Published online 14 October 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 4 **1**
https://doi.org/10.1093/nargab/lqab098

Characterizing batch effects and binding site-specific variability in ChIP-seq data

Mingxiang Teng [1,*], Dongliang Du[1], Danfeng Chen[2] and Rafael A. Irizarry[3]

[1]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612, USA, [2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA and [3]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA

Received June 16, 2021; Revised September 15, 2021; Editorial Decision October 04, 2021; Accepted October 05, 2021

ABSTRACT

Multiple sources of variability can bias ChIP-seq data toward inferring transcription factor (TF) binding profiles. As ChIP-seq datasets increase in public repositories, it is now possible and necessary to account for complex sources of variability in ChIP-seq data analysis. We find that two types of variability, the batch effects by sequencing laboratories and differences between biological replicates, not associated with changes in condition or state, vary across genomic sites. This implies that observed differences between samples from different conditions or states, such as cell-type, must be assessed statistically, with an understanding of the distribution of obscuring noise. We present a statistical approach that characterizes both differences of interests and these source of variability through the parameters of a mixed effects model. We demonstrate the utility of our approach on a CTCF binding dataset composed of 211 samples representing 90 different cell-types measured across three different laboratories. The results revealed that sites exhibiting large variability were associated with sequence characteristics such as GC-content and low complexity. Finally, we identified TFs associated with high-variance CTCF sites using TF motifs documented in public databases, pointing the possibility of these being false positives if the sources of variability are not properly accounted for.

INTRODUCTION

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has become a routine technique used to measure genome-wide epigenomic events such as transcription factor (TF) binding on DNA (1). However, due to non-specificity of the experimental protocol and other unwanted source of variability, distinguishing signals of interest from obscuring variability is critical toward accurately inferring the genomic locations of protein–DNA binding. Several genomic characteristics have been previously described to affect ChIP-seq measurements, including chromatin structure and GC content in a sample-specific manner (2–5). Current peak callers, the main ChIP-seq data processing tools used to detect protein binding locations, account for these effect separately for each sample (6). This is accomplished by modeling the effects, referred to as background or control signals (4,7–11). However, we find that individual ChIP-seq experiments provide limited power to fully describe all sources of variability, including batch effects. Recently, large consortia, such as the ENCODE project (12), have generated and made publicly available large numbers of ChIP-seq datasets across numerous human cell states and conditions (13). These datasets permit multi-sample analyses and bring into view other important source of unwanted variability.

We identified two main types of variability that if not accounted for can lead to false discoveries: batch effects and site-specific variability observed within biological replicates. These have been previously described for other high-throughput assays (14–18). Possible reasons for observing batch effects are the use of different experimental protocols by different laboratories and changes made to the assays by the manufacturer (2,4). We refer to these as experimental effects. Possible reasons for observing within-replicate site-specific variation include differences in chromatin structure, protein co-regulators and sequence characteristics such as GC-content (2–5). We refer to this as local chromatin related biological variability or, when there is no ambiguity, chromatin variability. Note that we have previously demonstrated that batch effects in ChIP-seq can be partially explained by GC-content biases (4). Here we describe sources of variability that are not fully explained with deterministic quantities such as GC-content.

Lacking the availability of predetermined deterministic covariates, such as the regions GC-content, we model the experimental and chromatin sources of variabilities with random effects rather than fixed effects to represent the variability induced by multiple factors. Specifically, we pose a mixed-effect hierarchical model with fixed effects used to model known conditions or states of interests, such as

*To whom correspondence should be addressed. Tel: +1 813 745 7734; Fax: +1 813 745 6107; Email: mingxiang.teng@moffitt.org

cell-type. We can leverage large ChIP-seq datasets to fit the model. Here, we demonstrate the usefulness of our approach by fitting our model to 211 CTCF ChIP-seq samples from 90 different cell-types measured across three laboratories obtained from the ENCODE project (12,13). We demonstrate that the experimental and chromatin variability is site-specific, that the DNA sequence composition of the CTCF binding sites drives both experimental and chromatin variability, with different sequence compositions affecting these two differently. This divergence of sequence composition underscores the critical role of protein co-regulators play in contributing to ChIP-seq variability by binding to specific sequence motifs co-localized with CTCF binding sites.

MATERIALS AND METHODS

Data acquisition and preprocessing

Human CTCF ChIP-seq data generated by the first production phase of ENCODE project (12) was downloaded from the UCSC data portal (https://genome.ucsc.edu/ENCODE/). Data from three production centers, comprising 94% of CTCF ChIP-seq by phase one ENCODE, were selected for downstream analysis. These centers are laboratories located at the Broad Institute (Broad), University of Texas at Austin (UTA) and University of Washington (UW). ChIP-seq reads were first aligned to human reference genome hg19 using Bowtie2 (19) with parameters '*-k 1 -N 0*'. Samples were removed if <1 million reads were found in the putative CTCF binding sites, resulting 211 ChIP-seq samples for downstream analysis. Here, CTCF binding sites were download from ENCODE encyclopedia (https://www.encodeproject.org/data/annotations/v2/). Only autosomal binding sites were considered to minimize sex biases. Sequencing reads were counted for binding sites using *featureCounts* (20) with read length extended to 150 bp and minimum read mapping quality set as 10. ChIP-seq read counts were then logarithmically transformed and quantile normalized using *voom* (21). In addition, INPUT data were downloaded for the small datasets involving nine cells from three laboratories.

Mixed-effect hierarchical model

Three levels of effects were considered in the model, including fixed effect for cell conditions and random effects for the experimental and chromatin variability. ChIP-seq signal Y for potential binding events thus was modeled as:

$$Y_{sijk} = \alpha_{si} + \beta_{sj} + \gamma_{sk} + \delta_{sijk}$$

where *s, i, j, k* indicate binding sites/locations, cell types, laboratories and technical replicates, respectively. Here the α_{si} are fixed effect for cell types indexed by i at location indexed by s, the β_{sj} are a normally distributed random effects for laboratory indexed by j at locations indexed by s with standard deviation σ, γ_{sk} are normally distributed random effects for biological replicates indexed by $_k$ for binding *site* indexed by s with standard deviation τ, and δ is measurement error with standard deviation ε. Here, σ and τ represented the experimental and chromatin variability, respectively.

