, BSgenome
= BSgenome, extend = extend, shift = shift, uniq = uniq)
save(gm_all_cgi,file="gm_all_cgi.RData")



samtools view -@ 16 -b -
L  /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed 0.5ng_1.bam >
0.5ng_1_promoter.bam
samtools view -@ 16  -b -L
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed 0.5ng_2.bam >
0.5ng_2_promoter.bam
samtools view -@ 16  -b -L
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed 5ng_1.bam >
5ng_1_promoter.bam
```

```
samtools view -@ 16  -b -L
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed 5ng_2.bam >
5ng_2_promoter.bam
samtools view -@ 16  -b -L
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed 10ng_1.bam >
10ng_1_promoter.bam
samtools view -@ 16  -b -L
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed 10ng_2.bam >
10ng_2_promoter.bam
samtools view -@ 16  -b -L
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed 100ng_1.bam >
100ng_1_promoter.bam
samtools view -@ 16  -b -L
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed 100ng_2.bam >
100ng_2_promoter.bam

#Remove header of  model-based-cpg-islands-hg19.txt
sed '1d' model-based-cpg-islands-hg19.txt > model_hg19_cpg.bed


samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_hg19_cpg.bed
0.5ng_1.bam > 0.5ng_1_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_hg19_cpg.bed
0.5ng_2.bam > 0.5ng_2_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_hg19_cpg.bed
5ng_1.bam > 5ng_1_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_hg19_cpg.bed
5ng_2.bam > 5ng_2_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_hg19_cpg.bed
10ng_1.bam > 10ng_1_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_hg19_cpg.bed
10ng_2.bam > 10ng_2_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_hg19_cpg.bed
100ng_1.bam > 100ng_1_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_hg19_cpg.bed
100ng_2.bam > 100ng_2_model_cpg.bam



samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_hg19_cpg.bed
0.5ng_1.bam > 0.5ng_1_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_hg19_cpg.bed
0.5ng_2.bam > 0.5ng_2_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_hg19_cpg.bed
5ng_1.bam > 5ng_1_UCSC_cpg.bam
```

```
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_hg19_cpg.bed
5ng_2.bam > 5ng_2_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_hg19_cpg.bed
10ng_1.bam > 10ng_1_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_hg19_cpg.bed
10ng_2.bam > 10ng_2_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_hg19_cpg.bed
100ng_1.bam > 100ng_1_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_hg19_cpg.bed
100ng_2.bam > 100ng_2_UCSC_cpg.bam
```

```
bamfile_gm_all_promoter=c("0.5ng_1_promoter.bam","0.5ng_2_promoter.bam",
"5ng_1_promoter.bam","5ng_2_promoter.bam",
"10ng_1_promoter.bam","10ng_2_promoter.bam",
"100ng_1_promoter.bam","100ng_2_promoter.bam")
bamfile_gm_all_model_cpg=c("0.5ng_1_model_cpg.bam","0.5ng_2_model_cpg.bam",
"5ng_1_model_cpg.bam","5ng_2_model_cpg.bam",
"10ng_1_model_cpg.bam","10ng_2_model_cpg.bam",
"100ng_1_model_cpg.bam","100ng_2_model_cpg.bam")
bamfile_gm_all_UCSC_cpg=c("0.5ng_1_UCSC_cpg.bam","0.5ng_2_UCSC_cpg.bam",
"5ng_1_UCSC_cpg.bam","5ng_2_UCSC_cpg.bam",
"10ng_1_UCSC_cpg.bam","10ng_2_UCSC_cpg.bam",
"100ng_1_UCSC_cpg.bam","100ng_2_UCSC_cpg.bam")
```

```
gm_all_promoter=lapply(X=bamfile_gm_all_promoter, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size = ws)
save(gm_all_promoter,file="gm_all_promoter.RData")
gm_all_model_cpg=lapply(X=bamfile_gm_all_model_cpg, FUN=MeDIPS.createSet,
BSgenome = BSgenome, extend = extend, shift = shift, uniq = uniq,window_size = ws)
save(gm_all_model_cpg,file="gm_all_model_cpg.RData")
gm_all_UCSC_cpg=lapply(X=bamfile_gm_all_UCSC_cpg, FUN=MeDIPS.createSet,
BSgenome = BSgenome, extend = extend, shift = shift, uniq = uniq,window_size = ws)
save(gm_all_UCSC_cpg,file="gm_all_UCSC_cpg.RData")
```

```
pdf('Calibration.pdf')
MeDIPS.plotCalibrationPlot(CSet = CS, main = "Calibration Plot", MSet
```

```
= gm_100ng[[1]], rpkm = TRUE, xrange = TRUE)



cor.matrix = MeDIPS.correlation(MSets = c(gm_0.5ng,gm_5ng,gm_10ng,gm_100ng), plot = T,
method = "pearson")
save(cor.matrix,file="cor.matrix.RData")


for (i in 1:length(gm_all)) {
png(paste("Calibration_",i,".png",sep=""))
MeDIPS.plotCalibrationPlot(CSet = CS, MSet = gm_all[[i]], rpkm = TRUE, xrange = TRUE)
dev.off()
}


for (i in 1:length(bamfile_gm_all)) {
pdf(paste("Saturation_",i,".pdf",sep=""))
sr=MeDIPS.saturation(file= bamfile_gm_all[i], BSgenome = BSgenome, extend = extend, shift
= shift, uniq = uniq,window_size = ws, nit=10, nrit=1, empty_bins=TRUE, rank=FALSE)
MeDIPS.plotSaturation(sr)
dev.off()
}


for (i in 1:length(bamfile_gm_all)) {
pdf(paste("SeqCoverage_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_gm_all[i], pattern = "CG", BSgenome = BSgenome,
extend = extend, shift = shift, uniq = uniq)
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="pie", cov.level = c(0,1, 2, 3, 4, 5))
dev.off()
}

for (i in 1:length(bamfile_gm_all_promoter)) {
pdf(paste("SeqCoverage_Promoter_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_gm_all_promoter[i], pattern = "CG", BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq)
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="pie", cov.level = c(0,1, 2, 3, 4, 5))
dev.off()
}

for (i in 1:length(bamfile_gm_all_model_cpg)) {
pdf(paste("SeqCoverage_model_cpg_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_gm_all_model_cpg[i], pattern = "CG", BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq)
```

```
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="pie", cov.level = c(0,1, 2, 3, 4, 5))
dev.off()
}

for (i in 1:length(bamfile_gm_all_UCSC_cpg)) {
pdf(paste("SeqCoverage_UCSC_cpg_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_gm_all_UCSC_cpg[i], pattern = "CG", BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq)
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="pie", cov.level = c(0,1, 2, 3, 4, 5))
dev.off()
}



for (i in 1:length(bamfile_gm_all)) {
pdf(paste("SeqCoverageHistogram_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_gm_all[i], pattern = "CG", BSgenome = BSgenome,
extend = extend, shift = shift, uniq = uniq)
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="hist", t = 15)
dev.off()
}

for (i in 1:length(bamfile_gm_all)) {
er = MeDIPS.CpGenrich(file = bamfile_gm_all[i], BSgenome = BSgenome,
extend = extend, shift = shift, uniq = uniq)
write.table(er, file = "CpG_Enrichment.bed", append=T, quote =F, sep="\t", row.names=F,
col.names=!file.exists("CpG_Enrichment.bed"))
}

multiBamSummary BED-file --BED
/home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_hg19.bed --bamfiles
0.5ng_1.sorted.bam 0.5ng_2.sorted.bam 5ng_1.sorted.bam 5ng_2.sorted.bam 10ng_1.sorted.bam
10ng_2.sorted.bam 100ng_1.sorted.bam 100ng_2.sorted.bam --labels 0.5ng_1 0.5ng_2 5ng_1
5ng_2 10ng_1 10ng_2 100ng_1 100ng_2 -out results_promoter.npz -p max

plotCorrelation --corData results_promoter.npz --plotFile heatmap_promoter.pdf --corMethod
pearson --whatToPlot heatmap --skipZeros --removeOutliers --outFileCorMatrix
pearson_cor_heatmap_promoter --plotTitle "Pearson Correlation of Read Counts"
```

Correlation

