


# MoCoLo: a testing framework for motif co-localization

Qi Xu , Imee M.A. del Mundo, Maha Zewail-Foote, Brian T. Luke, Karen M. Vasquez and Jeanne Kowalski

Corresponding authors. Jeanne Kowalski, Department of Oncology, Dell Medical School, University of Texas at Austin, Austin, TX, 78712, USA; Tel.: +15124955737; E-mail: Jeanne.Kowalski@austin.utexas.edu; Karen M. Vasquez, Dell Pediatric Research Institute, Division of Pharmacology and Toxicology, College of Pharmacy, The University of Texas at Austin, Austin, Texas, 78723, USA; Tel.: +15124953040; E-mail: karen.vasquez@austin.utexas.edu

## Abstract

Sequence-level data offers insights into biological processes through the interaction of two or more genomic features from the same or different molecular data types. Within motifs, this interaction is often explored via the co-occurrence of feature genomic tracks using fixed-segments or analytical tests that respectively require window size determination and risk of false positives from over-simplified models. Moreover, methods for robustly examining the co-localization of genomic features, and thereby understanding their spatial interaction, have been elusive. We present a new analytical method for examining feature interaction by introducing the notion of reciprocal co-occurrence, define statistics to estimate it and hypotheses to test for it. Our approach leverages conditional motif co-occurrence events between features to infer their co-localization. Using reverse conditional probabilities and introducing a novel simulation approach that retains motif properties (e.g. length, guanine-content), our method further accounts for potential confounders in testing. As a proof-of-concept, motif co-localization (MoCoLo) confirmed the co-occurrence of histone markers in a breast cancer cell line. As a novel analysis, MoCoLo identified significant co-localization of oxidative DNA damage within non-B DNA-forming regions that significantly differed between non-B DNA structures. Altogether, these findings demonstrate the potential utility of MoCoLo for testing spatial interactions between genomic features via their co-localization.

**Keywords:** co-localization testing; property-informed simulation; DNA motif

## INTRODUCTION

The increasing number of genomic datasets produced by high-throughput sequencing and prediction algorithms has revealed interactions between genomic features and biological processes [1–3]. Although these interactions take many forms, their concept, derivation and evaluation remain embedded in the frequency of ‘co-occurrence’. Co-occurrence describes an event in which two or more features are present, which can be tested for their appearance together more often than would be expected by chance [4]. On the other hand, ‘co-localization’ refers to an event in which two or more features are both present in the same spatial region/proximity. While co-localization requires co-occurrence, the latter does not imply the former. Herein, we focus upon sequence motif interaction by introducing a criterion that requires the occurrence of a genomic feature within another feature and vice-versa. We refer to this criterion as reciprocal sequence co-occurrence and define metrics that enable characterization of co-localization using it.

Historically, for testing the co-occurrence of events two general approaches are used, one based on a Fisher’s exact test and another based on Monte-Carlo simulation [4, 5]. Statistical models rely on strict assumptions that may not always be suitable for genomic analyses. For example, parametric tests assume an *a priori* distribution that is oftentimes based upon independent

events. These testing assumptions would be difficult to address since they involve finding the optimal model and parameters to characterize varying lengths of genomic regions that are often correlated between molecular features. While empirical methods may overcome strict modeling assumptions, they require simulations that take into account sequence properties (e.g. length, nucleotide content) to generate meaningful results. This type of sequence property-informed simulation often comes with the price of high computational costs and thus, may be difficult to achieve in the absence of an efficient algorithm.

Herein, we introduce motif co-localization (MoCoLo) as a framework for direct testing of sequence-level co-localization using empirical methods coupled with a property-informed simulation algorithm. A class of hypotheses is constructed for testing the random occurrence of one feature in another feature and vice-versa (i.e. reciprocal occurrence). For hypothesis testing, a simulation method is introduced that incorporates sequence properties to ensure that the simulated data is representative of the properties embedded in the observed data such that differences in occurrence due to confounding factors are minimized. We demonstrate the method with two case applications for testing genome-wide co-localization between sequence-level molecular features of the same data type using histone modifications, and between different data types using alternative DNA (i.e. non-B DNA) structure-forming motifs (e.g.

Qi Xu is a PhD student affiliated with the Department of Molecular Biosciences and the Department of Oncology, University of Texas at Austin, Austin, TX.

Imee M.A. del Mundo is a research associate at College of Pharmacy and Dell Pediatric Research Institute, University of Texas at Austin, Austin, TX.

Maha Zewail-Foote is a professor at the Department of Chemistry and Biochemistry, Southwestern University, Georgetown, TX.

Brian T. Luke is a scientist at the Frederick National Laboratory for Cancer Research, Maryland.

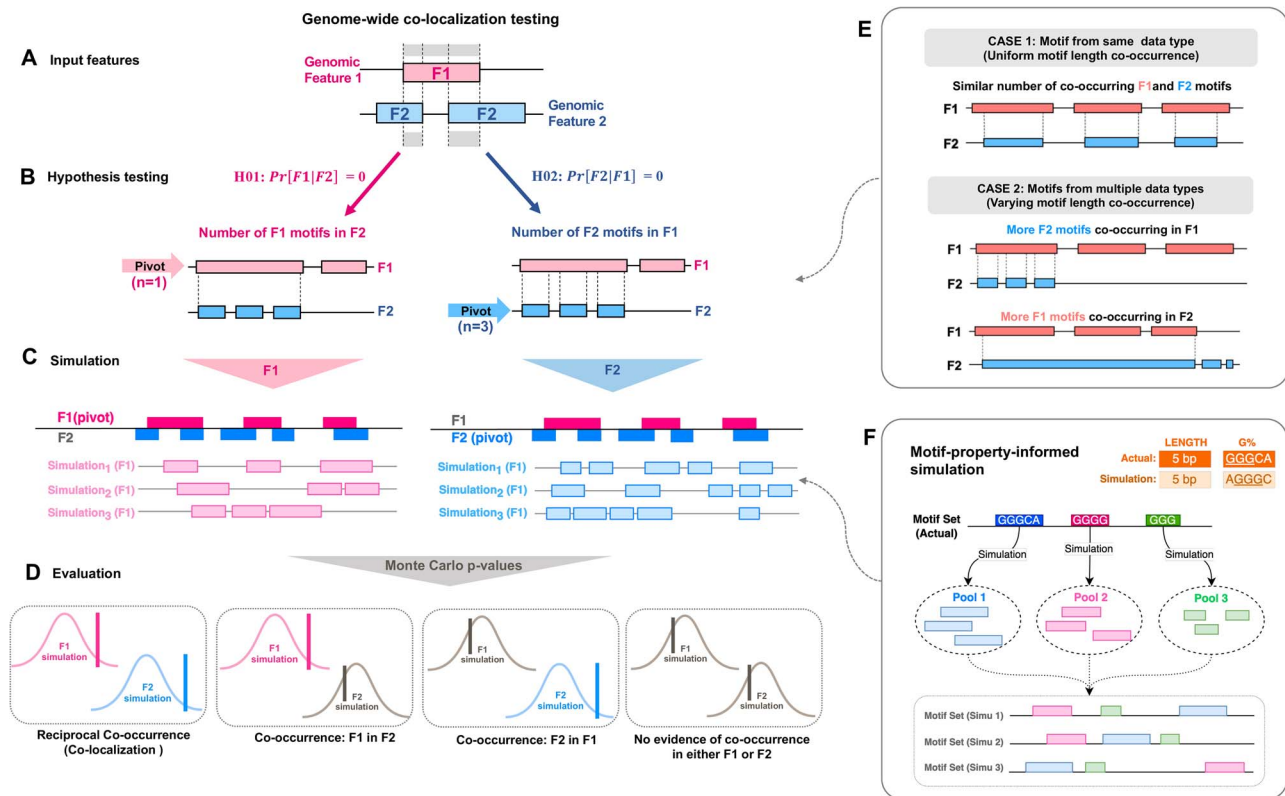
Karen M. Vasquez is a professor at the College of Pharmacy and Dell Pediatric Research Institute, University of Texas at Austin, Austin, TX.