Read counts in high-throughput sequencing data tend to follow over-dispersed Poisson distributions, with varied dispersion along expected read count (7). In order to consider variabilities between binding sites with small and large read counts equally, *voom* normalized ChIP-seq signals were selected to fit the mixed-effect model (21). Here, two components of the *voom* were considered: the logarithm values of ChIP-seq signal and their respective weights characterizing its statistical confidence. For each binding site, effects were estimated by weighted mixed-effects linear regression using *blmer* function in *R* package *blme* (22). Standard deviation of two random effect variables were extracted as the estimated variabilities. It is noted that chromatin variabilities were approximated as the sum of site-specific effects and random error here.

PWM score calculation

To assess the association between estimated standard deviations and binding site genuineness, we scanned CTCF motif enrichment in all analyzed CTCF binding sites. Specifically, we used the human CTCF motif from the JASPAR 2018 database (23) to define a position weight matrix (PWM) score sequence of the same size as the motif (24). We assigned a PWM score to each binding site by selecting the maximum PWM within the region that includes the binding site and its flanking 150 bp regions at both sides.

GC content calculation

GC content for narrow CTCF binding sites, ~150 bp, was calculated with a robust strategy as we described previously (4). In brief, to accommodate the sequencing reads that were partially located inside binding sites, GC content of 150 bp flanking regions were taken into consideration during the estimation of GC content for CTCF binding sites using a weighted approach.

4-mer motif frequency

All possible 4-mers were first generated not including complementary 4-mers. Their proportional frequencies within each CTCF binding site were then calculated by counting their appearances based on 4 bp sliding windows. Proportional frequencies were averaged across binding sites as the 4-mer frequencies for a given group, e.g. binding sites with high GC content and high chromatin variability. 4-mer motif enrichment was estimated by comparing their frequencies between high and low variability sites, with a difference deviated from three standard deviations from the mean difference identified as significant.

Canonical motif enrichment using publicly documented TFBSs

We first applied *DREME* motif discovery (25) to identify *de novo* motifs (Figure 5) enriched in high variability sites (top 5%) compared to the control sites (the remaining 95%) for each site group. Enriched motifs (*DREME E*-value < 0.001, Motif width >= 5 bp) were further compared with known transcription factor binding motifs documented in

public databases (JASPAR (23) and UniPROBE (26)) using Tomtom (27), in order to identify TFs associated with high ChIP-seq variabilities.

RESULTS

CTCF dataset

We collected a large set of CTCF binding ChIP-seq data for an insulator (28) that was analyzed in detail by the ENCODE project. This dataset was composed of 211 samples, the most of any transcription factors studied by ENCODE. It spanned 90 human cell types and three experimental laboratories. QC control, filtering, pre-processing and normalization were applied (see Materials and Methods section). To avoid the variability introduced by peak calling algorithms, we focused our analysis on CTCF 284 712 putative binding sites (Materials and Methods section) previously assembled by the ENCODE project (29). On average these sites were 150 base pair long.

Using exploratory data analysis, we identified regions associates with cell-type, laboratory and within-biological-replicate variability (Figure 1). These regions clearly demonstrate how not accounting for obscuring variation can lead to false positives. Sites that showed only strong cell-type effects exhibited a bimodal distribution across the cell-types, consistent with signals between associated with binding (on) and non-binding (off) events (Figure 1A and B). For a given site, all on-event samples showed comparable high signals while off-event samples held similar low signals. We saw this for both regions that were mostly on (Figure 1A) and cell-specific (Figure 1B). In contrast, sites that were strongly associated with the laboratory in which they were processed (Figure 1C and D) had a more continuous distribution, indicating potential batch-related variability induced by differences in experimental settings. Figure 1G highlighted such variability in confounding with cell type differences. Sites exhibiting strong unidentified biological variability (Figure 1E and F) also exhibited more of a continuous distribution, consistent with this being induced by differences in chromatin properties that span a wide range rather than discrete states. Figure 1H highlighted such variability across different replicates in different laboratories for two cell types.

Mixed-effect hierarchical model accounts for all source of variability

We fitted a random effects model with the cell-type effects represented with fixed effects and the experimental and chromatin variability represented with random effects. In this model, we accounted for the different levels of variability with site-specific standard deviations for each of the two types of effects. To fit this model, we used a weighted regression approach to account for different levels of variability for low and high signals (Materials and Methods section).

We fit this model to two datasets: the full QC-controlled 211 samples (Materials and Methods section) with imbalanced samples across different laboratories, and a balanced subset of 50 samples spanning 9 cell conditions and 3 laboratories. Most cell types were only profiled by one laboratory in the full dataset (imbalanced) while nearly all cell types were profiles by all laboratories in the smaller subset (balanced). Similar standard errors were estimated with the two datasets (Pearson correlation at 0.97 and 0.91 for experimental and chromatin variabilities, respectively), indicating the robustness of our proposed model and that we could use the balanced smaller subset to obtain precise estimates (Supplementary Figure S1A and B). If not further referred, the results presented in the rest of the paper are for the results of the balanced smaller subset.

The range of estimated standard errors for experimental variability were higher than that for chromatin variability (Supplementary Figure S1C). This indicates that batch effects were the dominant source of variability for a significant portion of binding sites, which is consistent with previous publications reporting on ChIP-seq and other sequencing techniques (4,16,30).

