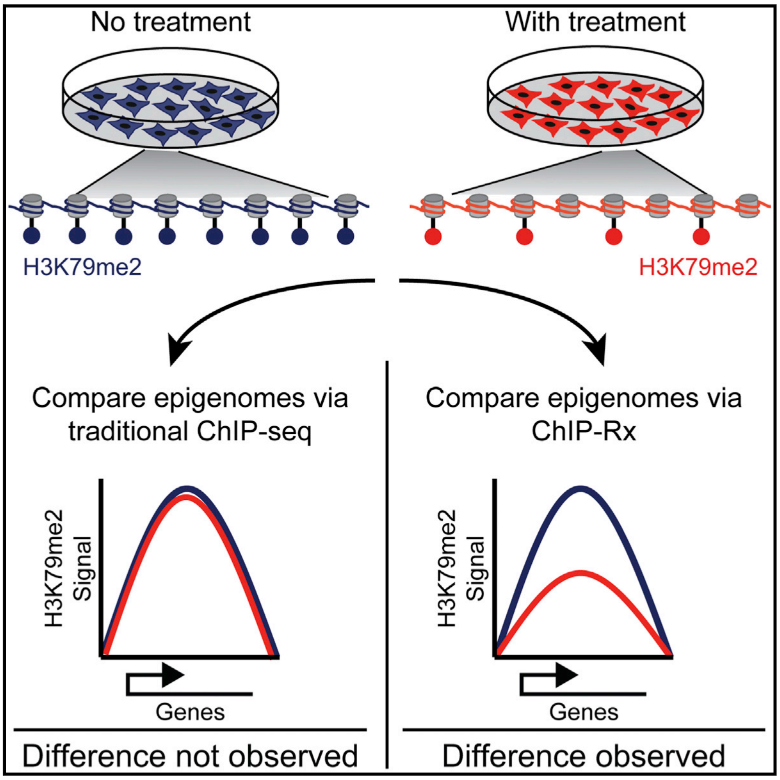
# Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome

Cell Reports

# Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome

## Graphical Abstract



## Highlights

ChIP-seq is a prevailing methodology to investigate and compare epigenomic states

Lack of an empirical normalization strategy has limited the usefulness of ChIP-seq

ChIP-Rx allows genome-wide quantitative comparisons of histone modification status

ChIP-Rx identifies graded epigenomic changes following chemical perturbations

## Authors

David A. Orlando, Mei Wei Chen, ..., James E. Bradner, Matthew G. Guenther

## Correspondence

dorlando@syros.com (D.A.O.),
mguenther@syros.com (M.G.G.)

## In Brief

The lack of an empirical methodology to enable normalization among chromatin immunoprecipitation coupled with massively parallel DNA sequencing (ChIP-seq) experiments has limited the precision and comparative utility of this technique. Orlando et al. describe a method, called ChIP with reference exogenous genome (ChIP-Rx), that allows one to perform genome-wide quantitative comparisons of histone modification status across cell populations using defined quantities of a reference epigenome. They use the method to detect disease-relevant epigenomic changes following drug treatment.

## Accession Numbers

GSE60104

Orlando et al., 2014, Cell Reports 9, 1163–1170
November 6, 2014 ©2014 The Authors
http://dx.doi.org/10.1016/j.celrep.2014.10.018

Cell Reports
Resource

# Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome

David A. Orlando,[1,*] Mei Wei Chen,[1] Victoria E. Brown,[1] Snehakumari Solanki,[1] Yoon J. Choi,[1] Eric R. Olson,[1] Christian C. Fritz,[1] James E. Bradner,[2,3] and Matthew G. Guenther[1,*]
[1]Syros Pharmaceuticals, 480 Arsenal Street, Watertown, MA 02472, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute
[3]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: dorlando@syros.com (D.A.O.), mguenther@syros.com (M.G.G.)

http://dx.doi.org/10.1016/j.celrep.2014.10.018

## SUMMARY

**Epigenomic profiling by chromatin immunoprecipitation coupled with massively parallel DNA sequencing (ChIP-seq) is a prevailing methodology used to investigate chromatin-based regulation in biological systems such as human disease, but the lack of an empirical methodology to enable normalization among experiments has limited the precision and usefulness of this technique. Here, we describe a method called ChIP with reference exogenous genome (ChIP-Rx) that allows one to perform genome-wide quantitative comparisons of histone modification status across cell populations using defined quantities of a reference epigenome. ChIP-Rx enables the discovery and quantification of dynamic epigenomic profiles across mammalian cells that would otherwise remain hidden using traditional normalization methods. We demonstrate the utility of this method for measuring epigenomic changes following chemical perturbations and show how reference normalization of ChIP-seq experiments enables the discovery of disease-relevant changes in histone modification occupancy.**

## INTRODUCTION

The ability to map genomic occupancy of transcriptional regulators, histone posttranslational modifications, and DNA methylation (epigenomic modifications) has enabled the elucidation of transcriptional mechanisms, genome organization, mapping of functional regulatory elements, and discovery of disease-associated chromatin markers (Badeaux and Shi, 2013; Barski et al., 2007; Lee and Young, 2013; Rivera and Ren, 2013; Zhou et al., 2011). Such targeted and large-scale epigenome mapping efforts have revealed chromatin regulatory proteins that are therapeutic targets for a wide variety of human diseases (Azad et al., 2013; Dawson and Kouzarides, 2012; Deshpande et al., 2012; Wee et al., 2014). Many of these chromatin regulators exhibit cell-type-selective, gene-selective, or disease-relevant effects, creating a critical need to study the chromatin modifications catalyzed by these regulators. Accurate quantification of both global and loci-specific chromatin modifications is needed to allow the discovery and characterization of epigenomic regulators and epigenome-modulating agents.

Traditional ChIP-seq methodologies are not inherently quantitative and therefore do not allow direct comparisons between samples derived from different cell types or between cells that have experienced a perturbation, such as a genomic alteration or chemical treatment. For example, if we employ the traditional reads per million (RPM) ChIP-seq normalization method, a cell population containing chromatin state ‘‘A’’ (a high level of histone posttranslational modification) will appear similar to a cell population containing chromatin state ‘‘B,’’ where 50% of the signal has been removed (Figure 1A), because the signal is quantified as a simple percentage of all mapped reads. Moreover, additional variables, such as variations in genome fragmentation, immunoprecipitation efficiency, or other experimental steps, frequently confound analysis. Efforts to correct for these variables have produced in silico normalization strategies, but an empirical method to enable direct and quantitative comparisons among epigenomic ChIP-seq data sets is still lacking (Bardet et al., 2012; Landt et al., 2012; Liang and Keleş, 2012; Liu et al., 2013; Nair et al., 2012). Because of the experimental and analytical restrictions of ChIP-seq, a robust normalization methodology is needed to quantify epigenome differences among varying cell populations, treatments, and genomic states.

## RESULTS

Here we present a method, called ChIP with reference exogenous genome (ChIP-Rx), that utilizes a constant amount of reference or ‘‘spike-in’’ epigenome, added on a per-cell basis, to allow direct comparison between two or more ChIP-seq samples (Figure 1B). Analogous methodologies have been applied in areas of gene-expression analysis that have revealed global transcriptional amplification upon normalization and in MethylC-seq, where bisulfate conversion rates have been normalized (Kanno et al., 2006; Krueger et al., 2012; Lin et al., 2012; Lovén et al., 2012; van de Peppel et al., 2003). These advancements have allowed standardization, precision, and a mechanistic

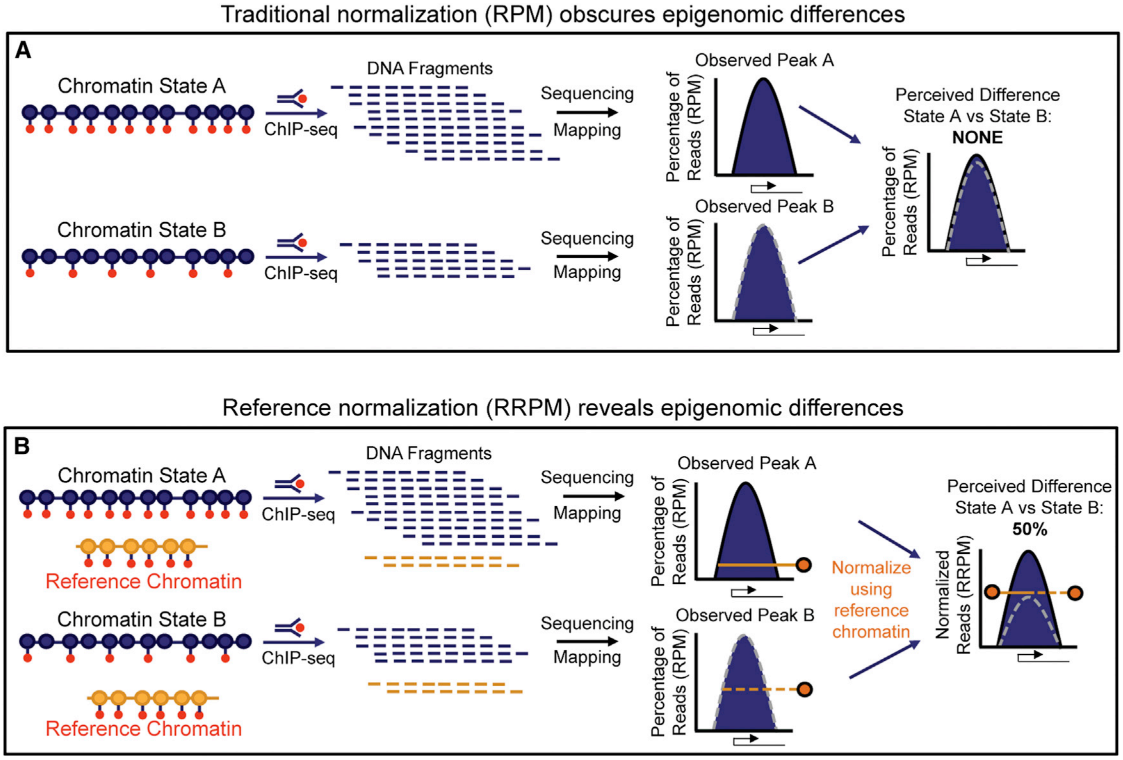

Traditional normalization (RPM) obscures epigenomic differences
A
Chromatin State A
ChIP-seq
DNA Fragments
Sequencing
Mapping
Observed Peak A
Percentage of Reads (RPM)
Perceived Difference State A vs State B: NONE
Chromatin State B
Observed Peak B
Reference normalization (RRPM) reveals epigenomic differences
B
Reference Chromatin
Normalize using reference chromatin
Perceived Difference State A vs State B: 50%
Normalized Reads (RRPM)

**Figure 1. Normalization and Interpretation of ChIP-Seq Data**

(A) Schematic representation of a typical ChIP-seq data workflow. Interrogation of a human epigenome (Blue circles, nucleosomes) with a full complement of histone modification (red circles, top) versus an epigenome with a half complement of histone modification (red circles, bottom). ChIP, sequencing, and mapping using reads per million (RPM) reveals ChIP-seq peaks (blue). A comparison of the peaks as a percentage of the total reads reveals little difference.