Jeanne Kowalski is a professor at the Department of Oncology, Dell Medical School, University of Texas at Austin, Austin, TX.

Received: October 26, 2023. Revised: January 8, 2024. Accepted: January 9, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



**Figure 1.** Overview of the MoCoLo framework. MoCoLo provides a simulation-based approach to test co-localization of two genomic features, integrating the processes of testing feature selection, property-informed simulation, and statistical evaluation. **(A)** Input. For testing co-localization, the input encompasses the genomic motif regions associated with features F1 and F2. **(B)** Hypothesis testing. A ‘pivot’ feature is designated for hypothesis testing, recognizing that differences between the two motif data types can affect testing results (see also E). The co-localization assessment uses the number of the overlapping pivot features in the other as metrics. **(C)** Simulation. The motif-property-informed simulations will be performed for each of the pivot motif group selected (see also F). It takes motif sequence characteristics into consideration to maintain the resemblance between the actual and the simulation groups. **(D)** Significance evaluation. MoCoLo determines the significance of co-localization by evaluating the two metrics reciprocally, incorporating Monte Carlo *P*-values in its results. If both hypothesis testing shows significant *P*-value, the two features are evaluated with ‘co-localization via reciprocal occurrence’. If only one side of tests shows significant *P*-value not the other, the two features have ‘co-occurrence of one in the other’ but not co-localization. **(E)** Motif type impact on co-localization testing. Case 1 showcases co-localization when the length distributions of motifs from two features are alike, often originating from the same data type. Case 2 illustrates a co-localization scenario where motifs from the two features have contrasting sequence lengths. Here, a motif from one feature might overlap with several motifs from the other feature. The chosen testing hypothesis and simulation method in such situations can yield different results. **(F)** Simulation design. The design of the simulation method in MoCoLo emphasizes a motif-property-informed approach. This includes simulating individual motifs, constructing simulation pools and assembling the simulated motif sets. Additionally, a ‘dynamic tolerance’ is utilized to enhance computation efficiency and ensure a close resemblance between the actual and simulated data.

G-quadruplex DNA, Z-DNA and mirror repeats) and 8-oxo-dG, an indicator of oxidative DNA damage.

## METHODS

### Overview of MoCoLo framework

MoCoLo is an approach to test for global, genome-wide reciprocal co-occurrence, i.e., co-localization. We describe our method within the context of two genomic features, feature 1 and feature 2 (F1, F2) (Figure 1A), each defined by varying lengths and numbers of motifs (M1, M2). Interest is in addressing the question of whether these two feature motif libraries are co-localized and if so, to describe their co-localization by genomic region. This study provides a simulation-based approach to test co-localization of two genomic features, integrating the processes of hypothesis testing metric selection, property-informed simulation and statistical evaluation.

### Reciprocal co-localization assessment

Our approach is designed for genome-wide reciprocal co-localization assessments (Figure 1A). Existing methods mostly test co-localization within the same genomic data type. While

examining the notion of co-localization between motifs derived from different molecular data types, attention must be paid to the differences in sequence composition that define each data type (Figure 1E). It is essential to consider the impact of difference in motif types on co-localization evaluation. In Case 1, similar motif length distributions, typically stemming from the same data type, might result in comparable counts of co-occurrence between two features (Figure 1E, top). Conversely, Case 2 depicts a situation where the motif lengths of the two features differ distinctly, potentially leading to one motif overlapping with multiple motifs from its counterpart (Figure 1E, bottom). Depending on the hypothesis and metric selected, these scenarios might produce varied results.

### Duo hypotheses and testing metric

Therefore, we introduce two hypotheses that are both necessary to infer co-localization between F1 and F2 motif libraries (Figure 1B). The first hypothesis, H01, tests genome-wide, whether the number of F1 motifs in F2 motifs is greater than expected by random chance. Likewise, H02, tests genome-wide, whether the number of F2 motifs in F1 motifs is greater than chance. The two statistics for testing each hypothesis are based on estimates of

conditional probabilities. A ‘pivot’ feature needs to be designated for hypothesis testing, recognizing the differences between the two motif data types. The co-localization assessment uses the number of the overlapping pivot features in the other as metrics.

### Sequence property-informed simulation

As an empirical method, MoCoLo simulates expected data under a specified null hypothesis and compare it to the actual observed data (Figure 1C). It offers a simulation method informed by sequence properties to closely retain the characteristics of each motif groups. Unlike typical methods that utilize random repositioning of regions, our method includes information on motif properties such as nucleotide composition in addition to motif length. The simulation method is developed by introducing new concepts such as simulation pool construction, motif sets assembling and dynamic tolerance, together to ensure a more nuanced simulation while maintaining the computational efficiency (Figure 1F).

### Testing hypotheses

We introduce two hypotheses that are both necessary to infer co-localization between F1 and F2 motif libraries in MoCoLo. The first hypothesis, H01, tests genome-wide, whether the number of F1 motifs in F2 motifs is greater than zero. The second hypothesis, H02, tests genome-wide, whether the number of F2 motifs in F1 motifs is greater than zero. Formally, we introduce the following two hypotheses:

$$H_{01} : p_{12} = 0 \text{ vs. } H_{01a} : p_{12} > 0$$

$$H_{02} : p_{21} = 0 \text{ vs. } H_{02a} : p_{21} > 0$$

where:

$$p_{12} = \Pr [F1|F2]$$

$$p_{21} = \Pr [F2|F1]$$

Below, we introduce two metrics for testing each hypothesis:

$$\hat{p}_{12} = \sum_{i=1}^{NF2} \sum_{j=1}^{NF1} \sum_{k=1}^{l(F_{1j})} I \{F_{1ijk} \subseteq F_{2i}\};$$

$$\hat{p}_{21} = \sum_{j=1}^{NF1} \sum_{i=1}^{NF2} \sum_{k=1}^{l(F_{2i})} I \{F_{2jik} \subseteq F_{1j}\}$$

where  $I\{\cdot\}$  is an indicator function, NF1 and NF2 are the number of motif libraries within features F1 and F2, respectively, and  $l(F_{1j})$  indicates the length of the  $j$ th motif from F1 feature with  $l(F_{2i})$  the length of the  $i$ th motif from F2 feature.

### Testing statistics

For gene-level overlap testing between two gene sets, denoted by G1 and G2, there exists options that are largely based on a Fisher exact test, with some popular choices being a Jaccard similarity coefficient and a hypergeometric distribution. If testing is two-sided, then we have no prior belief about direction and are simply testing whether the odds of success (‘overlap’) differs from 1 or not. On the other hand, one may be interested in a one-sided test of whether the odds of success (‘overlap of G1’) is greater (or less) in G2. In this context of a one-sided scenario, though not explicitly stated as such, one gene set is defined as fixed (i.e. ‘pivot’) that is compared against the other. We propose an analogous approach within a sequence context by introducing a feature variable pivot in which to conduct a (‘two-sided’) test of

association, the collection of which, H01: F1 in F2 and H02: F2 in F1 tests for co-localization association between features and the separation of which enables a ‘one-sided’ alternative. For pivot selection: we define ‘pivot selection’ as the choice of reference feature to derive evaluation metrics. For testing H01, we quantify the total number of F1 motifs in F2, and thus, F2 is the pivot feature. Likewise, for testing H02, we quantify the total number of F2 motifs in F1, and thus, F1 is the pivot feature. Hence, we can evaluate co-localization by the reciprocal sequence co-occurrence by exchanging reference and query feature motifs.

### Sequence property-informed simulation

Traditional brute force approaches simulate same-length genomic regions at random genome locations [6]. This step fulfills the length requirement in simulation. However, the composition of the motif sequences in these simulated regions needed to be further checked and only those with similar nucleotide compositions (e.g. similar %G) are retained to fulfill the composition requirement. This can be computationally intensive and inefficient due to the potential non-existence of same-length regions with matching composition, which may lead to infinite loop situations.