To demonstrate the utility of the mixed-effect model fit for separating signal from noise, we applied hierarchical clustering to the CTCF binding signals with high signal to noise ratios. Specifically, we filtered the top 500 CTCF binding sites ranked by the statistical significance of their across-cell-type variability and applied clustering to only those sites. This resulted in groups of samples with well recognized cell identities (Figure 2A). In contrast, when we applied the same clustering algorithm but after filtered to the top 500 sites based on their experimental variability ability, we clearly distinguished samples from the different labs (Figure 2B). As expected, if we filtered by sites with high chromatin variability but low for the other two sources of variability, it did not result in grouping of any obvious meaning (Figure 2C). In summary, we identified 57% of binding sites show high (upper quartile) cell effects, experimental variability or chromatin variability, among which 11%, 10% and 8% are dominated by only one effect (identified as lower half in the other two effects), respectively.

Variability summaries can detect false positives

To further assess our model's performance on all putative CTCF binding sites, we examined if the estimated variability summaries were informative in detecting false positive binding sites. We evaluated the enrichment of canonical CTCF motifs (23,31) on all CTCF binding sites using position weight matrix (PWM) score (32) (Materials and Methods section). Both experimental and chromatin variability estimates were found to be negatively associated with the PWM score across all binding sites, indicating that sites with high variance are more likely to be false positive detection of binding sites (Figure 3A and B). Specifically, while experimental variability showed a decreasing monotonic trend with PWM score (Pearson correlation -0.19), a bimodal distribution was observed for chromatin variability with the low variance cluster presenting relatively high PWM score. The bimodal feature implied that chromatin variability is sounded in protein binding sites regardless with high or low confidence, suggesting it as an intrinsic property of ChIP-seq data. It is also worth noting that, overall, general chromatin variability was higher than experimental variability for the majority of binding sites (Figure 3), although the range of latter is larger than the former (Supplementary Figure S1C). These together suggest that experimental dis-

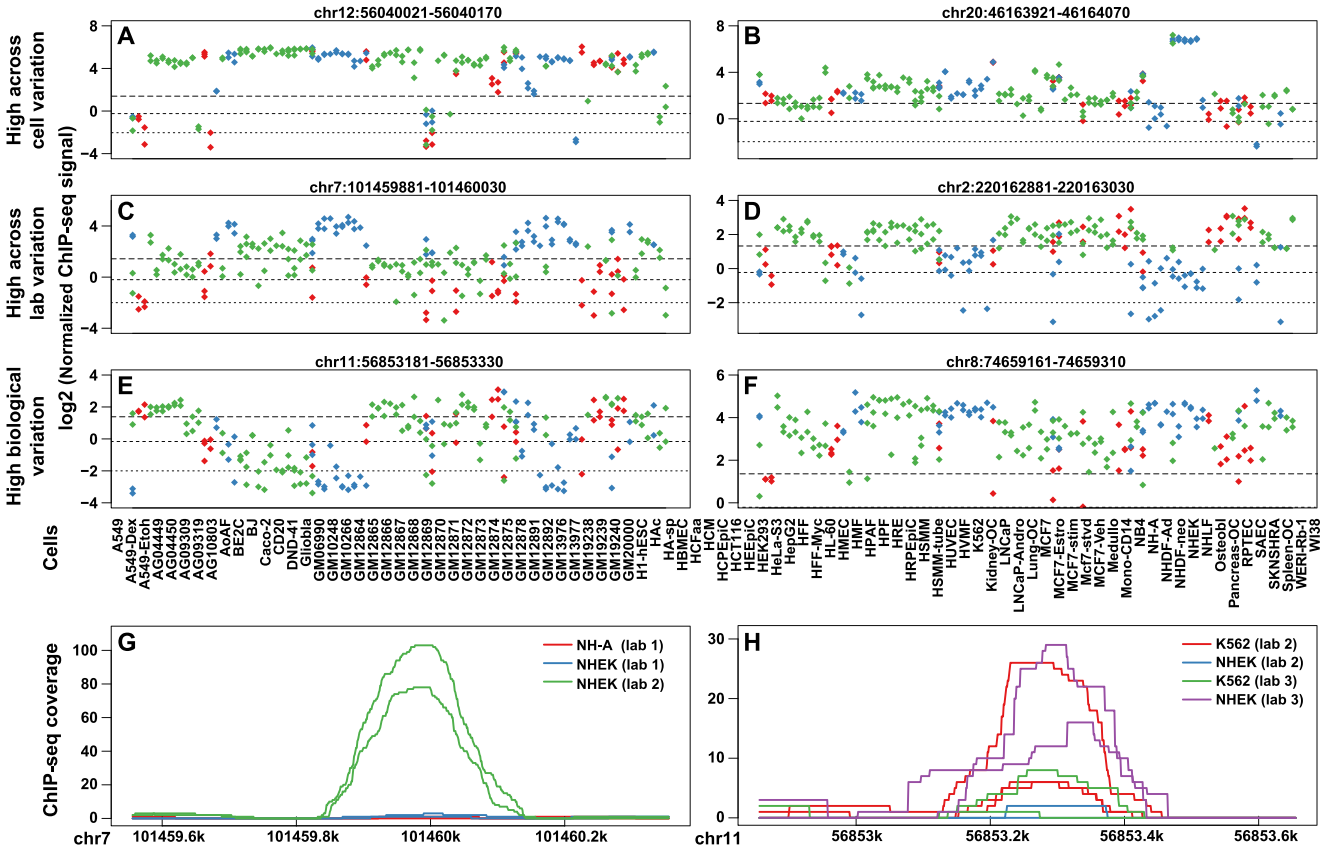

Figure 1. Illustrative CTCF binding sites demonstrating different types of variability in ChIP-seq signals. (**A** and **B**) Two CTCF binding sites showing strong cell-type effects. (**C** and **D**) Two sites showing strong batch effects associated to laboratories. (**E** and **F**) Two sites presenting high withing replicate variability across all cell-types. Cell types corresponding to each column of plotted dots are listed in the same order at the bottom. The same cell types under different library preparation or treatment protocols are labeled as different cell names with an abbreviation suffix of the protocols/treatments. Dot colors represent samples from different laboratories. Dashed lines at different styles represent the normalized ChIP-seq signals at 25, 50 and 75 percentiles. (**G** and **H**) ChIP-seq pileup signals of two example regions in (C and E) for selected cells. Signals of 800 bp centering the binding sites are shown. In each figure, replicates from the same laboratories are in the same colors.

