*Published online 27 July 2021*

*NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3* **1**
*https://doi.org/10.1093/nargab/lqab064*

# SpikChIP: a novel computational methodology to compare multiple ChIP-seq using spike-in chromatin

**Enrique Blanco [1], Luciano Di Croce [1,2,3,*] and Sergi Aranda [1,*]**

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona 08002, Spain and [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluis Companys 23, Barcelona 08010, Spain

Received February 19, 2021; Revised May 06, 2021; Editorial Decision June 28, 2021; Accepted July 01, 2021

## ABSTRACT

**In order to evaluate cell- and disease-specific changes in the interacting strength of chromatin targets, ChIP-seq signal across multiple conditions must undergo robust normalization. However, this is not possible using the standard ChIP-seq scheme, which lacks a reference for the control of biological and experimental variabilities. While several studies have recently proposed different solutions to circumvent this problem, substantial analytical differences among methodologies could hamper the experimental reproducibility and quantitative accuracy. Here, we propose a computational method to accurately compare ChIP-seq experiments, with exogenous spike-in chromatin, across samples in a genome-wide manner by using a local regression strategy (spikChIP). In contrast to the previous methodologies, spikChIP reduces the influence of sequencing noise of spike-in material during ChIP-seq normalization, while minimizes the overcorrection of non-occupied genomic regions in the experimental ChIP-seq. We demonstrate the utility of spikChIP with both histone and non-histone chromatin protein, allowing us to monitor for experimental reproducibility and the accurate ChIP-seq comparison of distinct experimental schemes. spikChIP software is available on GitHub (https://github.com/eblancoga/spikChIP).**

## INTRODUCTION

The development of chromatin immunoprecipitation (ChIP) coupled with the next-generation sequencing (seq) methodologies has been pivotal for characterizing the genomic distribution of a vast collection of chromatin-associated proteins, histone post-translational modifications (PTMs) and histone variants (1–4), and for building the cartography of functional elements of the human genome in the international collaborative efforts (5–9).

In its traditional scheme, ChIP-seq is essentially a semi-quantitative method that enables the researcher to determine the relative occupancy of one factor in a given genomic region, with respect to the rest of the genome. However, the semi-quantitative nature of the ChIP-seq, as well as, multiple sources of biological and technical variability hamper the direct comparison of ChIP signal strength between different conditions (e.g. cell types, metabolic states or pathological situations) (10). For instance, an increase in genomic occupancy of a chromatin factor could simply be the result of variability in the efficiency of immunoprecipitation between experiments. Moreover, while running the sequencer, a standard practice is to equilibrate the output DNA eluted after the immunoprecipitation by mixing equal proportions of barcoded libraries to run the samples in a multiplexed manner. Therefore, even a substantial global reduction of a histone variant occupancy per cell

*To whom correspondence should be addressed. Tel: +34 9 33160135; Fax: +34 9 33160099; Email: sergi.aranda@crg.eu
Correspondence may also be addressed to Luciano Di Croce. Email: luciano.dicroce@crg.eu

would remain hidden in a ChIP-seq experiment after normalizing by the total number of reads (10). Although the consistent replication of ChIP-seq experiments can reveal the biological tendency in the interacting strength of the chromatin factor, a robust normalizing strategy is required to accurately compare ChIP-seq results across different experimental conditions, such as different cell types and/or stimulus.

Several groups have pioneered different strategies based on the use of internal reference controls (spike-in), which provides a feasible solution to accurately normalize and compare ChIP-seq experiments (11–13). The spike-in strategy in ChIP-seq is based on the initial combination of a set of experimental samples with a fixed amount of exogenous material (e.g. cells or chromatin) from another species. As long as the amount of spike-in added is constant, the number of the reference reads after sequencing is expected to be similar. Therefore, the observable differences in the reads of the experimental samples across conditions can be exclusively attributed to biological variation. If the read numbers of the reference genome after sequencing are not the same, by any technical mean, a normalization factor can be easily calculated *ad hoc* to equilibrate the spike-in signal among samples. The same correction computed from spike-in reads is then used to normalize the experimental ChIP-seq, thus enabling the fair comparison of the ChIP-seq signal across the samples.

Despite this common scheme, the different proposed methodologies differ in the computational approach to correct the spike-in and, consequently, the experimental sample (Supplementary Table S1). For instance, the method named 'ChIP with reference exogenous genome' or ChIP-Rx (11) implements this correction by dividing the total number of mapped reads per million (RPM) from each experimental ChIP-seq (e.g. human) for the corresponding number of spike-in reads (e.g. *Drosophila melanogaster*). Alternatively, the Tag removal method proposes to use the total number of reference reads to compute a scaling factor for random removal of tags from samples with the higher number of reads (see Materials and Methods section) (13). Although initially appealing, these normalization methods present in our opinion important shortcomings, such as (i) the spike-in reads mapped not only along the read enriched regions (named ChIP peaks) but also over the non-enriched background regions are used for computing the correction factor, (ii) the correction factor is uniformly applied to all experimental reads in the actual experiment, treating both non-specific and specific signal loci with the same correction value and (iii) the computational removal of informative reads results in the loss of genomic coverage, thus impacting in downstream analyses.

Recently, with the aim of overcoming the limitations of the previous approaches, Guertin *et al.* proposed a third approach based on a linear local regression method (14). This approach computes a correction coefficient, defined by a linear regression model, for the systematic and gradual correction of the pre-defined ChIP-seq peaks from a reference parallel ChIP-seq. This parallel ChIP-seq is performed to profile the genomic occupancy of a pervasive chromatin factor (e.g. CTCF) whose genomic distribution is assumed to be unchanged between different cell types and/or treatments and which is clearly distinguishable from the experimental target (14). Once computed from the reference parallel ChIP-seq, the same coefficient is also used to correct a pre-defined subset of peaks in the experimental ChIP-seq. The conceptual improvements of this computational approach stem from the fact that the correction factor gradually increases along with the informative power (as number of reads) of the peaks. However, the addition of a computational step to pre-select the real signal loci in this strategy could introduce an additional bias step, as the consistency in the outcome of available peak calling tools is limited (15). In addition, this analysis would impede the genome-wide evaluation of the signal-to-noise ratio, thereby limiting the informative power of the ChIP-seq.

In order to overcome the abovementioned obstacles, we developed a novel computational method that performs the genome-wide normalization of ChIP-seq data adapting the spike-in control correction to the class of genomic region (Figure 1A). In our view, the spikChIP method offers several benefits, as (i) it normalizes ChIP-seq signal over the complete genome, and not just from a subset of selected regions; (ii) in order to compute the correction factor, the influence of the reads from background regions, although dominating over the total number of reads, is minimized; and (iii) the correction factor derived from the spike-in material is not uniformly applied over all the experimental ChIP signal, instead, it is increasingly and gradually applied from background to positive ChIP signal regions.

## MATERIALS AND METHODS

### Description of the spikChIP software

We developed spikChIP as a Perl script that performs the normalization of two or more ChIP-seq experiments with spike-in according to five distinct strategies of correction: raw, traditional, ChIP-Rx, tag removal and spikChIP. Source data, additional documentation and several examples of use are freely distributed in GitHub (https://github.com/eblancoga/spikChIP). In brief, spikChIP is a command-line program running in Linux and Mac OS-X environments that analyzes BAM files of reads previously aligned to a synthetic genome constituted of a sample genome (e.g. human) and a spike genome (e.g. fruit