To overcome these issues, we devised a novel optimal search strategy. As opposed to simultaneously simulating all motifs at once, instead, we simulated motifs individually and constructed a ‘simulation pool’ that tags traits of interest for matching by motif length and composition. We then randomly sample a motif set (as set of simulated motifs with defined traits) from this pool that can be readily matched as the ‘random’ counterpart of the actual data motif set. Considering that another region with the exact same traits as the test region may not exist in the genome, with this approach, we were able to avoid the infinite loop created by enabling a ‘dynamic tolerance’ that performs an automatic adjustment on the simulation tolerance.

### Data sources and processes

#### Histone data

The ChIP-seq data of H4K20me3 and H3K9me3 in the human MCF-7 breast cancer cell line was downloaded from the NCBI Gene Expression Omnibus (GSE143653) [7], which included ChIP-seq data for ChIP\_Input\_MCF7 (GSM4271438), H4K20me3\_BR\_MCF7 (GSM4271378) and H3K9me3\_BR\_MCF7 (GSM4271318).

#### 8-oxo-dG DIP-seq data

The OxIDIP-Seq data were downloaded from the GEO database (GSE100234) [8]. It contained the genome-wide distribution of 8-oxo-dG accumulation in human non-tumorigenic epithelial breast cells from the MCF10A human cell line. The processed peaks data were provided by the author in bed format.

#### Non-B DNA-forming motifs

Non-B DNA-forming motifs were extracted from the updated version Non-B DB v2.0 database (<https://nonb-abcc.ncifcrf.gov/>, human\_hg19) [9]. An update to correct the A-phased repeat motifs data was received from Frederick National Laboratory for Cancer Research. It includes 13,966,212 motifs covering seven types of non-B DNA structures: A-phased repeats (APR), G-quadruplex DNA (G4 DNA), Z-DNA, direct repeats (DR), inverted repeats (IR), mirror repeats (MR, also H-DNA) and short tandem repeats (STR).

### Function implementation

The functions `bedtools_shuffle` and `bedtools_random` from the `valr` package [10, 11] are utilized to sample genomic regions. The ‘within’ parameter is used to control whether to perform the

with-in chromosome simulation or not. The `bedtools_coverage` is utilized to quantify the overlapped regions between motifs from two genomic regions. Only with the length of overlapped region >0 are the two regions considered co-localized. The visualization functions are implemented with the `ggplot2` package [12] as well as the `ComplexHeatmap` package [13]. The significance annotation function in the visualization is from the `ggpubr` package [14].

## Statistical significance

For the evaluation of statistical significance in the co-localization testing, a Monte-Carlo-based *P*-value is computed. This is executed for each formulated hypothesis. The computation involves a systematic comparison between metrics derived from both simulated and observed datasets. Specifically, the assessment quantifies the proportion wherein the metrics extracted from the simulated datasets are consistently different from the corresponding metrics derived from the actual observed datasets.

## RESULTS

We applied MoCoLo to two case studies that focused on defining co-localization of different genomic and epigenomic features. In our first case study, we investigated the co-localization of two histone markers, H4K20me3 and H3K9me3 (same data type). Case 1 provides a straightforward example of testing co-localization with direct length-only simulation and underscores the importance of two hypothesis tests, as a proof-of-concept. The second case study probed into the co-localization of non-B DNA-forming sequences with 8-oxo-dG lesion sites (different data type). We hypothesized that the distribution of 8-oxo-dG and non-B DNA-forming sequences within the genome differs between motif features. Case 2 highlights the need for feature-informed simulation in the testing framework. Here, both length and percentage of guanine (%G) of sequences were maintained to be similar and thus, minimize their differential effect in testing.

### The same-data-type co-localization testing of histone markers in breast cancer (Case 1)

#### Background

Histone modifications play a significant role in regulating gene expression and maintaining genome stability. Among these modifications, H4K20me3 and H3K9me3 are well known for their roles in the formation of heterochromatin, a condensed form of chromosomal DNA associated with repression of gene expression. H4K20me3 plays roles in heterochromatin formation, gene expression repression [15] and genome stability regulation [16]. Similarly, H3K9me3 is also crucial for heterochromatin formation [17, 18]. Our primary objective was to ascertain the extent of co-localization between H4K20me3 and H3K9me3 in the MCF-7 human breast cancer cell line utilizing the MoCoLo method as a proof-of-concept (Figure 2A).

#### Co-localization testing

H4K20me3 and H3K9me3 are both histone modification data generated from CHIP-seq experiments, thus sharing a data type and displaying comparable peak length distributions (Figure 2B). For our co-localization analysis, we conducted tests bi-directionally: one approach simulated H4K20me3 regions ( $n=31,646$  regions) to establish the statistical distribution, and the alternate approach employed H3K9me3 regions ( $n=34,095$  regions). Same lengths were retained while simulating histone peak regions ( $n=100$ ). We then evaluated the test by using two metrics in terms of the overlapped H4K20me3 and the overlapped

H3K9me3. Both metrics showed significant differences in the observed group compared to the expected group, suggesting co-localization between these two histone markers. The count of overlapping regions is also assessed based on varying overlapping coverages (Figure 2C and D). In addition, we evaluated the co-localization at different genomic locations using the overlapped H4K20me3 as the evaluation metric. The results showed a higher number of overlapped regions in the observed group at exon, intergenic, intron, promoter-TSS (transcription start sites) and transcription termination sites regions (Figure 2E).

The initial dataset for this case study underwent analysis via the segment annotation tool, ChromHMM. This tool delineates genomic regions by highlighting co-occurrence states between H4K20me3 and H3K9me3 [19]. With MoCoLo we were able to formally test for co-localization between histone sites. Both approaches affirm the interaction between H4K20me3 and H3K9me3 sites, either in terms of co-occurrence using ChromHMM or co-localization using MoCoLo.

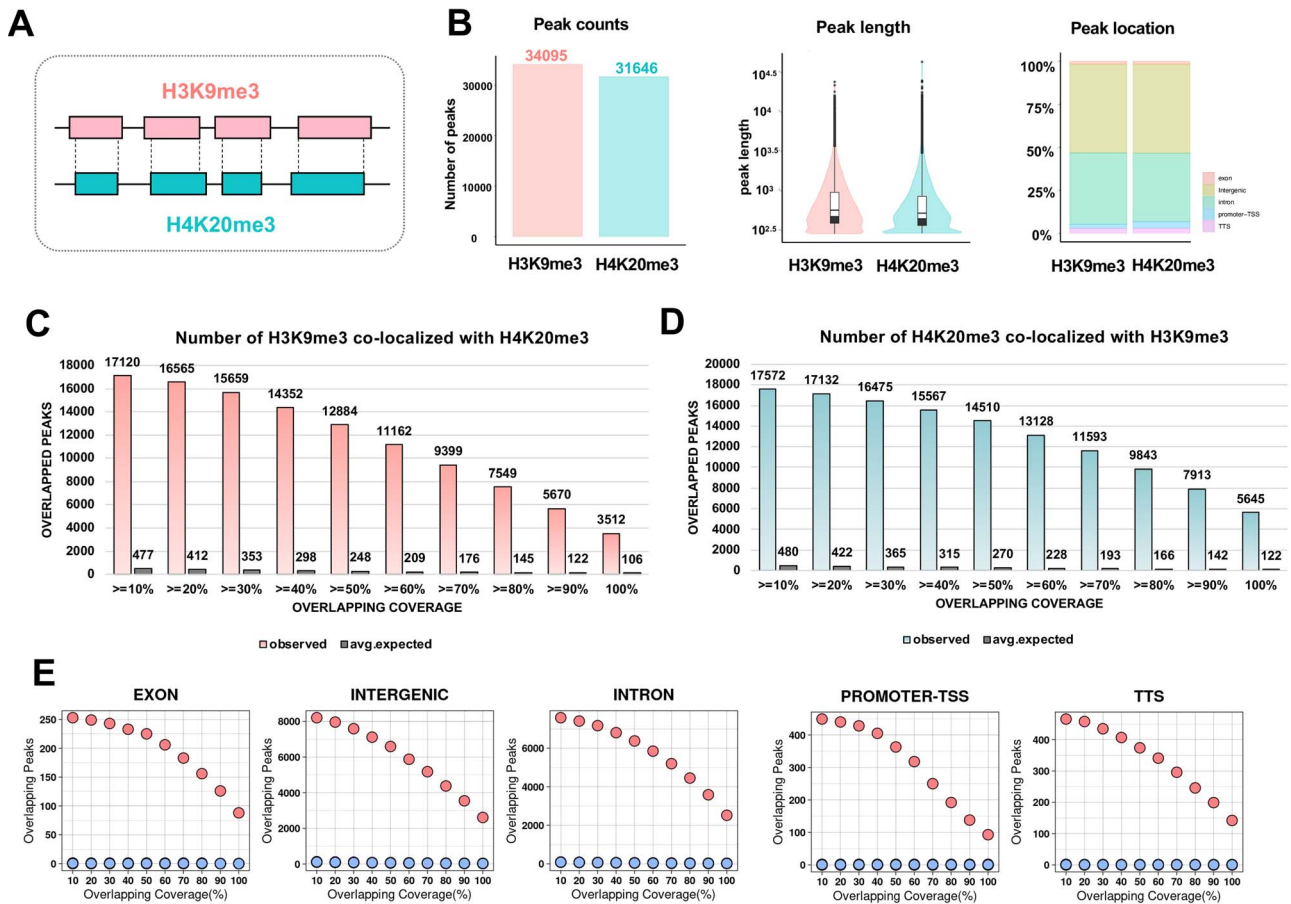
### The across-data-type co-localization testing of endogenous and exogenous features of genomic instability (Case 2)