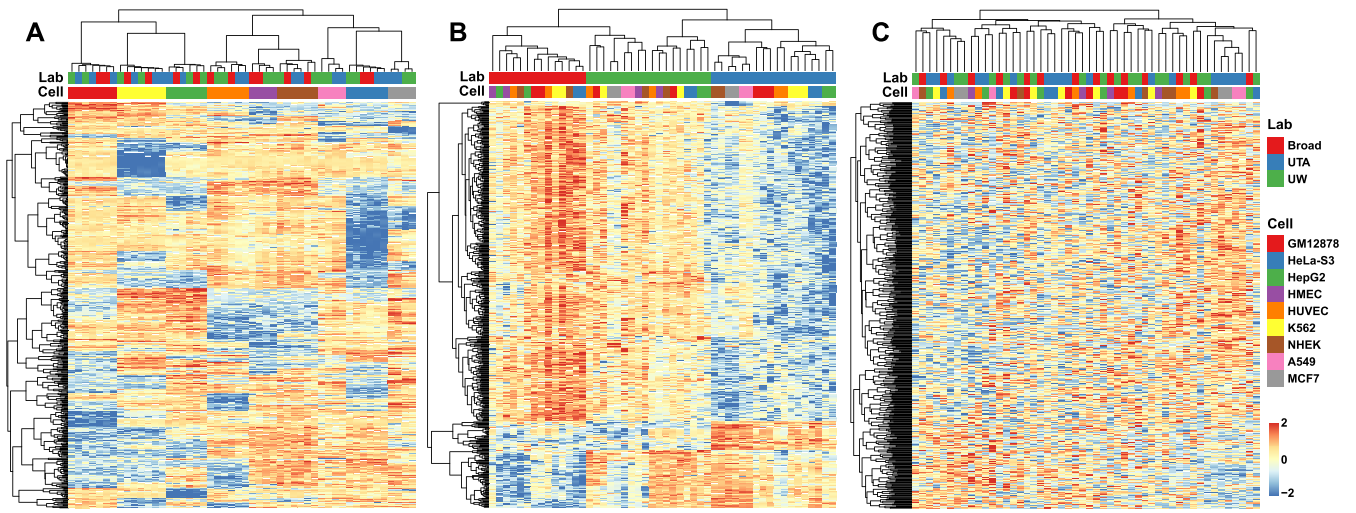

Figure 2. Heatmap of ChIP-seq signals based on CTCF binding sites filtered by high level of the three different types of variability. (**A**) Heatmap of sites with strongest cell-type effects but low variabilities (lower quartile). (**B**) Heatmap of sites with strongest experimental variability but low cell effects and chromatin variability. (**C**) Heatmap of sites with high levels of chromatin but low cell effects. Each heatmap contains 500 sites (rows) and 50 samples (columns). Colored bars at the top of heatmaps represent different laboratories and cell types.