```
library(MeDIPS)
library(BSgenome.Hsapiens.UCSC.hg19)
BSgenome="BSgenome.Hsapiens.UCSC.hg19"
uniq=TRUE
extend=300
shift=0
ws=100

bamfile_gm_5ng=c("/work/blueridge/yanzhu/zhenning/medips/gm_5ng_rep1.bam","/work/blueridge/yanzhu/zhenning/medips/gm_5ng_rep2.bam","/work/blueridge/yanzhu/zhenning/medips/gm_5ng_rep3.bam","/work/blueridge/yanzhu/zhenning/medips/gm_5ng_rep4.bam")

gm_5ng=lapply(X=bamfile_gm_5ng, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)

save(gm_5ng,file="/work/blueridge/yanzhu/zhenning/medips/gm_5ng.RData")

bamfile_gm_10ng=c("/work/blueridge/yanzhu/zhenning/medips/gm_10ng_rep1.bam","/work/blueridge/yanzhu/zhenning/medips/gm_10ng_rep2.bam","/work/blueridge/yanzhu/zhenning/medips/gm_10ng_rep3.bam","/work/blueridge/yanzhu/zhenning/medips/gm_10ng_rep4.bam")

gm_10ng=lapply(X=bamfile_gm_10ng, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)

save(gm_10ng,file="/work/blueridge/yanzhu/zhenning/medips/gm_10ng.RData")

bamfile_gm_100ng=c("/work/blueridge/yanzhu/zhenning/medips/gm_100ng_rep1.bam","/work/blueridge/yanzhu/zhenning/medips/gm_100ng_rep2.bam","/work/blueridge/yanzhu/zhenning/medips/gm_100ng_rep3.bam")

gm_100ng=lapply(X=bamfile_gm_100ng, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)

save(gm_100ng,file="/work/blueridge/yanzhu/zhenning/medips/gm_100ng.RData")

bamfile_gm_0.5ng=c("/work/blueridge/yanzhu/zhenning/medips/gm_0.5ng_rep1.bam","/work/blueridge/yanzhu/zhenning/medips/gm_0.5ng_rep2.bam","/work/blueridge/yanzhu/zhenning/medips/gm_0.5ng_rep3.bam")

gm_0.5ng=lapply(X=bamfile_gm_0.5ng, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)
```

```
save(gm_0.5ng,file="/work/blueridge/yanzhu/zhenning/medips/gm_0.5ng.RData")

bamfile_gm_1ng=c("/work/blueridge/yanzhu/zhenning/medips/gm_1ng_rep1.bam","/work/bluer
idge/yanzhu/zhenning/medips/gm_1ng_rep2.bam","/work/blueridge/yanzhu/zhenning/medips/g
m_1ng_rep3.bam")

gm_1ng=lapply(X=bamfile_gm_1ng, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)

save(gm_1ng,file="/work/blueridge/yanzhu/zhenning/medips/gm_1ng.RData")

bamfile_gm_5ng_input="/work/blueridge/yanzhu/zhenning/medips/gm_5ng_input.bam"

gm_5ng_input=MeDIPS.createSet(file = bamfile_gm_5ng_input, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)

save(gm_5ng_input,file="/work/blueridge/yanzhu/zhenning/medips/gm_5ng_input.RData")




bamfile_hb_5ng_hmedip=c("/work/blueridge/yanzhu/zhenning/medips/hb_5ng_hmedip_rep1.ba
m","/work/blueridge/yanzhu/zhenning/medips/hb_5ng_hmedip_rep2.bam")

hb_5ng_hmedip=lapply(X=bamfile_hb_5ng_hmedip, FUN=MeDIPS.createSet,
BSgenome = BSgenome, extend = extend, shift = shift, uniq =
uniq,window_size = ws)

save(hb_5ng_hmedip,file="/work/blueridge/yanzhu/zhenning/medips/hb_5ng_hmedip.RData")

bamfile_hb_50ng_hmedip=c("/work/blueridge/yanzhu/zhenning/medips/hb_50ng_hmedip_rep1.
bam","/work/blueridge/yanzhu/zhenning/medips/hb_50ng_hmedip_rep2.bam","/work/blueridge/
yanzhu/zhenning/medips/hb_50ng_hmedip_rep3.bam")

hb_50ng_hmedip=lapply(X=bamfile_hb_50ng_hmedip, FUN=MeDIPS.createSet,
BSgenome = BSgenome, extend = extend, shift = shift, uniq =
uniq,window_size = ws)

save(hb_50ng_hmedip,file="/work/blueridge/yanzhu/zhenning/medips/hb_50ng_hmedip.RData")

bamfile_hb_5ng_hmedip_inupt="/work/blueridge/yanzhu/zhenning/medips/hb_5ng_hmedip_inu
pt.bam"
```

```
hb_5ng_hmedip_inupt=MeDIPS.createSet(file =
bamfile_hb_5ng_hmedip_inupt, BSgenome = BSgenome, extend = extend,
shift = shift, uniq = uniq,window_size = ws)

save(hb_5ng_hmedip_inupt,file="/work/blueridge/yanzhu/zhenning/medips/hb_5ng_hmedip_inupt.RData")




CS = MeDIPS.couplingVector(pattern = "CG", refObj = gm_5ng[[1]])
save(CS,file="/work/blueridge/yanzhu/zhenning/medips/CS.RData")
cor.matrix = MeDIPS.correlation(MSets =
c(gm_5ng,gm_10ng,gm_100ng,gm_0.5ng,gm_1ng,gm_5ng_input,hb_5ng_hmedip,hb_50ng_hmedip,hb_5ng_hmedip_inupt),
plot = T, method = "pearson")
save(cor.matrix,file="/work/blueridge/yanzhu/zhenning/medips/cor.matrix.RData")


load("/work/blueridge/yanzhu/zhenning/medips/gm_5ng.RData")
load("/work/blueridge/yanzhu/zhenning/medips/gm_10ng.RData")
load("/work/blueridge/yanzhu/zhenning/medips/gm_100ng.RData")
load("/work/blueridge/yanzhu/zhenning/medips/gm_0.5ng.RData")
load("/work/blueridge/yanzhu/zhenning/medips/gm_1ng.RData")
load("/work/blueridge/yanzhu/zhenning/medips/gm_5ng_input.RData")


load("/work/blueridge/yanzhu/zhenning/medips/hb_5ng_hmedip.RData")
load("/work/blueridge/yanzhu/zhenning/medips/hb_50ng_hmedip.RData")
load("/work/blueridge/yanzhu/zhenning/medips/hb_5ng_hmedip_inupt.RData")


load("/work/blueridge/yanzhu/zhenning/medips/CS.RData")
load("/work/blueridge/yanzhu/zhenning/medips/cor.matrix.RData")




library(BSgenome.Mmusculus.UCSC.mm10)
BSgenome="BSgenome.Mmusculus.UCSC.mm10"
uniq=1
extend=300
shift=0
ws=250
```