#### Background

Genomic instability is a hallmark of cancer and other genetic diseases and can result from DNA damage from both exogenous and endogenous sources. Among the four DNA bases (A, T, C, G), guanine (G) has the lowest redox potential and thus has the highest propensity for oxidative damage [20–22]. The oxidative lesion, 8-oxo-dG, therefore serves as a ubiquitous marker of oxidative stress [23, 24] and is a pre-mutagenic lesion contributing to genome instability [20, 25–27]. Sequences that can adopt alternative (i.e. non-B) DNA structures are commonly enriched in guanines [20, 28–30]. Non-B DNA structures have also been shown to be co-localized with mutation hotspots in human cancer genomes [31, 32] and can stimulate the formation of DNA double-strand breaks also jeopardizing genome stability [33–35]. Further, 8-oxo-dG lesions have been shown to be enriched and/or refractory to repair in some types of non-B DNA (e.g. G4 DNA and Z-DNA) [36–41], suggesting that these lesions may accumulate within such structure-forming sequences. The separate occurrences of 8-oxo-dG and non-B DNA-forming sequences are not uniformly distributed across the genome. The non-random distribution of 8-oxo-dG [36] may be due to increased oxidative damage potential and/or varied repair efficiencies within the local environment. We examined the genome-wide co-localization of 8-oxo-dG and non-B DNA-forming regions and whether it differs between non-B DNA structures (Figure 3A), which include A-phased repeats (APR), G-quadruplex DNA (G4 DNA), Z-DNA, direct repeats (DR), inverted repeats (IR), mirror repeats (MR, also H-DNA) and short tandem repeats (STR).

#### Necessity of maintaining G-content in 8-oxo-dG region simulation

The accurate simulation of 8-oxo-dG regions is intrinsically tied to preserving the G-content. When randomizing positions of 8-oxo-dG regions, it is imperative to retain the inherent G-content since 8-oxo-dG is the oxidized form of guanine. Omitting this essential characteristic would lead to a misrepresentation in the simulation. From this standpoint, it becomes evident that the preservation of G-content is an important for the simulation step in this case.



**Figure 2.** MoCoLo evaluates the co-localization of two histone markers, H4K20me3 and H3K9me3 (Case 1). (A) The objective is to assess the significance of co-localization between the H4K20me3 and H3K9me3 histone markers. (B) Peak details for the H4K20me3 and H3K9me3 markers in the MCF-7 breast cancer cell line. Both markers, from the same data type, display comparable peak length distributions: H4K20me3 has 31,646 peaks, and H3K9me3 has 34,095 peaks. (C) and (D) Genome-wide mapping utilizes H4K20me3 and H3K9me3 as pivots to evaluate two distinct metrics. The count of overlapping regions is assessed based on varying overlapping coverages (defined by the minimum intersection size). (E) Regional mapping examines the number of overlapping H4K20me3 peaks in co-localization across various genomic domains, such as exons, intergenic areas, introns, promoter-TSS and TTS. There are more overlapped peaks in the observed group than the expected group. (top dots: observed; bottom dots: expected).

## Testing results

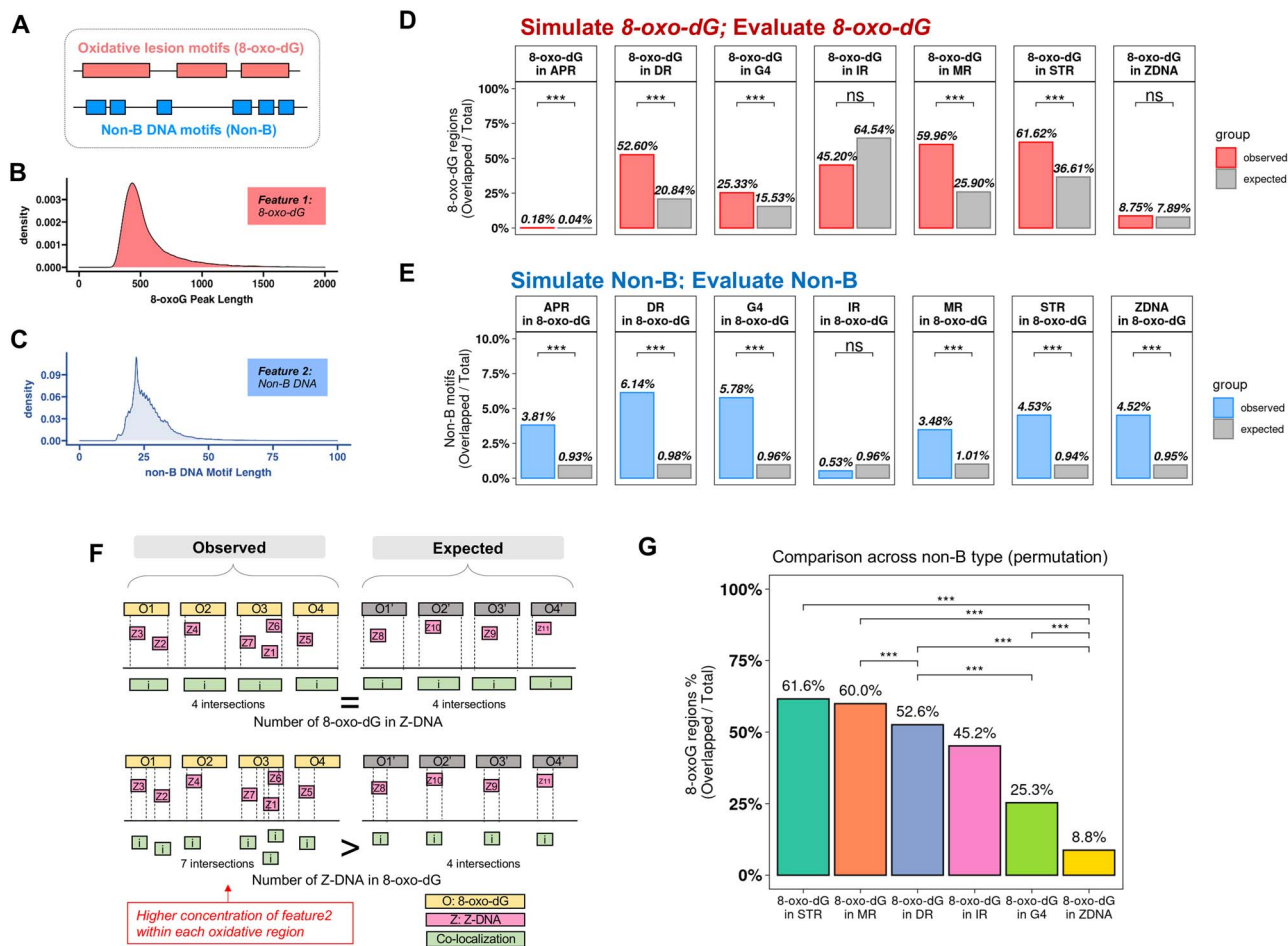
The length of 8-oxo-dG regions from DIP-seq (Figure 3B) and the length of non-B DNA motif (Figure 3C) show a distinct difference. Notably, 8-oxo-dG peaks detected from DIP-seq experiments were overall larger in length (median: ~500 bases) as compared to non-B DNA motifs (median: ~25 bases). This observation underscores the need of reciprocal hypothesis testing (Figure 1E). Further, the sequence property-informed simulation method from MoCoLo was applied to 8-oxo-dG peaks ( $n=50,027$ ) for genomic region simulation ( $n=100$ ) that retains guanine contents in addition to motif lengths.