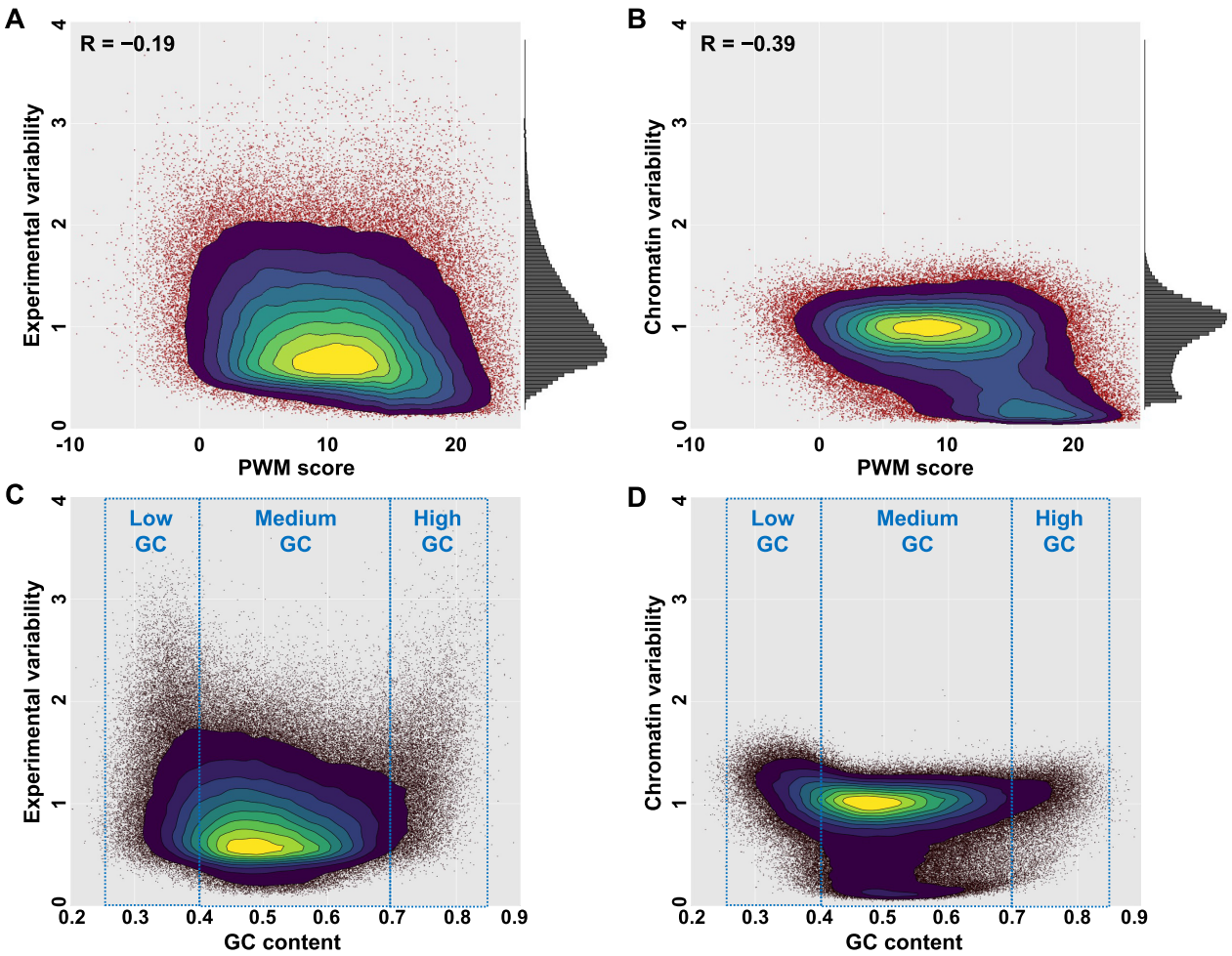

Figure 3. Variabilities associated to CTCF motif enrichment and GC-content bias. Both experimental variability (**A** and **C**) and chromatin variability (**B** and **D**) are showed. Pearson correlations with CTCF motif enrichment are labeled at the top left in (**A** and **B**). Different GC-content levels are highlighted with dotted squares in (**C** and **D**).

crepancies acted as local events instead of global by affecting a selected portion of binding sites when introducing ChIP-seq variability.

It is noted that the analysis above is based on CTCF binding sites annotated during phase 2 of ENCODE project. ENCODE has been improving their data processing protocol constantly to create more reliable annotations. For instance, their up-to-date annotation (v4 1.5.1) has incorporated IDR approach (33) to control reproducibility between sample replicates. We found that IDR analysis could decrease false positive detection of CTCF binding sites especially when chromatin variability is high (Supplementary Figure S2). In brief, we performed the same analysis with IDR-controlled sites as we did for all CTCF binding sites before. We compared the distributions of the estimated variabilities and found IDR-controlled sites resulted in a lower frequency of high chromatin variability. We didn't observe significant improvement in terms of accounting for experimental variability. This is expected as IDR analysis doesn't intend to control cross-sample/cell variance such as batch effects.

Variabilities enhanced by GC-content biases

We have previously reported that GC-content introduces significant bias in ChIP-seq signals (4). To understand the roles GC-content plays in the variability described here, we compared the estimated standard deviations to the GC-content of each CTCF binding sites (Figure 3C and D). Here, GC-content was calculated using a robust approach to ease the boundary effects of narrow CTCF binding sites (Materials and Methods section). Higher variability was observed on binding sites with extreme high or low GC-content. Such GC-content biases were statistically significant ($P < 10\text{^}(-16)$) for both types of variabilities when compared to medium GC binding sites, while inflated effects were observed at some sites with high experimental variability (Figure 3C and D). It has been hypothesized that GC-content bias is introduced mainly by PCR-amplification during sequencing library preparation (34,35). However, our analysis suggests that in addition to that, other factors, such as local chromatin characteristics could also contribute to GC-content biases, underlining the interplay between local sequence composition and ChIP-seq variability.

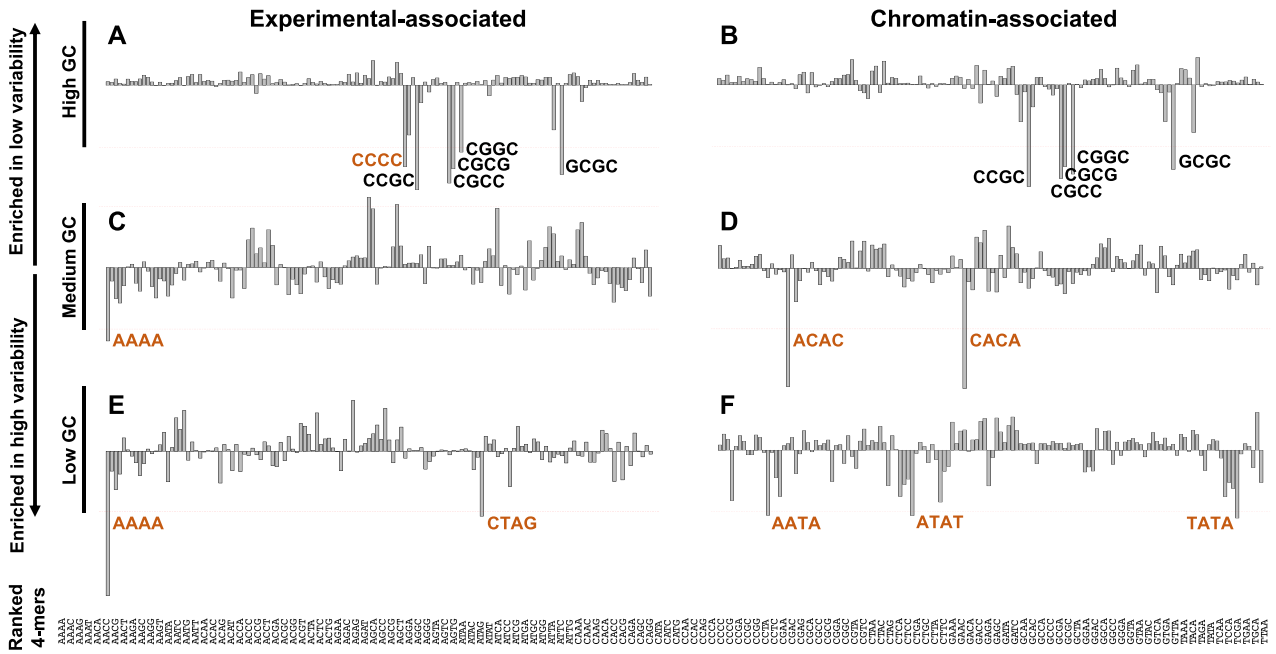

Figure 4. 4-mer enrichment in CTCF binding with high variabilities. Enriched 4-mers (bars deviated from *Mean – 3*SD* of overall difference, the dotted lines) were labeled with their names next to the plots. Subfigures are ordered by GC-content levels (high: **A** and **B**; medium: **C** and **D**; low: **E** and **F**) and variability type (experimental: **A**, **C** and **E**; chromatin: **B**, **D** and **F**). Names of the 4-mers corresponding to plotted bars are listed in the same order at the bottom.