(B) Schematic representation of a ChIP-seq data workflow with reference genome normalization. Interrogation of a human epigenome (Blue circles, nucleosomes) with a full complement of histone modification (red circles, top) versus an epigenome with a half complement of histone modification (red circles, bottom). A fixed amount of reference epigenome (orange, nucleosomes; red, histone modifications) is added to human cells in each condition. After ChIP, sequencing, and mapping, the ChIP sequence reads are normalized to the percentage of reference genome reads in the sample (reference-adjusted RPM [RRPM]). A comparison of ChIP-seq signals using normalized reads reveals a 50% difference between peaks. This method is called ChIP with reference exogenous genome (ChIP-Rx).

understanding of RNA transcription (van Bakel and Holstege, 2004; Jiang et al., 2011; Li et al., 2013); however, no cell-count-normalized methods have been applied to global correction of histone posttranslational modifications. Since a vast array of histone modifications have been described in eukaryotic cells that play roles in organismal development, maintenance of cell state, differentiation, and disease, including those associated with transcriptional processes, genome organization, DNA repair, and cell-cycle progression (Calo and Wysocka, 2013; Pastor et al., 2013; Rinn and Chang, 2012; Rivera and Ren, 2013; Tan et al., 2011; Tian et al., 2012), a quantitative method for comparing these key marks is needed. We reasoned that the *Drosophila melanogaster* genome would be a desirable exogenous reference for mammalian cells because the *Drosophila* genome is well studied and has a high-quality sequence assembly, there is minimal mapping of the *Drosophila* genome sequence to human or mouse genomes (>0.05%; Table S1; Supplemental Experimental Procedures), *Drosophila* cells are readily available in large quantities, and the *Drosophila* epigenome displays nearly all of the key histone modification marks reported in humans. Moreover, histone proteins are among the most conserved proteins from humans to yeast, indicating that available ChIP-quality antibodies would likely recognize both *Drosophila* and human chromatin (Sullivan et al., 2002; Wolffe and Pruss, 1996).

To determine the impact of mixing interspecies epigenomes, we tested whether the addition of a reference genome (*Drosophila* S2 cells) would inherently affect our ability to detect a histone modification within the test sample (human cells) using ChIP-seq. We compared Jurkat cells alone with Jurkat cells that had been mixed with *Drosophila* cells and analyzed the resulting histone H3 lysine-79 dimethyl (H3K79me2) ChIP-seq profiles (Figure S1; Table S2). We determined that mixing of *Drosophila* and human cells did not induce large-scale changes in the H3K79me2 profiles, as the profiles of these cell populations were highly correlated by total signal as well as enriched loci overlap (Pearson correlation = 0.96; Supplemental Experimental Procedures). Moreover, reads originating from human or *Drosophila* could be separated with 99% accuracy (Supplemental Experimental Procedures). Together, these results indicate that the addition of a reference genome did not impede our ability to detect histone mark occupancy.

We devised an experiment to test our ability to detect changes in histone modification occupancy throughout the human

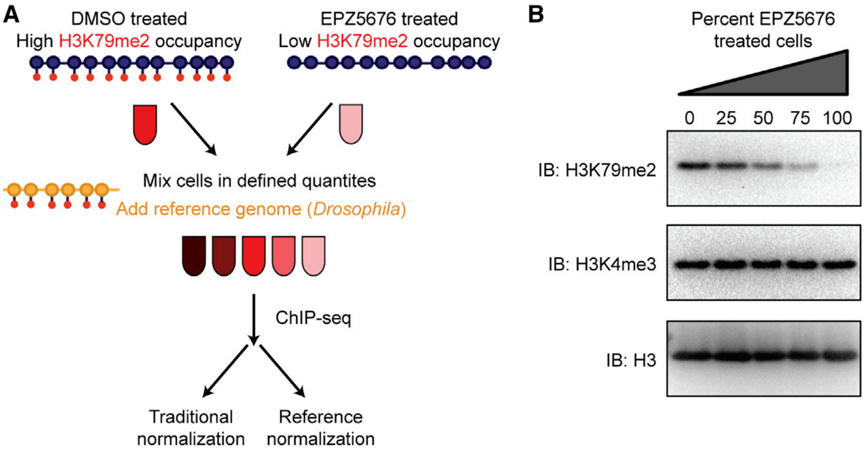


**Figure 2. Experimental Design of Differential H3K79me2 Detection**

(A) Schematic representation of differential H3K79me2 detection and normalization strategies. Two populations of cells were produced: a human epigenome (blue nucleosomes) with a full complement of H3K79me2 (red circles, top left) and a human epigenome (blue nucleosomes) with depleted H3K79me2 due to EPZ5676 exposure (top right). These cells were mixed in defined proportions in order to allow a dilution of total genomic histone modification (dark red to pink). Cell mixtures were subjected to ChIP-seq in the presence of the reference *Drosophila* epigenome (orange). ChIP-seq signals were calculated based on traditional or *Drosophila*-reference-normalized methods. See also Figure S1.

(B) Western blot validation of H3K79me2 depletion in Jurkat cells. Mixtures of 0%–100% EPZ5676-treated cells (0:100; 25:75; 50:50, 75:25; 100:0 proportions of [DMSO-treated:EPZ5676-treated] cells) were measured by immunoblot (IB) for the presence of H3K79me2, H3K4me3, or total histone H3 (loading control). Treated cells were exposed to 20 μM EPZ5676 for 4 days.
See also Table S1.

epigenome using ChIP-Rx (Figure 2A). We reasoned that an initial test of ChIP-Rx normalization should feature an epigenomic modification that could be readily removed and was not essential for cell viability in the model cell line. Using the selective DOT1L inhibitor EPZ5676 (Daigle et al., 2013), we depleted the Histone H3 lysine-79 dimethyl (H3K79me2) modification from Jurkat cell bulk histones (Figure 2B). The H3K79me2 modification is catalyzed by the DOT1L protein and is associated with the release of paused RNA Polymerase II and licensing of transcriptional elongation (van Leeuwen et al., 2002; Ng et al., 2002; Shanower et al., 2005; Steger et al., 2008). This modification is typically deposited within the 5′ regions of genes and its presence is not critical for Jurkat cell viability (Daigle et al., 2011, 2013; Schübeler et al., 2004; Steger et al., 2008). By mixing untreated cells (full H3K79me2) with EPZ5676-treated cells (H3K79me2 depleted), we created a set of cell populations with defined quantities of H3K79me2, as verified by immunodetection (Figure 2B). To each of these cell populations we added *Drosophila* cells at a ratio of one *Drosophila* cell per two human cells, which provided a constant "reference" amount of H3K79me2 per human cell. We performed ChIP-seq from samples consisting of 0:100, 25:75, 50:50, 75:25, and 100:0 proportions of EPZ5676 to DMSO-treated Jurkat cells. We then tested whether traditional ChIP-seq analysis methods would reveal the decrease in human per-cell H3K79me2 occupancy and, if not, whether the addition of the *Drosophila* epigenome would allow detection of H3K79me2 removal.

A key prediction of our normalization method is that as the global level of a histone modification is depleted in human cells, the percentage of total reads mapping to the reference *Drosophila* genome should increase. This is because the constant amount of reference genome added per human cell accounts for a greater percentage of total ChIP DNA fragments as human epitopes are lost (see the ratio of blue to orange DNA fragments in Figure 1B). To test this prediction, we interrogated cells with defined H3K79me2 levels (Figure 2B) by ChIP-Rx to measure genomic H3K79me2 occupancy. As a control, we also measured H3K4me3 occupancy, which is a histone modification that is not appreciably changed within our test cell populations (Figure 2B). As predicted, H3K79me2 depletion in Jurkat cells both reduced ChIP-Rx reads mapping to the human genome and increased reads mapping to the *Drosophila* genome (Figure 3A). We did not observe a similar change in the mapping ratio for H3K4me3 in the same samples, consistent with the finding that H3K4me3 was not preferentially removed from the human genome (Figure 3B). These results demonstrate that a reference genome can internally normalize the read count.

We next used the reference *Drosophila* genome to quantitatively normalize across experiments. To make ChIP-seq data quantitative on a per-cell basis, it is necessary to introduce a reference signal that is constant per cell, from which a normalization factor can be derived. Our ChIP-Rx protocol uses the signal from a fixed amount of *Drosophila* genome per human cell as this reference. We derived a normalization factor (see the Supplemental Experimental Procedures) for each experiment, such that the resulting *Drosophila* signal was equilibrated across all experiments (Table S2; Figure S2). Using traditional RPM normalization, the loci-specific ChIP-seq profiles and metagene profiles for H3K79me2 and H3K4me3 appear unchanged for the majority of samples (Figures 3C–3F), despite evidence that the H3K79me2 modification is progressively depleted (Figure 2B). After normalization with the *Drosophila* reference (normalized reference-adjusted RPM [RRPM]), a striking and gradated decrease in H3K79me2 signal across the samples is evident (Figures 3C, 3E, 3G, and S3). Normalization did not appreciably affect the metagene profiles of the control H3K4me3 experiments (Figures 3D, 3F, 3H, and S3). Repeat experiments produced the same result in all cases: normalization revealed a loss of H3K79me2 across the samples and H3K4me3 profiles were not significantly affected (Figure S4). These results indicate that normalization to a *Drosophila* reference is an effective method for quantitatively comparing multiple experiments and can reveal changes in histone modification that may not be apparent without proper normalization.

Having validated the ChIP-Rx methodology using standardized quantities of H3K79me2, we next tested our ability to

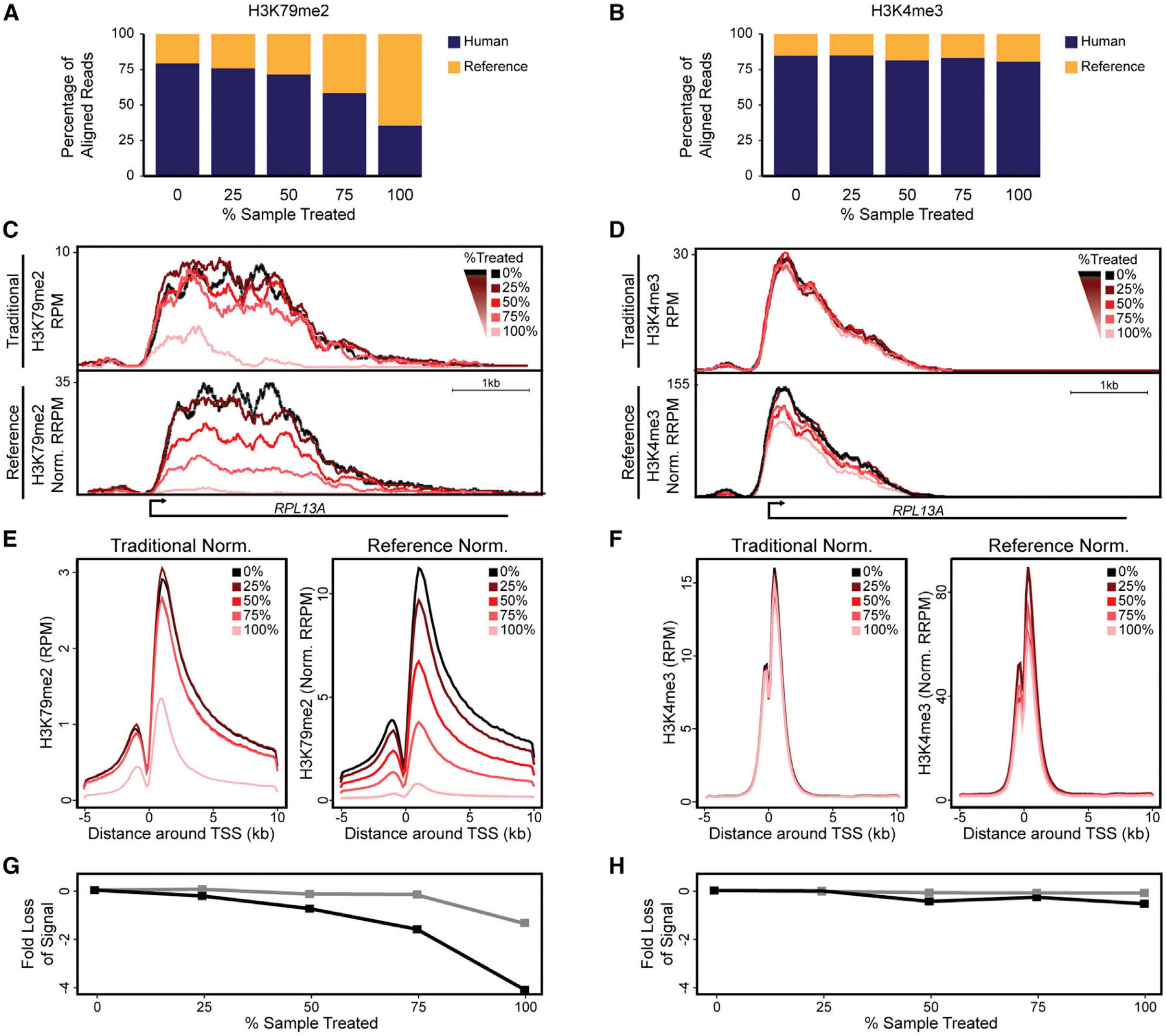


**Figure 3. ChIP-Rx Reveals Quantitative Epigenome Changes**

(A and B) Percentage of reads aligning to either test (human, blue) or *Drosophila* (reference, orange) genomes after H3K79me2 ChIP-Rx (A) or H3K4me3 ChIP-Rx (B). Samples containing 0%, 25%, 50%, 75%, or 100% EPZ5676 treated Jurkat cells were used as defined in Figure 2B.