We observed a significantly higher number of 8-oxo-dG regions co-localizing with five non-B DNA structures (MR, DR, STR, G4 DNA and APR) in the observed group (Supplementary Table 1). Conversely, for IR and Z-DNA, the 8-oxo-dG regions did not exhibit significant co-localization when compared to other random genomic regions (Figure 3D and Supplementary Figure 1A). Furthermore, when evaluating using the non-B DNA motif count as the metric, we identified a significantly higher number of six types of non-B DNA-forming motifs that co-localized in 8-oxo-dG regions compared to the simulated group. These motifs include MR, DR, STR, G4 DNA, Z-DNA and APR (Figure 3E and Supplementary Figure 1B).

The co-localization of APR-forming regions and 8-oxo-dG peak regions only indicate that APRs are located in proximity to the 8-oxo-dG region since A-tracts themselves do not contain guanines. This is because the 8-oxo-dG peaks from DIP-seq experiments are ~500 bp while the A-phased repeats are ~25 bp. Therefore, a 25-bp APR motif may co-localize within a 500-bp 8-oxo-dG region from DIP-seq peaks but does not mean that the one-base-specific oxidative guanine is located within the A-phased repeats themselves. The A-phased repeats are defined as three or more tracts of four to nine adenines or adenines followed by thymines, with centers separated by 11–12 nucleotides [9]. The difference in peak sizes between the two data sets reflects a limitation of the current experimental technology to detect 8-oxo-dG within relatively smaller peak regions (Supplementary Figure 2). It would be more fitting if the 8-oxo-dG sites can be detected in a narrower region or at single-base resolution.

## The dual hypothesis testing identified Z-DNA hotspots within 8-oxo-dG regions

Utilizing both ‘total overlapped 8-oxo-dG motifs’ and ‘total overlapped non-B DNA motifs’ as evaluative metrics bring clarity to the intricacies of feature co-localization, as exemplified by the Z-DNA case. ‘Total overlapped 8-oxo-dG motifs’ measures



**Figure 3.** MoCoLo evaluates the co-localization of 8-oxo-dG and non-B DNA-forming regions (Case 2). (A) The overview of the genome-wide mapping of 8-oxo-dG peaks and non-B DNA motifs. (B) and (C) The length distribution of 8-oxo-dG peaks (median, ~500 bases) and non-B DNA-forming motifs (median, ~25 bases). (D) The numbers of overlapped 8-oxo-dG regions (the observed) that co-localized with non-B DNA motifs by non-B DNA category. 8-oxo-dG shows significant co-localization with six non-B DNA types except IR and Z-DNA. (E) The numbers of overlapped motifs of each non-B DNA type that co-localized with 8-oxo-dG regions. Six non-B DNA types show significant co-localization of their structure forming region and 8-oxo-dG region except IR. (F) While testing the co-localization between Z-DNA and 8-oxo-dG, there is significantly higher frequency of overlapped Z-DNA in the observed group while there is no significant difference of overlapped 8-oxo-dG. The explanation is that there is a high enrichment of Z-DNA in certain 8-oxo-dG regions. Therefore, while counting Z-DNA, there are higher overlapped Z-DNA (bottom) while the overlapped 8-oxo-dG regions stay the same (top). The observation highlights the need and benefits of using two-metric evaluation of co-localization and the importance of pivot feature selection. (G) Comparative analyses of co-localization between different non-B DNA types and 8-oxo-dG. It investigates whether certain non-B DNA types exhibit higher co-localization with 8-oxo-dG compared to others. The evaluation of co-localization by using the number of overlapped 8-oxo-dG regions as the metric and the testing result across non-B DNA types.

the total count of 8-oxo-dG regions that overlapped with non-B DNA, providing insights into the oxidative damage sustained by these motifs. In contrast, the ‘total overlapped non-B DNA motif’ captures the number of non-B DNA motifs present within 8-oxo-dG regions, signifying their placement within oxidatively damaged DNA regions.

For 8-oxo-dG regions that are overlapped with Z-DNA, the total number of 8-oxo-dG is not significantly higher in the observed group than random (Figure 3D). However, when we determined the total overlapped Z-DNA motifs within the 8-oxo-dG peak regions, the number is significantly higher in the observed group ( $P < 0.001$ ) than by random chance (Figure 3E). While these results may appear conflicting, it indicates a high number of overlapped Z-DNA-forming regions within each oxidative region and suggests that Z-DNA may be more frequently affected by oxidative pressures marked by 8-oxo-dG (Figure 3F).

For comparison, we initially employed a simpler strategy that did not consider G-content, resulting in significant findings that suggested an overrepresentation of 8-oxo-dG regions overlapping

with Z-DNA. However, the result is potentially misleading due to the lack of differential G-content consideration, which is likely reflected in the result. By considering G-content, our testing showed that the occurrence of 8-oxo-dG regions overlapping with Z-DNA was not significantly higher than in control groups with similar G-content. This suggests that regions rich in G-content, which include Z-DNA, are not exclusively associated with 8-oxo-dG regions. This finding aligns with biological expectations and reflects a more accurate representation of the biological system under study. Thus, the MoCoLo framework helps to determine the validity of co-localization, supporting the rejection of one or both hypotheses when not substantiated.

### The post-testing comparison after co-localization testing

Comparing the co-localization of 8-oxo-dG and various non-B DNA types, MoCoLo provides additional statistical tests. The goal is to test the co-localization across genomic features. In this case,

the example is the non-B DNA motif, which is stratified into different types. This method is used to investigate whether a specific type of non-B DNA motif demonstrates a more pronounced co-localization with the 8-oxo-dG feature than its counterparts.

To evaluate the co-localization between each pair of non-B DNA types, we employ a permutation analysis ( $n = 100$ ). This involves reshuffling the non-B DNA motif regions across the paired non-B DNA types and conducting a subsequent co-localization analysis for each iteration to establish the null model. The count of overlapping 8-oxo-dG regions is utilized as the metric to compare co-localizations with oxidative regions across the seven non-B DNA categories. These counts of overlapped regions are then normalized (by dividing by the total count of 8-oxo-dG regions or the respective non-B DNA motif library sizes) to ensure comparability.

In terms of the overlapped 8-oxo-dG regions (Figure 3G), we observed significantly higher proportion of 8-oxo-dG regions to co-localize with MR (60.0%) than with DR (52.6%) and Z-DNA (8.8%). The co-localization of 8-oxo-dG and with STR (61.6%) and G4 (25.3%) are significantly higher than with the Z-DNA forming sequences. It also shows significantly higher frequency in DR than in G4 DNA and Z-DNA.