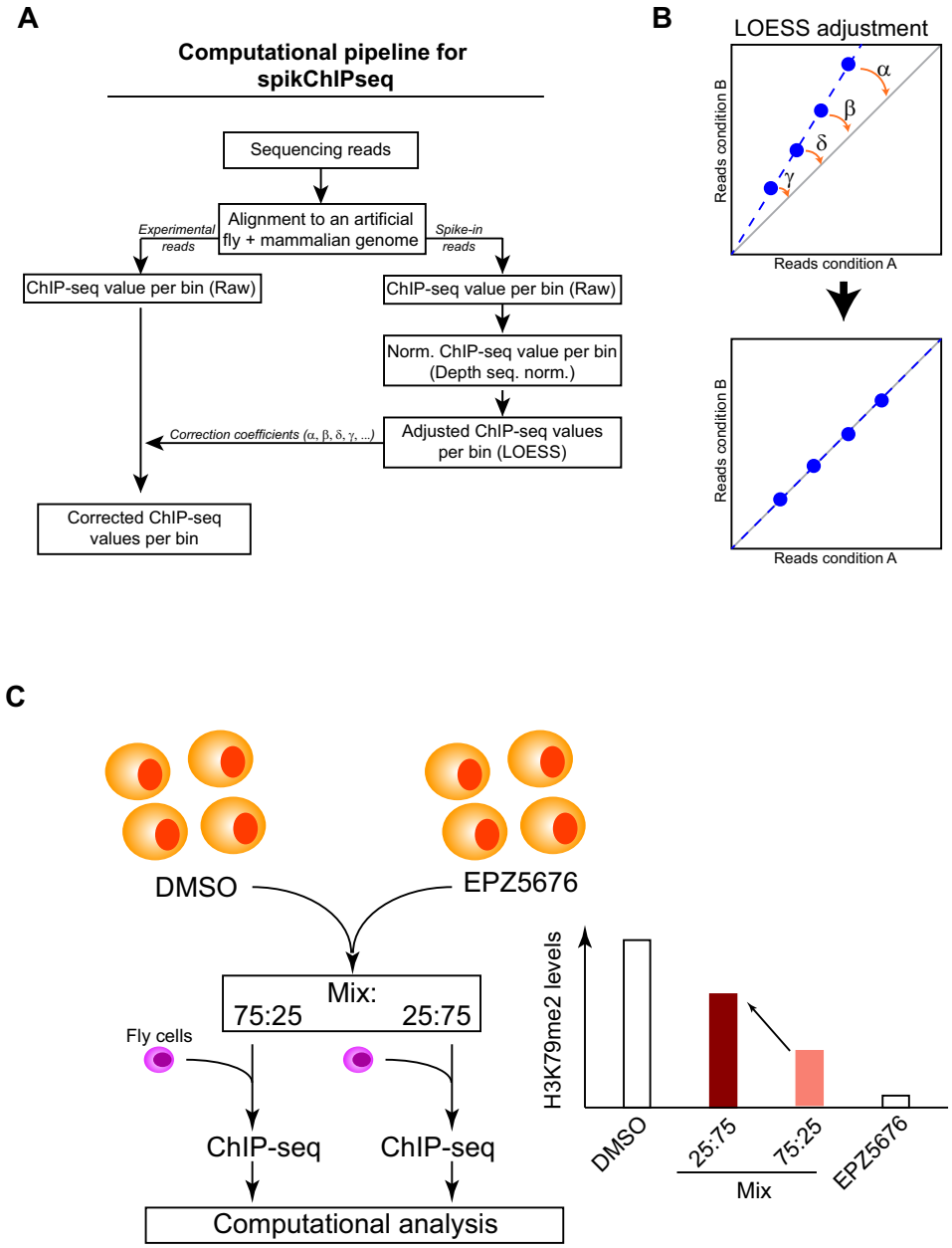


**Figure 1.** Computational pipeline for normalizing ChIP-seq data using spike-in ChIP-signal in a genome-wide manner. (**A**) Diagram summarizing the computational analysis for spikChIP normalization (see Materials and Methods section for details). (**B**) Scheme representing the locally estimated scatterplot smoothing (LOESS) normalization of the spike-in ChIP-seq data (α, β, δ, γ are distinct correction coefficients calculated from specific bins to normalize spike-in data, these coefficients are then used to normalize the experimental ChIP-seq). (**C**) Scheme representing the experimental approach undertaken in (11) to generate a gradual ChIP-seq signal for H3K79me2.

fly) to assign a normalized value of ChIP signal to each bin of both genomes. We implemented two alternative scoring schemes: average value and maximum value within a bin. By taking advantage of the location of ChIP-seq peaks previously computed in BED files, spikChIP is able to classify the bins of normalized values into bins of peaks and background. Users must provide a configuration file with information about the files of reads and peaks that correspond to each ChIP-seq experiment that will be analyzed. In addition, it is necessary to provide a file with the list of chromosome names that were used for the mapping and their sizes in both genomes. The output of spikChIP consists of a series of text files grouped by normalization strategy, which contain the value assigned to each bin of sample and spike-in genomes. File compression is implemented and additional options are available to reduce the final output storage space. Moreover, spikChIP generates boxplots of each corrected distribution of values in both sample and spike-in genomes to evaluate the performance of every strategy of normalization.

### Genome segmentation in bins of the same size

SpikChIP is able to normalize ChIP-seq experiments with spike-in from samples of virtually any organism. For that, users must provide a file with the list of chromosomes of the sample genome and the spike-in genome, together with their sizes. Such files can be easily retrieved from the UCSC genome browser (16). To analyze all the ChIP-seq experiments shown along this work, we downloaded the genome assembly information of human and the fruit fly (hg19 and dm3, respectively) from the following sites:

http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/chromInfo.txt.gz

```
chr1    249250621

(...)
```

http://hgdownload.soe.ucsc.edu/goldenPath/dm3/database/chromInfo.txt.gz

```
chr2L   23011544

(...)
```

We have implemented the option in spikChIP to allow users to select the size of the bins for the segmentation. Here, we performed the normalizations using a bin size of 1 kbp in all the cases shown along this article. Internally, spikChIP generated each segmentation file of non-overlapping bins of 1 kb in BED format using the following GAWK command (showing the command for fly only):

```
% gawk 'BEGIN{OFS = "\t";offset = 999;}{for(i = 1;i<$2-offset;i = i+offset+1) print
$1,i,i+offset;}' chromInfo.txt > fly_1Kb.bed

chr2L 1    1000

chr2L 1001  2000

chr2L 2001  3000

(...)
```

In total, after filtering the Het and M chromosome files out, we ended up with 3 095 665 bins for the human genome and 120 397 bins for the fruit fly genome.

### Synthesis of the human+fruit fly genome for mapping

First, we merged the full set of chromosomes of each genome assembly (hg19 for human and dm3 for *D. melanogaster*) in FASTA format into a single file (genome.fa). Next, to distinguish human from fly sequences after mapping, we appended the tag '_FLY' to the name of the fly chromosomes in the resulting FASTA file and in the chromInfo.txt file described above.

```
>chr1

(...)

>chr2L_FLY

(...)
```

Finally, we used the command bowtie-build from the BOWTIE suite (17) to generate the corresponding indexes for posterior mapping of the ChIP-seq raw data files.

### ChIP-seq raw data files and mapping

We defined our initial dataset of H3K79me2 (NCBI GEO series accession: GSE60104) with the samples Jurkat_K79_25%_R1 (GSM1465005) and Jurkat_K79_75%_R1 (GSM1465007) from (11). We constructed a second dataset of H3K79me2 by integrating the samples Jurkat_K79_0%_R1(GSM1465004), Jurkat_K79_50%_R1 (GSM1465006), and Jurkat_K79_100%_R1(GSM1465008) into the initial set. As a control, we gathered the following samples of H3K4me3 from (11): Jurkat_K4_0%_R1 (GSM1464999), Jurkat_K4_25%_R1 (GSM1465000), Jurkat_K4_50%_R1 (GSM1465001), Jurkat_K4_75%_R1 (GSM1465002) and Jurkat_K4_100%_R1 (GSM1465003). We constructed our H3K27me3 dataset (NCBI GEO series accession GSE64243) from the raw data of the samples PC9_control_H3K27me3_Dmspike (GSM1890165) and PC9_EZH2inh_H3K27me3_Dmspike (GSM1890166) from (13). Finally, to define the Estrogen Receptor-alpha (ER) dataset (NCBI GEO series accession GSE102882), we retrieved the raw data of the samples SLX-8047_1b_ER_none (GSM2747692), SLX-8047_1a_ER_Fulvestrant (GSM2747691), SLX-8047_2b_ER_none (GSM2747694) and SLX-8047_2a_ER_Fulvestrant (GSM2747693) from (14).