(C and D) Sequenced reads from H3K79me2 (C) and H3K4me3 (D) immunoprecipitations at the RPL13A gene locus in traditional reads per million (RPM,top) or reference-adjusted reads per million (RRPM, bottom; see Experimental Procedures). Color indicates the percentage of sample treated with EPZ5676. The gene model is shown below the track.

(E) Meta-gene profile of H3K79me2-occupied genes in Jurkat cells. Meta-gene profiles were produced with traditional RPM (left) or RRPM (right). Color indicates the percentage of Jurkat cell sample treated with EPZ5676 as in Figure 2B. Region −5 to +10 kb around the transcription start site (TSS) is shown. Meta-gene profile was derived from top 5,000 protein-coding genes as defined by total H3K79me2 signal in the 0% treated (untreated with EPZ5676) sample. A meta-gene profile representing all genes is shown in Figure S3.

(F) Meta-gene profile of H3K4me3-occupied genes in Jurkat cells. Meta-gene profiles were produced with traditional RPM (left) or RRPM (right). Color indicates the percentage of Jurkat cell sample treated with EPZ5676 as in Figure 2B. Region −5 to +10 kb around the transcription start site (TSS) is shown. Meta-gene profile was derived from top 5,000 protein-coding genes as defined by total H3K4me3 signal in the 0% treated (untreated with EPZ5676) sample. A meta-gene profile representing all genes is shown in Figure S3.

(G and H) Line graphs display the observed fold-change difference in average meta-gene signal across the −5 to +10 kb window around the TSS for each H3K79me2 (G) or H3K4me3 (H) ChIP sample (x axis) relative to the signal from the 0% treated population using traditional (gray) or reference (black) normalization. See also Figures S2–S4 and Table S2.

normalize between ChIP-seq experiments in a disease-relevant system. MV4;11 acute myelomonocytic leukemia cells are a DOT1L-inhibitor sensitive model of human mixed-lineage-linked leukemia, a disease characterized by reciprocal translocations of the mixed-lineage leukemia (MLL) gene (Daigle et al., 2011, 2013; Deshpande et al., 2012). We treated MV4;11 cells with DOT1L inhibitor and measured changes in H3K79me2 occupancy in the presence or absence of the reference *Drosophila*

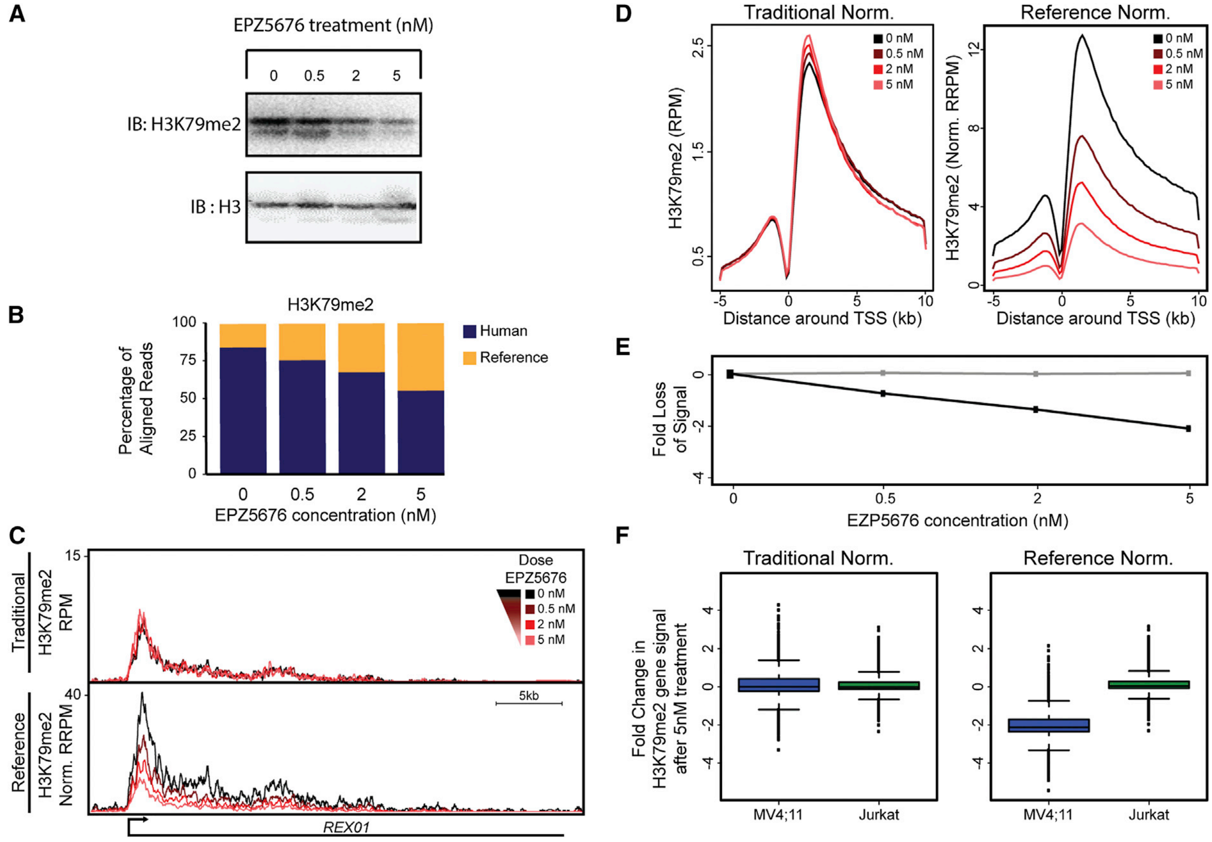


**Figure 4. ChIP-Rx Reveals Epigenomic Alterations in Disease Cells that Respond to Drug Treatment**

(A) Western blot showing the levels of H3K79me2 in MV4;11 cells after treatment for 4 days with increasing concentrations of EPZ5676.

(B) Percentage of H3K79me2 ChIP-seq reads aligning to either test (human, blue) or *Drosophila* (reference, orange) genomes after H3K79me2 ChIP-Rx from MV4;11 cells treated as in (A).

(C) Sequenced reads from H3K79me2 immunoprecipitations at the REXO1 gene locus in standard RPM (top) or RRPM (bottom) (see Experimental Procedures). Color indicates the concentration of EPZ5676 given to each sample. The gene model is shown below the track.

(D) Meta-gene profile of H3K79me2-occupied genes in MV4;11 cells. Meta-gene profiles were produced with traditional Reads Per Million (RPM, left) or Reference-adjusted Reads Per Million (RRPM, right). Color indicates the concentration of EPZ5676 used in each sample. The region −5 kb to +10 kb around the TSS is shown. Meta-gene profile was derived from top 5,000 protein-coding genes as defined by total H3K79me2 signal in the 0nM treated (untreated with EPZ5676) sample. A meta-gene profile representing all genes is shown in Figure S3.

(E) Line graph displays the observed fold-change difference in average meta-gene signal across the −5 to +10 kb window around the TSS for each H3K79me2 ChIP sample (x axis) relative to the signal from the 0 nM treated population using standard (gray) or reference (black) normalization.

(F) Box plots display the distribution of the observed fold change of H3K79me2 signal −5 kb to +10 kb around the TSS of all genes between the 0 nM and 5 nM treated samples (blue, MV4;11; green, Jurkat) for all genes using traditional (left) or reference-adjusted (right) normalization (see the Supplemental Experimental Procedures).

See also Figures S3 and S5 and Table S2.

epigenome (Figures 4A–4E, S5A, and S5B). EPZ5676 induced a dose-dependent decrease in bulk H3K79me2 (Figure 4A), but this result was masked when we quantified H3K79me2 occupancy using traditional normalization (Figures 4C and 4E). We observed a dose-dependent decrease in H3K79me2 genomic occupancy only after employing reference normalization (Figures 4C–4E). This unmasking of epigenomic effects may be critical for understanding the cell-type-selective effects of small-molecule epigenome modulators. For example, MV4;11 cells exhibit global H3K79me2 depletion at a low dose of EPZ5676, consistent with the known selectivity of the DOT1L inhibitor for leukemic cells carrying MLL translocations, but EPZ5676-insensitive Jurkat cells do not (Figures 4A and S5C; Daigle et al., 2013). Thus, normalizing to a reference exogenous genome rectifies the protein-level measurements and genome occupancy of modified histones, and reveals subtle epigenomic changes that may underlie or predict cellular responses to drugs (Figure 4). These results show that ChIP-Rx enables the discovery of epigenomic changes that can provide insight into disease and inform drug mechanisms.

## DISCUSSION

In summary, we have demonstrated that ChIP-Rx allows the discovery and quantification of dynamic epigenomic profiles across mammalian cells that would otherwise remain hidden using traditional normalization methods. A recent study employed a similar reference strategy for ChIP-seq normalization (Bonhoure et al., 2014); however, our method offers two crucial advantages that allow direct comparative epigenomic analysis. First, our method introduces the reference at the beginning of the experiment, thus normalizing for variation throughout the experiment, including chromatin fragmentation and immunoenrichment, both of which are critical for epitope and genome retrieval (Kidder et al., 2011; Meyer and Liu, 2014; Raha et al., 2010). Second, as was analogously shown for RNA expression correction (Lovén et al., 2012), our method introduces the reference "spike-in" on a per-cell basis as opposed to total chromatin, thus allowing the detection of unidirectional chromatin changes irrespective of variations in ploidy or gross chromatin. Thus, our method provides greater accuracy in determining epigenome changes that occur upon cell perturbation or exposure to small-molecule inhibitors as compared with current methods. Importantly, ChIP-Rx allows for the detection of subtle epigenomic changes, as opposed to qualitative occupancy calls, and thus advances the ChIP-seq methodology from a descriptive, binary readout to one that reveals gradated epigenomic changes. This is particularly important for the dose-ranging characterization of chemical tools and therapeutics targeting chromatin-associated complexes via genome-wide approaches. Application of this methodology to additional model systems, including mouse, rat, and zebrafish (Table S1), as well as additional histone modifications, including repressive (i.e., H3K27me3) and activating (i.e., H3K27ac) histone modifications, will enable far-reaching studies of comparative epigenomics.

We recommend the implementation of ChIP-Rx whenever quantitative or comparative epigenomic changes are under investigation. The method described here will be critical for understanding the global and site-selective epigenomic changes that occur in human disease, during cell-state changes, and especially the action of small-molecule inhibitors of chromatin-modulating proteins.