```
load("mouse_6w.RData")
load("mouse_16w.RData")
load("mouse_23w.RData")
cor.matrix = MeDIPS.correlation(MSets = c(mouse_6w, mouse_16w, mouse_23w), plot = F,
method = "pearson")
save(cor.matrix,file="cor.matrix.RData")

levelplot(full, main="Genome-wide correlations", xlab="", ylab="", col.regions=rgb.palette(120),
cuts=100, at=seq(0.5,1,0.01), font=2,
scales=list(font=2,x=list(labels=c("6wk_rep1","6wk_rep2","16wk_rep1","16wk_rep2","23wk_re
p1","23wk_rep2"),rot=45,font=2),y=list(labels=c( "6wk_rep1","6wk_rep2","16wk_rep1","16wk
_rep2","23wk_rep1","23wk_rep2" ),font=2)))
```

Correlation Heatmap

```
upper=cor.matrix
upper[lower.tri(upper)]<-0
upper_no_diag=cor.matrix
upper_no_diag[lower.tri(upper,diag=T)]<-0
lower=t(upper_no_diag)
full=upper+lower

full=read.table("pearson_cor_heatmap_promoter")
full=read.table("CorMatrix")
full<-data.matrix(full)

library(lattice)
rgb.palette <- colorRampPalette(c("blue", "yellow"), space = "rgb")
levelplot(full, main="Genome-wide correlations", xlab="", ylab="", col.regions=rgb.palette(120),
cuts=100, at=seq(0,1,0.01))


levelplot(full, main="Genome-wide correlations", xlab="", ylab="", col.regions=rgb.palette(120),
cuts=100, at=seq(0.49,1,0.01), font=2, scales=list(font=2,x=list(labels=c("0.5 ng_rep1","0.5
ng_rep2","5 ng_rep1","5 ng_rep2","10 ng_rep1","10 ng_rep2","100
ng_ref"),rot=45,font=2),y=list(labels=c("0.5 ng_rep1","0.5 ng_rep2","5 ng_rep1","5
ng_rep2","10 ng_rep1","10 ng_rep2","100 ng_ref"),font=2)))

levelplot(full, main="Genome-wide correlations", xlab="", ylab="", col.regions=rgb.palette(120),
cuts=100, at=seq(0.5,1,0.01), font=2,
scales=list(font=2,x=list(labels=c("gm_pol2_ENCODE","1000K_1F_S_rep1","1000K_1F_S_re
```

```
p2", "100K_1F_M_rep1", "100K_1F_M_rep2",
"100K_2F_M_rep1","100K_2F_M_rep2","50K_1F_M_rep1","50K_1F_M_rep2","50K_2F_M_r
ep1","50K_2F_M_rep2"),rot=45,font=2),y=list(labels=c
( "gm_pol2_ENCODE","1000K_1F_S_rep1","1000K_1F_S_rep2", "100K_1F_M_rep1",
"100K_1F_M_rep2",
"100K_2F_M_rep1","100K_2F_M_rep2","50K_1F_M_rep1","50K_1F_M_rep2","50K_2F_M_r
ep1","50K_2F_M_rep2" ) ,font=2)))
```

MA-Plot

```
cancer_vs_normal=MeDIPS.meth(MSet1 = mouse_cancer, MSet2 = mouse_normal, CSet =
CS_mouse, p.adj = "bonferroni", diff.method = "edgeR", MeDIP = T, CNV = F, minRowSum =
10)

save(cancer_vs_normal,file="/work/blueridge/yanzhu/zhenning/medips/cancer_vs_normal.RDat
a")

dm1=cancer_vs_normal$edgeR.adj.p.value<0.1
dm2=cancer_vs_normal$edgeR.p.value<0.01

pdf("MA_Plot.pdf")
smoothScatter(cancer_vs_normal$edgeR.logCPM,cancer_vs_normal$edgeR.logFC,ylim=c(-
4,4),xlim=c(-3,4), xlab="avg log methylation", ylab="methylation logFC", main="MA Plot")
abline(h=-4:4, col="grey",lty=2)
abline(v=c(-2,0,2,4), col="grey",lty=2)

points(cancer_vs_normal$edgeR.logCPM[dm2],cancer_vs_normal$edgeR.logFC[dm2],
col="orange",pch=".")
points(cancer_vs_normal$edgeR.logCPM[dm1],cancer_vs_normal$edgeR.logFC[dm1],
col="red",pch=4)

dev.off()
```

Bisulfite Validation

```
bsv=read.table("BisSeq_regions.txt",sep="\t", header=T)

columns=c("CF","edgeR.logFC","edgeR.p.value", "mouse_cancer_rep1.bam.rms",
"mouse_normal_rep1.bam.rms", "mouse_cancer_rep1.bam.rpkm",
"mouse_normal_rep1.bam.rpkm", "mouse_cancer_rep1.bam.counts",
"mouse_normal_rep1.bam.counts")
```

```
mr.roi=MeDIPS.selectROIs(results=cancer_vs_normal , rois=bsv[,c(3:5,1)],
columns=columns,summarize="avg")

if(all(mr.roi$ROI==bsv[,1])){
mr.roi=cbind(mr.roi, bsv)
}

BS=rep(c("BS_ad5","BS_n5"),2)
sample=rep(c("Tumor","Normal"),2)
method=c(rep("rms",2),rep("rkpm",2),rep("count",2))

pdf("BS_validation.pdf")

par(mfrow=c(2,2))
color=rep("black", dim(mr.roi)[1])
color[mr.roi$CF>= quantile(mr.roi$CF,.75)]="red"
color[mr.roi$CF<= quantile(mr.roi$CF,.25)]="blue"

for(i in 1:4){
  val=mr.roi[,c(columns[i+3],BS[i]) ]
  val=val[!is.na(rowSums(val)),]
  co=cor(val[,1],val[,2])
  plot(val[,1],val[,2], sub=paste("correlation:", round(co,3)), main=sample[i],
ylim=c(0,100),ylab=paste(sample[i], "BiSulfit [%]"), xlab=paste(sample[i], "MeDIP-Seq",
method[i]), col=color)
  abline(lm(val[,2]~val[,1]))
}

legend("bottomright", legend=c("lower quantile CF", "25%-75% CF", "upper quantile CF"),
fill=c("blue", "black", "red"))
dev.off()
```

Comparing rrbs %methylation to MeDIP rkpm and rms at promoter region

```
library(MeDIPS)
library(BSgenome.Hsapiens.UCSC.hg19)
BSgenome="BSgenome.Hsapiens.UCSC.hg19"
uniq=TRUE
extend=300
shift=0
ws=250

bamfile_b2_9="b2_9.bam"
b2_9=MeDIPS.createSet(file = bamfile_b2_9, BSgenome = BSgenome, extend = extend, shift =
shift, uniq = uniq,window_size = ws)
```

```
CS = MeDIPS.couplingVector(pattern = "CG", refObj = b2_9)
result=MeDIPS.meth(MSet1 = b2_9, CSet = CS_mouse, MeDIP = T)

bsv=read.table("percmeth_rrbs_ref_prom_ID.txt", header=T)
columns=c("CF","b2_9.bam.rms","b2_9.bam.rpkm","b2_9.bam.counts")
mr.roi=MeDIPS.selectROIs(results=result, rois=bsv[,c(2:4,1)],
columns=columns,summarize="avg")
if(all(mr.roi$ROI==bsv[,1])){mr.roi=cbind(mr.roi, bsv)}