Low sequence complexity is associated with high variability

Because GC-content bias is confounded with the estimated ChIP-seq variability, we stratified the CTCF binding sites into three GC-content groups for downstream analysis. Superficially we divided the sites into groups with 0.25-0.4, 0.4-0.7 and 0.7-0.85 GC proportions, respectively (Figure 3C and D). We further explored the roles of genomic sequence composition in ChIP-seq variability by comparing high to low variability sites in each group.

We defined high variability sites as the top 5% in each GC-content group, and compared them with the remaining 95% of low variability sites (control). The 5% threshold was chosen to reflect a clear increase among each binding site group (Supplementary Figure S3). We counted the frequencies of all 4-mers for CTCF binding sites and calculated the differences of the average 4-mer frequencies between high and low variability sites (Materials and Methods section). Several 4-mers showed strong enrichment in high variability sites, suggesting their roles in increasing ChIP-seq variability (Figure 4). Note that 4-mers hardly showed enrichment in the 95% low variability sites. When summarizing the enriched 4-mers across different binding site groups (different GC-content levels and types of variabilities), we found that they were characterized by simple sequence compositions, usually containing only 1 or 2 nucleotides (18 out of 19 in total enriched 4-mers). This suggests that low sequence complexity acts as a key characteristic for the high variability sites.

We further compared the 4-mer enrichment in sites with high levels of both experimental and chromatin variability. Some 4-mers showed variability-specific enrichment, while other 4-mers were shared by both groups (Figure 4). Specifically, sites with high experimental variability were highly enriched with low complexity 4-mers across all GC-content levels (e.g., AAAA and CCCC). In contrast, sites with high chromatin variability were uniquely enriched with palindrome 4-mers or 2 bp repeats (i.e. ATAT, TATA, ACAC and CACA). This suggests that while low sequence complexity is a universal feature across high variability ChIP-seq sites, experimental covariates tend to be more sensitive to extremely low complexity genomic regions than chromatin-associated covariates.

Protein co-localization plays a role in high variability sites

Several studies have reported the confounding effects of co-localized proteins on ChIP-seq studies (3,5). We thus extended the above sequence composition analysis to study the enrichment of canonical transcription factor binding sites (TFBS) (23) at high variability sites of CTCF binding (Materials and Methods section). The sequence of sites associated with canonical motifs showed low complexity among high variabilities, similar to the results on 4-mer analysis (Figure 5A).

We then generated a list of potentially co-localized proteins for which motifs were enriched in the high variability sites for each GC-content group (Materials and Methods section, Supplementary Table S1). Unique and shared motifs were found between different types of variability (Figure 5A). In summary, we identified a list of known TFs that were previously reported to co-localize with CTCF binding sites, e.g. YY1 (36) and SMAD3 (37), indicating their co-colocalization could increase variability levels in CTCF

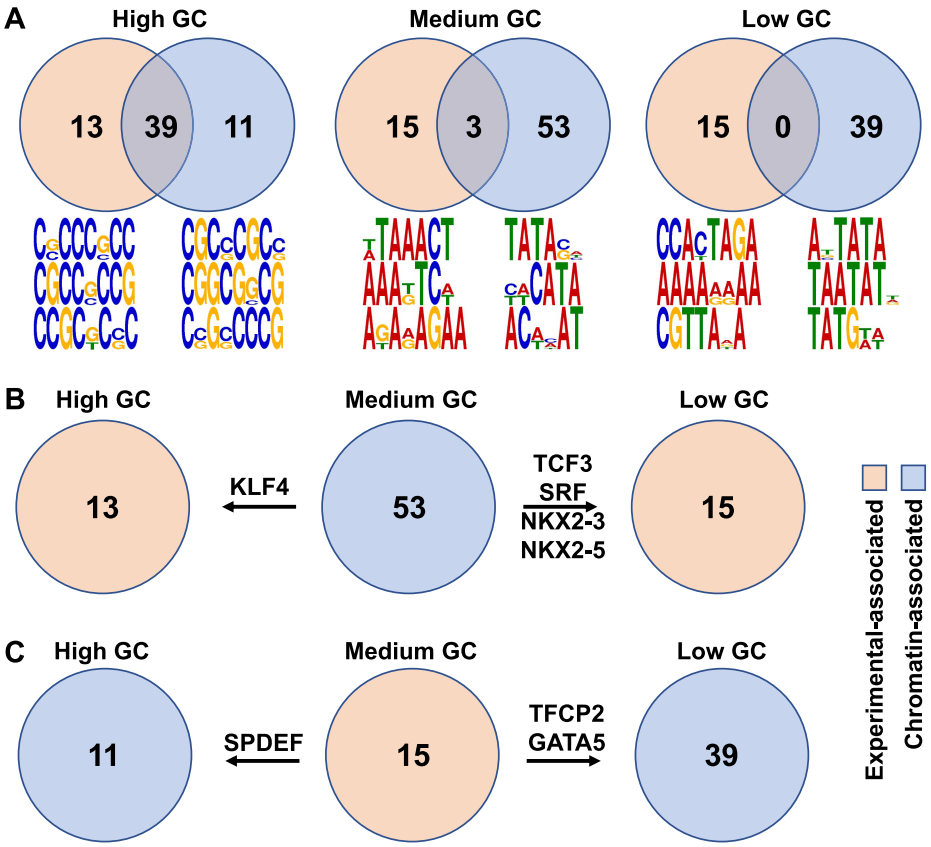

Figure 5. Transcription factors associated to high ChIP-seq variabilities. (**A**) Venn diagram listing the numbers of enriched TFs in high ChIP-seq chromatin variability sites. Top three associated motifs (limited to 8 bp long) are also listed with each group. (**B**) Enriched TFs shifting from chromatin variability sites to experimental variability sites with GC-content decreasing and increasing. Names are listed for the shifted TFs. (**C**) Similar to (**B**) but with shifting direction reversed.