Next, for the samples of each dataset, we used BOWTIE (17) to map the FASTQ files of reads over the human+fly genome indexes described above (BOWTIE parameters -p 4 -t -m 1 -S). Finally, we used SAMTOOLS (18) to filter the unaligned reads (option -F 0×4) out and, by using the '_FLY' tag, the mapped reads corresponding to the fly spike-in control were then separated from the human experimental ones into two different BAM files.

### ChIP-Rx and Tag removal normalization of the ChIP-seq data values

For each experiment, spikChIP counted the number of reads within each human and fly bin (average or maximum value, separately) using the function recoverChIPlevels of SeqCode (https://github.com/eblancoga/seqcode). Next, to assign the final value of one bin depending on the normalization method, spikChIP employed the following transformations (formulated below): (i) absolute values, in millions of reads, which were used as raw values; (ii) absolute values divided by the total number of human+fly reads per sample, for the traditional normalization; (iii) absolute values divided by the total number of fly reads per sample, for the ChIP-Rx normalization and (iv) down-sampling of a fraction of human reads from the experiment with more abundance of spike-in reads respecting the same proportion of fly reads that was observed between both conditions, for the Tag removal normalization.

(Raw)

Let X be the number of reads counted on a particular bin B, the raw value per bin was calculated as:

$$\text{Raw (B)} = X / 10^6.$$

(Standard ChIP-seq normalization)

Let X be the number of reads counted on a particular bin B and N be the total number of human and fly mapped reads, the traditional value per bin was calculated as:

$$\text{Standard (B)} = X / N.$$

(ChIP-Rx)

Let X be the number of reads counted on a particular bin B and F be the total number of fly mapped reads, the ChIP-Rx value per bin was calculated as:

$$\text{ChIP-Rx (B)} = X / F.$$

(Tag removal)

Let M and N (M < N) the number of spike-in reads mapped over the fruit fly genome for the two conditions studied on each dataset presented here. We calculated a normalization factor as:

$$R = M / N.$$

Next, we used this R factor to down-sample human reads from the experiment with more abundance of spike-in reads (N), using the command samtools view -h -b -s R -o sample_adjusted.bam sample_original.bam of the SAMTOOLS (18).

### Calculation of local regression normalization of the ChIP-seq data values (SpikChIP)

Local regression methods, in contrast to canonical linear regression techniques, are able to apply different fitting models to particular subsets or segments of the data, which are identified by the nearest neighbor algorithm. Inspired in a similar treatment proposed for RNA-seq normalization of the RPKMs of spike-in controls (19), we applied the LOESS function (LOcally Estimated Scatterplot Smoothing) from the R library affy to the traditional normalization values, to perform the local regression of data. We instructed the loess function normalize.loess to use the adjustment on the values in the fly spike-in genome as a subset to guide the normalization of the human values. A pseudo-count of 0.1 was added to each value before running the normalization function.

More in detail, we concatenated the fly and the human files of bins containing the traditionally normalized values in both conditions 25:75 and 75:25. The first 120 397 lines of this file corresponded to the fly bins (used as a subset to guide the LOESS) and the rest of the lines to the human bins (normalized in the LOESS method by the corrections to adjust the previous subset of bins). Once the normalization was performed, we separated again the human bins from the fly bins into two different files per ChIP-seq experiment. This procedure can be easily generalized to more than two samples, as the LOESS function is applicable to multiple conditions.

### Discrimination of bins associated to peaks and to background

MACS2 with the –broad option (20) was used to identify the list of ChIP-seq peaks along both genomes in the 25:75 and 75:25 conditions for the H3K79me2 dataset, and for the Control and EZH2 inhibitor conditions of the Tag removal dataset. As the reference set of peaks for further analysis at each case, we selected the sample in which a higher number of peaks was reported (25:75 and Control, respectively) To distinguish between bins that contain ChIP-seq peaks and bins that constitute the background, we calculated the overlap between MACS peaks and the coordinates of the segmentation bins at human and fly chromosomes described above with the function matchpeaks of the SeqCode package (https://github.com/eblancoga/seqcode). A similar procedure with MACS2 peak calling was adapted for the rest of datasets: H3K27me3 (broad option), H3K4me3 (default, no broad peaks) and ER (default, no broad peaks).

### Generation of profiles for the UCSC genome browser

Resulting files of normalized values can be converted into BedGraph profiles to upload in genome browsers with a series of simple bash commands. The following spikChIP output line

```
#bin_info       corrected_value_1       corrected_value_2

chr1*1*1001    0.0997764035926034    0.100224097481314
```

can be translated into the next BedGraph lines (one per condition/profile)

```
chr1  1    1001  0.0997764035926034
```

and

```
chr1  1    1001  0.100224097481314
```

with the independent commands below:

```
zcat results/FINAL_EXAMPLE_SPIKCHIP_1000_avg_normalized_sample.txt.gz | sed 's/\*/ /g' |
gawk '{print $1,$2,$3,$4}'

zcat results/FINAL_EXAMPLE_SPIKCHIP_1000_avg_normalized_sample.txt.gz | sed 's/\*/ /g' |
gawk '{print $1,$2,$3,$5}'
```

## RESULTS

### Benchmarking spikChIP: a novel local regression method for comparative analysis of ChIP-seq in a genome-wide manner

Our approach, inspired by the spike-in-based RNA-seq quantification methods (19,21), shares conceptual similarities with the recent linear local correction approach (14) and consists in the application of a local regression, in this case, over the read counts of all the genome-wide bins determined along the chromosomes (see Materials and Methods section). Thereby, our method can introduce a distinct correction factor to each bin in the genome, depending on its class. First, a local regression (LOESS) is computed from the read counts on bins in the spike-in genome in order to accommodate the two ChIP-seq conditions compared into the same best-fit line (Figure 1A and B). Next, the values from the real experiment (experimental reads) are corrected following the previous local normalization calculated using the spike-in bins (Figure 1A). Under this approach, the adjustment on a region containing a true ChIP-seq signal is expected to be substantially higher than the change computed for bins with a background signal.

To assess the accuracy of our proposal, we compared the performance of spikChIP with the ChIP-Rx (11) and Tag removal (13) strategies on a reference dataset. We took advantage of the available ChIP-seq data published by Guenther *et al*. that included fly material as spike-in control (11). In this study, the authors artificially generated a pre-defined ChIP signal gradient for the di-methylation of lysine-79 of histone H3 (H3K79me2), a histone mark deposited on the initial 5′-end of genes and that typically constitutes a small fraction (∼2.8%) of total histone H3 in cancer cells (22). To achieve a controlled range of distinct conditions, they mixed different proportions of Jurkat cells that had been untreated or treated with a selective inhibitor for the H3K79-methyltranferase DOT1L (EPZ5676). The mixture aims to reflect the global change in the average H3K79me2 level per cell. Finally, a constant amount of fly cells was used as an internal reference control for normalization. Once mixed, the sample and the spike-in material are captured using the same antibody against H3K79me2, since the epitope is highly conserved between human and flies.

For our benchmarking, we initially selected two intermediate conditions: (i) the 25:75 (DMSO:EPZ5676) proportion, which has higher levels of H3K79me2 and (ii) the 75:25 proportion, with lower levels of H3K79me2 (Figure 1C). As indicated in our computational pipeline (Figure 1A), first we mapped the resulting sequencing reads to an artificial genome in which we included the human and the fruit fly chromosomes (see Materials and Methods section). Next, we segmented the genomes of the sample (human) and the spike-in control (fly) into bins of 1 kbp. The selection of this size was based on finding a balance between ChIP-seq resolution, and the computational memory demand and running time required for a more compact bin segmentation. After separating the mapped reads into human and fly, we calculated the corresponding ChIP-seq value of H3K79me2 in both conditions within all bins from both genome segmentations (see Materials and Methods section). These initial values, which were not corrected by any normalization method, were considered to be the raw value (Figure 1A). We then employed the spike-in raw reads to compute the normalization based on the ChIP-Rx, Tag removal, or our spikChIP approaches (Figure 2A). For ChIP-Rx and Tag removal correction, we used the total number of aligned fly reads to calculate a normalizing factor (ChIP-Rx) or scaling factor (Tag removal), to equilibrate the ChIP signal or the number of reference reads, respectively, as previously suggested (11,13). For spikChIP, we first normalized the spike-in sample for the total number of reads and then applied the LOESS correction in the fly bins to the best fit-line (Figure 2A and B). As when applying spikChIP, the normalization factors applied over the reference genome (bins of 1 kbp) depend on the signal strength, our method results in a slight improved equilibration of spike-in ChIP signal between samples (see slope on Figure 2).