BS=rep(c("BS"),2)
sample=rep(c("Cao_100ng"),2)
method=c(rep("rms",1),rep("rkpm",1),rep("count",1))

pdf("BS_validation.pdf")

par(mfrow=c(2,1))
color=rep("black", dim(mr.roi)[1])
color[mr.roi$CF>= quantile(mr.roi$CF,.75)]="red"
color[mr.roi$CF<= quantile(mr.roi$CF,.25)]="blue"

for(i in 1:2){
  val=mr.roi[,c(columns[i+1],BS[i]) ]
  val=val[!is.na(rowSums(val)),]
  co=cor(val[,1],val[,2])
  plot(val[,1],val[,2], sub=paste("correlation:", round(co,3)), main=sample[i],
ylim=c(0,100),ylab=paste(sample[i], "BiSulfit [%]"), xlab=paste(sample[i], "MeDIP-Seq",
method[i]), col=color)
  abline(lm(val[,2]~val[,1]))
}
legend("bottomright", legend=c("lower quantile CF", "25%-75% CF", "upper quantile CF"),
fill=c("blue", "black", "red"))
dev.off()
```

mouse data

```
library(MeDIPS)
library(BSgenome.Mmusculus.UCSC.mm10)
BSgenome="BSgenome.Mmusculus.UCSC.mm10"
uniq=1
extend=300
shift=0
ws=250
```

```
bamfile_mouse_6w=c("6w_1.bam","6w_2.bam")
mouse_6w=lapply(X=bamfile_mouse_6w, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)
save(mouse_6w,file="mouse_6w.RData")

bamfile_mouse_16w=c("16w_1.bam","16w_2.bam")
mouse_16w=lapply(X=bamfile_mouse_16w, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)
save(mouse_16w,file="mouse_16w.RData")

bamfile_mouse_23w=c("23w_1.bam","23w_2.bam")
mouse_23w=lapply(X=bamfile_mouse_23w, FUN=MeDIPS.createSet, BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq,window_size =
ws)
save(mouse_23w,file="mouse_23w.RData")

CS = MeDIPS.couplingVector(pattern = "CG", refObj = mouse_23w)
save(CS,file="CS.RData")

-----------------------------------------------------------------------------------------------------------------
-----------------------



mouse_16WT_vs_mouse_6WN_FDR=MeDIPS.meth(MSet1 = mouse_16w, MSet2 =
mouse_6w,CSet = CS, p.adj = "fdr", diff.method = "edgeR",
MeDIP = T, CNV = F,minRowSum = 10)
save(mouse_16WT_vs_mouse_6WN_FDR,file="mouse_16WT_vs_mouse_6WN_FDR.RData")

dm1=mouse_16WT_vs_mouse_6WN_FDR$edgeR.adj.p.value<0.1
dm2=mouse_16WT_vs_mouse_6WN_FDR$edgeR.p.value<0.01

write.table(mouse_16WT_vs_mouse_6WN_FDR[which(dm1),],
file="mouse_16WT_vs_mouse_6WN_FDR.tsv", sep="\t", quote=F,
row.names=F, col.names=T)

write.table(mouse_16WT_vs_mouse_6WN_FDR[which(dm2),],
file="mouse_16WT_vs_mouse_6WN_Pvalue.tsv", sep="\t", quote=F,
row.names=F, col.names=T)
```

```
mouse_23WT_vs_mouse_6WN_FDR=MeDIPS.meth(MSet1 = mouse_23w, MSet2 =
mouse_6w,CSet = CS, p.adj = "fdr", diff.method = "edgeR",
MeDIP = T, CNV = F,minRowSum = 10)
save(mouse_23WT_vs_mouse_6WN_FDR,file="mouse_23WT_vs_mouse_6WN_FDR.RData")

dm1=mouse_23WT_vs_mouse_6WN_FDR$edgeR.adj.p.value<0.1
dm2=mouse_23WT_vs_mouse_6WN_FDR$edgeR.p.value<0.01

write.table(mouse_23WT_vs_mouse_6WN_FDR[which(dm1),],
file="mouse_23WT_vs_mouse_6WN_FDR.tsv", sep="\t", quote=F,
row.names=F, col.names=T)

write.table(mouse_23WT_vs_mouse_6WN_FDR[which(dm2),],
file="mouse_23WT_vs_mouse_6WN_Pvalue.tsv", sep="\t", quote=F,
row.names=F, col.names=T)




mouse_23WT_vs_mouse_16WT_FDR=MeDIPS.meth(MSet1 = mouse_23w, MSet2 =
mouse_16w,CSet = CS, p.adj = "fdr", diff.method = "edgeR",
MeDIP = T, CNV = F,minRowSum = 10)
save(mouse_23WT_vs_mouse_16WT_FDR,file="mouse_23WT_vs_mouse_16WT_FDR.RData
")

dm1=mouse_23WT_vs_mouse_16WT_FDR$edgeR.adj.p.value<0.1
dm2=mouse_23WT_vs_mouse_16WT_FDR$edgeR.p.value<0.01

write.table(mouse_23WT_vs_mouse_16WT_FDR[which(dm1),],
file="mouse_23WT_vs_mouse_16WT_FDR.tsv", sep="\t", quote=F,
row.names=F, col.names=T)

write.table(mouse_23WT_vs_mouse_16WT_FDR[which(dm2),],
file="mouse_23WT_vs_mouse_16WT_Pvalue.tsv", sep="\t", quote=F,
row.names=F, col.names=T)
```

```
mouse_16WT_vs_mouse_6WN_Pvalue.s=MeDIPS.selectSig(results
=mouse_16WT_vs_mouse_6WN_FDR,p.value = 0.01, adj = F, ratio = NULL,
bg.counts = NULL, CNV = F)

mouse_16WT_vs_mouse_6WN_Pvalue.s.m=MeDIPS.mergeFrames(frames=mouse_16WT_vs_
mouse_6WN_Pvalue.s,
distance = 1)

mouse_16WT_vs_mouse_6WN_Pvalue.s.gain=mouse_16WT_vs_mouse_6WN_Pvalue.s[which
(mouse_16WT_vs_mouse_6WN_Pvalue.s[,
grep("logFC", colnames(mouse_16WT_vs_mouse_6WN_Pvalue.s))] > 0),]

mouse_16WT_vs_mouse_6WN_Pvalue.s.loss=mouse_16WT_vs_mouse_6WN_Pvalue.s[which(
mouse_16WT_vs_mouse_6WN_Pvalue.s[,
grep("logFC", colnames(mouse_16WT_vs_mouse_6WN_Pvalue.s))] < 0),]

mouse_16WT_vs_mouse_6WN_Pvalue.s.gain.m=MeDIPS.mergeFrames(frames=mouse_16WT
_vs_mouse_6WN_Pvalue.s.gain,
distance = 1)

mouse_16WT_vs_mouse_6WN_Pvalue.s.loss.m=MeDIPS.mergeFrames(frames=mouse_16WT
_vs_mouse_6WN_Pvalue.s.loss,
distance = 1)



write.table(mouse_16WT_vs_mouse_6WN_Pvalue.s.m, file =
"mouse_16WT_vs_mouse_6WN_Pvalue.s.m.bed", quote =F, sep="\t", row.names=F,
col.names=F)

write.table(mouse_16WT_vs_mouse_6WN_Pvalue.s.gain.m, file =
"mouse_16WT_vs_mouse_6WN_Pvalue.s.gain.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

write.table(mouse_16WT_vs_mouse_6WN_Pvalue.s.loss.m, file =
"mouse_16WT_vs_mouse_6WN_Pvalue.s.loss.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)



mouse_16WT_vs_mouse_6WN_FDR.s=MeDIPS.selectSig(results
=mouse_16WT_vs_mouse_6WN_FDR,p.value = 0.1, adj = T, ratio = NULL,
bg.counts = NULL, CNV = F)
```