The testing extension provides an alternative perspective to subgroups of genomic regions inherent to a singular genomic feature. Additionally, this approach melds both permutation (resampling within paired non-B DNA types) and bootstrap (simulation of the 8-oxo-dG region) methodologies. This provides more insights in the co-localization and helps us understand how endogenous damage in the DNA and its structures are linked.

## Property-informed simulation ensures G-content retention in 8-oxo-dG simulations

### Simulation design

A straightforward way to simulate genomic regions is to randomly place all regions independently. While this satisfies length considerations, ensuring compositional accuracy, like matching nucleotide compositions, becomes challenging. The simulation here is not simply simulating the sequence. It uses a genome-wide search to find genomic regions with similar sequence properties to the actual motif (Figure 4A). Currently there is not a computation-effective workflow existing to simulate genomic regions with both length and G-content. To counter these inefficiencies, we introduced a new search strategy for simulation in MoCoLo (Figure 1F). Instead of a collective simulation of all motifs, motifs are simulated individually, populating a 'simulation pool' tagged by motif traits such as length and composition. Within each analysis where multiple simulations are needed, those simulated regions that meet requirements are stored in memory to form a simulation pool. From this pool, we then select a motif set that mirrors our actual dataset. A built-in 'dynamic tolerance' mechanism ensures efficient matching, preventing infinite loops by automatically adjusting the simulation tolerance, especially when an exact genome match is elusive.

### G-content variability

For 8-oxo-dG regions, the G-content distribution presents two distinct peaks, approximately at 12.5% and 30.0%. A comparative analysis between simulations—with and without G-content restrictions—demonstrates the necessity of retain %G while simulating 8-oxo-dG regions. The property-informed simulation method in MoCoLo successfully preserves the dual-peak distribution, along with maintaining an identical length distribution (Figure 4B, left). In contrast, neglecting G-content in simulation retains only length distribution (Figure 4B, right).

## Simulation parameters

The selection of parameters plays a pivotal role in simulation. We can observe a minor shift in the G-content distribution, which reflects the simulation tolerance (Figure 4B, left-top). Property-informed simulation in MoCoLo features 'dynamic tolerance'. It is mainly regulated by two parameters: 'starting tolerance (start)' and 'incremental step (step)'. Using the %G simulation as an example, the starting tolerance can vary from zero, indicating that the simulated motif should precisely reflect the %G of the actual motif, to one, which suggests no %G restrictions. In scenarios where the starting tolerance is excessively restrictive, the algorithm autonomously increases the tolerance in pre-defined increments determined by the 'incremental step'. The specific values assigned to 'starting tolerance' and 'incremental step' dictate the characteristics of the simulated groups, subsequently affecting their resemblance to the actual data (Figure 4C). While using restrictive parameters ideally improves similarity, it might inversely affect computational efficiency, resulting in extended running time. Thus, users need to balance between efficiency and precision.

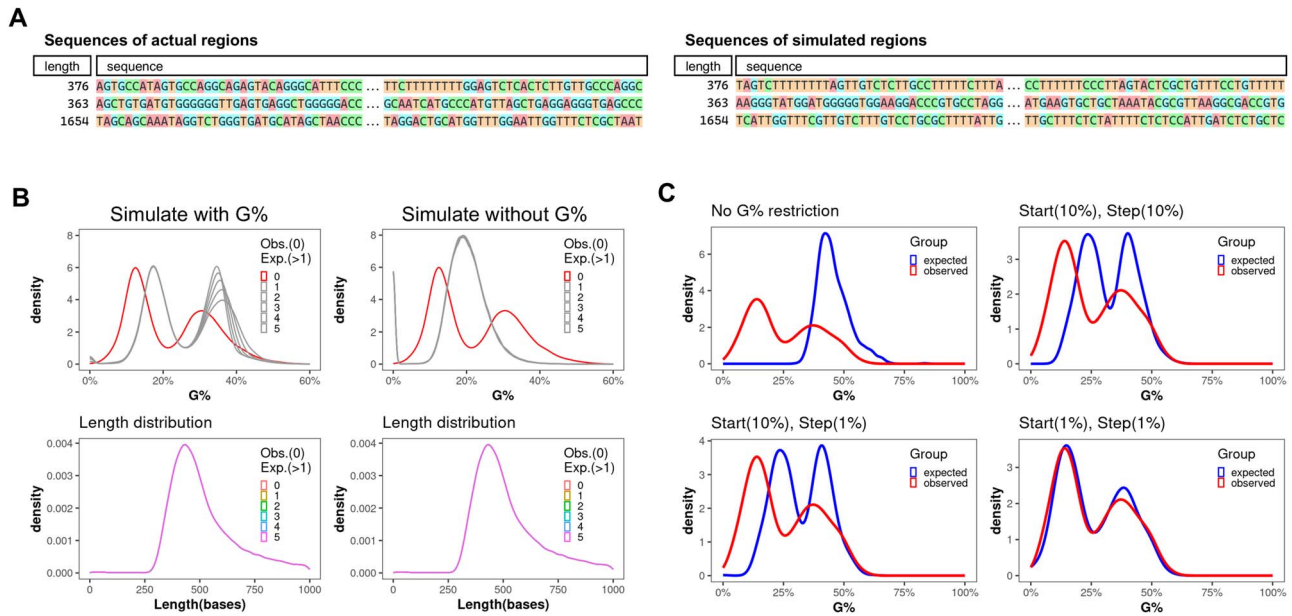
## DISCUSSION

We introduce MoCoLo, a testing framework for genomic co-localization, which has several key innovations and advantages. First, MoCoLo employs a unique approach to co-localization testing that directly probes for genomic co-localization with duo-hypotheses testing. This means that MoCoLo can deliver more detailed and nuanced insights into the interplay between different genomic features. Second, MoCoLo features a novel method for informed genomic simulation, taking into account intrinsic sequence properties such as length and guanine-content. This simulation method enables us to identify genome-wide co-localization of 8-oxo-dG sites and non-B DNA-forming regions, providing a deeper understanding of the interactions between these genomic elements.

## Biological significance

When applied to real-world data, MoCoLo revealed the significant co-localization of H4K20me3 and H3K9me3, vital for heterochromatin formation, in the MCF-7 breast cancer cell line. This aligns with recent findings that underscore the role of histone modifications in regulating gene expression and chromatin structure, which are particularly critical in cancer genomics. Studies have shown that histone modifications can serve as markers for transcriptional repression or activation and are often altered in cancer cells, affecting gene expression patterns crucial for tumor progression and metastasis [42, 43]. The MoCoLo framework, by highlighting the interaction between these modifications, provides a novel angle from which to view chromatin dynamics and their implications in cancer biology. In addition, histone epigenetic marks have been shown to predict somatic mutations, suggesting a complex interplay between chromatin organization and genomic stability. Extending this, it would be intriguing to investigate whether the interplay between non-B DNA motifs and histone marks could influence somatic mutagenesis [44–46].