Then, we used the same correction factors to normalize the human experimental bins (Figures 3 and 4). An appropriately analytical normalization using spike-in should display a qualitative and quantitative difference between both experimental ChIP signals in ChIP-seq peaks of H3K79me2 (∼2% of total read, Figure 4A), while keeping their background levels equilibrated, respectively. As shown in Figure 3, either using the ChIP-Rx, Tag removal or spikChIP normalization, ChIP signal strength on enriched loci (Figure 3, the genomic region marked with a gray box) is remarkably different between the 25:75 and the 75:25 samples. However, after a detailed inspection over the background regions, we consistently found a sustained increase also in such areas when using the ChIP-Rx or Tag removal correction (Figure 3, vertical scaling). These genome-wide changes observed in the target occupancy over the background are misleading since the genomic distribution of H3K79me2 must be considered to be equivalent between both samples, resulting from mixing the same samples with different proportions. Instead, when using spikChIP, the ChIP signal at background regions remains equivalent between the samples (Figure 3).

Indeed, the genome-wide difference in the ChIP signal (25:75 versus 75:25) over the bins belonging to H3K79me2 peaks is consistently increased, either using ChIP-Rx, Tag removal normalization or the spikChIP approach (Figure 4B and C). However, the ChIP-Rx and Tag removal normalizations result in a disproportionate correction over the non-enriched background genomic bins (∼20–25% difference in the median of ChIP-seq signal, Figure 4B and C). Strikingly, the ChIP-seq

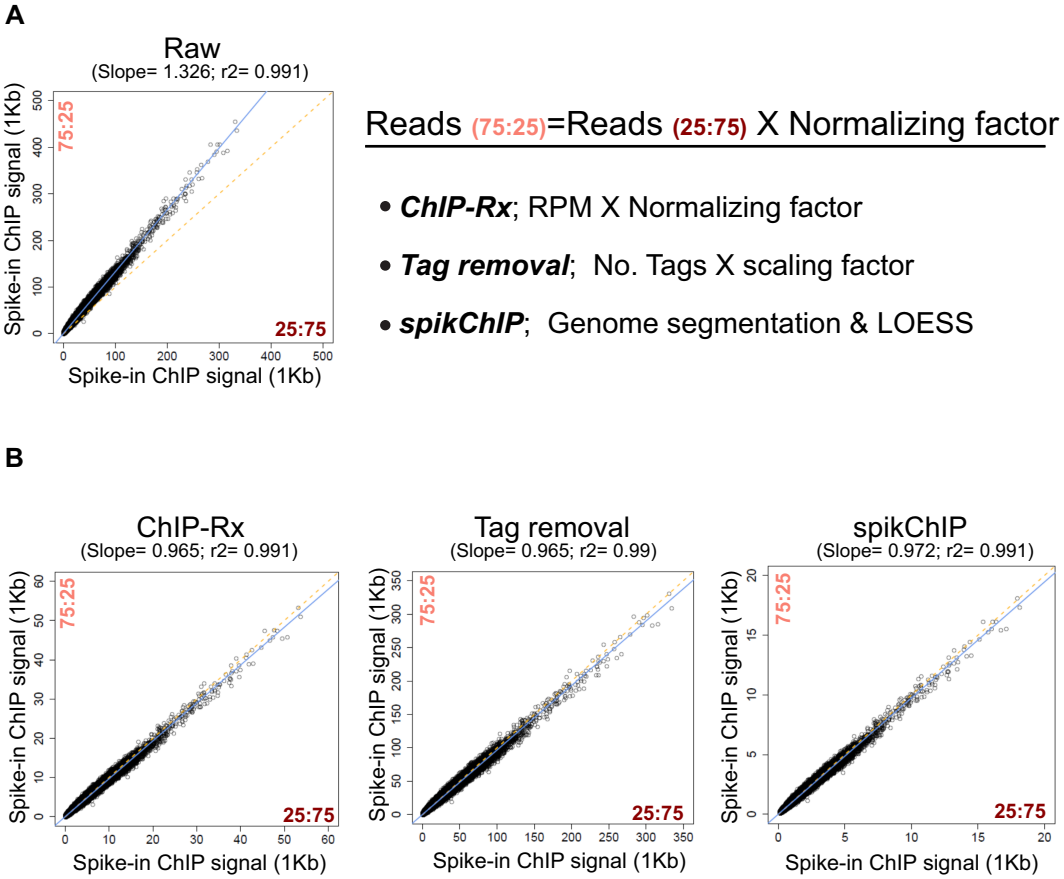


**Figure 2.** ChIP-Rx, Tag removal and spikChIP normalization of the reference genome. The Drosophila genome is segmented in 1 kb bins. For each bin (dot), the average ChIP-seq signal for H3K79me2 on 75:25 ($y$-axis: 8 554 870 reads) versus 25:75 ($x$-axis: 6 226 309 reads) is represented before normalization (**A**, raw) and (**B**) after ChIP-Rx, Tag removal and spikChIP normalizations. Linear regression line is depicted in blue and the $y = x$ line in dotted orange.

signal at background regions remains equilibrated after spikChIP correction (<5% difference in the median of ChIP-seq signal) between both conditions (Figure 4B and C). These results indicate that spikChIP enables the normalization of ChIP-seq experiments, with exogenous spike-in, in a genome-wide manner, by applying a gradual and progressive correction of ChIP-seq signal from background to enriched read regions.

### SpikChIP application to multiple samples and distinct histone marks

Under certain circumstances, the comparison of ChIP-seq data will involve more than two samples. To fulfill this need, we have generalized the spikChIP method to compare any number of ChIP-seq experiments and generate the resulting normalization values by applying the same local regression function to multiple samples at once. As shown in Supplementary Figure S1a, after spikChIP normalization we detected gradual reduction of H3K79me2 ChIP signal from untreated control to EPZ5676-treated cells, including intermediate samples with different proportion of these cells (25:75, 50:50 and 75:25) up to a total of five samples. While the gradual ChIP signal reduction is observed at enriched peak loci, background regions remain equilibrated when applying spikChIP, but not when using ChIP-Rx or Tag removal strategies. Importantly, by normalizing H3K4me3 ChIP data from these cells, a histone mark unaffected by the EPZ5676 treatment (11), we detected equivalent ChIP-seq signal levels in both peak and background regions (Supplementary Figure S1b).

To rule out that this major improvement in the detection of H3K79me2 changes over peaks and not over background regions is exclusive of such a particular histone mark, we set out to additionally evaluate the performance of spikChIP, using an independent second dataset (13). Enhancer of zeste homolog 2 (EZH2) is a histone-lysine N-methyltransferase enzyme responsible for the methylation of lysine 27 on histone H3 (H3K27) (23). GSK126 is a potent selective inhibitor of EZH2, which induces a global reduction of H3K27me3 in cancer cell lines (24). In particular, Egan and collaborators showed a global reduction of H3K27me3 in PC9 lung adenocarcinoma cells treated 5 days with 1 μM of GSK126 (13). The authors performed chromatin preparation from both control and GSK126 treated cells, which was mixed with an equal