```
mouse_16WT_vs_mouse_6WN_FDR.s.m=MeDIPS.mergeFrames(frames=mouse_16WT_vs_mouse_6WN_FDR.s,
distance = 1)

mouse_16WT_vs_mouse_6WN_FDR.s.gain=mouse_16WT_vs_mouse_6WN_FDR.s[which(mouse_16WT_vs_mouse_6WN_FDR.s[,
grep("logFC", colnames(mouse_16WT_vs_mouse_6WN_FDR.s))] > 0),]

mouse_16WT_vs_mouse_6WN_FDR.s.loss=mouse_16WT_vs_mouse_6WN_FDR.s[which(mouse_16WT_vs_mouse_6WN_FDR.s[,
grep("logFC", colnames(mouse_16WT_vs_mouse_6WN_FDR.s))] < 0),]

mouse_16WT_vs_mouse_6WN_FDR.s.gain.m=MeDIPS.mergeFrames(frames=mouse_16WT_vs_mouse_6WN_FDR.s.gain,
distance = 1)

mouse_16WT_vs_mouse_6WN_FDR.s.loss.m=MeDIPS.mergeFrames(frames=mouse_16WT_vs_mouse_6WN_FDR.s.loss,
distance = 1)



write.table(mouse_16WT_vs_mouse_6WN_FDR.s.m, file =
"mouse_16WT_vs_mouse_6WN_FDR.s.m.bed", quote =F, sep="\t", row.names=F,
col.names=F)

write.table(mouse_16WT_vs_mouse_6WN_FDR.s.gain.m, file =
"mouse_16WT_vs_mouse_6WN_FDR.s.gain.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

write.table(mouse_16WT_vs_mouse_6WN_FDR.s.loss.m, file =
"mouse_16WT_vs_mouse_6WN_FDR.s.loss.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)


-------------------------------------------------------------------------------------------------------------------------

mouse_23WT_vs_mouse_6WN_Pvalue.s=MeDIPS.selectSig(results
=mouse_23WT_vs_mouse_6WN_FDR,p.value = 0.01, adj = F, ratio = NULL,
bg.counts = NULL, CNV = F)

mouse_23WT_vs_mouse_6WN_Pvalue.s.m=MeDIPS.mergeFrames(frames=mouse_23WT_vs_mouse_6WN_Pvalue.s,
distance = 1)
```

```
mouse_23WT_vs_mouse_6WN_Pvalue.s.gain=mouse_23WT_vs_mouse_6WN_Pvalue.s[which
(mouse_23WT_vs_mouse_6WN_Pvalue.s[,
grep("logFC", colnames(mouse_23WT_vs_mouse_6WN_Pvalue.s))] > 0),]

mouse_23WT_vs_mouse_6WN_Pvalue.s.loss=mouse_23WT_vs_mouse_6WN_Pvalue.s[which(
mouse_23WT_vs_mouse_6WN_Pvalue.s[,
grep("logFC", colnames(mouse_23WT_vs_mouse_6WN_Pvalue.s))] < 0),]

mouse_23WT_vs_mouse_6WN_Pvalue.s.gain.m=MeDIPS.mergeFrames(frames=mouse_23WT
_vs_mouse_6WN_Pvalue.s.gain,
distance = 1)

mouse_23WT_vs_mouse_6WN_Pvalue.s.loss.m=MeDIPS.mergeFrames(frames=mouse_23WT
_vs_mouse_6WN_Pvalue.s.loss,
distance = 1)



write.table(mouse_23WT_vs_mouse_6WN_Pvalue.s.m, file =
"mouse_23WT_vs_mouse_6WN_Pvalue.s.m.bed", quote =F, sep="\t", row.names=F,
col.names=F)

write.table(mouse_23WT_vs_mouse_6WN_Pvalue.s.gain.m, file =
"mouse_23WT_vs_mouse_6WN_Pvalue.s.gain.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

write.table(mouse_23WT_vs_mouse_6WN_Pvalue.s.loss.m, file =
"mouse_23WT_vs_mouse_6WN_Pvalue.s.loss.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)




mouse_23WT_vs_mouse_6WN_FDR.s=MeDIPS.selectSig(results
=mouse_23WT_vs_mouse_6WN_FDR,p.value = 0.1, adj = T, ratio = NULL,
bg.counts = NULL, CNV = F)

mouse_23WT_vs_mouse_6WN_FDR.s.m=MeDIPS.mergeFrames(frames=mouse_23WT_vs_m
ouse_6WN_FDR.s,
distance = 1)

mouse_23WT_vs_mouse_6WN_FDR.s.gain=mouse_23WT_vs_mouse_6WN_FDR.s[which(mo
use_23WT_vs_mouse_6WN_FDR.s[,
grep("logFC", colnames(mouse_23WT_vs_mouse_6WN_FDR.s))] > 0),]
```

```
mouse_23WT_vs_mouse_6WN_FDR.s.loss=mouse_23WT_vs_mouse_6WN_FDR.s[which(mo
use_23WT_vs_mouse_6WN_FDR.s[,
grep("logFC", colnames(mouse_23WT_vs_mouse_6WN_FDR.s))] < 0),]

mouse_23WT_vs_mouse_6WN_FDR.s.gain.m=MeDIPS.mergeFrames(frames=mouse_23WT_v
s_mouse_6WN_FDR.s.gain,
distance = 1)

mouse_23WT_vs_mouse_6WN_FDR.s.loss.m=MeDIPS.mergeFrames(frames=mouse_23WT_v
s_mouse_6WN_FDR.s.loss,
distance = 1)



write.table(mouse_23WT_vs_mouse_6WN_FDR.s.m, file =
"mouse_23WT_vs_mouse_6WN_FDR.s.m.bed", quote =F, sep="\t", row.names=F,
col.names=F)

write.table(mouse_23WT_vs_mouse_6WN_FDR.s.gain.m, file =
"mouse_23WT_vs_mouse_6WN_FDR.s.gain.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

write.table(mouse_23WT_vs_mouse_6WN_FDR.s.loss.m, file =
"mouse_23WT_vs_mouse_6WN_FDR.s.loss.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

----------------------------------------------------------------------------------------------------------
----------

mouse_23WT_vs_mouse_16WT_Pvalue.s=MeDIPS.selectSig(results
=mouse_23WT_vs_mouse_16WT_FDR,p.value = 0.01, adj = F, ratio = NULL,
bg.counts = NULL, CNV = F)

mouse_23WT_vs_mouse_16WT_Pvalue.s.m=MeDIPS.mergeFrames(frames=mouse_23WT_vs
_mouse_16WT_Pvalue.s,
distance = 1)

mouse_23WT_vs_mouse_16WT_Pvalue.s.gain=mouse_23WT_vs_mouse_16WT_Pvalue.s[whic
h(mouse_23WT_vs_mouse_16WT_Pvalue.s[,
grep("logFC", colnames(mouse_23WT_vs_mouse_16WT_Pvalue.s))] > 0),]

mouse_23WT_vs_mouse_16WT_Pvalue.s.loss=mouse_23WT_vs_mouse_16WT_Pvalue.s[whic
h(mouse_23WT_vs_mouse_16WT_Pvalue.s[,
grep("logFC", colnames(mouse_23WT_vs_mouse_16WT_Pvalue.s))] < 0),]

mouse_23WT_vs_mouse_16WT_Pvalue.s.gain.m=MeDIPS.mergeFrames(frames=mouse_23W
```

```
T_vs_mouse_16WT_Pvalue.s.gain,
distance = 1)

mouse_23WT_vs_mouse_16WT_Pvalue.s.loss.m=MeDIPS.mergeFrames(frames=mouse_23WT
_vs_mouse_16WT_Pvalue.s.loss,
distance = 1)



write.table(mouse_23WT_vs_mouse_16WT_Pvalue.s.m, file =
"mouse_23WT_vs_mouse_16WT_Pvalue.s.m.bed", quote =F, sep="\t", row.names=F,
col.names=F)

write.table(mouse_23WT_vs_mouse_16WT_Pvalue.s.gain.m, file =
"mouse_23WT_vs_mouse_16WT_Pvalue.s.gain.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