## EXPERIMENTAL PROCEDURES

### Human Cell Lines, Growth, and Treatment

Jurkat cells were obtained from ATCC and maintained in RPMI (Life Technologies) supplemented with 10% fetal bovine serum (FBS; Life Technologies) at 5% $CO_2$ in 37°C. MV4;11 cells were obtained from ATCC and maintained in RPMI (Life Technologies) supplemented with 10% FBS at 5% $CO_2$ in 37°C.

Jurkat cells were treated with DMSO or EPZ5676 (Selleck Chemicals, catalog number S7062) at 5 nM or 20 μM for 4 days, and MV4;11 cells were treated with DMSO or EPZ5676 at 0.5 nM, 2 nM, or 5 nM for 4 days. Live-cell numbers were quantified using the Countess cell counter (Life Technologies).

At harvest, cells were crosslinked with 1% formaldehyde by addition of 1/10 volume of fresh 11% formaldehyde solution (11% formaldehyde 0.1 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 50 mM HEPES) and incubation at room temperature for 8 min. Crosslinking reactions were quenched with a 1/20 volume of 2.5 M glycine for 1–5 min and cells were pelleted. The cells were then washed three times with ice-cold PBS. Washed cell pellets were flash frozen and stored at −80°C.

### Preparation of *Drosophila* S2 Cells

*Drosophila* S2 cells (ATCC catalog number CRL-1963; Biovest part number OO.763/OO.627) were cultured in Schneider's *Drosophila* media (Life Technologies catalog number 21720-024) supplemented with 10% FBS to attain a density of 0.5–0.6 × $10^6$ cells/ml. Cell culture and scale-up to 2 L was performed by Biovest International.

At harvest, cells were crosslinked with 1% formaldehyde by addition of a 1/10 volume of fresh 11% formaldehyde solution (11% formaldehyde 0.1 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 50 mM HEPES) and incubation at room temperature for 8 min. Crosslinking reactions were quenched with a 1/20 volume of 2.5 M glycine for 1–5 min and cells were pelleted. The cells were then washed three times with ice-cold PBS. Washed cell pellets were flash frozen and stored at −80°C at 1 × $10^8$ cells per aliquot.

### ChIP-Rx

For each ChIP-Rx experiment, a 2:1 ratio of human:*Drosophila* cells was used. This corresponds to 20 million crosslinked human cells and 10 million crosslinked S2 cells (Jurkat experiments) or 15 million crosslinked human cells and 7.5 million crosslinked S2 cells (MV4;11 experiments).

S2 cells were added to human cells at the beginning of the ChIP-Rx workflow (during nuclei isolation). Once *Drosophila* S2 and human cells were combined, the sample was treated as a single ChIP-seq sample throughout the experiment until completion of DNA sequencing.

Briefly, frozen, crosslinked human and *Drosophila* cells were resuspended in parallel in cold Lysis Buffer 1 (140 mM NaCl, 1 mM EDTA, 50 mM HEPES, 10% glycerol, 0.5% NP-40, 0.25% Triton-X-100), incubated 10 min at 4°C, and pelleted. Both human and *Drosophila* cell samples were resuspended in parallel in Lysis Buffer 2 (10 mM TRIS [pH 8.0], 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA), incubated for 10 min at 4°C, and combined to the desired cell number ratios (two human cells per one *Drosophila* cell). The composite cell nuclei was then pelleted and resuspended in sonication buffer (10 mM TRIS [pH 8.0], 1 mM EDTA, 0.1% SDS).

Composite samples (human + *Drosophila* S2) in sonication buffer were sonicated using a Covaris E220 sonication water bath for 5 min. Sheared chromatin was diluted 1:1 in 2× dilution buffer (300 mM NaCl, 2 mM EDTA, 50 mM TRIS [pH 8.0], 1.5% Triton-X, 0.1% SDS) and incubated with either H3K79me2 (Abcam 3594)- or H3K4me3 (Millipore 07-473)-conjugated Protein G Dynal beads (Invitrogen) overnight (8–16 hr, rotating) at 4°C, and then washed two times with wash buffer 1 (50 mM HEPES, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.75% Triton-X, 0.1% SDS, 0.05% DOC), two times with high-salt wash buffer (50 mM HEPES, 500 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.75% Triton-X, 0.1% SDS, 0.05% DOC), and one time with TE-NaCl buffer (10 mM Tris [pH 8.0], 1 mM EDTA, 50 mM NaCl). Samples were eluted from beads for 1 hr at 65°C in elution buffer (50 mM TRIS [pH 8.0], 10 mM EDTA, 1% SDS) and supernatant reverse-crosslinked at 65°C for 6–16 hr. Samples were diluted 1:1 with TE buffer (50 mM TRIS [pH 8.0], 1 mM EDTA) and treated with RNase A (0.2 mg/ml) for 2 hr at 37°C and then Proteinase K (0.2 mg/ml) for 2 hr at 55°C. DNA was isolated by phenol-chloroform extraction and ethanol precipitation. For detailed protocols see the Supplemental Experimental Procedures and Guenther et al. (2008).

### Library Construction, Sequencing, and Data Collection

Libraries were constructed with the Illumina Tru-Seq library preparation kit using a target fragment size of 200–400 bp and multiplexing barcodes. Libraries were sequenced using Illumina HiSeq 2000 with single-end reads for 40 cycles. Sequences were demultiplexed and aligned using Bowtie2 against a "genome" that combines the human hg19 genome and the *Drosophila* dm3 genome (see the Supplemental Experimental Procedures). Individual accession numbers and read statistics available in Table S2.

### Western Blots

Cells were harvested from all treatment groups and lysed with Triton extraction buffer (PBS containing 0.5% Triton X-100 [v/v], cOmplete Protease Inhibitors [Roche]) for 10 min with rotation. Nuclei were collected and acid extracted with 0.2 N HCl overnight. Histone proteins were collected from the supernatant and immunoblotted for H3K4me3 (Millipore 07-473), H3K79me2 (Abcam 3594), and histone H3 (Abcam 1791).

### Determination of the Normalization Factor

A complete description of the basis and derivation of the ChIP-Rx normalization factor is provided in the Supplemental Experimental Procedures. In brief, we derived a normalization constant, $\alpha$, such that after normalization the signal per-reference cell (β) is the same across all samples. The total ChIP-seq signal derived from reference cells is simply the count of reads (in millions) aligning to the *Drosophila* genome, which we represent as $N_d$. Because the percentage of reference cells as a fraction of the total number of cells is constant and we assume that the epigenome of the reference cells does not vary appreciably, we can derive α as

$$\alpha * N_d = \beta$$

Because β is a constant, we can simply rewrite this as

$$\alpha * N_d = 1$$

or

$$\alpha = \frac{1}{N_d},$$

multiplying the read counts by α produces a normalized read count in normalized RRPM.

## ACCESSION NUMBERS

The raw sequencing data reported in this work have been deposited in the NCBI Gene Expression Omnibus under accession number GSE60104.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.10.018.

## AUTHOR CONTRIBUTIONS

D.A.O., M.G.G., J.E.B., C.C.F., and E.R.O. designed and analyzed the research. M.W.C. conducted ChIP-seq and ChIP-Rx experiments. All other experiments were performed by M.W.C., V.E.B., Y.J.C., S.S., and M.G.G. D.A.O. performed the computational analysis. M.G.G. and D.A.O. wrote the manuscript.

## ACKNOWLEDGMENTS

We thank Richard Young, Matthew Eaton, Cindy Collins, Charles Lin, Jakob Loven, Tony Lee, Jason Marineau, Michael McKeown, and Peter Rahl for helpful discussions and comments. We also thank Thomas Volkert and the Whitehead Genome Technology Core for high-throughput sequencing. J.E.B. is a founder of Syros Pharmaceuticals. D.A.O., M.W.C., V.E.B., S.S., Y.J.C., E.R.O., C.C.F., and M.G.G. are employees of Syros Pharmaceuticals.

Received: August 19, 2014
Revised: September 24, 2014
Accepted: October 9, 2014
Published: October 30, 2014

## REFERENCES

Azad, N., Zahnow, C.A., Rudin, C.M., and Baylin, S.B. (2013). The future of epigenetic therapy in solid tumours—lessons from the past. Nat Rev Clin Oncol *10*, 256–266.

Badeaux, A.I., and Shi, Y. (2013). Emerging roles for chromatin as a signal integration and storage platform. Nat. Rev. Mol. Cell Biol. *14*, 211–224.

Bardet, A.F., He, Q., Zeitlinger, J., and Stark, A. (2012). A computational pipeline for comparative ChIP-seq analyses. Nat. Protoc. *7*, 45–61.

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

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., and Delorenzi, M.; CycliX Consortium (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. *24*, 1157–1168.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.

Daigle, S.R., Olhava, E.J., Therkelsen, C.A., Majer, C.R., Sneeringer, C.J., Song, J., Johnston, L.D., Scott, M.P., Smith, J.J., Xiao, Y., et al. (2011). Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor. Cancer Cell *20*, 53–65.

Daigle, S.R., Olhava, E.J., Therkelsen, C.A., Basavapathruni, A., Jin, L., Boriack-Sjodin, P.A., Allain, C.J., Klaus, C.R., Raimondi, A., Scott, M.P., et al. (2013). Potent inhibition of DOT1L as treatment of MLL-fusion leukemia. Blood *122*, 1017–1025.

Dawson, M.A., and Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. Cell *150*, 12–27.

Deshpande, A.J., Bradner, J., and Armstrong, S.A. (2012). Chromatin modifications as therapeutic targets in MLL-rearranged leukemia. Trends Immunol. *33*, 563–570.

Guenther, M.G., Lawton, L.N., Rozovskaia, T., Frampton, G.M., Levine, S.S., Volkert, T.L., Croce, C.M., Nakamura, T., Canaani, E., and Young, R.A. (2008). Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. Genes Dev. *22*, 3403–3408.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.

Kanno, J., Aisaki, K.I., Igarashi, K., Nakatsu, N., Ono, A., Kodama, Y., and Nagao, T. (2006). "Per cell" normalization method for mRNA measurement by quantitative PCR and microarrays. BMC Genomics *7*, 64.

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

Krueger, F., Kreck, B., Franke, A., and Andrews, S.R. (2012). DNA methylome analysis using short bisulfite sequencing data. Nat. Methods *9*, 145–151.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. *22*, 1813–1831.

Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237–1251.

Li, Y., Wang, H., Muffat, J., Cheng, A.W., Orlando, D.A., Lovén, J., Kwok, S.-M., Feldman, D.A., Bateup, H.S., Gao, Q., et al. (2013). Global transcriptional and translational repression in human-embryonic-stem-cell-derived Rett syndrome neurons. Cell Stem Cell *13*, 446–458.

Liang, K., and Keleş, S. (2012). Normalization of ChIP-seq data with control. BMC Bioinformatics *13*, 199.

Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. Cell *151*, 56–67.

Liu, B., Yi, J., Sv, A., Lan, X., Ma, Y., Huang, T.H., Leone, G., and Jin, V.X. (2013). QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions. BMC Genomics *14* (*Suppl 8*), S3.

Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., and Young, R.A. (2012). Revisiting global gene expression analysis. Cell *151*, 476–482.

Meyer, C.A., and Liu, X.S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat. Rev. Genet. *15*, 709–721.

Nair, N.U., Sahu, A.D., Bucher, P., and Moret, B.M.E. (2012). ChIPnorm: a statistical method for normalizing and identifying differential regions in histone modification ChIP-seq libraries. PLoS ONE *7*, e39573.

Ng, H.H., Feng, Q., Wang, H., Erdjument-Bromage, H., Tempst, P., Zhang, Y., and Struhl, K. (2002). Lysine methylation within the globular domain of histone

H3 by Dot1 is important for telomeric silencing and Sir protein association. Genes Dev. *16*, 1518–1527.

Pastor, W.A., Aravind, L., and Rao, A. (2013). TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. Nat. Rev. Mol. Cell Biol. *14*, 341–356.

Raha, D., Hong, M., and Snyder, M. (2010). ChIP-Seq: a method for global identification of regulatory elements in the genome. Curr. Protoc. Mol. Biol. *Chapter 21*, 1–14.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Rivera, C.M., and Ren, B. (2013). Mapping human epigenomes. Cell *155*, 39–55.

Schübeler, D., MacAlpine, D.M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D.E., O'Neill, L.P., Turner, B.M., Delrow, J., et al. (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev. *18*, 1263–1271.

Shanower, G.A., Muller, M., Blanton, J.L., Honti, V., Gyurkovics, H., and Schedl, P. (2005). Characterization of the grappa gene, the Drosophila histone H3 lysine 79 methyltransferase. Genetics *169*, 173–184.

Steger, D.J., Lefterova, M.I., Ying, L., Stonestrom, A.J., Schupp, M., Zhuo, D., Vakoc, A.L., Kim, J.-E., Chen, J., Lazar, M.A., et al. (2008). DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. Mol. Cell. Biol. *28*, 2825–2839.

Sullivan, S., Sink, D.W., Trout, K.L., Makalowska, I., Taylor, P.M., Baxevanis, A.D., and Landsman, D. (2002). The Histone Database. Nucleic Acids Res. *30*, 341–342.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., et al. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell *146*, 1016–1028.

Tian, Z., Tolić, N., Zhao, R., Moore, R.J., Hengel, S.M., Robinson, E.W., Stenoien, D.L., Wu, S., Smith, R.D., and Paša-Tolić, L. (2012). Enhanced top-down characterization of histone post-translational modifications. Genome Biol. *13*, R86.

van Bakel, H., and Holstege, F.C.P. (2004). In control: systematic assessment of microarray performance. EMBO Rep. *5*, 964–969.

van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D., and Holstege, F.C.P. (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Rep. *4*, 387–393.

van Leeuwen, F., Gafken, P.R., and Gottschling, D.E. (2002). Dot1p modulates silencing in yeast by methylation of the nucleosome core. Cell *109*, 745–756.

Wee, S., Dhanak, D., Li, H., Armstrong, S.A., Copeland, R.A., Sims, R., Baylin, S.B., Liu, X.S., and Schweizer, L. (2014). Targeting epigenetic regulators for cancer therapy. Ann. N Y Acad. Sci. *1309*, 30–36.

Wolffe, A.P., and Pruss, D. (1996). Hanging on to histones. Chromatin. Curr. Biol. *6*, 234–237.

Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat. Rev. Genet. *12*, 7–18.

**Cell Reports, Volume *9***

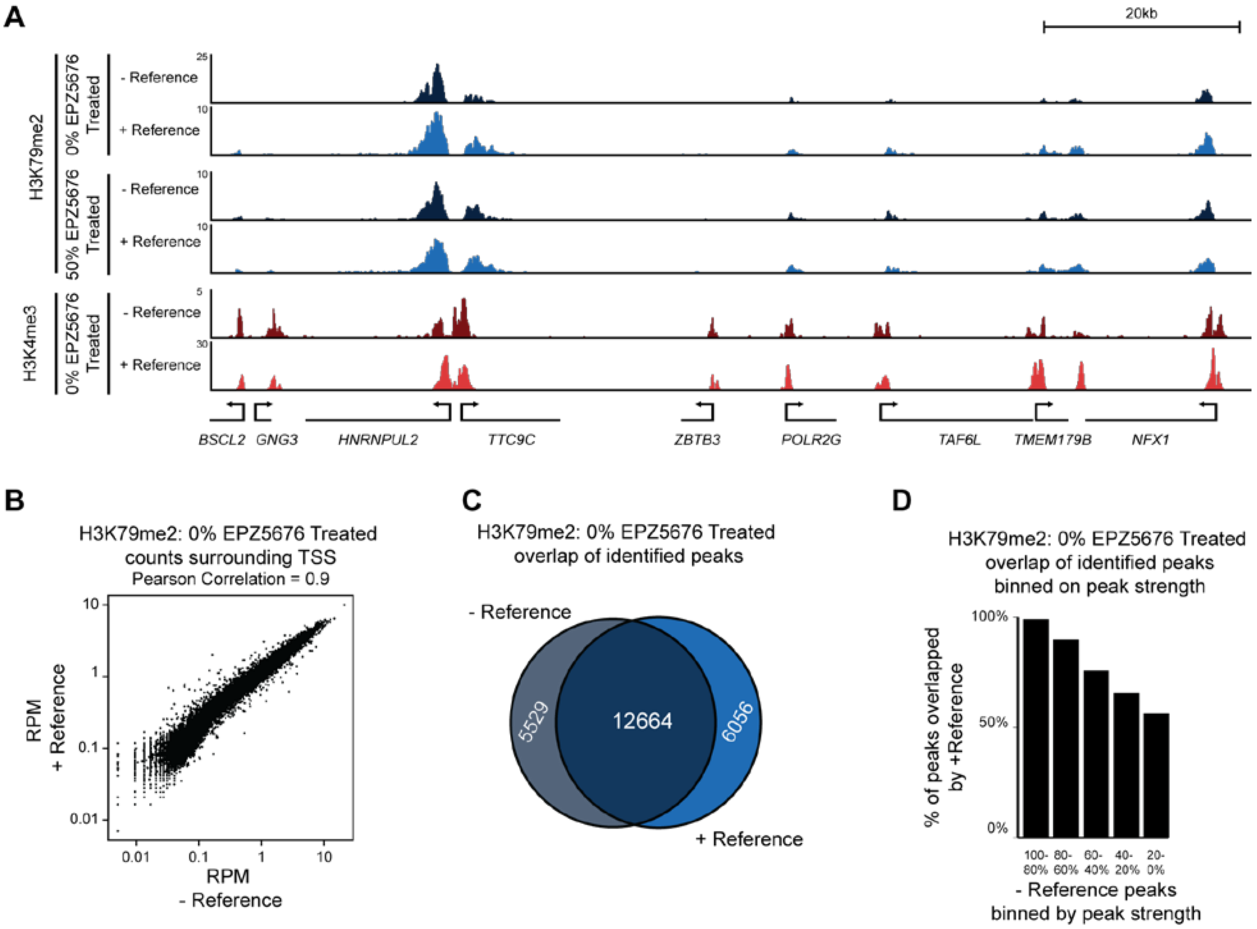
**Supplemental Information**

# Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome

**David A. Orlando, Mei Wei Chen, Victoria E. Brown, Snehakumari Solanki, Yoon J. Choi, Eric R. Olson, Christian C. Fritz, James E. Bradner, and Matthew G. Guenther**

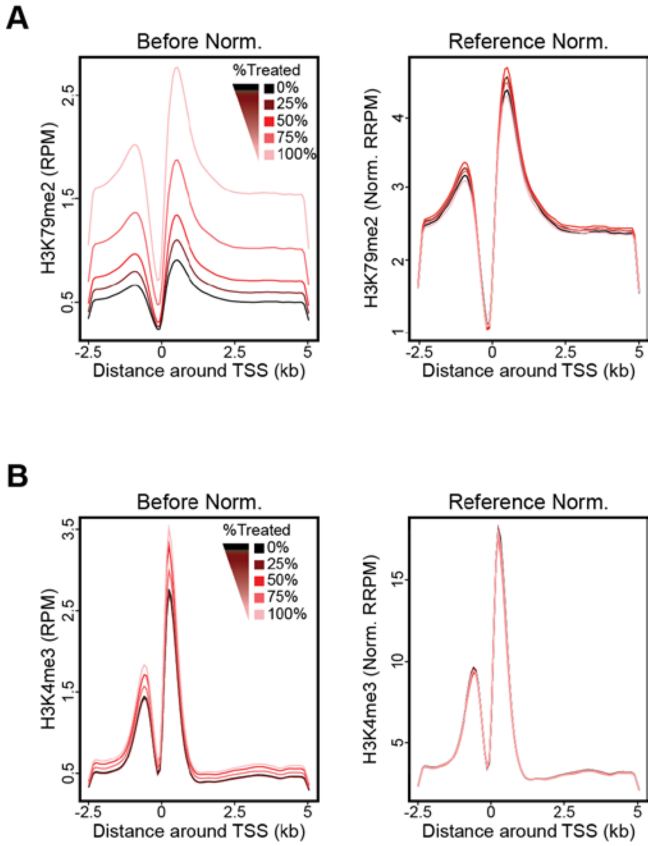
## Supplemental Figures:

**Figure S1:** ChIP-Seq profiles of samples with and without *drosophila* reference are highly similar, related to Figure 2.



**(A)** ChIP-Seq signal from pairs of samples without (-reference, darker color) or including *drosophila* S2 reference cells (+reference, lighter color). (Top) H3K79me2 in samples not treated with EPZ5676. (Middle) H3K79me2 in samples with 50% of the population treated with 20µM EPZ5676 (Selleck Chemicals) for 4 days. (Bottom) H3K4me3 in samples not treated with EPZ5676. Color indicates histone mark (blue=H3K79me2, red=H3K4me3). Y-axis is in traditional RPM. Gene models are shown below the track. (**B**) Scatterplot of RPM densities for a region -5kb/+10kb around each TSS for H3K79me2 samples not treated with EPZ5676 that included *drosophila* S2 reference cells (y-axis) or did not (x-axis). See "Effect of *drosophila* reference cells on human cell ChIP-seq" section for details. **(C)** Overlap of peaks identified in the H3K79me2 untreated samples by MACS2 at a p-value of 1x10$^{-9}$. **(D)** Overlap of peaks binned by peak strength. Peaks identified in the H3K79me2 untreated sample without reference were binned into quantiles based on MACS2 pileup value (x-axis). The percentage of peaks in each bin overlapped by peaks from the corresponding H3K79me2 sample with reference included is shown in the y-axis.

**Figure S2:** ChIP-seq signal from *drosophila* reference before/after normalization, related to Figure 3.



Meta-gene profiles across all *drosophila* protein coding genes in a region -2.5kb to +5kb around the TSS for H3K79me2 (**A**) or H3K4me3 (**B**). (Left) Read counts normalized by total aligned reads (human+*drosophila*) per-million. (Right) Read counts normalized by reference normalization derived constant (RRPM). Color indicates percentage of sample treated with EPZ5676.