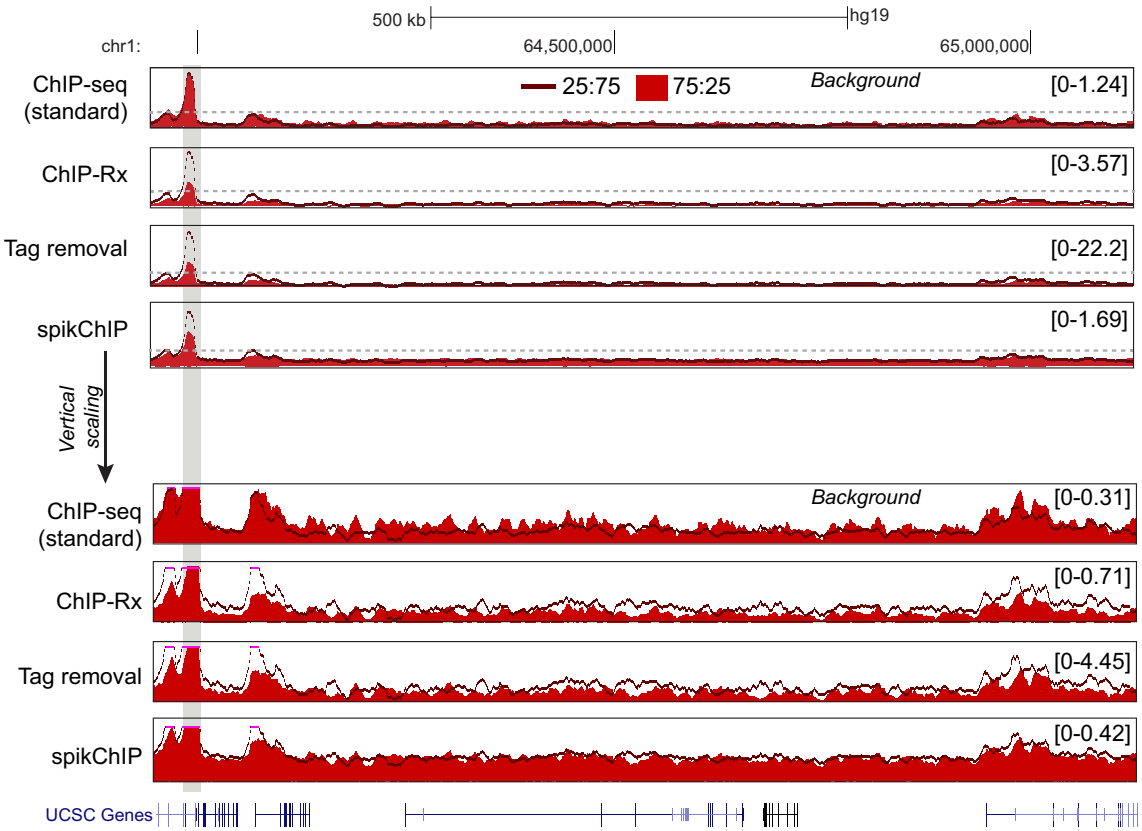


**Figure 3.** UCSC epigenome track. Overlayed ChIP-seq tracks, at different scales, from H3K79me2 on 25:75 and 75:25 samples, normalized by standard ChIP-seq normalization (ChIP-seq), ChIP-Rx, Tag removal or spikChIP (assembly: hg19, coordinates chr1:63,939,496–65,118,855). Using standard normalization, the global reduction of H3K79me2 is undetectable after normalizing by the total number of reads (standard normalization). While, after spike-in normalization, global changes in H3K79me2 occupancy, as the results of mixing samples, are detected. Note the differences in ChIP-seq signal over the background zone when applying ChIP-Rx or Tag removal, while remain equilibrated when applying spikChIP. Dashed lines indicate the limit for the vertical scaling.

amount of *Drosophila* chromatin, as spike-in. Then, ChIP was performed, using a mixture of two specific antibodies against the H3K27me3 and *D. melanogaster*-specific histone variant His2Av (H2Av), and the resulting material was processed by massive parallel sequencing (13). Similarly to the previous H3K79me2 dataset, we processed the read counts of H3K27me3 ChIP-seq in both conditions using ChIP-Rx, Tag removal and spikChIP normalization strategies (Supplementary Figure S2). Nevertheless, the impact of spikChIP on equilibrating the reference ChIP signal between both samples was striking for the gradual normalization applied on genomic bins (Supplementary Figure S2; *P*-value for the remaining differences in the spike-in ChIP signal distribution after computational normalization, Wilcoxon test; ChIP-Rx *P*-value<2.2e-16; Tag removal *P*-value = 9.611e-14; SpikChIP *P*-value = 0.01547). When the normalization factors were applied to the experimental ChIP, we clearly detected the reduction in H3K27me3 ChIP-seq signal upon GSK126 inhibition over the enriched peak regions (Supplementary Figure S3). However, while ChIP-Rx and Tag removal normalization showed an equivalent reduction of the signal in either peaks and background regions (∼81–85%), the use of spikChIP enabled to detect a greater effect of GSK126 over peaks with respect to non-enriched regions (69% and 23%, respectively, Supplementary Figure S3), thus affecting the biological interpretation of the results.

### SpikChIP for comparing transcription factor occupancy across samples

Since its development, the ChIP-seq technique has become one of the most widely-used methods in molecular biology (25), and the universal method to delineate the genome-wide maps of the distribution of histone, as well as non-histone proteins (e.g. transcription factors [TF] and chromatin remodelers). In addition to the analysis with histone PTMs, we envisioned that spikChIP could also be applied to accurately compare the genomic occupancy of non-histone proteins, such as TF that typically display a precise occupancy over their DNA binding sites. To test our hypothesis, we used available ChIP-seq data for a ligand-activated TF, the estrogen receptor-alpha (ER) (14). This dataset includes ChIP-seq data from breast cancer cells MCF-7 treated with a selective estrogen receptor (ER) degrader, fulvestrant, thus allowing to compare ER genomic