In addition, we were able to perform a genomic mapping between non-B DNA-forming regions and oxidatively damaged (8-oxo-dG) regions. Our results show significant co-localization of five types of non-B DNA-forming sequences within regions of 8-oxo-dG lesions. Our findings regarding G4 DNA is also consistent with a previous report showing significant enrichment of potential G4 DNA structures within 8-oxo-dG peaks compared to



**Figure 4.** Property-informed simulation with dynamic tolerance maintains G-content of motif sequence. **(A)** The examples of property-informed simulation that retain the properties of motif sequence in terms of length and G-content. **(B)** The distribution of G-content of 8-oxo-dG region includes two G-content peaks for 8-oxo-dG regions occur  $\sim 12.5\%$  and  $30.0\%$ . In the figure legend, ‘Obs. (0)’ denotes the data observed from experimental results. ‘Exp. (>1)’ refers to the expected distributions from multiple simulations, with each number from ‘1’ to ‘5’ representing a distinct simulation iteration. G-content focused simulations underline the significance of %G for 8-oxo-dG. Overlooking G-content captures only length variation, whereas MoCoLo maintains both dual-peak G-content and length distribution, with a minor G-content shift hinting at the simulation’s tolerance. In the figure legend, 0 represents the actual data and 1–5 represent the simulation group. **(C)** The flexibility of the simulation is primarily influenced by two hyper-parameters: ‘starting tolerance (start)’ and ‘incremental step (step).’ The range for starting tolerance spans from zero—denoting an exact match to the %G of the original motif—to one, indicating no constraints on %G. If the starting tolerance is too stringent, the algorithm automatically adjusts the tolerance using defined increments set by the ‘incremental step.’ The chosen values for ‘starting tolerance’ and ‘incremental step’ shape the attributes of the simulated groups, influencing their similarity to the real data. Top-left: An absence of %G constraint results in notable differences between simulated and actual groups; Bottom-right: Low start/step values result in heightened congruence between simulation and actual data, at the price of longer simulation time.

randomly distributed regions in the human genome, as predicted by sequence-based G4 DNA models [8]. Our observations about the high density of Z-DNA in 8-oxo-dG-containing regions complement the growing body of literature that indicates the involvement of non-canonical DNA structures in the regulation of gene expression and the maintenance of genome integrity [47]. By leveraging MoCoLo’s capabilities to compare the co-localization status of different non-B DNA types, we contribute to a more nuanced understanding of how these structures interact with oxidative lesions. The differences in co-localization between the non-B DNA types further underscore the complexity of the genomic architecture and its implications for cellular processes [48]. Future investigations across various cancer cell lines could expand upon these insights and validate the generality of our findings in the broader context of cancer genomics and epigenetics.

## Potential applications

The potential applications of MoCoLo are wide-ranging due to its fundamental role in mapping the complex network of genomic regulation. For example, it can elucidate the concerted actions of transcription factors and histone modifications, which are pivotal in gene expression regulation [49]. This interaction is especially relevant when considering the modulation of gene expression across various cell lines and pathological states. MoCoLo’s ability to analyze genomic sequence motifs further aids in determining transcription factor binding preferences, which are often influenced by sequences like AT- or GC-rich promoters and CpG islands, and how these features contribute to transcription

initiation and silencing based on methylation patterns [50]. By enabling the analysis of reciprocal co-occurrence, MoCoLo provides a robust framework for researchers to investigate the co-localization of diverse genomic motifs—ranging from TF binding sites [51] and CpG islands [52] to splice sites and miRNA binding sites [53]. The implications of this analysis extend from predicting TF binding events to deciphering the mechanisms of gene network regulation, exploring the evolution of gene expression control and identifying biomarkers for various diseases. The sequence-informed simulation aspect of MoCoLo, in particular, offers a refined approach to studying the co-localization of sequence-specific motifs, thereby enriching our understanding of the genomic architecture.

## Method comparison

There exist several strategies to indicate associations and co-occurrences in genomic studies (Table 1): Monte-Carlo-Based Approaches. The design of MoCoLo relies on the principles of Monte-Carlo tests, which are non-parametric models that offer wide test statistics and randomization strategies. These tests, while affording flexibility, come with the inherent challenge of being computationally intensive, demanding precise customization. The degree to which data characteristics are preserved in a null model can significantly influence the conclusions drawn from Monte-Carlo simulations. In an endeavor to perfect these simulations, MoCoLo employs a property-informed simulation technique to uphold sequence properties. An innovative feature introduced is the ‘dynamic tolerance’ in simulations, which modulates the tolerance level of sequence property differences



**Table 1:** Overview of method comparison across different testing strategies

Strategy	Bin-based	Analytical	Empirical
<b>Method</b>	ChromHMM	Bedtools	MoCoLo
<b>Testing</b>	Co-occurrence	Association	Co-localization
<b>Aspect of analysis</b>	Genomic annotation	Genomic Association	Genomic Co-localization
<b>Statistical method</b>	Hidden Markov model (Bernoulli distribution)	Fisher's Exact test (Binomial)	Probability-based
<b>Data resolution</b>	200 bp (user-defined bins)	Dynamic	Dynamic
<b>Pros</b>	- Scalable to multiple features - Designed for chromatin state inference and annotation	- Embedded within Bedtools suite. - Computationally efficient	- Property-informed simulation: Retains sequence properties in simulations for testing. - Dynamic tolerance: Efficient computational cost.
<b>Cons</b>	- Bin size bias for differing feature lengths. - Limited output without direct association testing or P-values.	- Background estimation can affect results. - Assumptions may oversimplify complex systems.	- Require computation resources as an empirical method

between the observed and the simulated groups. The art of formulating a research question in Monte Carlo testing methods plays a pivotal role, as it directly corresponds to the chosen test statistic. A case in point would be the analysis of co-localization of two genomic features, F1 and F2. The query might revolve around whether F1 appears within F2 more than what random chance would suggest. Interestingly, such a proposition can also be viewed from an asymmetric perspective, mandating a diverse test statistic. In order to address both perspectives in a unified framework, MoCoLo introduces dual hypotheses for infer co-localization between F1 and F2 motifs and offers two distinct metrics to test each hypothesis.

**Approaches based on fixed-window segmentation.** A prevalent approach in analyzing the co-occurrence of genomic elements involves segmenting them into multiple pre-defined window sizes, allowing for the calculation of statistics at the window level. Chromatin annotation tools such as ChromHMM, can be used to indicate the co-occurrence of two genomic features (the emission probability of a chromatin state). However, using a single fixed resolution during analysis may not be intuitive to decide resolutions especially when the two features in the testing have distinct length distribution. These tools, despite the output (in terms of chromatin state annotations), can certainly be used as a foundation to study the co-localization of two genomic features. There are challenges existing such as (i) setting up bin-sizes, (ii) restricted by statistical models, (iii) no direct testing significant P-value provided in the output, as the primary objective of segmentation tools is not to test co-localization but to infer the co-occurrence in chromatin states.

**Analytical test-based approaches.** Basic analytical tests often rely on a straightforward null model, like that of Fisher's exact test. When utilizing these tests, it's crucial to assess if the data aligns with the null model and to understand the test's resilience against any misalignments. Adopting an overly simplistic null model can lead to decreased P-values, heightening the chances of false positives. One implementation, Bedtools [35] provides an implementation that can calculate the number of overlaps and the number of intervals unique to each feature. But it requires to infer the number that are not present in each feature as the universal background. Constructing the control set demands meticulous attention when using analytical tests rooted in a universe of regions. Any disparities between the case and control

data sets in attributes such as genomic variability and aggregation could compromise the test's assumptions, potentially resulting in false positives. Recent methods that mine enriched n-wise combinations of genomic features have emerged to explore genomic overlaps by discerning patterns of intersection across multiple genomic datasets [54]. By expanding MoCoLo to include such n-wise overlap strategies, a deeper, more granular analysis of genomic feature co-localization may be explored as a future direction.

In summary, the main advantages of MoCoLo lie in its ability to handle dynamic and sequence-property-informed inputs, its reciprocal hypotheses testing, flexible simulation and its comprehensive output that allows for a more precise understanding of genomic feature co-localization.

#### Key Points

- MoCoLo framework provides a novel method for analyzing spatial interactions of genomic features at sequence-level using reciprocal co-occurrence.
- Property-informed simulation in MoCoLo minimizes confounding factors, enabling robust genome-wide co-localization assessments.
- Through case studies, MoCoLo demonstrated its utility in unveiling significant co-localizations, aiding in deeper molecular understanding.

## SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/bib>.

## FUNDING

This work was supported by grant [RR160093 to support J.K.], research funds from the Department of Oncology, Dell Medical School [to J.K.] and the National Institutes of Health [CA093729 to K.M.V]. Support for this work was partially funded by the Southwestern University's Garey Endowed Chair in Chemistry [to M.Z.F.]; and in part with Federal funds from the National Cancer

Institute, National Institutes of Health, Department of Health and Human Services [No. 75N91019D00024 to B.T].