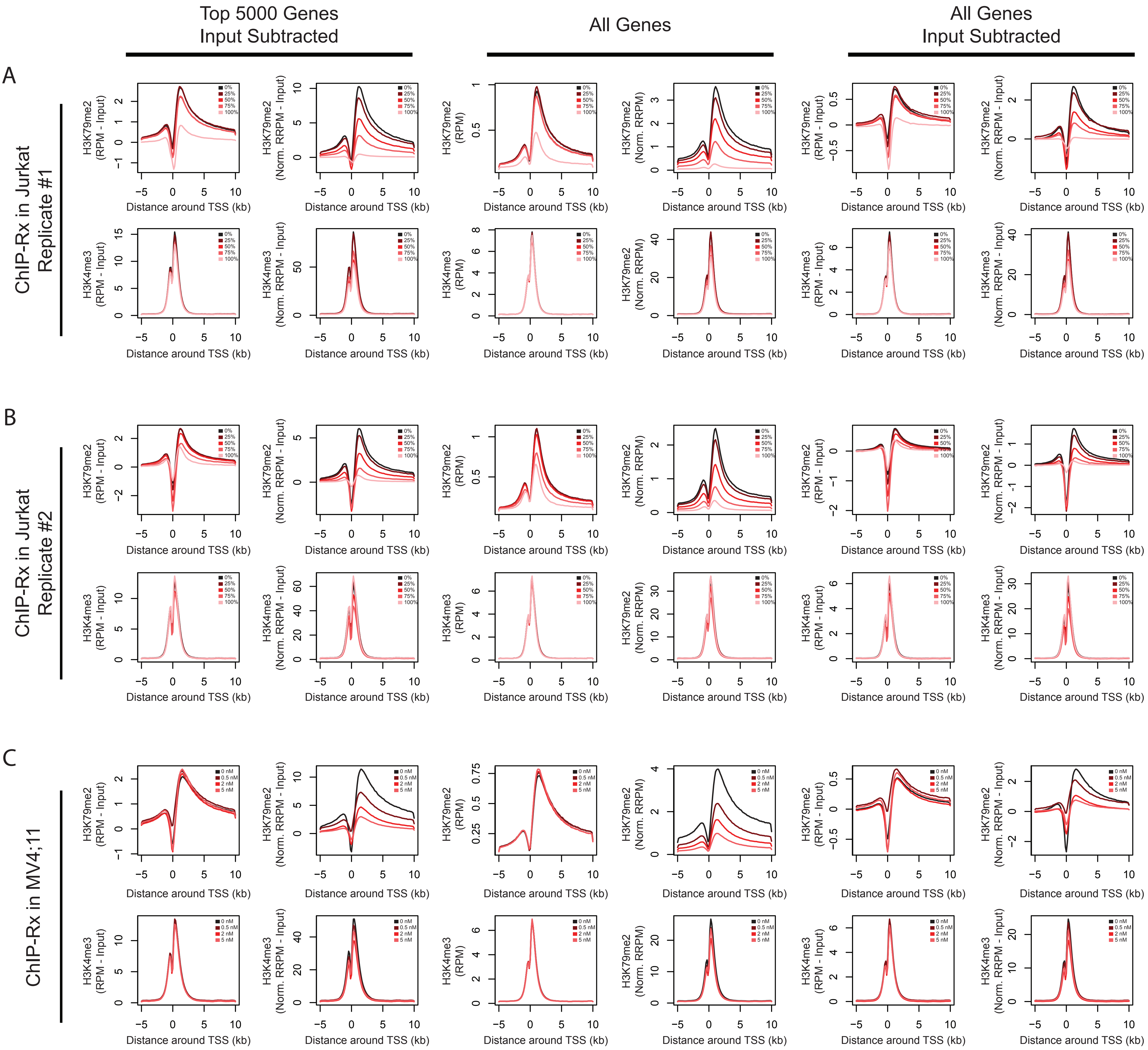
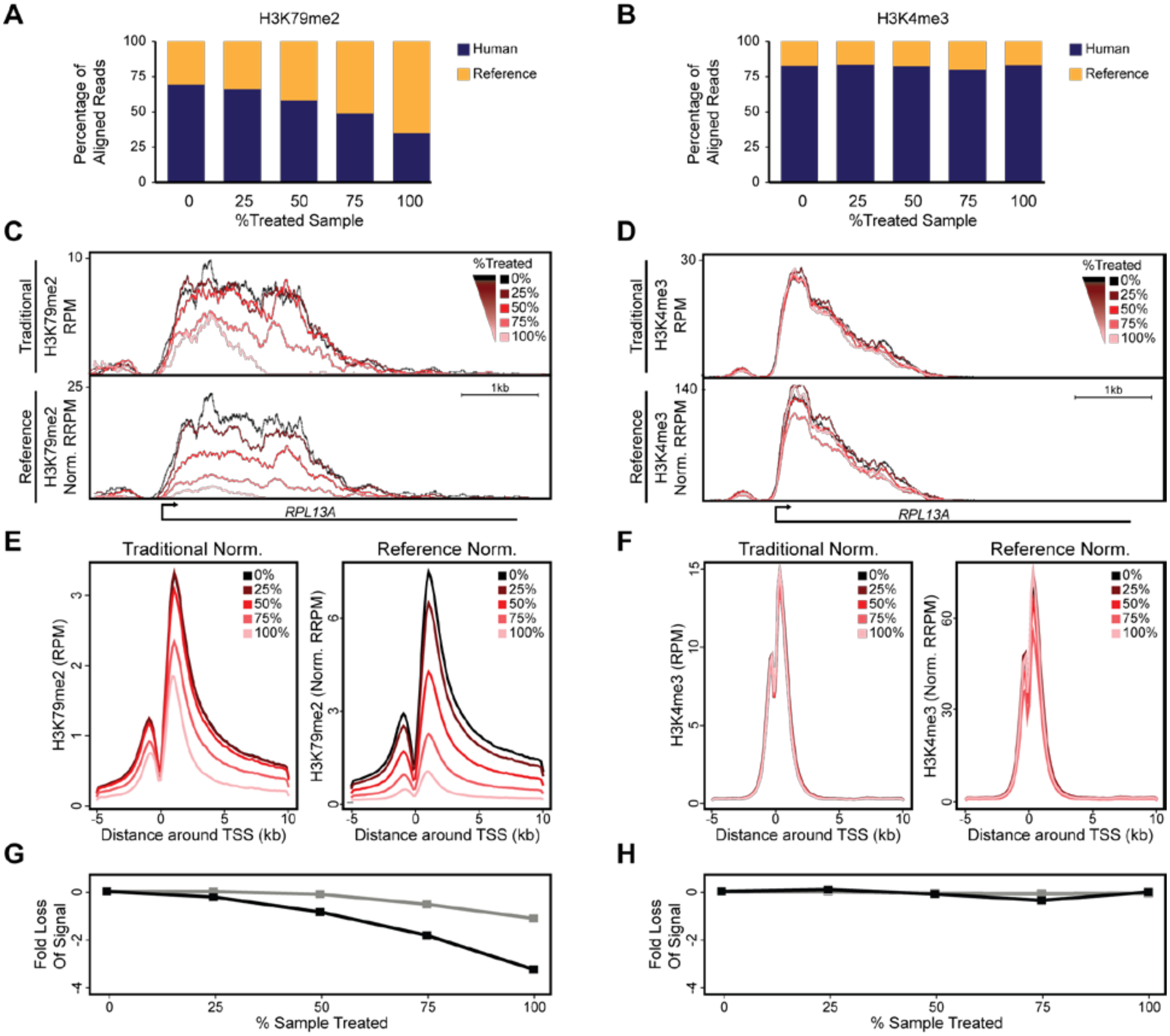

Top 5000 Genes
Input Subtracted
All Genes
All Genes
Input Subtracted
A
ChIP-Rx in Jurkat
Replicate #1
B
ChIP-Rx in Jurkat
Replicate #2
C
ChIP-Rx in MV4;11
H3K79me2
(RPM - Input)
H3K79me2
(Norm. RRPM - Input)
H3K79me2
(RPM)
H3K79me2
(Norm. RRPM)
H3K4me3
(RPM - Input)
H3K4me3
(Norm. RRPM - Input)
H3K4me3
(RPM)
H3K79me2
(Norm. RRPM)
0%
25%
50%
75%
100%
0 nM
0.5 nM
2 nM
5 nM
Distance around TSS (kb)
-5
0
5
10

**Figure S3:** ChIP-Rx metagenes using input subtraction and/or including all genes, related to Figure 3 and Figure 4.

**(A)** Metagene profiles of H3K79me2 (top) or H3K4me3 (bottom) samples from Figure 3 in the region -5kb to +10kb around the TSS for traditional (Left) and reference (Right) normalized data. (Far Left) Input subtracted metagenes for top 5000 protein-coding genes defined by total H3K79me2 or H3K4me3 are shown. (Middle) Metagenes for all protein coding genes are shown. (Far Right) Input subtracted metagenes for all protein coding genes are shown. Color indicates percentage of sample treated with EPZ5676. **(B)** Meta-gene profiles of H3K79me2 (top) or H3K4me3 (bottom) samples from Figure S4 in the region -5kb to +10kb around the TSS for traditional (Left) and reference (Right) normalized data. (Far Left) Input subtracted metagenes for top 5000 protein-coding genes defined by total H3K79me2 or H3K4me3 are shown. (Middle) Metagenes for all protein coding genes are shown. (Far Right) Input subtracted metagenes for all protein coding genes are shown. Color indicates percentage of sample treated with EPZ5676. **(C)** Meta-gene profiles of H3K79me2 (top) or H3K4me3 (bottom) samples from Figure 4 in the region -5kb to +10kb around the TSS for traditional (Left) and reference (Right) normalized data. (Far Left) Input subtracted metagenes for top 5000 protein-coding genes defined by total H3K79me2 or H3K4me3 are shown. (Middle) Metagenes for all protein coding genes are shown. (Far Right) Input subtracted metagenes for all protein coding genes are shown. Color indicates concentration of EPZ5676 given to each sample.

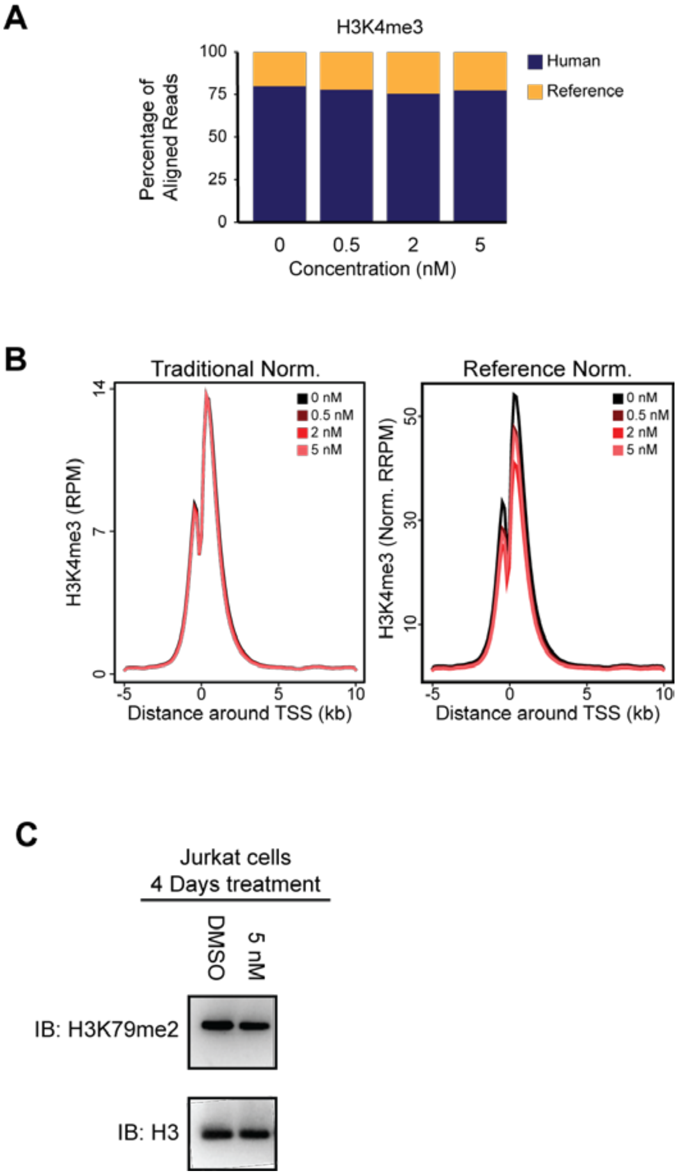
**Figure S4:** Control normalization using a reference epigenome reveals quantitative occupancy changes in a replicate experiment, related to Figure 3.



Percentage of ChIP-seq reads from H3K79me2 **(A)** or H3K4me3 **(B)** experiments aligning to human (Blue) or reference (Orange). ChIP-Seq signal for H3K79me2 **(C)** and H3K4me3 **(D)** at the RPL13A gene locus in standard Reads Per Million (RPM; Top) or Reference-adjusted Reads Per Million (RRPM; Bottom). Color indicates percentage of sample treated with EPZ5676. Gene model is shown below the track. Meta-gene profile across the top 5000 protein-coding genes as defined by total H3K79me2 signal **(E)** or total H3K4me3 **(F)** signal in 0% treated sample (no EPZ5676). Region shown is -5kb to +10kb around the TSS for traditional (Left) and reference (Right) normalized data. Line graphs display the observed fold-change difference in mean signal across the -5 to +10kb window for each H3K79me2 **(G)** or

H3K4me3 **(H)** sample (x-axis) relative to the signal from the 0% treated population using standard (Gray) or reference (Black) normalization.