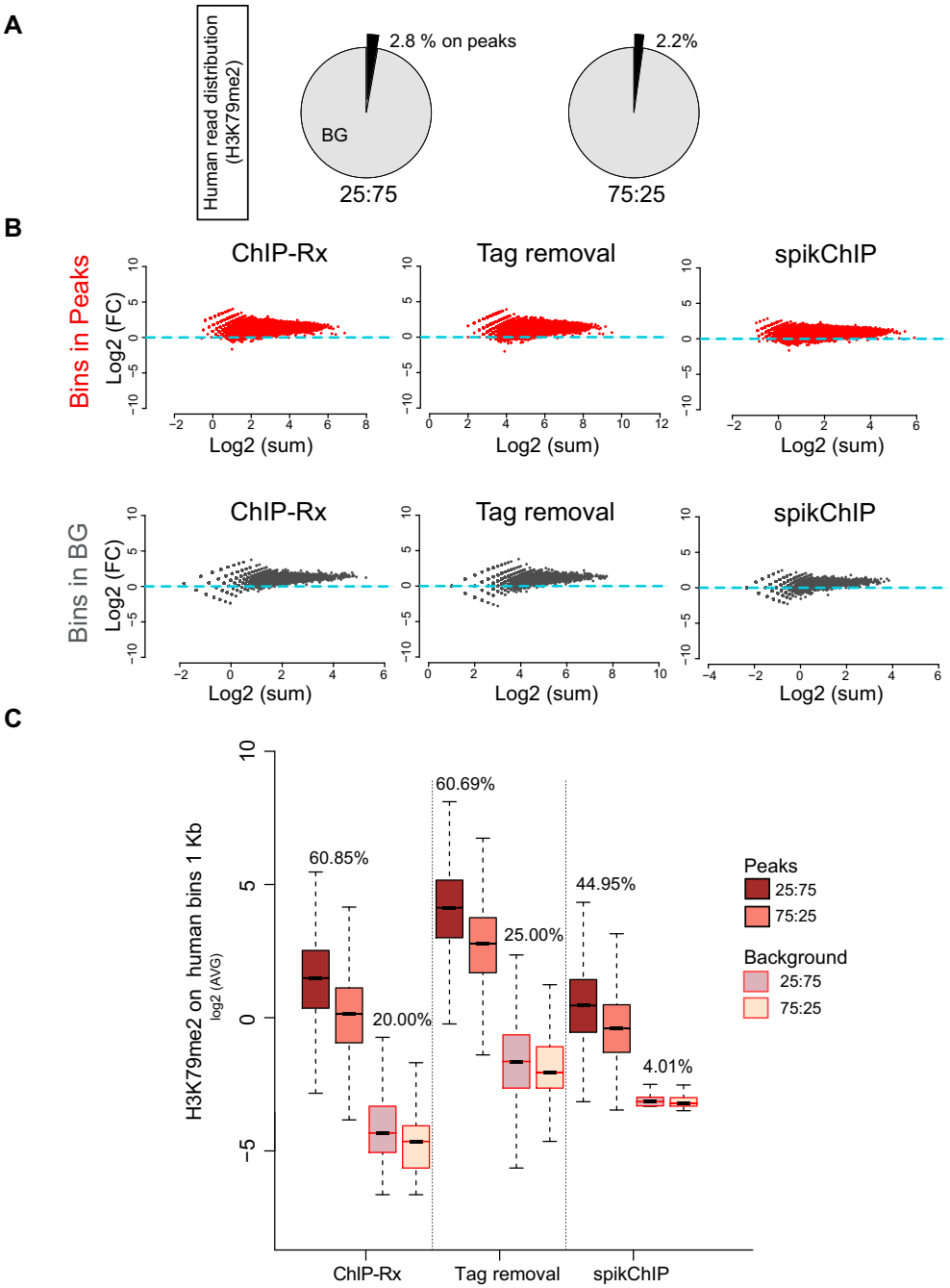


**Figure 4.** Experimental H3K79me2 ChIP-seq normalization by ChIP-Rx, Tag removal and spikChIP. (**A**) Distribution of the total experimental human reads on enriched peaks and non-enriched background regions in the 25:75 and 75:25 samples. Note the low percentage of reads that fall into enriched regions. (**B**) The scatter plots representing the Log2_fold change (25:75 versus 75:25) and the Log2_sum (25:75 and 75:25) ChIP-seq after ChIP-Rx, Tag removal and spikChIP normalization. Each dot represents a 1 kb bin. All bins belonging to peak regions according to MACS2 are shown (red dots) together with a comparable number of bins randomly selected from background regions (gray dots). (**C**) Boxplots representing the distribution of H3K79me2 ChIP-seq signal after ChIP-Rx, Tag removal and spikChIP normalization. The percentage of the difference between the medians of boxplots is indicated.

occupancy between untreated control and ER-depleted cells (14). Finally, untreated and fulvestrant-treated cells were pre-mixed with fly chromatin as spike-in reference captured with the specific H2Av antibody, therefore enabling to test the ability of spikChIP to normalize ChIP-seq from non-histone proteins.

We process the ChIP-seq data similarly to histone PTMs, for ChIP-Rx, Tag removal and spikChIP normalization. As in the case of H3K27me3 dataset, the spike-in signal adjustment resulted in a more similar distribution between samples when applying spikChIP (*P*-value for the remaining differences in the spike-in ChIP signal distribution after computational nor-

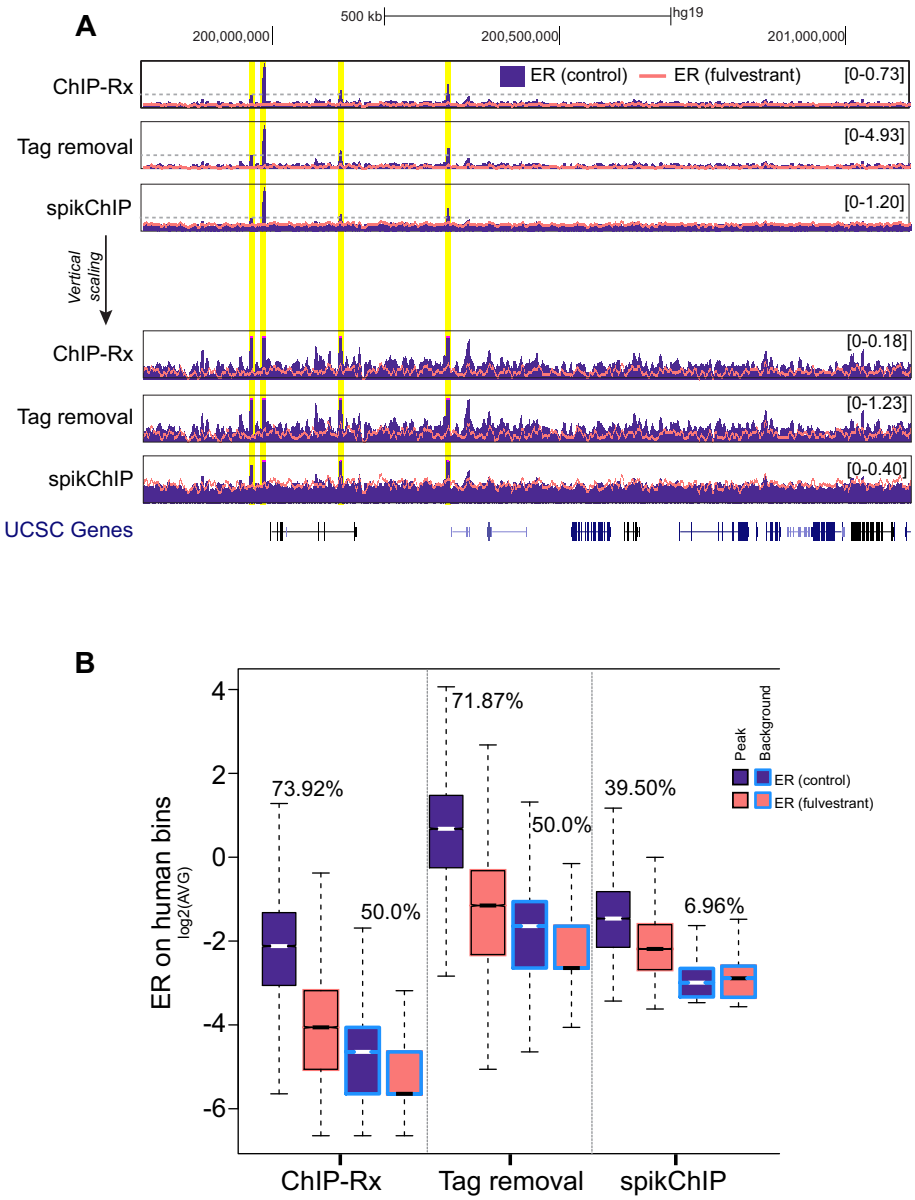


**Figure 5.** Experimental ER ChIP-seq normalization by ChIP-Rx, Tag removal and spikChIP. (**A**) Overlayed ChIP-seq tracks, at different scales, from ER ChIP-seq on untreated and fulvestrant-treated cells. ChIP-seq data were normalized by ChIP-Rx, Tag removal or spikChIP (assembly: hg19, coordinates chr1:199,775,015–201,113,512). Note the differences in ChIP-seq signal over the non-occupied genomic regions when applying ChIP-Rx or Tag removal, while remain equilibrated when applying spikChIP. (**B**) Boxplots representing the distribution of ER ChIP-seq signal after ChIP-Rx, Tag removal and spikChIP normalization. The percentage of the difference between the medians of boxplots is indicated. Dashed lines indicate the limit for the vertical scaling.

malization, Wilcoxon test; ChIP-Rx $P$-value = 3.446e-15; Tag removal $P$-value<2.2e-16; SpikChIP $P$-value = 0.02066). Then, after the normalization of the experimental ChIP, as shown in the supertrack from the genome browser (Figure 5A), and in a quantitative manner in the boxplots (Figure 5B), we could appreciate the loss of ChIP-seq signal both in ER-occupied as well as over the non-occupied loci on fulvestrant-treated cells, when normalizing the data using ChIP-Rx and Tag removal strategy. Instead, spikChIP normalization results in an evident loss of signal of ER occupancy over enriched regions, while as expected, the background signal remains equilibrated between untreated and fulvestrant-treated cells (Figure 5A and B). Overall, this result supports the application of spikChIP to compare binding occupancy of both histone and non-histone proteins across multiple samples. To our knowledge, the distinct gradual normalization from background to positive ChIP signal regions has not previously been considered in any of the existing normalization techniques. We thus believe our proposal presents a major technical advance for the genome-wide normalization and for comparison of ChIP-seq experiments.

## DISCUSSION

Limitations in the original ChIP-seq scheme, due to potential biases introduced by the technical variabilities, have required the development of strategies to implement an internal reference control across samples for further comparisons (11–13). Over the last 7 years, different alternative strategies have been proposed based on the spike-in concept to deal with the problem of comparability between multiple ChIP-seq samples. The introduction of exogenous spike-in material in the original ChIP-seq scheme enables (i) to monitor the quality of the ChIP-seq experiment and the technical variability along the experimental procedure and (ii) to computationally normalize the sequencing data for comparative analysis of ChIP-seq experiment in different experimental and/or biological conditions.

