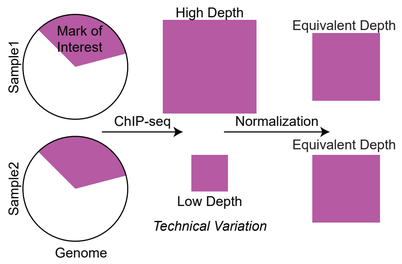
ChIP-seq Spike-In Normalization

The Good, the Bad, and the Ugly

Aaron Bogutz, Kentaro Mochizuki, Louis Lefebvre, Matthew Lorincz

Molecular Epigenetics Group, Life Sciences Institute, University of British Columbia

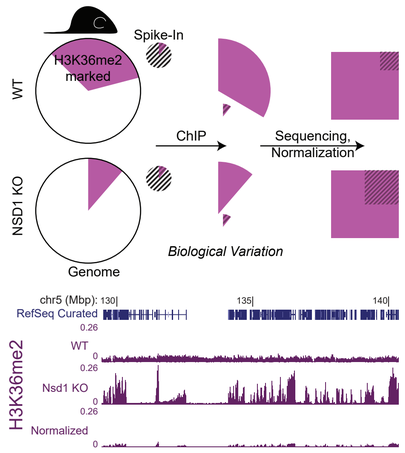
Why Perform Spike-In Normalization?



Next Generation Sequencing (NGS) produces millions or billions of short reads which can be aligned to a reference genome. Chromatin Immunoprecipitation followed by Sequencing (ChIP-seq) utilizes this technology to produce a genome-wide exploration of regions enriched for a given epigenetic mark. The number of reads corresponding to a genomic location will indicate the strength of signal. In order to compare across samples sequenced to varying depths and with varying efficiencies of mapping, metrics such as Counts per Million (CPM - equation (1)) are widely used. Normalization of total read count allows quantifications performed on different samples to be directly compared.

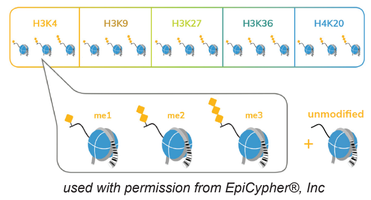
$$Counts\ Per\ Million\ (CPM) = \frac{Calculated\ Reads * 1M}{Total\ Mapped\ Reads} \quad (1)$$

If the absolute abundance of a given mark shifts between samples, however, this normalization can result in analysis artifacts. For example, NSD1 is the methyltransferase responsible for dimethylation of Histone 3 Lysine 36 (H3K36me2). In a knockout of *Nsd1*, ChIP-seq for H3K36me2 shows a distinct pattern of gain over some genomic regions. Inclusion of a Spike-In, a defined quantity of DNA copurified in ChIP-seq, allows for relative quantification of total levels of this mark, which are in fact greatly reduced in this background. Upon further analysis, the residual H3K36me2 signal is only found at H3K36me3 positive regions, likely low levels of dimethylation due to incomplete activity of SETD2 which normally trimethylates this residue.



I. Synthetic Barcoded Histones

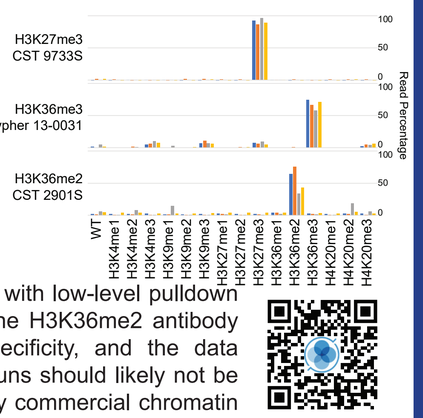
Epicypher produces synthetic histones bearing specific posttranslational modifications wrapped with barcoded DNA. By including a panel of modifications prior to ChIP-seq, antibody specificity and relative quantification are made possible.



used with permission from EpiCypher®, Inc

Antibody Specificity

Quantification of reads corresponding to individual marks in the panel of synthetic histones used allows a direct readout of antibody specificity. For example, the H3K27me3 antibody used shows ~90% specificity, whereas the H3K36me3 antibody averages around 70% specificity with low-level pulldown of other trimethylated lysines. The H3K36me2 antibody shows much more variable specificity, and the data produced from these ChIP-seq runs should likely not be used. Epicypher has tested many commercial chromatin antibodies using this technology and the results are available at chromatinantibodies.com[5] (see QR).

