Dovetail GENOMICS
Part of Cantata Bio, LLC

EPIGENETICS

Application Note

Demystifying Hi-C Data Normalization: A Guide for Genomic Researchers

Key Takeaways

- Hi-C analysis workflows often include built-in methods for individual sample matrix balancing.
- The KR method, present in both Juicer and Cooler tools, provides a fast, robust method for individual feature calling while preserving topology.
- Between-sample normalization methods are primarily used for differential region detection when comparing different sample conditions.
- Multidimensional matrix formats allow on-the-fly normalization, preserving raw counts for current between-sample analysis tools. topology features.

Introduction

Genomic research has come a long way since the draft of the human genome was first published over two decades ago, driven by the continuous advancements in next-generation sequencing (NGS) technologies. One of the most transformative methods to emerge from these developments is Hi-C, a library preparation technique that provides unique insights into the three-dimensional organization of the genome. However, analyzing Hi-C data presents its own set of challenges, particularly regarding data normalization. In this comprehensive guide, we'll delve into the intricacies of Hi-C data normalization, providing you with a deep understanding of its importance and the various tools and methods available for this critical step in the analysis pipeline.

Hi-C: Unraveling Genomic 3D Structure

Hi-C, short for High-throughput Chromosome Conformation Capture, is a groundbreaking technique that uses proximity ligation technology to capture the three-dimensional interactions of DNA sequences. This approach not only reveals primary sequence data but also provides valuable insights into how DNA physically interacts in a spatial context. Just like standard whole-genome shotgun sequencing, Hi-C data captures genetic alterations, including single nucleotide variations (SNVs), small insertions/deletions (indels), copy number variations (CNVs), and structural variations (SVs). However, what sets Hi-C apart is its ability to uncover topological features, such as chromosomal territories, active/inactive compartmentalization, topologically associated domains (TADs), and chromatin loops.

The Hi-C methodology can be broken down into five core steps:

- Chromatin Crosslinking: Chromatin is crosslinked to "lock" chromatin interactions