Current strategies for ChIP-seq normalization using spike-in diverge in (i) the type of biological material to be used as a reference sample for normalization, (ii) the capturing method of the spike-in material and (iii) the computational method used to analyze the sequencing data (Supplementary Table S1). It is important to mention that the adoption of different strategies can introduce inconsistencies to the final results, as each method presents its own benefits and limitations. In addition, the existence of several alternatives (Supplementary Table S1) might complicate decision-making for researchers about when and how to apply a normalization method for comparative analysis of ChIP-seq. With the aim of providing a practical and unified reference framework for comparative ChIP-seq analysis, here we propose the addition of exogenous xenogenic fly cells, whenever possible, and the use of a second antibody against a fly-specific histone variant, as a best practice for comparative ChIP-seq normalization for mammalian genomes (Figure 6).

We recommend using fly material as the spike-in because (i) the genome sequence has been extensively assembled; (ii) the fly chromatin has been largely characterized at epigenetic levels; (iii) the evolutionary distance between fly and mammalian genomes is sufficient to allow an unambiguous alignment of the reads (11,13) and (iv) fly cells are relatively easy to culture with standard tissue culture procedures and instruments. The addition of xenogenic cells of the spike-in material enables the whole procedure to be monitored from the beginning, thereby minimizing the impact of technical variabilities for the biological interpretation. Further, by mixing cells, it is possible to tackle eventual changes in genomic ploidy (e.g. due to genomic instability or differences in cell cycle progression), thereby providing an estimation of the average ChIP-seq signal per cell. This quantitative estimation is not possible when mixing fragmented chromatin. However, the option of mixing cells is only available when the number of cells can be evaluated accurately (e.g. cells growing on a dish), or when the experimental sample and spike-in material are fragmented with the same settings. On the contrary, when the number of cells in the sample is uncertain (e.g. animal tissue samples) or when the experimental sample and spike-in material require different settings for fragmentation, the addition of the fragmented spike-in material at the chromatin level is a more appropriate option. The use of a second antibody, for a fly-specific histone variant (H2Av) to capture the spike-in material, avoids the cross-reactivity constraint of the experimental antibody and to reduce any potential variability due to competition between the spike-in control and the experimental material, which usually exceeds the amount of spike-in material by far. Moreover, the genomic occupancy profile of the fly-specific H2Av is already characterized (13), thus providing a control point for assessing ChIP-seq performance.

These guidelines take into consideration that (i) spike-in material should be present in all samples at equal amounts at the earliest step during the ChIP-seq procedure and (ii) spike-in material should be present in a low-enough quantity (giving a significantly lower number of reads as that of the experimental reads) to not interfere with the actual ChIP-seq experiment yet still give an accurate normalization in the final sequencing data. As previously reported, the number of spike-in reads in the final sequencing step should be at least one million reads, and approximately, 2%–5% of the experimental genome, to minimize the changes in overall material used for ChIP-seq (11,13). This final amount of reads can be influenced by the ratio of the mixture as well as by the quality of the antibody and/or the abundance of the target in the experimental condition. Taking into account these considerations, and the relative ratio between the size of the fly genome and the two most widely used mammalian experimental models (mouse and human), we recommend the use of different final mixtures (Figure 6).

Finally, at the computational level, spikChIP methodology has shown to be very effective for the precise comparison of ChIP signals across samples without a pre-defined selection of the loci. Indeed after normalization, spikChIP generates consistently a tight adjustment in the spike-in ChIP signal across the genome between each pair of ChIP-seq conditions, independently of the experimental dataset. Moreover, SpikChIP is able to correct for possible technical bias and to compute a local correction factor, thereby minimizing the impact of the correction over non-occupied genomic regions. According to proteomic quantification, certain histone modifications are present in a relatively high abundance over the total histone H3. For instance, H3K9me2 and H3K27me2 typically represent 30–40% of total histone H3 in human cancer cell lines (22,26). In the case of H3K27me2 mark, this percentage can increase up to 70% of total H3 in mouse embryonic stem cells, and it displays a pervasive distribution over large genomic regions (27). In contrast, H3K27me3, which represents only 7% of histone H3 in mouse embryonic stem cells, preferentially accumulates over CpG-dense promoters of transcriptionally repressed genes (27). Thus, while global reduction of H3K27me2 would be observed over large genomic regions, H3K27me3 overall