## AUTHOR CONTRIBUTIONS

Methodology Conceptualization: JK, QX; Case 2 Biological Design and Conceptualization: MZF, IDM, KMV; Analysis: JK, QX, BTL; Original Draft: JK, QX; Editing of Original Draft: KMV, MZF, IDM; Funding Acquisition: JK, KMV.

## DATA AVAILABILITY

The MoCoLo is available under a GPL-3.0 license in the Kowalski Lab GitHub repository. The code can be accessed at <https://github.com/kmlabdms/MoCoLo>.

## REFERENCES

- Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51.
- Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.
- Birney E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;**447**:799–816.
- Kanduri C, Bock C, Gundersen S, et al. Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics* 2019;**35**:1615–24.
- Ferkingstad E, Holden L, Sandve GK. Monte Carlo null models for genomic data. *Statistical Science* 2015;**30**:59–71.
- Heger A, Webber C, Goodson M, et al. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* 2013;**29**:2046–8.
- Gopi LK, Kidder BL. Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains. *Nat Commun* 2021;**12**:1419.
- Amente S, Di Palo G, Scala G, et al. Genome-wide mapping of 8-oxo-7, 8-dihydro-2'-deoxyguanosine reveals accumulation of oxidatively-generated damage at DNA replication origins within transcribed long genes of mammalian cells. *Nucleic Acids Res* 2019;**47**:221–36.
- Cer RZ, Donohue DE, Mudunuri US, et al. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* 2013;**41**:D94–100.
- Riemyndy KA, Sheridan RM, Gillen A, et al. Valr: reproducible genome interval analysis in R. *F1000Research* 2017;**6**:1025.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.
- Wickham H. ggplot2. *Wiley Interdisciplinary Rev Comput Stat* 2011;**3**:180–5.
- Gu Z. Complex heatmap visualization. *Imeta* 2022;**1**:e43.
- Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0, 2023. <https://rpkgs.datanovia.com/ggpubr/>.
- Karachentsev D, Sarma K, Reinberg D, Steward R. PR-Set7-dependent methylation of histone H4 Lys 20 functions in repression of gene expression and is essential for mitosis. *Genes Dev* 2005;**19**:431–5.
- Schotta G, Sengupta R, Kubicek S, et al. A chromatin-wide transition to H4K20 monomethylation impairs genome integrity and programmed DNA rearrangements in the mouse. *Genes Dev* 2008;**22**:2048–61.
- Fischle W, Wang Y, Jacobs SA, et al. Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev* 2003;**17**:1870–81.
- Lachner M, Sengupta R, Schotta G, Jenuwein T. Trilogies of histone lysine methylation as epigenetic landmarks of the eukaryotic genome. *Cold Spring Harbor symposia on quantitative biology* 2004;**69**:209–218.
- Gopi LK, Kidder BL. Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains. *Nat Commun* 2021;**12**:1419.
- Bacolla A, Temiz NA, Yi M, et al. Guanine holes are prominent targets for mutation in cancer and inherited disease. *PLoS Genet* 2013;**9**:e1003816.
- Steenken S, Jovanovic SV. How easily Oxidizable is DNA? One-electron reduction potentials of adenosine and guanosine radicals in aqueous solution. *J Am Chem Soc* 1997;**119**:617–8.
- Kasai H, Tanooka H, Nishimura S. Formation of 8-hydroxyguanine residues in DNA by X-irradiation. *Gan* 1984;**75**:1037–9.
- van Loon B, Markkanen E, Hubscher U. Oxygen as a friend and enemy: how to combat the mutational potential of 8-oxoguanine. *DNA Repair (Amst)* 2010;**9**:604–16.
- Klaunig JE, Kamendulis LM. The role of oxidative stress in carcinogenesis. *Annu Rev Pharmacol Toxicol* 2004;**44**:239–67.
- Kompella P, Vasquez KM. Obesity and cancer: a mechanistic overview of metabolic changes in obesity that impact genetic instability. *Mol Carcinog* 2019;**58**:1531–50.
- Shibutani S, Takeshita M, Grollman AP. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature* 1991;**349**:431–4.
- Markkanen E. Not breathing is not an option: how to deal with oxidative DNA damage. *DNA Repair (Amst)* 2017;**59**:82–105.
- Del Mundo IMA, Vasquez KM, Wang G. Modulation of DNA structure formation using small molecules. *Biochim Biophys Acta Mol Cell Res* 2019;**1866**:118539.
- Wang G, Vasquez KM. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair (Amst)* 2014;**19**:143–51.
- Zhao J, Bacolla A, Wang G, Vasquez KM. Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci* 2010;**67**:43–62.
- Bacolla A, Tainer JA, Vasquez KM, Cooper DN. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res* 2016;**44**:5673–88.
- Xu Q, Kowalski J. NBBC: a non-B DNA burden explorer in cancer. *Nucleic Acids Res* 2023;**51**:W357–64.
- Wang G, Vasquez KM. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc Natl Acad Sci* 2004;**101**:13448–53.
- Wang G, Christensen LA, Vasquez KM. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc Natl Acad Sci* 2006;**103**:2677–82.
- Wang G, Carbajal S, Vijg J, et al. DNA structure-induced genomic instability in vivo. *JNCI: J Nat Cancer Inst* 2008;**100**:1815–7.
- Ohno M, Miura T, Furuichi M, et al. A genome-wide distribution of 8-oxoguanine correlates with the preferred regions for recombination and single nucleotide polymorphism in the human genome. *Genome Res* 2006;**16**:567–75.
- Chan K, Sterling JF, Roberts SA, et al. Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. *PLoS Genet* 2012;**8**:e1003149.

38. Clark DW, Phang T, Edwards MG, et al. Promoter G-quadruplex sequences are targets for base oxidation and strand cleavage during hypoxia-induced transcription. *Free Radic Biol Med* 2012;**53**:51–9.
39. Chan K, Gordenin DA. Clusters of multiple mutations: incidence and molecular mechanisms. *Annu Rev Genet* 2015;**49**:243–67.
40. Ding Y, Fleming AM, Burrows CJ. Sequencing the mouse genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J Am Chem Soc* 2017;**139**:2569–72.
41. Wu J, McKeague M, Sturla SJ. Nucleotide-resolution genome-wide mapping of oxidative DNA damage by click-code-Seq. *J Am Chem Soc* 2018;**140**:9783–7.
42. Yang Y, Zhang M, Wang Y. The roles of histone modifications in tumorigenesis and associated inhibitors in cancer therapy. *J Nat Cancer Center* 2022;**2**:277–90.
43. Audia JE, Campbell RM. Histone modifications and cancer. *Cold Spring Harb Perspect Biol* 2016;**8**(4):a019521.
44. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 2012;**488**:504–7.
45. Polak P, Karlič R, Koren A, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* 2015;**518**:360–4.
46. Georgakopoulos-Soares I, Morganella S, Jain N, et al. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res* 2018;**28**:1264–71.
47. Matos-Rodrigues G, Hisey JA, Nussenzweig A, Mirkin SM. Detection of alternative DNA structures and its implications for human disease. *Mol Cell* 2023;**83**:3622–41.
48. Wang G, Vasquez KM. Dynamic alternative DNA structures in biology and disease. *Nat Rev Genet* 2023;**24**:211–34.
49. Boeva V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Front Genet* 2016;**7**:24.
50. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev* 2011;**25**:1010–22.
51. Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* 2012;**13**:613–26.
52. Issa J-P. CpG island methylator phenotype in cancer. *Nat Rev Cancer* 2004;**4**:988–93.
53. Zhang F, Wang D. The pattern of microRNA binding site distribution. *Genes (Basel)* 2017;**8**:296.
54. Ferré Q, Capponi C, Puthier D. OLOGRAM-MODL: mining enriched n-wise combinations of genomic features with Monte Carlo and dictionary learning. *NAR Genom Bioinform* 2021;**3**:lqab114.