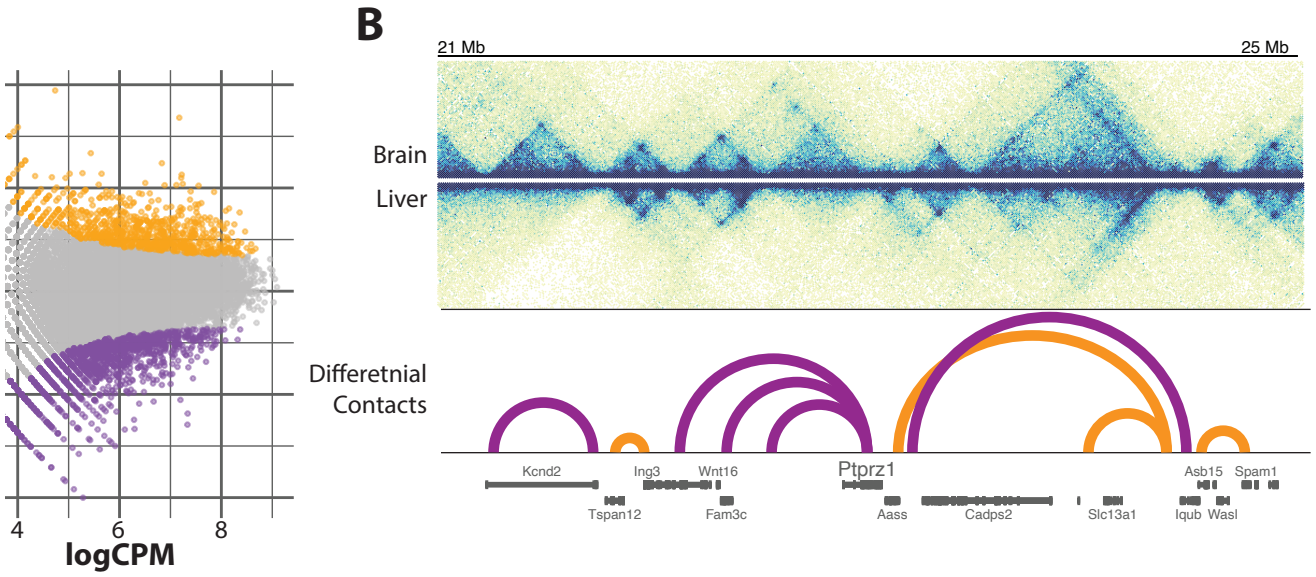

Figure 1 Multiple-sample normalization enables robust statistical detection of differential interactions.
A) By applying concepts from gene expression, in this case, loess normalization, we can detect differences between brain and liver tissues using standard statistical models. B) When visualizing these interactions, the resulting differential contacts mirror what we can visually detect in the contact matrices.

in their original 3D positions.

- Chromatin Fragmentation: Chromatin is fragmented to create free ends that will be used for ligation.
- Ligation of Free Ends: The free ends are ligated together.
- Library Generation: A sequencing-compatible library is generated from the ligated fragments.
- Paired-End Sequencing: Paired-end sequencing is performed on a compatible NGS system.

The resulting paired-end data contains 3D structural information. Paired-end reads that map at a distance from each other in linear sequence space indicate genomic regions found in close proximity in 3D space. This 3D genomic structure plays a vital role in gene regulation by controlling access to regulatory elements within the genome.

The Need for Hi-C Data Normalization

Hi-C data, like other NGS datasets, is not immune to biases and limitations. One unique challenge specific to Hi-C data is the exponential decay of interaction frequency with the distance between two genomic regions. In simpler terms, regions that are close in genomic coordinate space are more likely to form chimeric ligation products, leading to distortions in the data. To address this probabilistic dependency and other systematic biases, data normalization becomes essential in Hi-C data analysis.

Understanding Normalization Approaches

Normalization approaches for Hi-C data can be categorized into two main types: explicit and implicit. Explicit approaches aim to directly account for individual biases, including GC content, fragmentation, mappability, and enzyme cut sites. They rely on the assumption that these biases are well understood and can be accurately accounted for. The two primary explicit normalization methods are:

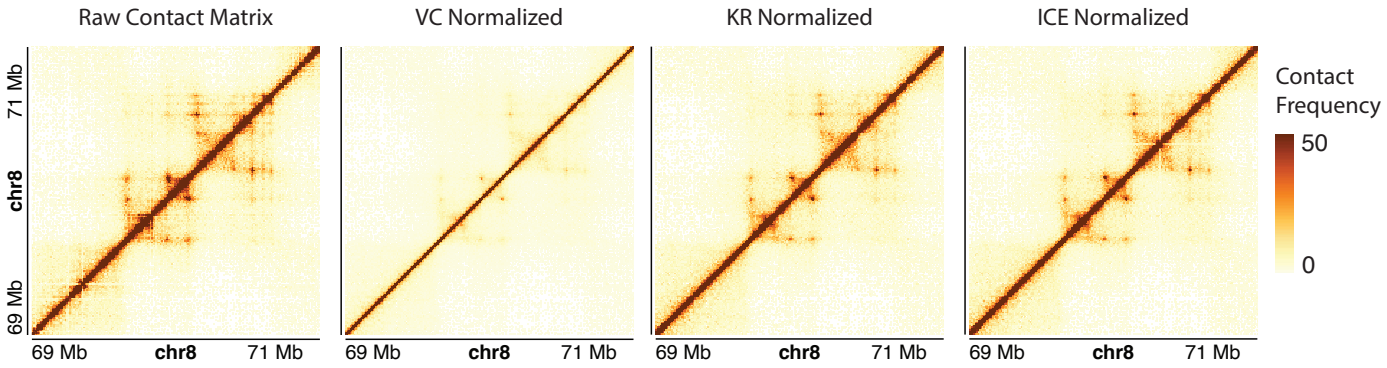

Table 2 Visual Comparison of Normalization Results.Each contact matrix depicts the same 2Mb region on chromosome 8 of a Micro-C library that was sequenced with 800 million read pairs. The matrix was subjected to different normalization approaches and plotted in R. The scale bar is consistently maintained across each normalization approach to better visualize the impact of normalization. This image clearly demonstrates the challenges associated with using coverage alone as a normalization strategy,