**Figure S5:** H3K4me3 Control metagenes for Figure 4 (MV4;11 ChIP-Rx experiment), related to Figure 4.



**A)** Percentage of ChIP-seq reads from H3K4me3 experiments aligning to human (blue) or reference (orange). **B)** Meta-gene profile across the top 5000 protein-coding genes as defined by total H3K4me3 signal in 0nM sample. Region -5kb to +10kb around the TSS is shown for H3K4me3 for traditional (Left) and reference (Right) normalized data. **C)** Western blot validation of unchanging H3K79me2 levels in Jurkat cells following 5nM treatment with EPZ5676 for 4 days. H3K79me2 or total histone H3 (control) levels were measured by immunoblot.

## Supplemental Tables:

**Table S1:** Overlap of *drosophila* genome with other organisms

| **Organism** | **% of *drosophila* 50mer's aligning** | **Build** |
|---|---|---|
| *Drosophila melanogaster* | 95.20% | dm3 |
| *Mus musculus* | 0.35% | mm10 |
| *Rattus norvegicus* | 0.34% | rn5 |
| *Homo sapiens* | 0.32% | hg19 |
| *Gorilla gorilla* | 0.31% | gorGor3 |
| *Danio rerio* | 0.30% | danRer7 |
| *Vicugna pacos* | 0.24% | vicPac1 |
| *Arabidopsis lyrata* | 0.15% | Phytozome_9.0_107 |
| *Schmidtea mediterranea* | 0.15% | ASM18107v1 |
| *Candida albicans* | 0.09% | Ca19 |
| *Arabidopsis thaliana* | 0.08% | tair10 |
| *Saccharomyces cerevisiae* | 0.07% | sacCer3 |
| *Saccharomyces castellii* | 0.06% | ASM23734v1 |
| *Caenorhabditis elegans* | 0.05% | ce6 |
| *Toxoplasma gondii* | 0.05% | ToxoDB-10.0_TgondiiGT1 |
| *Schizosaccharomyces pombe* | 0.04% | ASM294v2 |

Every 50bp segment of the *drosophila* genome was aligned against the genome of various model organisms (rows) using (Langmead et al., 2009) (-e 70 -k 1 -n 2 --best –chunkmbs 200). The percentage of the 50bp segments (Column 2), and the particular genome build aligned against (Column 3) are shown.

**Table S2:** Datasets used in this paper, their normalization factors and additional statistics

List of ChIP-seq datasets generated for this manuscript (rows). For each dataset, the sample number, title, GEO accession ID, histone modification profiled, organism, cell-type are given. Where appropriate, the ratio of human:*drosophila* cells, the percentage of human sample comprising EPZ5676 treated cells, and the concentration of EPZ5676 used is listed. The read length, number of reads mapping to human or *drosophila*, and well as the derived traditional (Reads per million [human] mapping reads) or reference based normalization factors (Reference-adjusted Reads Per Million, RRPM) are listed.

The total number of sequenced reads, as well as number of uniquely or duplicate reads aligning to human or *drosophila* are given.

## Extended Experimental Procedures:

### ChIP-Rx (Chromatin ImmunoPrecipitation with Reference exogenous genome).

*Human cell lines, growth and treatment*

Jurkat cells were obtained from ATCC and maintained in RPMI (Life Technologies) supplemented with 10% FBS (Life Technologies) at 5% $CO_2$ in 37°C. MV4;11 cells were obtained from ATCC and maintained in RPMI (Life Technologies) supplemented with 10% FBS at 5% $CO_2$ in 37°C.

Jurkat cells were treated with DMSO or EPZ5676 (Selleck Chemicals; Cat #S7062) at 5 nM or 20 µM for 4 days, and MV4;11 cells were treated with DMSO or EPZ5676 at 0.5 nM, 2 nM, or 5 nM for 4 days. Live cell numbers were quantified using Countess cell counter (Life Technologies).

At harvest, cells were crosslinked with 1% formaldehyde by addition of 1/10 volume of fresh 11% formaldehyde solution (11% formaldehyde 0.1M NaCl, 1mM EDTA, 0.5mM EGTA, 50mM HEPES pH 8.0) and incubation at room temperature for 8 min. Crosslinking reactions were quenched with 1/20 volume of 2.5M Glycine for 1-5 min and cells pelleted. Cells were then washed 3x with ice cold PBS. Washed cell pellets were flash frozen and stored at -80$^{O}$C.

*Preparation of drosophila S2 cells*

*Drosophila* S2 cells (ATCC catalog number CRL-1963; Biovest part #OO.763/OO.627) were cultured in Schneider's drosophila media (Life Technologies catalog number 21720-024) supplemented with 10% FBS to attain a density of 0.5-0.6 x $10^6$ cells/ml. Cell culture and scale up to 2 liters was performed by Biovest International, Minneapolis, MN.

At harvest, cells were crosslinked with 1% formaldehyde by addition of 1/10 volume of fresh 11% formaldehyde solution (11% formaldehyde 0.1M NaCl, 1mM EDTA, 0.5mM EGTA, 50mM HEPES pH 8.0) and incubation at room temperature for 8 min. Crosslinking reactions were quenched with 1/20 volume of 2.5M Glycine for 1-5 min and cells pelleted. Cells were then washed 3x with ice cold PBS. Washed cell pellets were flash frozen and stored at -80$^{O}$C at 1 x $10^8$ cells per aliquot. Cross-linking was performed by Biovest International, Minneapolis, MN.

*Chromatin Immunoprecipitation – combining query and reference (S2) cells*

For each ChIP-Rx experiment a 2:1 ratio of human:*drosophila* cells was used. This corresponds to 20 million crosslinked human cells and 10 million crosslinked S2 cells (Jurkat experiments) or 15 million crosslinked human cells and 7.5 million crosslinked S2 cells (MV4;11 experiments). Note: Given our observed alignment rates, it may be possible to use larger ratios of human to *drosophila* cells (i.e. 5:1) in future experiments if desired (See *Expected percentage of read alignments* section, below). Note:

conditions and time of crosslinking should be the same for interrogated human cells and reference *drosophila* S2 cells.

S2 cells were added to human cells at the beginning of the ChIP-Rx workflow (During nuclei isolation). Once drosophila S2 and human cells were combined, the sample was treated as a single ChIP-seq sample throughout the experiment to completion of DNA sequencing. Briefly, frozen, crosslinked human and drosophila cells were resuspended in parallel in cold lysis buffer 1 (140mM NaCl; 1mM EDTA; 50mM HEPES pH 7.5; 10% Glycerol; 0.5% NP-40; 0.25% Triton-X-100) containing protease inhibitors (COmplete, Roche catalog #11697498001). Human and drosophila cells were incubated in lysis buffer 1 for 10 min at 4°C, and pelleted at 2,000 RPM x 5 min. Both human and *drosophila* cell samples were resuspended in parallel in lysis buffer 2 (10mM TRIS pH 8.0; 200mM NaCl; 1mM EDTA; 0.5mM EGTA), incubated for 10 min at 4°C, and combined to the desired cell number ratios (2 human cells per 1 drosophila cell). The composite cell nuclei was then pelleted at 2,000 RPM x 5 min. Composite cell pellets were resuspended in sonication buffer (10mM TRIS pH 8.0; 1mM EDTA; 0.1% SDS).

### *Chromatin Immunoprecipitation – sonication and immunoenrichment*

Standard ChIP was performed as described previously, with some modification (Guenther et al., 2008). Briefly, composite samples (Human + S2) in sonication buffer (Described above; 10nM Tris pH 8.0, 1mM EDTA, 0.1% SDS) were sonicated using a Covaris E220 sonication waterbath for 5 minutes. Sheared chromatin, the bulk of which was sized 200-2000bp was diluted 1:1 in 2x dilution buffer (300mM NaCL; 2mM EDTA; 50mM TRIS pH 8.0; 1.5% Triton-X; 0.1% SDS) and incubated with either H3K79me2 (abcam 3594) or H3K4me3 (Millipore 07-473) -conjugated Protein G Dynal beads (Invitrogen) for overnight (8-16 hours) rotation at 4°C, and then washed 2x with wash buffer 1 (50mM HEPES pH 7.5, 140mM NaCl, 1mM EDTA, 1mM EGTA, 0.75% Triton-X, 0.1% SDS, 0.05% DOC), 2x with high-salt Wash buffer (50mM HEPES pH 7.5, 500mM NaCl, 1mM EDTA, 1mM EGTA, 0.75% Triton-X, 0.1% SDS, 0.05% DOC) and 1x with TE-NaCl buffer (10mM Tris pH 8.0, 1mM EDTA, 50mM NaCl). The samples were eluted from beads for 1 hour at 65°C in elution buffer (50mM TRIS pH 8.0; 10mM EDTA; 1% SDS) and supernatant reverse-crosslinked at 65°C for 6-16 hours. Samples were diluted 1:1 with TE buffer (50mM TRIS pH 8.0; 1mM EDTA) and treated with RNAse A (0.2mg/ml) for 2 hours at 37°C and then Proteinase K (0.2mg/ml) for 2 hour at 55°C. DNA was isolated by phenol-chloroform extraction and ethanol precipitation.

### *Library construction, sequencing and data collection*

Libraries were constructed with Illumina Tru-Seq library preparation kit using a target fragment size of 200-400bp and multiplexing barcodes. Libraries were sequenced using Illumina HiSeq 2000 with single-end reads for 40 cycles. Sequences were de-multiplexed and aligned using Bowtie2 against a “genome” which combines the human hg19 genome and the *drosophila* dm3 genome (see Alignment of ChIP-seq samples with *drosophila* reference epigenome). Data for this manuscript has been deposited in

GEO with the accession number GSE60104. Individual accession numbers and read statistics available in Table S2.

## Calculation of *drosophila* genome overlap with other genomes

To determine the overlap of the *drosophila* genome with other genomes, we created a FASTQ with “reads” representing the *drosophila* genome. This FASTQ was generated by sliding a 50bp window across the *drosophila* genome (build dm3) in 1bp steps and using the sequence in each window as a 50bp “read”. This generated a FASTQ with 168,735,787 “reads”, representing every 50bp segment of the *drosophila* genome.

This FASTQ was aligned against each tested organism’s genome using (Langmead et al., 2009) (parameters: -e 70 -k 1 -n 2 --best –chunkmbs 200). The percentage of the *drosophila* genome that aligns to each organism, as well as the genome build used are described in Table S1. We find that there is a very small amount of genomic overlap between *drosophila* and the tested genomes (0.04% - 0.35%). However as overlap is not exactly 0%, this implies there may be a small set of regions where ChIP-Seq signal may cross-map between organisms. It is also important to note that due to the presence of unresolved base-pairs (i.e. N’s) in the dm3 genome build, there are ~5% of 50bp “reads” that do not align to the *drosophila* genome using the specified bowtie parameters.

## Alignment of ChIP-seq samples without *drosophila* reference epigenome

Sequenced reads from the human ChIP-seq experiment in pure Jurkat cells (Table S2, Sample #2-6), were aligned to the human genome (hg19) using (Langmead and Salzberg, 2012)(v 2.0.5) with the default (--sensitive) parameters. Sequenced reads from the H3K79me2 ChIP-seq experiment in pure S2 cells (Table S2, Sample #1), were aligned to the *drosophila* genome (dm3) using (Langmead and Salzberg, 2012)(v 2.0.5) with the default (--sensitive) parameters.