ChIP-seq protocols. A number of other proteins were also highlighted to associate with high variability sites, although we did not identify a mechanism for these (Supplementary Table S1). A possible explanation of co-factor induced variability might attribute to the varied effects from co-factors due to their presence or alterations across experiments.

Finally, we compared TFBS enrichment across different GC-content binding sites groups. We observed a more complex pattern of enrichment for TFBSs (Figure 5B and C). For example, TCF3 was found to be enriched in sites with high chromatin variability sites in medium GC-content, while it was enriched in high experimental variability sites for low GC-content binding sites. This highlights the roles of combinatorial effects between GC-content and protein co-localizers in contributing to ChIP-seq variability.

DISCUSSION

We proposed a mixed-effect hierarchical model to estimate differences across states or conditions of interest and deconvolve variability into experimental- and chromatin-associated from ChIP-seq data. We demonstrated the utility of this approach on a CTCF ChIP-seq dataset. We found that low sequence complexity underlined high variability, with different patterns of low complexity corresponding to different sources of this variability. By studying the sequence motifs of the high variability sites, we identified the combinatorial roles of GC-content bias and different co-localized proteins in increasing ChIP-seq variability. Our approach is applicable for large sets of ChIP-seq data from the same TF. Currently, there are not many such datasets which is why we only applied the approach to CTCF. We expect the availability of such datasets to increase.

We modeled experimental and chromatin effects as random effects to represent broad factors affecting ChIP-seq data from experimental preparation and site-specific effects, respectively, and estimated cell conditions as fixed effects. However, due to the complexity of bias source factors, it is hard to clearly define the types of effects for some bias factors. For example, different antibodies trigger differences at the experimental preparation and the interplay with local chromatin structure, suggesting their contribution to both experimental and chromatin variabilities. As a result, although we termed 'experimental' and 'chromatin' to represent batch effects and site-specific variability separately, caution should be taken to precisely interpret the estimated variabilities. Particularly, we selected laboratories as the batch representative for a list of confounded factors, including library preparation protocol, applied antibody, sequencing instruments and cell culture-related lab effects etc. The estimated experimental variabilities thus showed clear

association with laboratory information, although they reflected the combination of effects from multiple confounding factors. Similarly, chromatin variability reflects another combination of site-specific effects such as local chromatin structure, motif characteristics, cell culture-related replicate effects, co-factors and their interplay with antibodies etc. Fixed cell effects reflect a combination of biological effects such as bio-sample information, cell types, drug treatment and DNA variants (38–40) etc. Due to their confounding with cell conditions, their effects were estimated as part of the fixed cell effect in our model. We highlighted DNA variant effects on CTCF binding using an example (Supplementary Figure S4) demonstrating different alleles (evaluated by Strelka2 (41)) are associated with different levels of CTCF binding on 14 LCL cell lines from the UW laboratory.

ChIP-seq INPUT data has been successfully applied to remove local structural biases in peak calling (7–9). However, there is ongoing debate regarding the inclusion of INPUT profiles in comparing peak differences across samples (42–45). We didn't incorporate INPUT data in our model to simplify estimation and avoid introducing additional biases due to library preparation differences between ChIP and INPUT data. In practice, studies have demonstrated that INPUT differences are usually not strong enough to dominate the real ChIP-seq difference (42–44). For sanity check, we analyzed INPUT data separately with the same model we applied for CTCF data. We estimated the experimental and chromatin variabilities at CTCF binding sites in INPUT dataset and compared them with those estimated with CTCF dataset (Supplementary Figure S5). We didn't see significant co-occurrence of variability although some binding sites do show high variability in both CTCF and INPUT, suggesting potential improvement at these sites if INPUT data are accounted for. In addition, we didn't see in INPUT data the bimodal distributed chromatin variability as presented in CTCF data. One explanation is that CTCF binding shows low chromatin variability when the local motif is highly preferred, while INPUT has no preference and shows high randomness.

We have described an association between sequence composition and ChIP-seq variability. However, further experiments are needed to better interpret these findings. For instance, although PWM scores are widely accepted to measure protein binding affinities for a given genomic sequences, they do not completely determine binding events. For example, CTCF was found to be one of the boundary markers for topologically associating domains (TAD) on chromatin recently (46,47). The biological meaning of CTCF ChIP-seq variability could be better described using TAD information once 3D chromatin interaction data (48) become available. In addition, future work is needed to fully capture the nature of low complexity sequences triggering ChIP-seq variabilities, with the aid of tools such as SEG program (49).

DATA AVAILABILITY

The data underlying this article are available in UCSC Data Portal at https://genome.ucsc.edu/ENCODE/.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Dr Brooke Fridley, Dr Aik-Choon Tan and Dr Ann Chen for the informative discussion.