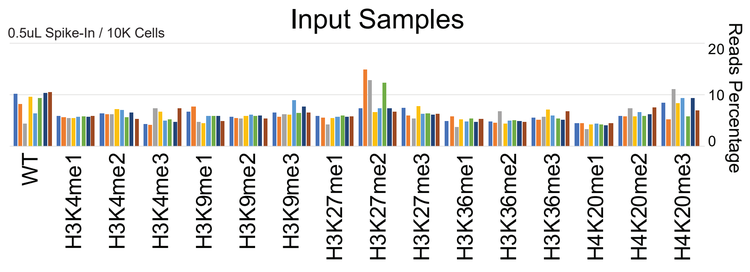


Alignment and Quantification

Sequential alignment to the reference genome and the smaller barcode sequences is sufficient for quantification, as no reads are predicted to align to both. Duplicates should not be discarded, however, as the small size of the barcoded DNA greatly increases duplication rates relative to the reference genome.

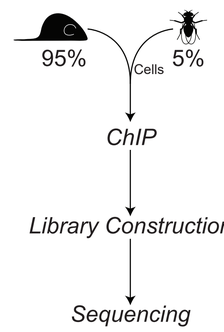
Caveats

The availability of synthetic histones marked by a particular posttranslational modification of interest is not guaranteed, and newer approaches such as Cut&Run and Cut&Tag necessitate alternative technologies (also in production from EpiCypher). In addition, variability in quantities of barcodes from input samples has been observed, with potential knock-on effects for quantification pipelines.

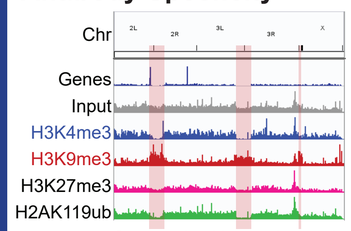


II. Exogenous Cells

Inclusion of defined numbers of cells from another species (*Drosophila* being a common choice) prior to ChIP-seq allows pulldown of DNA from both species. Antibody specificity can be qualitatively inferred by comparison with other datasets. The efficiency of pulldown of DNA from the Spike-In species should be equivalent across samples, and the ratio of Spike-In to reference reads makes relative quantification possible.



Antibody Specficity



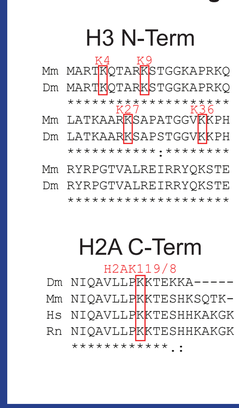
The distribution of many epigenetic marks in model organisms such as *Drosophila* is relatively well characterized, and comparison of data from Spike-In samples to these reference sets or to known distribution can provide confidence that antibodies are behaving as predicted.

Alignment and Quantification

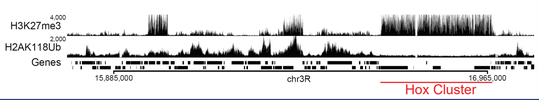
A concatenated genome is recommended for alignment of data following Spike-In from another species. Regions homologous between the two species will result in a lower mapping quality of reads in the region, as they will map both to the Reference and the Spike-In genomes. By removing reads with low mapping quality, the possibility of reads from the Spike-In affecting your data will be minimized. BAM files can be deconvoluted by the inclusion of prefixes to chromosome names in concatenated genomes.