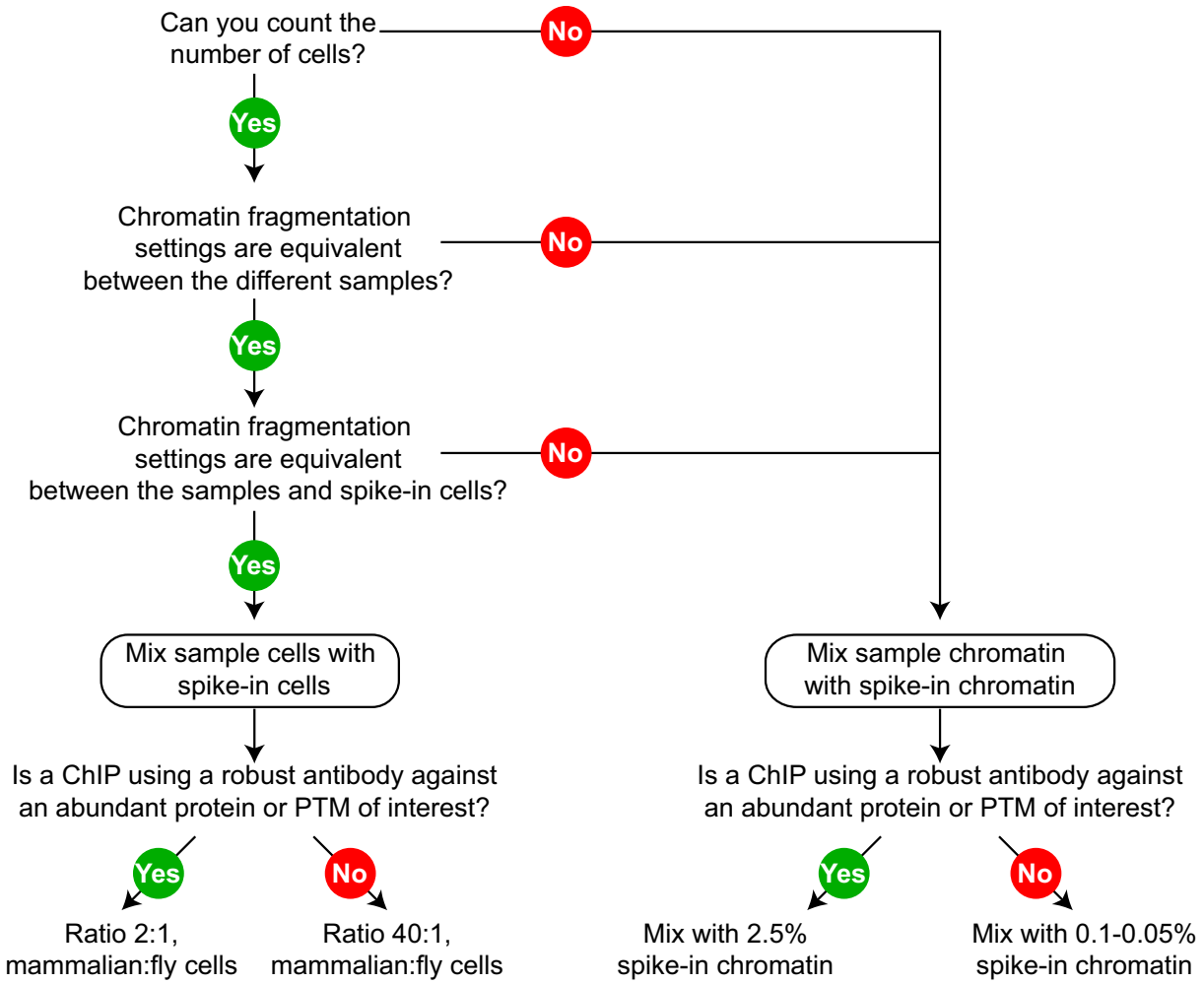


**Figure 6.** Flow chart to select the most appropriate experimental strategy for a spike-in controlled ChIP-seq experiment. We have designed a practical guide under the form of a decision tree to systematically implement a consistent protocol for the addition of the spike-in material when performing ChIP-seq experiments. In our roadmap, we state some key questions that are relevant for deciding which type of spike-in to add (e.g. fly cells or fragmented chromatin). From top to the bottom; in case we can count the number of our experimental cells, and the chromatin fragmentation settings of our experimental samples and the spike-in reference are the same, we recommend to pre-mix sample and spike-in cells. This would enable to monitor all the ChIP procedure from the beginning and will allow us to evaluate the average spike-in signal per cell equivalent among the samples. In case that any of these experimental premises fail (e.g. number of experimental cells is uncertain, or the settings to fragment the chromatin from the samples and spike-in is very different), we should consider the preparation of the experimental and spike-in chromatin separately. The amount of spike-in material to mix will depend on the yield of the experimental ChIP. The yield will depend on the robustness of the antibody, as well as in the abundance of the protein to capture. If the protein of interest is not abundant or occupies a small fraction of the genome (e.g. H3K4me3 mark or transcription factors), we must consider to mix spike-in reference in a low proportion, to avoid, after sequencing, a large number of reads belonging to the reference genome.

reduction would be confined (in the absence of a genomic redistribution) to discrete CpG-rich promoters. In the case of our benchmarking histone mark (H3K79me2), these specific changes of ChIP signal only over enriched regions are expected to be even more evident since this histone mark represents a small fraction of all histone H3 in cancer cells (<3% (22)) and displays a characteristic accumulation at 5′-end of transcriptionally active genes (11). Remarkably, in our benchmark the potential changes in the redistribution of H3K79me2 across the experiments are negligible, as in the experimental scheme untreated and EPZ5676-treated cells are mixed in different proportions to simulate the gradual reduction of the H3K79me2 mark. Finally, in contrast to histone PTMs, transcription factors are predominantly allocated within the DNA-binding site, thus displaying a characteristic discrete genomic localization. We found that using spikChIP, while correcting the ChIP-seq signal over the enriched loci, the background ChIP signal of both histone and non-histone proteins is equilibrated, thus increasing the biological interpretation while comparing ChIP-seq across samples. To conclude, we strongly believe that the systematic use of spike-in references in the ChIP-seq experiments will provide a more precise picture of the dynamics of the epigenome in different conditions.

## DATA AVAILABILITY

The source code for SpikChIP is available in an online repository (https://github.com/eblancoga/SpikChIP). The datasets reanalyzed in this study are deposited in Gene Expression Omnibus (GEO) repository under the accession numbers GSE60104, GSE64243 and GSE102882. We have deposited all normalized tracks generated by spikChIP in this UCSC session: https://genome.ucsc.edu/s/DiCroceLab/spikChIP_NAR%2DGB_2021.

## SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

## ACKNOWLEDGEMENTS

We specially thank Dr.Cecilia Ballare, Dr. Pedro Vizan and Dr. François Le Dily, as well as all the members of the Di Croce laboratory for critical reading of the manuscript and insightful discussions and V.A. Raker for scientific editing.

*Authors' Contributions:* E.B. conceived and performed the bioinformatics analysis of deep sequencing data and contributed to writing the manuscript. L.D.C. contributed to data analysis and interpretation and to writing the manuscript. S.A. conceived and planned this project, performed data analysis and interpretation, and wrote the manuscript with input from the coauthors.

## FUNDING

Spanish of Economy, Industry and Competitiveness (MEIC) (BFU2016-75008-P, and PID2019-108322GB-100), "Fundación Vencer El Cancer" (VEC), the European Regional Development Fund (FEDER), and from AGAUR to L.D.C. The Ramon y Cajal program of the Ministerio de Ciencia, Innovación y Universidades and the European Social Fund under the reference number RYC-2018-025002-I, and the Instituto de Salud Carlos III-FEDER (PI19/01814), to S.A. We acknowledge the funding support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme / Generalitat de Catalunya.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Albert,I., Mavrich,T.N., Tomsho,L.P., Qi,J., Zanton,S.J., Schuster,S.C. and Pugh,B.F. (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature*, **446**, 572–576.
2. 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.
3. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
4. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
5. Consortium,E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
6. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
7. Consortium,E.P., Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
8. Stunnenberg,H.G., International Human Epigenome,C. and Hirst,M. (2016) The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, **167**, 1145–1149.
9. Skipper,M., Eccleston,A., Gray,N., Heemels,T., Le Bot,N., Marte,B. and Weiss,U. (2015) Presenting the epigenome roadmap. *Nature*, **518**, 313.
10. Chen,K., Hu,Z., Xia,Z., Zhao,D., Li,W. and Tyler,J.K. (2015) The overlooked fact: fundamental need for Spike-In control for virtually all genome-wide analyses. *Mol. Cell. Biol.*, **36**, 662–667.
11. Orlando,D.A., Chen,M.W., Brown,V.E., Solanki,S., Choi,Y.J., Olson,E.R., Fritz,C.C., Bradner,J.E. and Guenther,M.G. (2014) Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep.*, **9**, 1163–1170.
12. Bonhoure,N., Bounova,G., Bernasconi,D., Praz,V., Lammers,F., Canella,D., Willis,I.M., Herr,W., Hernandez,N. and Delorenzi,M. (2014) Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Res.*, **24**, 1157–1168.
13. Egan,B., Yuan,C.C., Craske,M.L., Labhart,P., Guler,G.D., Arnott,D., Maile,T.M., Busby,J., Henry,C., Kelly,T.K. *et al.* (2016) An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon EZH2 Inhibition. *PLoS One*, **11**, e0166438.
14. Guertin,M.J., Cullen,A.E., Markowetz,F. and Holding,A.N. (2018) Parallel factor ChIP provides essential internal control for quantitative differential ChIP-seq. *Nucleic Acids Res.*, **46**, e75.

15. Thomas,R., Thomas,S., Holloway,A.K. and Pollard,K.S. (2017) Features that define the best ChIP-seq peak calling algorithms. *Brief. Bioinform.*, **18**, 441–450.
16. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. *et al.* (2021) The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.
17. 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.
18. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* **25**, 2078–2079.
19. Loven,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.
20. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
21. Taruttis,F., Feist,M., Schwarzfischer,P., Gronwald,W., Kube,D., Spang,R. and Engelmann,J.C. (2017) External calibration with Drosophila whole-cell spike-ins delivers absolute mRNA fold changes from human RNA-Seq and qPCR data. *BioTechniques*, **62**, 53–61.
22. Leroy,G., Dimaggio,P.A., Chan,E.Y., Zee,B.M., Blanco,M.A., Bryant,B., Flaniken,I.Z., Liu,S., Kang,Y., Trojer,P. *et al.* (2013) A quantitative atlas of histone modification signatures from human cancer cells. *Epigenetics Chromatin*, **6**, 20.
23. Aranda,S., Mas,G. and Di Croce,L. (2015) Regulation of gene transcription by Polycomb proteins. *Sci. Adv.*, **1**, e1500737.
24. McCabe,M.T., Ott,H.M., Ganji,G., Korenchuk,S., Thompson,C., Van Aller,G.S., Liu,Y., Graves,A.P., Della Pietra,A. 3rd, Diaz,E. *et al.* (2012) EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature*, **492**, 108–112.
25. Aranda,S., Shi,Y. and Di Croce,L. (2016) Chromatin and epigenetics at the forefront: finding clues among peaks. *Mol. Cell. Biol.*, **36**, 2432–2439.
26. Huang,H., Lin,S., Garcia,B.A. and Zhao,Y. (2015) Quantitative proteomic analysis of histone modifications. *Chem. Rev.*, **115**, 2376–2418.
27. Ferrari,K.J., Scelfo,A., Jammula,S., Cuomo,A., Barozzi,I., Stutzer,A., Fischle,W., Bonaldi,T. and Pasini,D. (2014) Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. *Mol. Cell*, **53**, 49–62.