write.table(mouse_23WT_vs_mouse_16WT_Pvalue.s.loss.m, file =
"mouse_23WT_vs_mouse_16WT_Pvalue.s.loss.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)



mouse_23WT_vs_mouse_16WT_FDR.s=MeDIPS.selectSig(results
=mouse_23WT_vs_mouse_16WT_FDR,p.value = 0.1, adj = T, ratio = NULL,
bg.counts = NULL, CNV = F)

mouse_23WT_vs_mouse_16WT_FDR.s.m=MeDIPS.mergeFrames(frames=mouse_23WT_vs_
mouse_16WT_FDR.s,
distance = 1)

mouse_23WT_vs_mouse_16WT_FDR.s.gain=mouse_23WT_vs_mouse_16WT_FDR.s[which(
mouse_23WT_vs_mouse_16WT_FDR.s[,
grep("logFC", colnames(mouse_23WT_vs_mouse_16WT_FDR.s))] > 0),]

mouse_23WT_vs_mouse_16WT_FDR.s.loss=mouse_23WT_vs_mouse_16WT_FDR.s[which(m
ouse_23WT_vs_mouse_16WT_FDR.s[,
grep("logFC", colnames(mouse_23WT_vs_mouse_16WT_FDR.s))] < 0),]

mouse_23WT_vs_mouse_16WT_FDR.s.gain.m=MeDIPS.mergeFrames(frames=mouse_23WT_
vs_mouse_16WT_FDR.s.gain,
distance = 1)

mouse_23WT_vs_mouse_16WT_FDR.s.loss.m=MeDIPS.mergeFrames(frames=mouse_23WT_
vs_mouse_16WT_FDR.s.loss,
```

```
distance = 1)


write.table(mouse_23WT_vs_mouse_16WT_FDR.s.m, file =
"mouse_23WT_vs_mouse_16WT_FDR.s.m.bed", quote =F, sep="\t", row.names=F,
col.names=F)

write.table(mouse_23WT_vs_mouse_16WT_FDR.s.gain.m, file =
"mouse_23WT_vs_mouse_16WT_FDR.s.gain.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

write.table(mouse_23WT_vs_mouse_16WT_FDR.s.loss.m, file =
"mouse_23WT_vs_mouse_16WT_FDR.s.loss.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

-----------------------------------------------------------------------------------------------------------------
----------------------


intersectBed -wa -wb -a mouse_23WT_vs_mouse_6WN_Pvalue.s.m.bed -b
mouse_16WT_vs_mouse_6WN_Pvalue.s.m.bed > 23_6vs16_6_Pvalue.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_6WN_Pvalue.s.gain.m.bed -b
mouse_16WT_vs_mouse_6WN_Pvalue.s.gain.m.bed >
23_6vs16_6_Pvalue.hyper.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_6WN_Pvalue.s.loss.m.bed -b
mouse_16WT_vs_mouse_6WN_Pvalue.s.loss.m.bed >
23_6vs16_6_Pvalue.hypo.bed


intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_Pvalue.s.m.bed -b
mouse_16WT_vs_mouse_6WN_Pvalue.s.m.bed > 23_16vs16_6_Pvalue.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_Pvalue.s.gain.m.bed
-b mouse_16WT_vs_mouse_6WN_Pvalue.s.gain.m.bed >
23_16vs16_6_Pvalue.hyper.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_Pvalue.s.loss.m.bed
-b mouse_16WT_vs_mouse_6WN_Pvalue.s.loss.m.bed >
23_16vs16_6_Pvalue.hypo.bed
```

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_Pvalue.s.m.bed -b mouse_23WT_vs_mouse_6WN_Pvalue.s.m.bed > 23_6vs23_16_Pvalue.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_Pvalue.s.gain.m.bed -b mouse_23WT_vs_mouse_6WN_Pvalue.s.gain.m.bed > 23_6vs23_16_Pvalue.hyper.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_Pvalue.s.loss.m.bed -b mouse_23WT_vs_mouse_6WN_Pvalue.s.loss.m.bed > 23_6vs23_16_Pvalue.hypo.bed

---

intersectBed -wa -wb -a mouse_23WT_vs_mouse_6WN_FDR.s.m.bed -b mouse_16WT_vs_mouse_6WN_FDR.s.m.bed > 23_6vs16_6_FDR.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_6WN_FDR.s.gain.m.bed -b mouse_16WT_vs_mouse_6WN_FDR.s.gain.m.bed > 23_6vs16_6_FDR.hyper.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_6WN_FDR.s.loss.m.bed -b mouse_16WT_vs_mouse_6WN_FDR.s.loss.m.bed > 23_6vs16_6_FDR.hypo.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_FDR.s.m.bed -b mouse_16WT_vs_mouse_6WN_FDR.s.m.bed > 23_16vs16_6_FDR.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_FDR.s.gain.m.bed -b mouse_16WT_vs_mouse_6WN_FDR.s.gain.m.bed > 23_16vs16_6_FDR.hyper.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_FDR.s.loss.m.bed -b mouse_16WT_vs_mouse_6WN_FDR.s.loss.m.bed > 23_16vs16_6_FDR.hypo.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_FDR.s.m.bed -b mouse_23WT_vs_mouse_6WN_FDR.s.m.bed > 23_6vs23_16_FDR.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_FDR.s.gain.m.bed -b

```
mouse_23WT_vs_mouse_6WN_FDR.s.gain.m.bed > 23_6vs23_16_FDR.hyper.bed

intersectBed -wa -wb -a mouse_23WT_vs_mouse_16WT_FDR.s.loss.m.bed -b
mouse_23WT_vs_mouse_6WN_FDR.s.loss.m.bed > 23_6vs23_16_FDR.hypo.bed


mouse_16WT_vs_mouse_6WN_FDR=MeDIPS.meth(MSet1 = mouse_16WT, MSet2 =
mouse_6WN,CSet = CS, p.adj = "fdr", diff.method = "edgeR",
MeDIP = T, CNV = F,minRowSum = 10)
save(mouse_16WT_vs_mouse_6WN_FDR,file="mouse_16WT_vs_mouse_6WN_FDR.RData")
dm1=mouse_16WT_vs_mouse_6WN_FDR$edgeR.adj.p.value<0.1
dm2=mouse_16WT_vs_mouse_6WN_FDR$edgeR.p.value<0.01

write.table(mouse_16WT_vs_mouse_6WN_FDR[which(dm1),],
file="mouse_16WT_vs_mouse_6WN_FDR.tsv", sep="\t", quote=F,
row.names=F, col.names=T)

write.table(mouse_16WT_vs_mouse_6WN_FDR[which(dm2),],
file="mouse_16WT_vs_mouse_6WN_Pvalue.tsv", sep="\t", quote=F,
row.names=F, col.names=T)


pdf("MA_Plot_16WT_vs_6WN_FDR.pdf")
smoothScatter(mouse_16WT_vs_mouse_6WN_FDR$edgeR.logCPM,mouse_16WT_vs_mouse_
6WN_FDR$edgeR.logFC,ylim=c(-8,8),xlim=c(0,6),
xlab="avg log methylation", ylab="methylation logFC", main="MA Plot")
abline(h=-8:8, col="grey",lty=2)
abline(v=c(0,2,4,6), col="grey",lty=2)
points(mouse_16WT_vs_mouse_6WN_FDR$edgeR.logCPM[dm2],mouse_16WT_vs_mouse_6
WN_FDR$edgeR.logFC[dm2],
col="orange",pch=".")
points(mouse_16WT_vs_mouse_6WN_FDR$edgeR.logCPM[dm1],mouse_16WT_vs_mouse_6
WN_FDR$edgeR.logFC[dm1],
col="red",pch=4)
dev.off()

sum(mouse_16WT_vs_mouse_6WN_FDR$edgeR.p.value<0.01,na.rm=TRUE)
sum(mouse_16WT_vs_mouse_6WN_FDR$edgeR.adj.p.value<0.1,na.rm=TRUE)




m_cgi=read.table("/home/yanzhu/blueridge/model-based-cpg-islands/mm10/model-based-cpg-
islands-mm10.txt",
header=T)
m_exon=read.table("/home/yanzhu/blueridge/model-based-cpg-
islands/mm10/RefSeq_mm10_known_genes_exons.txt")
```

```
m_intron=read.table("/home/yanzhu/blueridge/model-based-cpg-
islands/mm10/RefSeq_mm10_known_genes_introns.txt")
m_prom=read.table("/home/yanzhu/blueridge/model-based-cpg-
islands/mm10/refGene_Prom_2000_500_mm10.txt")
m_exon$nr=1:dim(m_exon)[1]
m_prom$nr=1:dim(m_prom)[1]
m_intron$nr=1:dim(m_intron)[1]

results_CGI=MeDIPS.selectROIs(results=mouse_16WT_vs_mouse_6WN_FDR, rois=m_cgi)
results_exon=MeDIPS.selectROIs(results=mouse_16WT_vs_mouse_6WN_FDR, rois=m_exon)
results_intron=MeDIPS.selectROIs(results=mouse_16WT_vs_mouse_6WN_FDR,
rois=m_intron)
results_prom=MeDIPS.selectROIs(results=mouse_16WT_vs_mouse_6WN_FDR, rois=m_prom)
results_CGI_prom=MeDIPS.selectROIs(results=results_CGI, rois=m_prom)

reg=c("all regions", "Introns", "Exons","Promoter" , "CpG Islands" ,
"CpG Island \nPromoter")
DMR_rel=matrix(NA, 6,2, dimnames=list(reg, c("hypomethylated",
"hypermethylated")))

f=!is.na(mouse_16WT_vs_mouse_6WN_FDR$edgeR.p.value)
DMR_rel[1,1]=sum(mouse_16WT_vs_mouse_6WN_FDR$edgeR.p.value[f]<.01 &
mouse_16WT_vs_mouse_6WN_FDR$edgeR.logFC[f]<0)/dim(mouse_16WT_vs_mouse_6WN_
FDR)[1]
DMR_rel[1,2]=sum(mouse_16WT_vs_mouse_6WN_FDR$edgeR.p.value[f]<.01 &
mouse_16WT_vs_mouse_6WN_FDR$edgeR.logFC[f]>0)/dim(mouse_16WT_vs_mouse_6WN_
FDR)[1]

f=!is.na(results_intron$edgeR.p.value)
DMR_rel[2,1]=sum(results_intron$edgeR.p.value[f]<.01
&results_intron$edgeR.logFC[f]<0)/dim(results_intron)[1]
DMR_rel[2,2]=sum(results_intron$edgeR.p.value[f]<.01
&results_intron$edgeR.logFC[f]>0)/dim(results_intron)[1]

f=!is.na(results_exon$edgeR.p.value)
DMR_rel[3,1]=sum(results_exon$edgeR.p.value[f]<.01
&results_exon$edgeR.logFC[f]<0)/dim(results_exon)[1]
DMR_rel[3,2]=sum(results_exon$edgeR.p.value[f]<.01
&results_exon$edgeR.logFC[f]>0)/dim(results_exon)[1]

f=!is.na(results_prom$edgeR.p.value)
DMR_rel[4,1]=sum(results_prom$edgeR.p.value[f]<.01
&results_prom$edgeR.logFC[f]<0)/dim(results_prom)[1]
DMR_rel[4,2]=sum(results_prom$edgeR.p.value[f]<.01
&results_prom$edgeR.logFC[f]>0)/dim(results_prom)[1]
```

```
f=!is.na(results_CGI$edgeR.p.value)
DMR_rel[5,1]=sum(results_CGI$edgeR.p.value[f]<.01
&results_CGI$edgeR.logFC[f]<0)/dim(results_CGI)[1]
DMR_rel[5,2]=sum(results_CGI$edgeR.p.value[f]<.01
&results_CGI$edgeR.logFC[f]>0)/dim(results_CGI)[1]

f=!is.na(results_CGI_prom$edgeR.p.value)
DMR_rel[6,1]=sum(results_CGI_prom$edgeR.p.value[f]<.01
&results_CGI_prom$edgeR.logFC[f]<0)/dim(results_CGI_prom)[1]
DMR_rel[6,2]=sum(results_CGI_prom$edgeR.p.value[f]<.01
&results_CGI_prom$edgeR.logFC[f]>0)/dim(results_CGI_prom)[1]

pdf("Fraction_DMR_16WT_vs_6WN_FDR.pdf")
barplot(t(DMR_rel)*100,  main="Fraction of differentially methylated
regions", beside=T, ylab="Fraction of 250 pb windows [%]")
legend("topleft", c("hypomethylated", "hypermethylated"), fill=grey.colors(2))
dev.off()
```

```
mouse_16WT_vs_mouse_6WN.s=MeDIPS.selectSig(results
=mouse_16WT_vs_mouse_6WN,p.value = 0.01, adj = F, ratio = NULL,
bg.counts = NULL, CNV = F)

mouse_16WT_vs_mouse_6WN.s.m=MeDIPS.mergeFrames(frames=mouse_16WT_vs_mouse_
6WN.s,
distance = 1)

mouse_16WT_vs_mouse_6WN.s.gain=mouse_16WT_vs_mouse_6WN.s[which(mouse_16WT_
vs_mouse_6WN.s[,
grep("logFC", colnames(mouse_16WT_vs_mouse_6WN.s))] > 0),]

mouse_16WT_vs_mouse_6WN.s.loss=mouse_16WT_vs_mouse_6WN.s[which(mouse_16WT_
vs_mouse_6WN.s[,
grep("logFC", colnames(mouse_16WT_vs_mouse_6WN.s))] < 0),]

mouse_16WT_vs_mouse_6WN.s.gain.m=MeDIPS.mergeFrames(frames=mouse_16WT_vs_mo
use_6WN.s.gain,
distance = 1)

mouse_16WT_vs_mouse_6WN.s.loss.m=MeDIPS.mergeFrames(frames=mouse_16WT_vs_mo
use_6WN.s.loss,
```

```
distance = 1)




write.csv(mouse_16WT_vs_mouse_6WN.s.m, file =
"mouse_16WT_vs_mouse_6WN.s.m.csv", quote =F, row.names=F)

write.csv(mouse_16WT_vs_mouse_6WN.s.gain.m, file =
"mouse_16WT_vs_mouse_6WN.s.gain.m.csv", quote =F, row.names=F)

write.csv(mouse_16WT_vs_mouse_6WN.s.loss.m, file =
"mouse_16WT_vs_mouse_6WN.s.loss.m.csv", quote =F, row.names=F)



write.table(mouse_16WT_vs_mouse_6WN.s.m, file =
"mouse_16WT_vs_mouse_6WN.s.m.bed", quote =F, sep="\t", row.names=F,
col.names=F)

write.table(mouse_16WT_vs_mouse_6WN.s.gain.m, file =
"mouse_16WT_vs_mouse_6WN.s.gain.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

write.table(mouse_16WT_vs_mouse_6WN.s.loss.m, file =
"mouse_16WT_vs_mouse_6WN.s.loss.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)




mouse_16WT_vs_mouse_6WN_FDR.s=MeDIPS.selectSig(results
=mouse_16WT_vs_mouse_6WN_FDR,p.value = 0.1, adj = T, ratio = NULL,
bg.counts = NULL, CNV = F)

mouse_16WT_vs_mouse_6WN_FDR.s.m=MeDIPS.mergeFrames(frames=mouse_16WT_vs_m
ouse_6WN_FDR.s,
distance = 1)

mouse_16WT_vs_mouse_6WN_FDR.s.gain=mouse_16WT_vs_mouse_6WN_FDR.s[which(mo
```

```
use_16WT_vs_mouse_6WN_FDR.s[,
grep("logFC", colnames(mouse_16WT_vs_mouse_6WN_FDR.s))] > 0),]

mouse_16WT_vs_mouse_6WN_FDR.s.loss=mouse_16WT_vs_mouse_6WN_FDR.s[which(mo
use_16WT_vs_mouse_6WN_FDR.s[,
grep("logFC", colnames(mouse_16WT_vs_mouse_6WN_FDR.s))] < 0),]

mouse_16WT_vs_mouse_6WN_FDR.s.gain.m=MeDIPS.mergeFrames(frames=mouse_16WT_v
s_mouse_6WN_FDR.s.gain,
distance = 1)

mouse_16WT_vs_mouse_6WN_FDR.s.loss.m=MeDIPS.mergeFrames(frames=mouse_16WT_v
s_mouse_6WN_FDR.s.loss,
distance = 1)




write.csv(mouse_16WT_vs_mouse_6WN_FDR.s.m, file =
"mouse_16WT_vs_mouse_6WN_FDR.s.m.csv", quote =F, row.names=F)

write.csv(mouse_16WT_vs_mouse_6WN_FDR.s.gain.m, file =
"mouse_16WT_vs_mouse_6WN_FDR.s.gain.m.csv", quote =F, row.names=F)

write.csv(mouse_16WT_vs_mouse_6WN_FDR.s.loss.m, file =
"mouse_16WT_vs_mouse_6WN_FDR.s.loss.m.csv", quote =F, row.names=F)



write.table(mouse_16WT_vs_mouse_6WN_FDR.s.m, file =
"mouse_16WT_vs_mouse_6WN_FDR.s.m.bed", quote =F, sep="\t", row.names=F,
col.names=F)

write.table(mouse_16WT_vs_mouse_6WN_FDR.s.gain.m, file =
"mouse_16WT_vs_mouse_6WN_FDR.s.gain.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)

write.table(mouse_16WT_vs_mouse_6WN_FDR.s.loss.m, file =
"mouse_16WT_vs_mouse_6WN_FDR.s.loss.m.bed", quote =F, sep="\t",
row.names=F, col.names=F)
```

QC
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_mm10.bed 6w_1.bam > 6w_1_promoter.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_mm10.bed 6w_2.bam > 6w_2_promoter.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_mm10.bed 16w_1.bam > 16w_1_promoter.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_mm10.bed 16w_2.bam > 16w_2_promoter.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_mm10.bed 23w_1.bam > 23w_1_promoter.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/refGene_Prom_2000_2000_mm10.bed 23w_2.bam > 23w_2_promoter.bam

samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_mm10_cpg.bed 6w_1.bam > 6w_1_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_mm10_cpg.bed 6w_2.bam > 6w_2_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_mm10_cpg.bed 16w_1.bam > 16w_1_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_mm10_cpg.bed 16w_2.bam > 16w_2_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_mm10_cpg.bed 23w_1.bam > 23w_1_model_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/model_mm10_cpg.bed 23w_2.bam > 23w_2_model_cpg.bam

samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_mm10_cpg.bed 6w_1.bam > 6w_1_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_mm10_cpg.bed 6w_2.bam > 6w_2_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_mm10_cpg.bed 16w_1.bam > 16w_1_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_mm10_cpg.bed 16w_2.bam > 16w_2_UCSC_cpg.bam
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_mm10_cpg.bed 23w_1.bam > 23w_1_UCSC_cpg.bam

```
samtools view -@ 16 -b -L /home/yanzhu/blueridge/refGene_Prom/UCSC_mm10_cpg.bed
23w_2.bam > 23w_2_UCSC_cpg.bam


bamfile_mouse_all=c("6w_1.bam", "6w_2.bam", "16w_1.bam", "16w_2.bam", "23w_1.bam",
"23w_2.bam")
bamfile_mouse_all_promoter=c("6w_1_promoter.bam", "6w_2_promoter.bam",
"16w_1_promoter.bam", "16w_2_promoter.bam", "23w_1_promoter.bam",
"23w_2_promoter.bam")
bamfile_mouse_all_model_cpg=c("6w_1_model_cpg.bam", "6w_2_model_cpg.bam",
"16w_1_model_cpg.bam", "16w_2_model_cpg.bam", "23w_1_model_cpg.bam",
"23w_2_model_cpg.bam")
bamfile_mouse_all_UCSC_cpg=c("6w_1_UCSC_cpg.bam", "6w_2_UCSC_cpg.bam",
"16w_1_UCSC_cpg.bam", "16w_2_UCSC_cpg.bam", "23w_1_UCSC_cpg.bam",
"23w_2_UCSC_cpg.bam")


for (i in 1:length(bamfile_mouse_all)) {
pdf(paste("Saturation_",i,".pdf",sep=""))
sr=MeDIPS.saturation(file= bamfile_mouse_all[i], BSgenome = BSgenome, extend = extend,
shift = shift, uniq = uniq,window_size = ws, nit=10, nrit=1, empty_bins=TRUE, rank=FALSE)
MeDIPS.plotSaturation(sr)
dev.off()
}


for (i in 1:length(bamfile_mouse_all)) {
pdf(paste("SeqCoverage_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_mouse_all[i], pattern = "CG", BSgenome = BSgenome,
extend = extend, shift = shift, uniq = uniq)
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="pie", cov.level = c(0,1, 2, 3, 4, 5))
dev.off()
}

for (i in 1:length(bamfile_mouse_all_promoter)) {
pdf(paste("SeqCoverage_Promoter_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_mouse_all_promoter[i], pattern = "CG", BSgenome =
BSgenome, extend = extend, shift = shift, uniq = uniq)
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="pie", cov.level = c(0,1, 2, 3, 4, 5))
dev.off()
}

for (i in 1:length(bamfile_mouse_all_model_cpg)) {
pdf(paste("SeqCoverage_model_cpg_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_mouse_all_model_cpg[i], pattern = "CG", BSgenome
= BSgenome, extend = extend, shift = shift, uniq = uniq)
```

```
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="pie", cov.level = c(0,1, 2, 3, 4, 5))
dev.off()
}

for (i in 1:length(bamfile_mouse_all_UCSC_cpg)) {
pdf(paste("SeqCoverage_UCSC_cpg_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_mouse_all_UCSC_cpg[i], pattern = "CG", BSgenome
= BSgenome, extend = extend, shift = shift, uniq = uniq)
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="pie", cov.level = c(0,1, 2, 3, 4, 5))
dev.off()
}



for (i in 1:length(bamfile_mouse_all)) {
pdf(paste("SeqCoverageHistogram_",i,".pdf",sep=""))
cr = MeDIPS.seqCoverage(file = bamfile_mouse_all[i], pattern = "CG", BSgenome = BSgenome,
extend = extend, shift = shift, uniq = uniq)
MeDIPS.plotSeqCoverage(seqCoverageObj=cr, type="hist", t = 15)
dev.off()
}

for (i in 1:length(bamfile_mouse_all)) {
er = MeDIPS.CpGenrich(file = bamfile_mouse_all[i], BSgenome = BSgenome,
extend = extend, shift = shift, uniq = uniq)
write.table(er, file = "CpG_Enrichment.bed", append=T, quote =F, sep="\t", row.names=F,
col.names=!file.exists("CpG_Enrichment.bed"))
}
```