FUNDING

Moffitt Cancer Center (Department Pilot Project Fund to M.T.); National Cancer Institute [P30CA076292 to Moffitt BBSR]; National Institute of General Medical Sciences [R35GM131802 to R.A.I.].
Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Chen,Y., Negre,N., Li,Q., Mieczkowska,J.O., Slattery,M., Liu,T., Zhang,Y., Kim,T.K., He,H.H., Zieba,J. *et al.* (2012) Systematic evaluation of factors influencing chip-seq fidelity. *Nat. Methods*, **9**, 609–614.
3. Teytelman,L., Thurtle,D.M., Rine,J. and van Oudenaarden,A. (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 18602–18607.
4. Teng,M. and Irizarry,R.A. (2017) Accounting for GC-content bias reduces systematic errors and batch effects in chip-seq data. *Genome Res.*, **27**, 1930–1938.
5. Worsley Hunt,R. and Wasserman,W.W. (2014) Non-targeted transcription factors motifs are a systemic component of chip-seq datasets. *Genome Biol.*, **15**, 412.
6. Bailey,T., Krajewski,P., Ladunga,I., Lefebvre,C., Li,Q., Liu,T., Madrigal,P., Taslim,C. and Zhang,J. (2013) Practical guidelines for the comprehensive analysis of chip-seq data. *PLoS Comput. Biol.*, **9**, e1003326.
7. 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.
8. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of chip-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
9. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of chip-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
10. Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Wong,W.H. (2008) An integrated software system for analyzing chip-chip and chip-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.
11. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from chip-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.
12. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
13. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
14. Leek,J.T. (2014) svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.*, **42**, e161.
15. Li,S., Labaj,P.P., Zumbo,P., Sykacek,P., Shi,W., Shi,L., Phan,J., Wu,P.Y., Wang,M., Wang,C. *et al.* (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.*, **32**, 888–895.
16. Zhang,Y., Parmigiani,G. and Johnson,W.E. (2020) ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.*, **2**, lqaa078.

17. Risso,D., Ngai,J., Speed,T.P. and Dudoit,S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**, 896–902.
18. Rasnic,R., Brandes,N., Zuk,O. and Linial,M. (2019) Substantial batch effects in TCGA exome sequences undermine pan-cancer analysis of germline variants. *BMC Cancer*, **19**, 783.
19. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**, 357–359.
20. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
21. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
22. Chung,Y., Rabe-Hesketh,S., Dorie,V., Gelman,A. and Liu,J. (2013) A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, **78**, 685–709.
23. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Cheneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.
24. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
25. Bailey,T.L. (2011) DREME: motif discovery in transcription factor chip-seq data. *Bioinformatics*, **27**, 1653–1659.
26. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.
27. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
28. West,A.G., Gaszner,M. and Felsenfeld,G. (2002) Insulators: many functions, many mechanisms. *Genes Dev.*, **16**, 271–288.
29. Roadmap Epigenomics,C., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
30. Hicks,S.C., Townes,F.W., Teng,M. and Irizarry,R.A. (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, **19**, 562–578.
31. 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.
32. Tan,G. and Lenhard,B. (2016) TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics*, **32**, 1555–1556.
33. Li,Q., Brown,J.B., Huang,H. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.
34. Kebschull,J.M. and Zador,A.M. (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.*, **43**, e143.
35. Aird,D., Ross,M.G., Chen,W.S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. *Genome Biol.*, **12**, R18.
36. Donohoe,M.E., Zhang,L.F., Xu,N., Shi,Y. and Lee,J.T. (2007) Identification of a ctcf cofactor, Yy1, for the x chromosome binary switch. *Mol. Cell*, **25**, 43–56.
37. Van Bortle,K., Peterson,A.J., Takenaka,N., O'Connor,M.B. and Corces,V.G. (2015) CTCF-dependent co-localization of canonical smad signaling factors at architectural protein binding sites in d. melanogaster. *Cell Cycle*, **14**, 2677–2687.
38. Kasowski,M., Grubert,F., Heffelfinger,C., Hariharan,M., Asabere,A., Waszak,S.M., Habegger,L., Rozowsky,J., Shi,M., Urban,A.E. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.
39. Kilpinen,H., Waszak,S.M., Gschwind,A.R., Raghav,S.K., Witwicki,R.M., Orioli,A., Migliavacca,E., Wiederkehr,M., Gutierrez-Arcelus,M., Panousis,N.I. *et al.* (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, **342**, 744–747.
40. Kasowski,M., Kyriazopoulou-Panagiotopoulou,S., Grubert,F., Zaugg,J.B., Kundaje,A., Liu,Y., Boyle,A.P., Zhang,Q.C., Zakharia,F., Spacek,D.V. *et al.* (2013) Extensive variation in chromatin states across humans. *Science*, **342**, 750–752.
41. Kim,S., Scheffler,K., Halpern,A.L., Bekritsky,M.A., Noh,E., Kallberg,M., Chen,X., Kim,Y., Beyter,D., Krusche,P. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
42. Lun,A.T. and Smyth,G.K. (2016) csaw: a bioconductor package for differential binding analysis of chip-seq data using sliding windows. *Nucleic Acids Res.*, **44**, e45.
43. Lun,A.T. and Smyth,G.K. (2014) De novo detection of differentially bound regions for chip-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Res.*, **42**, e95.
44. Ross-Innes,C.S., Stark,R., Teschendorff,A.E., Holmes,K.A., Ali,H.R., Dunning,M.J., Brown,G.D., Gojis,O., Ellis,I.O., Green,A.R. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.
45. Tu,S., Li,M., Chen,H., Tan,F., Xu,J., Waxman,D.J., Zhang,Y. and Shao,Z. (2021) MAnorm2 for quantitatively comparing groups of chip-seq samples. *Genome Res.*, **31**, 131–145.
46. Zuin,J., Dixon,J.R., van der Reijden,M.I., Ye,Z., Kolovos,P., Brouwer,R.W., van de Corput,M.P., van de Werken,H.J., Knoch,T.A., van,I.W.F. *et al.* (2014) Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 996–1001.
47. Nanni,L., Ceri,S. and Logie,C. (2020) Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries. *Genome Biol.*, **21**, 197.
48. Kempfer,R. and Pombo,A. (2020) Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.*, **21**, 207–226.
49. Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.