Caveats

Not all species feature the same epigenetic marks at the same frequency, or at all. For instance, H3K36me2 is very abundant in the mouse genome, with more than 50% of all nucleosomes bearing the mark in peri-implantation embryos as quantified by LC-MS/MS[3]. **In *Drosophila*, however, there is scant evidence for this mark being present, and relative pulldown of the two genomes highly favours the mouse.**



Antibody specificity is also not guaranteed to hold true between different species analyzed. *Drosophila*-Mouse H3 N-Terminal tail conservation is very high, but H2A conservation particularly surrounding the site of ubiquitylation by PRC1 is much lower. The antibody used in recent *Drosophila* studies is the same as that used in mammalian studies and shows poor correlation with H3K27me3[4]. **The lack of enrichment at known PRC1 targets suggests that this antibody lacks H2AK118ub specificity in *Drosophila*.**



Data Normalization Options

I. Subsample to Equivalent Spike Depth

1. Quantify libraries for Spike-In depth
2. Downsample libraries to equivalent Spike-In depth
3. Quantify raw reads from downsampled libraries

Pros:
- Straightforward

Cons:
- Throwing away useful reads
- Dependent on minimum depth library
- Scales meaningless

Dobrinic 2021[1]

Quantitation

Reads

II. Reads per Spike-In

1. Quantify libraries for Spike-In depth
2. Quantify reads / Spike-In reads

Pros:
- Somewhat straightforward
- Somewhat intuitive

Cons:
- Scales become significantly different than traditional methods

Shirane 2020[2]

Quantitation

$$\frac{Reads * 1M}{Spike\text{-}in\ Reads}$$

III. Read Ratio Normalization

1. Quantify libraries for Spike-In depth
2. Quantify libraries for Reference depth
3. Calculate ratio of Reference to Spike-In reads
4. Calculate Normalization Factor (NF - equation (2))
5. Calculate RPM/CPM as usual
6. Apply Normalization Factor to quantitation

Pros:
- Direct comparisons of normalized vs unnormalized data possible
- Scales intuitive
- Normalization factor can be applied to other types of quantitation
- Takes input into account

Cons:
- More complicated
- Dependent on input equivalency

Github Scripts[6]

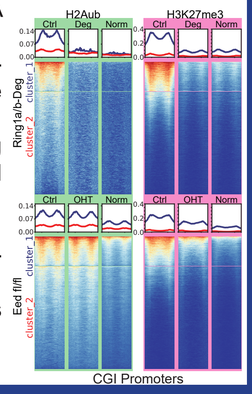
Quantitation

$$CPM * NF$$

$$NF = \frac{\frac{Ref_{IP\text{-}exp}}{Spike_{IP\text{-}exp}}}{\frac{Ref_{input\text{-}exp}}{Spike_{input\text{-}exp}}} \Big/ \frac{\frac{Ref_{IP\text{-}ctrl}}{Spike_{IP\text{-}ctrl}}}{\frac{Ref_{input\text{-}ctrl}}{Spike_{input\text{-}ctrl}}} \quad (2)$$

Best Practices

- Aim for a suitable number of Spike-In reads (usually 1-5% of total). Too many reads will decrease the number of reads available for your experimental samples, whereas too few reads will result in noise amplification through normalization.
- Be aware of effective genome size when calculating Spike-In amount, especially when adding cells. The *Drosophila* genome, for instance, is less than one tenth the size of the mouse genome, so many more *Drosophila* cells are required to get equivalent DNA pulldown.
- Select an appropriate Spike-In. Choice should reflect the minimum reduction in mappability while still revealing antibody efficiency.
- Spend time analyzing your data – simply running pipelines and ignoring intermediate quality control steps could result in artifact propagation.
- Check unnormalized vs normalized data routinely. Altered genomic patterns should be evident in experimental conditions even prior to normalization. If this isn't seen, then any normalization effects are potential artifacts.



Citations

1. PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency. Dobrinić P, Szczurek AT, Klose RJ. *Nat Struct Mol Biol.* 2021 Oct;28(10)
2. NSD1-deposited H3K36me2 directs de novo methylation in the mouse male germline and counteracts Polycomb-associated silencing. Shirane K, Miura F, Ito T, Lorincz MC. *Nat Genet.* 2020 Oct;52(10)
3. Nucleome programming is required for the foundation of totipotency in mammalian germline development. Nagano M et al. *EMBO J.* 2022 Jul 4;41(13)
4. Widespread activation of developmental gene expression characterized by PRC1-dependent chromatin looping. Loubiere V, Papadopoulos GL, Szabo Q, Martinez AM, Cavalli G. *Sci Adv.* 2020 Jan.
5. https://chromatinantibodies.com/
6. https://github.com/abogutz/NormalizationScripts