## Alignment of ChIP-seq samples with *drosophila* reference epigenome

The genome sequences for human (hg19) and *drosophila* (dm3) were concatenated to produce a combined genome sequence. To avoid chromosome name duplication, all *drosophila* chromosome names had the “_dm3” suffix added to them. A custom Bowtie2 library was built from this combined genome sequence using the “bowtie2-build” command. All sequenced reads from samples that had reference added (Table S2, Samples 7-50) were aligned against this custom library using (Langmead and Salzberg, 2012)(v2.0.5) with default parameters (--sensitive).

The resulting SAM files were then split, such that reads aligning to human chromosomes were placed in one file and reads aligning to *drosophila* chromosomes (those ending in _dm3) were placed in another. Finally, the *drosophila* SAM files were modified to remove the “_dm3” suffix to allow the alignments to be visualized in a browser which uses standard *drosophila* chromosome names.

The number of reads mapping to human or *drosophila* for each experiment are listed in Table S2.

## Effect of *drosophila* reference cells on human cell ChIP-seq

To determine if adding *drosophila* S2 reference cells to a human cells would negatively affect human ChIP-Seq data we analyzed 3 pairs of ChIP-Seq data wherein one sample included S2 reference cells and the other sample did not: 1) H3K79me2 in Jurkat cells not treated with EPZ5676 (Table S2: Samples 7,3), 2) H3K79me2 in Jurkat cells where 50% of the population was treated with EPZ5676 (Table S2: Samples 9,2), and 3) H3K4me3 in Jurkat cells not treated with EPZ5676 (Table S2: Samples 12,4). Reads derived from ChIP-seq samples were aligned and the data was visualized in the IGV genome browser. ChIP-seq profiles for all datasets in a representative region of the genome is shown in Figure S1A. By visual inspection it is clear that the profiles within a pair are highly similar, indicating that the addition of *drosophila* S2 reference cells does not impair the ability to perform ChIP-Seq.

We then performed a more detailed analysis of the H3K79me2 data pair from untreated Jurkat cells (Table S2: Samples 7,3). We calculated the RPM/bp signal for each dataset in a region -5/+10kb around all protein coding genes (Figure S1B). The pearson correlation was 0.967. We further took each of these -5kb/10kb regions and calculated the RPM/bp each 100bp window individually. This gave a vector of 2,969,500 counts (150 100bp windows in each of 19797 genes). The pearson correlation of these vectors was 0.901.

We next used MACS2 (Liu, T; MACS v2) to identify high-confidence peaks in each sample, using the parameters “-f BAM –g hs –keep-dup 1 p 1e-09”. We found a very similar number of peaks identified, with 24,802 peaks identified in the sample without *drosophila* S2 reference cells, and *22*,608 peaks in the sample with *drosophila* S2 reference cells added. Taken together, there were 24,249 unique regions identified between the two samples, with the majority of peaks (12,664) identified in both samples (Figure S1C). We also calculated the correlation of ChIP-Seq signal across the union of identified peaks. Each of the 24,249 peaks were binned into 100bp windows and the RPM/bp in each window was calculated. The 24,249 peaks resulted in 27,377,121 100bp windows and the Pearson correlation across these windows was 0.814.

To determine if the addition of reference S2 cells may impair the ability to detect peaks with low signal, we analyzed the overlap of peaks, binning by peak strength. We divided the 24,802 peaks from the no reference sample into 5 quantiles (100-80%, 80-60%, 60-40%, 40-20%, 20-0%) based on macs2 pileup values. We then calculated the percentage of peaks in each bin that are overlapped by peaks identified in the sample with reference cells added (Figure S1D). We find that for the strongest peaks there is nearly perfect overlap (99% for 100-80% and 90% for 80-60%). Even at the lowest level of signal (20-0%) we still detect more than half (56%) of called peaks. Taken together, these results suggest that the inclusion of *drosophila* S2 reference cells does not induce gross technical artifacts in ChIP-Seq data.

**Calculation of human/*drosophila* read separation accuracy**

To determine the ability to separate human and *drosophila* reads we analyzed H3K79me2 ChIP-seq data from pure populations of Jurkat cells and *drosophila* S2 cells respectively (Samples 3 and 1 in Table S2). We then combined the FASTQs and aligned the combined FASTQ against the combined human/*drosophila* genome (see Alignment of ChIP-seq samples with reference epigenome). For each mapped read we recorded the read’s organism of origin (i.e. which sample that read came from) and the read's organism of alignment. In our dataset, we had a total of 38,801,425 mapped reads, 38,721,726 of which had the same organism of origin and alignment, and 79,699 of which mapped to the incorrect organism. Thus, we were able to separate reads with an accuracy of 99.795%.

### Expected percentage of read alignments

At the outset of experiments, we sought to spike-in enough *drosophila* reference cells to obtain at least 2-5% of the endpoint ChIP-Rx reads as mapping to *drosophila*. To make as few assumptions as possible, we chose to calculate our spike-in ratio based solely on genome size. The human genome is ~20x larger than the *drosophila* genome, so we chose a ratio of 2:1 human to *drosophila* cells (40:1 = 2.5%). However the resulting percentages of ChIP-seq reads mapping to the *drosophila* genome were significantly higher than 1/40 (See Table S2 for percentages, or Figure 3A/B or 4A for examples). For the histone modification ChIP-seq experiments this is expected as the histone marks each occupy a different percentage of the differently sized *drosophila* and human genomes. Consistent with this, directional change in organismal alignment percentages as H3K79me2 is depleted is exactly as expected: As the global level of a histone modification is depleted in human cells, the percentage of total reads mapping to the reference *drosophila* genome increases (Figure 3). We did note some variation in the percentages between replicates but this is most likely due to replicate noise and the fact that the replicate experimental procedures occurred on different days.

We further analyzed the alignment rates (Human vs *drosophila*) for the ChIP input samples containing 2:1 human:*drosophila* cells. We find that these are closer to the theoretical predicted values than the immunoenriched histone modification ChIP-Seq, but still somewhat higher than expected (1:20 actual vs 1:40 predicted for experiments in Figure 1; Table S2). This could be due to a variety of known biological, technical or experimental factors including differential genomic composition of interspersed and tandem repeat elements affecting read mapping or heterochromatin structure affecting sonication among others. It is interesting to note that within a single experiment, the ChIP inputs are highly consistent (average SD < 2%). Given our observed alignment rates, it may be possible to use larger ratios of human to *drosophila* cells (ie 5:1) in future experiments if desired.

### Derivation of normalization factor

To make ChIP-seq data quantitative on a per-cell basis, it is necessary to introduce a reference signal that is constant per cell, from which a normalization factor can be derived. Our ChIP-Rx (Chromatin ImmunoPrecipitation with Reference exogenous genome) protocol uses the signal from a fixed amount of *drosophila* genome per human cell as this reference. We derived a normalization factor, α, for each experiment, such that the resulting *drosophila* signal was equilibrated across all experiments. The mathematical derivation of the normalization factor α is as follows:

Let:

- $\alpha$ = the normalization factor
- $\beta$ = The reference signal from the reference cells

- Nd = The total number of reads (in millions) from a sample aligning to the reference genome
- r = The percentage of the sample (by cell-number) comprised of reference cells

Then, because we the reference signal should always be the same, we can write

$$\beta = \alpha \frac{Nd}{r}$$

And, since β is always the same, we can arbitrarily set it to any value, and for convenience we use 1.

$$1 = \alpha \frac{Nd}{r}$$

Then we can solve for α.

$$\alpha = \frac{r}{Nd}$$

Since r is the same for all experiments in this work, we can further simplify to

$$\alpha = \frac{1}{Nd}$$

Therefore the normalization constant used in the paper is 1 over the number of reads mapping to *drosophila* per million (Reference-adjusted Reads Per Million, RRPM). Traditional normalization would use the value 1 over the number of reads mapping to human per million (Reads Per Million, RPM).

The value of the derived normalization constants, as well as the number of reads mapping to the human and *drosophila* genomes are detailed in Table S2.

**Visualization of ChIP-seq data**

To create the browser tracks shown in Figure 3, 4, and S1 we created bedgraph files from each human SAM file (see Alignment of ChIP-seq samples with reference epigenome) using the macs2 (Liu) "pileup" command with an extension size of 200 (--extsize 200). The values in the bedgraphs were then scaled by either the traditional or reference derived normalization factor with a floor set at 0.1 after scaling. Finally these normalized/scaled bedgraphs were converted into IGV compatible TDF files using "igvtools toTDF" to produce traditional or reference normalized TDF files. These TDF files were then visualized in the IGV(Thorvaldsdóttir et al., 2013) (v.2.3) browser to produce the gene tracks shown.

**Creation of metagenes and calculation of average signal loss**

The set of all ~19,800 human gene TSS's were extracted from Gencode v19 (Harrow et al., 2012) using all protein-coding genes with level 1 or 2. To create the human meta-genes shown in Figures 3,4, S3, S4, and S5 the region -5kb to +10kb around each TSS was separated into 100 equally sized bins and the number of ChIP-seq reads mapping into each bin was calculated (reads were extended 200bp in the direction of alignment). These counts were then scaled by either the traditional or reference derived normalization constant, and finally the column-wise average was visualized. For the metagenes labeled as “Top 5000 genes” (Figure 3,4, S3 [far left], S4, S5), the set of genes analyzed was not all known genes. Rather each gene was scored by total H3K79me2 or H3K4me3 signal in the 0% or 0nM sample within each experiment. The top 5000 genes based on this signal were then selected and used to create the metagene profiles. For the “All Genes” metagenes in Figure S3 (middle, far right), all 19,800 TSS’s (above) were used in calculating the column-wise average.

To create the input-subtracted metagenes shown in Figure S3 (far right, far left), we first needed to calculate the number of reads over ChIP input for each bin for each experiment. First, we counted the number of reads in each bin for the matching ChIP input experiment. Because the read-depth is different between the experiments, we converted the reads counts in each bin to RPM for both the histone mark and the ChIP input. Next, we subtracted the ChIP input RPM from the histone RPM bin count. Finally, we converted back to “raw” read count over ChIP input by multiplying the input subtracted histone RPM count by 1/RPM (reversing the RPM scaling). This matrix of counts over background can now be used as above. The counts were then scaled by either the traditional or reference derived normalization constant, and the column-wise average was visualized. The same set of genes were used for the top 5,000 input-subtracted metagenes as for the non-input-subtracted metagenes in all cases.

The fold-change in signal for a given experiment was calculated as the $\log_2$ of the mean of the metagene signal for that experiment over the mean of the metagene signal for the reference experiment (0% treated for Figure 3, S4, 0nM for Figure 4).

The *drosophila* metagenes shown in Figure S2 were constructed in the same manner as the human metagenes described above, with two differences: 1) Reads were binned into 100 equally spaced bins in the region -2.5kb to +5kb around the TSS. 2) The gene set used was all 16,700 genes which were derived from the gene entries downloaded from FlyBase (r.5.55)(St Pierre et al., 2014).

**Calculation of fold-change in H3K79me2 signal following EPZ5676 treatment**

The total number of ChIP-seq reads mapping into the region -5kb to +10kb around the TSS of every human gene was calculated (See Creation of metagenes and calculation of average signal loss for details of gene set) for the 0nM and 5nM treated samples in MV4;11 and Jurkat cells (Table S2, Samples 37,40,49 and 50). These counts were then scaled by either the traditional or reference derived normalization constant and the $\log_2$ fold change between 5nM/0nM was calculated and visualized in the boxplots.

## References

Guenther, M.G., Lawton, L.N., Rozovskaia, T., Frampton, G.M., Levine, S.S., Volkert, T.L., Croce, C.M., Nakamura, T., Canaani, E., and Young, R.A. (2008). Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. Genes Dev. *22*, 3403–3408.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

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

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

Liu, T. MACS v2: https://github.com/taoliu/MACS/.

St Pierre, S.E., Ponting, L., Stefancsik, R., McQuilton, P., and FlyBase Consortium (2014). FlyBase 102--advanced approaches to interrogating FlyBase. Nucleic Acids Res. *42*, D780–788.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.