- Yaffe and Tanay's Probabilistic Model: A pioneering probabilistic model designed to account for known biases.
- HiCNorm: An algorithm built upon similar principles to Yaffe and Tanay's model.

Implicit approaches, on the other hand, make use of the assumption of "equal loci visibility." They assume that cumulative bias is captured within the sequencing depth of each bin of the contact matrix. Some common implicit methods include:

- Sequential Component Normalization (SCN)
- Iterative Correction and Eigenvector Decomposition (ICE)
- Knight and Ruiz (KR) Method
- ChromoR: Utilizes a Bayesian approach.
- Binless: A relatively new algorithm offering a hybrid approach.
- Vanilla Coverage (VC) and Square Root Supplement (VCSQ): Simpler implicit algorithms, but less widely used.

The choice between explicit and implicit methods depends on the level of understanding of biases and the complexity of the genome being studied. Explicit methods require more user-defined parameters, making them suitable for less-studied organisms where biases may be poorly understood. Implicit methods, with their simplicity, are often preferred for well-studied genomes like human and mouse.

Pre-processing Pipelines and Data Formats

In Hi-C data analysis, several pre-processing pipelines and data formats are commonly used:

- Juicer: An all-in-one pre-processing pipeline developed by the Aiden lab. It generates *.hic files, which are widely supported by downstream computational tools and visualization software.
- Cooler: A more recent entrant that uses the HDF5 data structure, offering computational advantages. Cooler generates *.cool files and is increasingly adopted for downstream analyses.
- HiCExplorer: A suite of tools for end-to-end Hi-C data analysis, supporting both its native matrix format and *.cool files.

These pipelines offer multiple options for data normalization, adding flexibility to your analysis.

Comparing Two or More Conditions

Genomic insights are often derived from comparing multiple samples or conditions. When it comes to Hi-C data, methods for between-sample comparisons are limited. MultiHiCcompare uses a loess regression approach, adapted from the gene expression

literature. A more recent approach is BNBC, which performs matrix smoothing on individual matrix "bands" before batch correction using ComBat. These methods often ignore systematic, sequence-dependent biases shared among all sample conditions.

Unified Approach for Within- and Between-Sample Normalization

One crucial consideration in Hi-C data analysis is whether to normalize both within samples and between samples. The answer is remarkably straightforward – no. Multidimensional matrix formats, such as *.hic and *.cool files, store correction weights independently of raw interaction counts. This feature allows on-the-fly normalization during downstream feature calling, preserving the raw counts. Hence, a single, normalized contact matrix can be generated while retaining access to the raw counts necessary for between-sample analysis tools.

Choosing the Right Approach

As the field of Hi-C data analysis continues to evolve, researchers often grapple with the choice of normalization method and pipeline. To date, no single "gold standard" method has emerged. Various studies have compared different algorithms, with Rao et al. opting for the KR method due to its computational efficiency. However, KR may falter with sparse contact matrices, in which case ICE, a robust balancing method, can be used. Overall, SCN, KR, and ICE strategies tend to perform similarly, with only minor differences at lower resolutions.

In practical terms, the choice between these approaches often depends on the tools and pipelines you are using, as many provide KR and ICE as common normalization methods. The field is rapidly advancing, and newer algorithms like Binless may influence the consensus in the future.

For more information about Dovetail Genomics Bioinformatics Services, visit: