Robust estimation of Hi-C contact matrices by fused lasso reveals preferential insulation of super-enhancers by strong TAD boundaries and a synergistic role in cancer

Yixiao Gong[1,2,+], Charalampos Lazaris[1,2,+], Aurelie Lozano[3], Prabhanjan Kambadur[4], Panagiotis Ntziachristos[5], Iannis Aifantis[1,2,*], Aristotelis Tsirigos[1,2,6,*]

[1] Department of Pathology, NYU School of Medicine, New York, NY 10016, USA

[2] Laura and Isaac Perlmutter Cancer Center and Helen L. and Martin S. Kimmel Center for Stem Cell Biology, NYU School of Medicine, New York, NY 10016, USA

[3] Center for Computational and Statistical Learning, IBM T.J. Watson Research Center, NY 10598, USA

[4] Bloomberg LP, 731 Lexington Avenue, New York City, NY, USA

[5] Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

[6] Applied Bioinformatics Laboratories, NYU School of Medicine, NY 10016, USA

[+] Equal contribution

* To whom correspondence should be addressed. Tel: +16465012693; Email: Aristotelis.Tsirigos@nyumc.org; Correspondence may also be addressed to Iannis Aifantis. Tel: +1 212 263 9898, Fax: +1 212 263 9210, E-mail: Ioannis.Aifantis@nyumc.org

ABSTRACT

The metazoan genome is compartmentalized in megabase-scale areas of highly interacting chromatin known as topologically associating domains (TADs), typically identified by computational analyses of Hi-C sequencing data. TADs are demarcated by boundaries that are largely conserved across cell types and even across species, although, increasing evidence suggests that the seemingly invariant TAD boundaries may exhibit plasticity and their insulating strength can vary. However, a genome-wide characterization of TAD boundary strength in mammals is still lacking. A systematic classification and characterization of TAD boundaries may generate new insights into their function. In this study, we use fused two-dimensional lasso as a machine-learning method to first improve Hi-C contact matrix reproducibility, and, subsequently, categorize TAD boundaries based on their strength. We demonstrate that increased boundary strength is associated with elevated CTCF levels and that TAD boundary insulation scores may differ across cell types. Intriguingly, we observed that super-enhancer elements are preferentially insulated by strong boundaries. Furthermore, a pan-cancer analysis revealed that strong TAD boundaries and super-enhancer elements are frequently co-duplicated. Taken together, our findings suggest that super-enhancers insulated by strong TAD boundaries may be exploited, as a functional unit, by cancer cells to promote oncogenesis.

INTRODUCTION

The advent of proximity-based ligation assays has allowed us to probe the three-dimensional chromatin organization at an unprecedented resolution [1, 2]. Hi-C, a high-throughput chromosome conformation variant, has enabled genome-wide identification of chromatin-chromatin interactions [3]. Hi-C has revealed that the metazoan genome is organized in areas of active and inactive chromatin known as A and B compartments respectively [3]. These are further compartmentalized in super-TADs [4], topologically associating domains (TADs) [5–7] and sub-TADs [8], as well as gene neighbourhoods [9]. Several algorithms have been already developed to reveal this hierarchical chromatin organization, including Directionality Index (DI)

[5], Armatus [10], TADtree [11], Insulation Index (Crane) [12], IC-Finder [13] and others. However, none of these studies has systematically explored the properties of the hierarchical organization of TADs. Additionally, although TADs are seemingly invariant, mounting evidence suggests that TAD boundaries can vary in strength, ranging from permissive TAD boundaries that allow more inter-TAD interactions to more rigid (strong) boundaries that clearly demarcate adjacent TADs [14]. Recent studies have shown that in *Drosophila*, exposure to heat-shock caused local changes in certain TAD boundaries resulting in TAD merging [15]. A recent study showed that during motor neuron (MN) differentiation in mammals, TAD and sub-TAD boundaries in the *Hox* cluster are not rigid and their plasticity is linked to changes in gene expression during differentiation [16]. It has also been demonstrated that boundary strength is positively associated with the occupancy of structural proteins including CCCTC-binding factor (CTCF) [5]. Despite the fact that there is a handful of studies demonstrating that TAD boundaries can vary in strength in organisms like *Drosophila*, no study has yet addressed the issue of boundary strength in mammals and how it may be related to potential boundary disruptions and aberrant gene activation in cancer. Here we introduce a new method based on fused two-dimensional lasso [17] in order to: (a) robustly estimate Hi-C contact matrices, (b) categorize TAD boundaries based on their insulating strength, (c) characterize TAD boundaries in terms of CTCF binding and other functional elements, and (d) investigate potential genetic alterations of TAD boundaries in cancer. We anticipate that our study will help generate new insights into the significance of TAD boundaries.

MATERIALS AND METHODS

Comprehensive re-analysis of published high-resolution Hi-C datasets

In order to develop a method that successfully handles variation in Hi-C data and improves reproducibility, we carefully selected our Hi-C datasets to represent technical variation due to the execution of the experiments by different laboratories and/or the usage of different restriction enzymes. We identified publicly available human Hi-C datasets that fulfilled the following criteria: (i) availability of two biological replicates and (ii) sufficient sequencing depth

to robustly identify topologically-associating domains (TADs) as described in our TAD calling benchmark study [18]. Specifically, we ensured that our datasets included samples with at least ~40 million intra-chromosomal read pairs and that the Hi-C experiment was performed in biological replicates, either by using one restriction enzyme (HindIII or MboI) (H1 cells and their derivatives [19], K562, KBM7 and NHEK cells [20] and in-house generated CUTLL-1), or two enzymes (HindIII or MboI) (GM12878 [20], IMR90 [5, 21]), in order to examine the consistency of predicted Hi-C interactions across different enzymes. All datasets were then comprehensively re-analysed using our HiC-bench platform [18]. Quality assessment analysis revealed that the samples varied considerably in terms of total numbers of reads, ranging from ~150 million reads to more than 1.3 billion (**Supplementary Figure 1a**). Mappable reads were over 96% in all samples. The percentages of total accepted reads corresponding to *cis* (ds-accepted-intra, dark green) and *trans* (ds-accepted-inter, light green) (**Supplementary Figure 1b**) also varied widely, ranging from ~17% to ~56%. Duplicate read pairs (*ds-duplicate-intra* and *ds-duplicate-inter*; red and pink respectively), non-uniquely mappable (*multihit*; light blue), single-end mappable (*single-sided*; dark blue) and unmapped reads (*unmapped*; dark purple) were discarded. Self-ligation products (*ds-same-fragment*; orange) and reads mapping too far (*ds-too-far*; light purple) from restriction sites or too close to one another (*ds-too-close*; orange) were also discarded. Only double-sided uniquely mappable *cis* (*ds-accepted-intra*; dark green) and *trans* (*ds-accepted-inter*; light green) read pairs were used for downstream analysis. Despite the differences in sequencing depth and in the percentages of useful reads across samples, all samples had enough useful reads for TAD. However, due to the wide differences in sequencing depth, and to ensure fair comparisons of Hi-C matrices in this study, all datasets were down-sampled such that the number of usable intra-chromosomal reads pairs was ~40 million for each replicate. Finally, to study the effect of sequencing depth, we also resampled at ~80 and ~120 million read pairs, by limiting our evaluation to those samples that had adequate sequencing depth.

Scaled Hi-C contact matrices

Hi-C contact matrices were scaled by: (a) the total number of (usable) intra-chromosomal read pairs, and (b) the "effective length" of the corresponding pair of interacting bins [22]. More specifically, the scaled Hi-C count corresponding to interactions between the Hi-C matrix bins *i*,*j* (y_{ij}) is defined by the formula:

$$y_{ij} = \frac{x_{ij}}{eff_i \cdot eff_j \cdot N}$$

where x_{ij} is the original number of interactions between the bins *i* and *j*, eff_i the effective length for the bin *i*, eff_j the effective length for the bin j, and *N* is the total number of read pairs.

Distance-normalized Hi-C contact matrices

Genomic loci that are further apart in terms of linear distance on DNA tend to give fewer interactions in Hi-C maps than loci that are closer. For intra-chromosomal interactions, this effect of genomic distance should be taken into account. Consequently, the interactions were distance-normalized using a *z*-score that was calculated taking into account the mean Hi-C counts for all interactions at a given distance *d* and the corresponding standard deviation. Thus, the *z*-score for the interaction between the Hi-C contact matrix bins *i* and *j* (z_{ij}) is given the following equation:

$$z_{ij} = \frac{y_{ij} - \mu(d)}{\sigma(d)}$$

where y_{ij} corresponds to the number of interactions between the bins *i* and *j*, $\mu(d)$ to the mean (expected) number of interactions for distance d=|j-i| and $\sigma(d)$ is the corresponding standard deviation of the mean.

Fused two-dimensional lasso

While our naïve scaling approach successfully increased the cross-enzyme and same-enzyme correlation of Hi-C matrices, we sought to improve the correlation even further. We used two-dimensional lasso, an optimization machine learning technique widely used to analyse noisy datasets, especially images [17]. This technique is very-well suited for

identifying topological domains based on contact maps generated by Hi-C sequencing experiments for two reasons: (a) Hi-C datasets are inherently noisy, and (b) topological domains are continuous DNA segments of highly interacting loci that would represent solid squares along the diagonal of Hi-C contact matrices. Topological domains map to squares of different length along the diagonal of the Hi-C contact matrix, but they are not solid as they contain several gaps, i.e. scattered regions on those squares that show little or no interaction. Two-dimensional fused lasso [23] addresses the issue by penalizing differences between neighbouring elements in the contact matrix. This is achieved by the penalty parameter λ (lambda), as described in the equation:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda \sum_{(i,j) \in E} \left| \beta_i - \beta_j \right| ,$$

where y is the original (i.e. observed) contact matrix, and $\hat{\beta}$ is the estimated contact matrix such that the objective function described above in minimized. In the interest of computational efficiency, we also applied one-dimensional lasso on the Hi-C contact matrices in order to estimate the matrices for high values of λ ($\lambda >> 1$) and obtain the full hierarchy of TAD boundaries. Using one-dimensional lasso instead of the two-dimensional version had no negative impact on the correlations of Hi-C contact matrices between replicates (**Supplementary Figure 2**).

Calculation of same-enzyme and cross-enzyme correlations

We calculated two types of correlation for Hi-C matrices, to evaluate the performance of our method: (a) same-enzyme correlation which corresponds to all the Hi-C replicates prepared with the same restriction enzyme, (b) cross-enzyme correlation which corresponds to all the sample pairs where the same Hi-C sample was prepared with two different enzymes (e.g HindIII/MboI). Pearson correlation coefficients were calculated either on the filtered, ICE-corrected [24] or scaled Hi-C contact matrices (Pearson) or the distance-normalized ones (Pearson/*z*-score).

TAD boundary "ratio" insulation score

Given a potential TAD boundary, we denote the "upstream" TAD to the left of the boundary as *L*, and the "downstream" TAD to the right as *R*. The cross-TAD interactions between *L* and *R* are denoted as *X*. The "ratio" insulation score is defined as follows:

$$\text{ratio} = \text{intra}_{max}/\text{inter}$$

where:

$$\text{intra}_{max} = \max(\text{mean}(L), \text{mean}(R))$$

$$\text{inter} = \text{mean}(X)$$

For more details, see [18].

Classification of boundaries based on fused two-dimensional lasso

We applied two-dimensional fused lasso to categorize TAD boundaries based on their strength. The rationale behind this categorization is that topological domains separated by more "permissive" (i.e. weaker) boundaries [25] will tend to fuse into larger domains when lasso is applied, compared to TADs separated by well-defined, stronger boundaries. We indeed applied this strategy and categorized boundaries into multiple groups ranging from the most permissive to the strongest boundaries. The boundaries that were lost when λ value was increased from 0 to 0.25, fall in the first category (λ=0), the ones lost when λ was increased to 0.5, in the second (λ=0.2) etc.

Association of CTCF levels with boundary strength

We obtained CTCF ChIP-sequencing data for the cell lines utilized in this study (with the exception of KBM7 for which no publicly available dataset was available) and we uniformly re-processed all data using HiC-bench [18]**.** Total CTCF levels (i.e. aggregated peak intensities from potentially multiple CTCF peaks) at each TAD boundary were calculated and their normalized distributions for each boundary category (weak to strong) were plotted in boxplots in order to demonstrate the association of increased boundary strength with increased levels of CTCF binding. We performed this analysis separately for TSS-only and non-TSS CTCF

187 binding sites. The rationale behind this separate analysis was based on the observation that
188 several TAD boundaries, especially strong boundaries, contain TSSs.
189
190 **Association of boundary strength with super-enhancers**
191 Super-enhancers were called using H3K27ac ChIP-seq data from GEO, ENCODE and in-
192 house generated data. Reads were first aligned with Bowtie2 v2.3.1 [26] and then HOMER
193 v4.6 [27] was used to call super-enhancers, all with standard parameters. For each super-
194 enhancer in each sample, we identified the corresponding TAD and its TAD boundaries. We
195 then calculated (per sample) the percentage of super-enhancers that are surrounded by
196 boundaries belonging in each boundary category, demonstrating that most super-enhancers
197 are insulated by strong boundaries.
198
199 **RESULTS**
200 **Analysis workflow**
201 The overall workflow, including our benchmark strategy and downstream analysis, is
202 summarized in **Figure 1a**. Our analysis starts with unprocessed Hi-C contact matrices
203 ("filtered" matrices). We then generate processed Hi-C matrices using both ICE "correction"
204 and our naïve "scaling" approach. Then, fused two-dimensional lasso is applied either on the
205 actual matrices or, alternatively, on the distance-normalized matrices. Matrix reproducibility
206 between biological replicates is assessed across samples for a variety of parameters, for
207 example, resolution, distance between interacting loci, sequencing depth, etc. Finally,
208 downstream analysis, involves the characterization of TAD boundaries based on their
209 insulating strength, the enrichment in CTCF binding, proximity to repeat elements and super-
210 enhancers, and, finally, their genetic alterations in cancer.
211
212 **Assessment of same-enzyme and cross-enzyme reproducibility of Hi-C contact**
213 **matrices**

Hi-C is prone to biases and multiple algorithms have been developed for Hi-C bias correction, including probabilistic modelling methods [22], Poisson or negative binomial normalization [28] and the widely popular Iterative Correction and Eigenvalue decomposition method (ICE) [24], which assumes “equal visibility” of genomic loci. A similar iterative method named Sequential Component Normalization was introduced by Cournac *et al.* [29]. Additional efficient correction methods have been developed to handle high-resolution Hi-C datasets [30]. However, estimating highly reproducible Hi-C contact maps remains a challenging task [31], especially at high resolutions, as we also demonstrate below. Specifically, we focused on multiple factors that may play an important role on reproducibility: first, we separately considered biological replicates of Hi-C libraries generated with the same or different restriction enzymes; second, we studied the impact of Hi-C matrix resolution (i.e. bin size); third, we assessed reproducibility as a function of the distance of interacting loci pairs. Pearson correlation coefficients were calculated for each pair of replicates (same- or cross-enzyme) on Hi-C contact matrices estimated by three methods: (i) naïve filtering (i.e. matrix generation by simply using double-sided accepted intra-chromosomal read pairs from **Supplementary Figure 1a**), (ii) iterative correction (ICE) which has already been demonstrated to improve cross-enzyme correlation, and (iii) our own “naïve” scaling method that only corrects for effective length bias (see Methods for details). Importantly, correlations were computed both on the actual matrices, but also on the distance-normalized matrices (see Methods for details), as Hi-C interactions are typically concentrated around the diagonal of the Hi-C contact matrix, and values are dropping exponentially as the distance between the interacting pairs is increasing (**Supplementary Figure 1c**). Distance-normalized matrices account for the expected Hi-C read count as a function of distance and may therefore reveal real distal interactions. The results of our benchmark analysis are summarized in **Figure 1b**: the left panel summarizes the correlations between replicates generated by the same restriction enzyme, whereas the right panel the correlations between replicates generated by a different restriction enzymes. In both scenarios, as expected, correlations drop quickly as finer resolutions (from 100kb to 20kb) are considered, especially in the distance-normalized matrices. The same conclusion applies for increasing

distance (from 2Mb to 10Mb) between interacting loci, demonstrating that long-range interactions require ultra-deep sequencing (beyond what is currently available in most of the datasets in this study) in order to be detected reliably. To elaborate on this point, we repeated the analysis after retaining only those samples with two replicates of at least 70 million or 110 million usable intra-chromosomal reads and resampling them down to 80 million or 120 million per replicate (**Supplementary Figure 3** and **Supplementary Figure 4** respectively). Both conclusions hold true with the new sequencing depth and are independent of the Hi-C contact matrix estimation method. Finally, bias-correction methods (ICE and our scaling approach) indeed improved cross-enzyme correlation over the naïve filtering method (**Figure 1b**). Interestingly, this improvement came at the expense of lower correlations in the same-enzyme case. More specifically, we observed that the largest the gain in cross-enzyme correlations, the greater the loss in same-enzyme correlations (ICE method) (**Figure 1b**).

Fused lasso improves same-enzyme and cross-enzyme correlations of Hi-C contact matrices

Motivated by the poor performance of all methods at fine resolutions and by the observation of a surprising trade-off between cross-enzyme and same-enzyme correlations when correcting for enzyme-related biases, we applied fused two-dimensional lasso [23], to obtain improved estimates of Hi-C contact matrices. Briefly, two-dimensional fused lasso introduces a parameter λ which penalizes differences between neighboring values in the Hi-C contact matrix (see Methods for details). The effect of parameter λ is demonstrated in **Figure 2a** where we show an example of the application of fused two-dimensional lasso on a Hi-C contact matrix focused on an 8Mb locus on chromosome 8 for different values of parameter λ. To evaluate the performance of fused lasso, we calculated same-enzyme and cross-enzyme Pearson correlations between Hi-C contact matrices generated from different replicates. Pearson correlation coefficients were calculated either for iteratively-corrected (ICE) or scaled Hi-C contact matrices (at different λ values) and compared to the naïve filtering approach. The results are summarized in **Figure 2b**. Increasing λ improves correlation independent of

resolution, restriction enzyme and bias-correction method, demonstrating the robustness of our approach. Similarly, fused two-dimensional lasso improves the reproducibility of distance-normalized matrices as demonstrated in **Figure 2c**. In all cases, as the value of λ increases, the relative improvements in correlation are diminished. This observation can guide the selection of λ, however a minimum of two replicates per sample are necessary to compute the correlation and implement this strategy. Instead, we propose the use of degrees of freedom as described in [23]. As demonstrated in **Supplementary Figure 2b**, the degrees of freedom are rapidly decreasing for small values of λ, and quickly reaching a plateau with a moderate increase in λ.

Fused lasso reveals a TAD hierarchy linked to TAD boundary strength

After demonstrating that parameter λ improves reproducibility of Hi-C contact matrices independent of the bias-correction method, we hypothesized that increased values of λ may also define distinct classes of TADs with different properties. For this reason, we now allowed λ to range from 0 to 5 (after a finite value of λ, λ>>5, the entire Hi-C matrix attains a constant value independent of the value of λ). For efficient computation, we used a one-dimensional approximation of the two-dimensional lasso solution (see Methods for details and **Supplementary Figure 2**). We then identified TADs at multiple λ values using HiC-bench, and we observed that the number of TADs is monotonically decreasing with the value of λ (**Figure 3a**), suggesting that by increasing λ, we are effectively identifying larger TADs encompassing smaller TADs detected at lower λ values. Equivalently, certain TAD boundaries "disappear" as λ is increased. Therefore, we hypothesized that TAD boundaries that disappear at lower values of λ are weaker (i.e. lower insulation score), whereas boundaries that disappear at higher values of λ are stronger (i.e. higher insulation score). To test this hypothesis, we identified the TAD boundaries that are "lost" at each value of λ, and generated the distributions of the insulation scores for each λ. As insulation score, we used the Hi-C "ratio" score (see Methods), which was shown to outperform other TAD calling methods [18]. Indeed, as hypothesized, TAD boundaries lost at higher values of parameter λ are associated

with higher TAD insulation scores (**Figure 3b**). We then stratified TAD boundaries into five classes (numbers 1 through 5 in **Figure 3b**; zero corresponds to lack of boundary) according to their strength, independently in each Hi-C dataset used in this study. A heatmap representation including all TAD boundaries and their associated class across all samples is depicted in **Figure 3c**. Unbiased hierarchical clustering correctly grouped replicates and related cell types independent of enzyme biases or batch effects related to the lab that generated the Hi-C libraries, suggesting that TAD boundary strength can be used to distinguish cell types. Equivalently, this finding suggests that, although TAD boundaries have been shown to be largely invariant across cell types, a certain subset of TAD boundaries may exhibit varying degrees of strength in different cell types. As expected, TAD boundary strength was found to be positively associated with CTCF levels, suggesting that stronger CTCF binding confers stronger insulation. Since we noticed that several TAD boundaries contain TSSs, this analysis was done separately for all CTCF peaks (data not shown) and TSS-only CTCF peaks (**Figure 3d**). Both approaches revealed the same trend, with the exception of the class of strongest boundaries, where CTCF levels in TSS regions were significantly higher compared to non-TSS regions, suggesting that the strongest boundaries are formed by CTCF-mediated loops at gene promoters. SINE elements have also been shown to be enriched at TAD boundaries [5], and besides confirming this finding, we now demonstrate that Alu elements (the most abundant type of SINE elements) are enriched at stronger TAD boundaries (**Supplementary Figure 5**, top-left panel). A comprehensive analysis of all major repetitive element subtypes can be found in **Supplementary Figure 5**.

Super-enhancers are preferentially localized within TADs demarcated by at least one strong boundary

We then explored what type of functional elements are localized within TADs demarcated by strong TAD boundaries. Specifically, we tested super-enhancers identified in matched samples (see Methods for details). Super-enhancers are key regulatory elements thought to be defining cell identity [9, 32], and are usually found near the center of TADs [33]. Our

326 analysis determined that they are significantly more frequently localized within TADs insulated
327 by at least one strong TAD boundary (**Figure 3e**). Further analysis revealed that, in many
328 cases, super-enhancers are insulated by strong boundaries both in the upstream and
329 downstream directions (~3 times more likely compared to a strong/weak boundary
330 combination). We then mined the tissue-based map of the human proteome [34], a collection
331 of ubiquitously expressed genes as well as genes of variable tissue-specificity. Remarkably,
332 our analysis revealed that the genes closest to strong boundaries are significantly enriched in
333 the class of ubiquitously expressed genes (**Supplementary Figure 6a**). However, and
334 consistent with previous findings, tissue-specific genes are more enriched further away from
335 the TAD boundaries, in the vicinity of super-enhancers. Taken together, our findings suggest
336 that, because of their significance in gene regulation, super-enhancers should only target
337 genes confined in the "correct" TAD or neighborhood, while remaining strongly insulated from
338 genes in adjacent TADs. This is conceivably achieved by the strong TAD boundaries we have
339 identified in this study. At the same time, ubiquitously expressed genes are insulated from
340 enhancer elements in adjacent TADs by the same strong TAD boundaries in order to maintain
341 proper expression levels, unaffected by regulation from enhancer elements in adjacent TADs.
342
343 **Strong TAD boundaries are co-duplicated with super-enhancers and oncogenes in**
344 **cancer**
345 To further investigate the importance of variable boundary strength, we asked whether TAD
346 boundaries are prone to genetic alterations in cancer. To this end, we mined structural variants
347 released by the International Cancer Genome Consortium (ICGC) [35]. A summary of the
348 reported variant types across all cancer types available on ICGC, is presented in
349 **Supplementary Figure 6b**. First, for each focal (up to 1Mb) deletion event, we identified the
350 TAD boundaries closest to the breakpoints, and calculated the frequency of deletions by
351 boundary strength. We observed that the frequency of deletions monotonically decreased with
352 increasing boundary strength (**Figure 4a**). This suggests that strong TAD boundaries are less
353 frequently lost in cancer, as they may "safeguard" functional elements that are necessary for

proliferation. By contrast, the frequency of tandem duplications (up to 1Mb) increased with increasing boundary strength (**Figure 4b**). Both results were robust to various cutoffs on the sizes of the structural variants, within the usual range of TAD sizes (from 250kb to 2.5Mb). Then, to further clarify the connection between super-enhancers, strong TAD boundaries and cancer, we studied tandem duplication events where super-enhancers (obtained from the largest available collection of super-enhancers [36]) are co-duplicated with adjacent strong boundaries. As demonstrated in **Figure 4c**, super-enhancers are indeed co-duplicated with strong TAD boundaries. This suggests that, in cancer, not only are strong boundaries protected from deletions, but they are also co-duplicated with super-enhancer elements. Intriguingly, this observation raises the possibility that pairs of super-enhancers and oncogenes represent functional entities encapsulated by strong boundaries, that are frequently duplicated or perhaps amplified in malignancies. Such an example of an oncogene is shown in **Figure 4d**: *MYC*, a well-known oncogene that is typically overexpressed in cancer, is localized next to a strong TAD boundary and is co-duplicated with the boundary as well as with several proximal super-enhancers. Taken together, these observations highlight the importance of strong TAD boundaries in the context of cancer.

DISCUSSION

Multiple recent studies have revealed that the metazoan genome is compartmentalized in boundary-demarcated functional units known as topologically associating domains (TADs). TADs are highly conserved across species and cell types. A few studies, however, provide compelling evidence that specific TADs, despite the fact that they are largely invariant, exhibit some plasticity. Given that TAD boundary disruption has been recently linked to aberrant gene activation and multiple disorders including developmental defects and cancer, categorization of boundaries based on their strength and identification of their unique features becomes of particular importance. In this study, we developed a method based on fused two-dimensional lasso in order to categorize TAD boundaries based on their strength. We demonstrated that our method: (a) improves the correlation of Hi-C contact matrices irrespective of the Hi-C bias

correction method used, (b) reveals multiple levels of chromatin organization and (c) successfully identifies boundaries of variable strength and that strong predicted boundaries exhibit certain expected features, such as elevated CTCF levels and increased insulating capacity. By performing an integrative analysis of estimated boundary strength with super-enhancers in matched samples, we observed that super-enhancers are preferentially insulated by strong boundaries, suggesting that super-enhancers and strong boundaries may represent a biologically relevant entity. Motivated by this observation, we examined the frequency of structural alterations involving strong boundaries and super-enhancers. We found that not only strong boundaries are “protected” from deletions, but, more importantly, they are co-duplicated together with super-enhancers. Recently, it has been shown that genetic or epigenetic alterations near enhancers may lead to aberrant activation of oncogenes [37–40]. Our results, expand on these studies and highlight a synergistic role of super-enhancers and TAD boundaries in cancer.

AUTHOR CONTRIBUTIONS

YG and CL performed computational analyses and generated figures. AT, AL and PK conceived this study. PN performed the CUTLL-1 Hi-C experiments. PN and IA offered biological insights and helped with the interpretation of Hi-C data. AT designed and implemented the method. CL and AT wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We would like to thank all members of the Tsirigos and Aifantis Laboratories for critical evaluation of the manuscript. We would like to thank the Applied Bioinformatics Laboratories (ABL) at the NYU School of Medicine for providing bioinformatics support and helping with the analysis and interpretation of the data. This work has used computing resources at the NYU

High Performance Computing Facility (HPCF). We also thank the Genome Technology Center (GTC) for expert library preparation and sequencing. This shared resource is partially supported by the Cancer Center Support Grant, P30CA016087, at the Laura and Isaac Perlmutter Cancer Center.

FUNDING

The study was supported by the American Cancer Society [RSG-15-189-01-RMC to AT] and a Leukemia & Lymphoma Society New Idea Award [8007-17 to AT]. NYU Genome Technology Center (GTC) is a shared resource, partially supported by the Cancer Center Support Grant [P30CA016087] at the Laura and Isaac Perlmutter Cancer Center.

TABLE AND FIGURES LEGENDS

Figure 1. (a) Overall workflow and benchmark strategy, **(b)** Comparison of Hi-C contact matrices between biological replicates generated from Hi-C library using the same or different restriction enzyme; Hi-C matrices were estimated using three methods (naïve filtering, iterative correction and simple scaling); assessment was performed using Pearson correlation on the actual or distance-normalized Hi-C matrices at resolutions ranging from 100kb to 20kb and maximum distances of 2Mb, 6Mb and 10Mb between interacting pairs

Figure 2. Fused two-dimensional lasso improves reproducibility of Hi-C contact matrices. (a) Example of application of fused two-dimensional lasso on a Hi-C contact matrix focused on a 8Mb locus on chromosome 8 for different values of parameter λ (top panel=original matrix; bottom panel=distance-normalized matrix), **(b)** Hi-C contact matrix correlations are improved by increasing the value of fused lasso parameter λ both for matrices estimated by ICE as well as by our simple scaling method. **(c)** Hi-C contact matrix correlations of distance-normalized matrices. Correlations of Hi-C contact matrices generated by the naïve filtering method are marked by the red line in each panel.

436

437 **Figure 3. Classification and characterization of TAD boundaries according to insulation**
438 **score. (a)** Number of TADs for λ values ranging from 0 to 5, **(b)** TAD boundaries lost at higher
439 values of parameter λ are associated with higher TAD insulation scores, **(c)** Heatmap
440 representation of TAD boundary insulation strength across samples; hierarchical clustering
441 correctly groups replicates and related cell types independent of enzyme biases or batch
442 effects related to the lab that generated the Hi-C libraries, **(d)** TAD boundary strength is
443 associated with CTCF levels, **(e)** Fraction of super-enhancer elements in the vicinity of
444 boundaries of variable strength. The gradient of blue corresponds to λ values with darker blue
445 denoting higher λ value.

446

447 **Figure 4. Pan-cancer analysis of strong vs weak TAD boundaries. (a)** Schematic of pan-
448 cancer analysis (left panel) and classification of focally deleted boundaries in cancer according
449 to their strength (right panel), **(b)** Schematic of pan-cancer analysis (left panel) and
450 classification of focally duplicated boundaries in cancer according to their strength (right panel),
451 **(c)** Schematic of pan-cancer analysis (left panel) and co-duplications of TAD boundaries with
452 super-enhancers in cancer (right panel), **(d)** Snapshot of the *MYC* locus: a strong boundary
453 (black bar) is frequently co-duplicated with *MYC* and potential super-enhancers in cancer
454 patients (highlighted area). IGV tracks from top to bottom: boundary score (gray), strong
455 boundaries (black bars), super-enhancer track from SEA (blue bars), RefSeq genes,
456 duplication frequency (red graph) and ICGC patient tandem duplications (red bars).

457

458 **Supplementary Figure 1. Quality assessment of Hi-C datasets. (a)** Counts of Hi-C read
459 pairs in various read categories: dark and light green indicate read pairs that were not
460 designated as artifacts and can be used in downstream analyses, **(b)** Percentages of Hi-C
461 reads in each category, **(c)** Average scaled Hi-C read pair count as a function of distance
462 between interacting loci.

463

Supplementary Figure 2. Fused one-dimensional lasso improves reproducibility of distance-normalized Hi-C contact matrices. (a) distance-normalized Hi-C contact matrix correlations are improved by increasing the value of fused lasso parameter λ both for matrices estimated by ICE as well as by our simple scaling method; correlations of distance-normalized Hi-C contact matrices generated by the naïve filtering method are marked by the red line in each panel. The gradient of blue corresponds to λ values with darker blue denoting higher λ value, **(b)** degrees of freedom as a function of λ.

Supplementary Figure 3. Comparison of Hi-C contact matrices between biological replicates generated from Hi-C library using the same restriction enzyme. Three methods (naïve filtering, iterative correction and simple scaling) were used for estimation. Assessment was performed using Pearson correlation on the actual or distance-normalized Hi-C matrices at resolutions ranging from 100kb to 20kb and maximum distances of 2Mb, 6Mb and 10Mb between interacting pairs. Only samples with approximately 80 million usable intra-chromosomal reads were considered.

Supplementary Figure 4. Comparison of Hi-C contact matrices between biological replicates generated from Hi-C library using the same restriction enzyme. Three methods (naïve filtering, iterative correction and simple scaling) were used for estimation. Assessment was performed using Pearson correlation on the actual or distance-normalized Hi-C matrices at resolutions ranging from 100kb to 20kb and maximum distances of 2Mb, 6Mb and 10Mb between interacting pairs. Only samples with approximately 120 million usable intra-chromosomal reads were considered.

Supplementary Figure 5. Normalized numbers of repeat elements in proximity to boundaries of certain boundary strength. Darker blue in the blue colour gradient denotes higher boundary strength.

Supplementary Figure 6. (a) Fraction of ubiquitous genes as well as genes of increasing tissue-specificity in the vicinity of boundaries of variable strength, **(b)** Distribution of somatic structural alterations in the ICGC database.

REFERENCES

1. Dekker J, Marti-Renom MA, Mirny LA: **Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data.** *Nat Rev Genet* 2013, **14**:390–403.

2. Schmitt AD, Hu M, Ren B: **Genome-wide mapping and analysis of chromosome architecture.** *Nat Rev Mol Cell Biol* 2016, **17**:743–755.

3. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326**:289–293.

4. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DCA, Aitken S, Xie SQ, Morris KJ, Itoh M, Kawaji H, Jaeger I, Hayashizaki Y, Carninci P, Forrest ARR, FANTOM Consortium, Semple CA, Dostie J, Pombo A, Nicodemi M: **Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation.** *Mol Syst Biol* 2015, **11**:852.

5. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485**:376–380.

6. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen N, Dekker J, Heard E: **Spatial partitioning of the regulatory landscape of the X-inactivation centre.** *Nature* 2012, **485**:381–385.

7. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the Drosophila genome.** *Cell* 2012, **148**:458–472.

8. Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong C-T, Hookway TA, Guo C, Sun Y, Bland MJ, Wagstaff W, Dalton S, McDevitt TC, Sen R, Dekker J, Taylor J, Corces VG: **Architectural protein subclasses shape 3D organization of genomes during lineage commitment.** *Cell* 2013, **153**:1281–1295.

9. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, Young RA: **Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes.** *Cell* 2014, **159**:374–387.

10. Filippova D, Patro R, Duggal G, Kingsford C: **Identification of alternative topological domains in chromatin.** *Algorithms Mol Biol* 2014, **9**:14.

11. Weinreb C, Raphael BJ: **Identification of hierarchical chromatin domains.** *Bioinformatics* 2016, **32**:1601–1609.

12. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ: **Condensin-driven remodelling of X chromosome topology during dosage compensation.** *Nature* 2015, **523**:240–244.

13. Haddad N, Vaillant C, Jost D: **IC-Finder: inferring robustly the hierarchical organization of chromatin folding.** *Nucleic Acids Res* 2017.

14. Cubeñas-Potts C, Corces VG: **Topologically Associating Domains: An invariant framework or a dynamic scaffold?** *Nucleus* 2015, **6**:430–434.

15. Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong C-T, Cubeñas-Potts C, Hu M, Lei EP, Bosco G, Qin ZS, Corces VG: **Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing.** *Mol Cell* 2015, **58**:216–231.

16. Narendra V, Bulajić M, Dekker J, Mazzoni EO, Reinberg D: **CTCF-mediated topological boundaries during development foster appropriate gene regulation.** *Genes Dev* 2016, **30**:2657–2662.

17. Friedman J, Hastie T, Höfling H, Tibshirani R: **Pathwise coordinate optimization**. *Ann Appl Stat* 2007, **1**:302–332.

18. Lazaris C, Kelly S, Ntziachristos P, Aifantis I, Tsirigos A: **HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking.** *BMC Genomics* 2017, **18**:22.

19. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao Y, Liang J, Zhao H, Lobanenkov VV, Ecker JR, Thomson JA, Ren B: **Chromatin architecture reorganization during stem cell differentiation.** *Nature* 2015, **518**:331–336.

20. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell* 2014, **159**:1665–1680.

21. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B: **A high-resolution map of the three-dimensional chromatin interactome in human cells.** *Nature* 2013, **503**:290–294.

22. Yaffe E, Tanay A: **Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.** *Nat Genet* 2011, **43**:1059–1065.

564 23. Tibshirani RJ, Taylor J: **The solution path of the generalized lasso**. *The Annals of*
565 *Statistics* 2011, **39**:1335–1371.

566 24. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker
567 J, Mirny LA: **Iterative correction of Hi-C data reveals hallmarks of chromosome**
568 **organization.** *Nat Methods* 2012, **9**:999–1003.

569 25. Rocha PP, Raviram R, Bonneau R, Skok JA: **Breaking TADs: insights into**
570 **hierarchical genome organization.** *Epigenomics* 2015, **7**:523–526.

571 26. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods*
572 2012, **9**:357–359.

573 27. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H,
574 Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-**
575 **regulatory elements required for macrophage and B cell identities.** *Mol Cell* 2010,
576 **38**:576–589.

577 28. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS: **HiCNorm: removing biases in Hi-C**
578 **data via Poisson regression.** *Bioinformatics* 2012, **28**:3131–3133.

579 29. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J: **Normalization of a**
580 **chromosomal contact map.** *BMC Genomics* 2012, **13**:436.

581 30. Knight PA, Ruiz D: **A fast algorithm for matrix balancing**. *IMA Journal of Numerical*
582 *Analysis* 2013, **33**:1029–1047.

583 31. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S: **Comparison of**
584 **computational methods for Hi-C data analysis.** *Nat Methods* 2017.

585 32. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI,
586 Young RA: **Master transcription factors and mediator establish super-enhancers at key**
587 **cell identity genes.** *Cell* 2013, **153**:307–319.

588 33. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, Ren B: **A**
589 **compendium of chromatin contact maps reveals spatially active regions in the human**
590 **genome.** *Cell Rep* 2016, **17**:2042–2059.

591 34. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson
592 Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto
593 CA-K, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist P-H, Berling H,
594 Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K,
595 Forsberg M, et al.: **Proteomics. Tissue-based map of the human proteome.** *Science*
596 2015, **347**:1260419.

597 35. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J,
598 Whitty B, Wong-Erasmus M, Yao L, Kasprzyk A: **International Cancer Genome**
599 **Consortium Data Portal--a one-stop shop for cancer genomics data.** *Database (Oxford)*
600 2011, **2011**:bar026.

601 36. Wei Y, Zhang S, Shang S, Zhang B, Li S, Wang X, Wang F, Su J, Wu Q, Liu H, Zhang Y:
602 **SEA: a super-enhancer archive.** *Nucleic Acids Res* 2016, **44**:D172–9.

603 37. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H,
604 Opitz JM, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas
605 SA, Osterwalder M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A, Mundlos S:
606 **Disruptions of topological chromatin domains cause pathogenic rewiring of gene-**
607 **enhancer interactions.** *Cell* 2015, **161**:1012–1025.

608 38. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov
609 AO, Suvà ML, Bernstein BE: **Insulator dysfunction and oncogene activation in IDH**
610 **mutant gliomas.** *Nature* 2016, **529**:110–114.

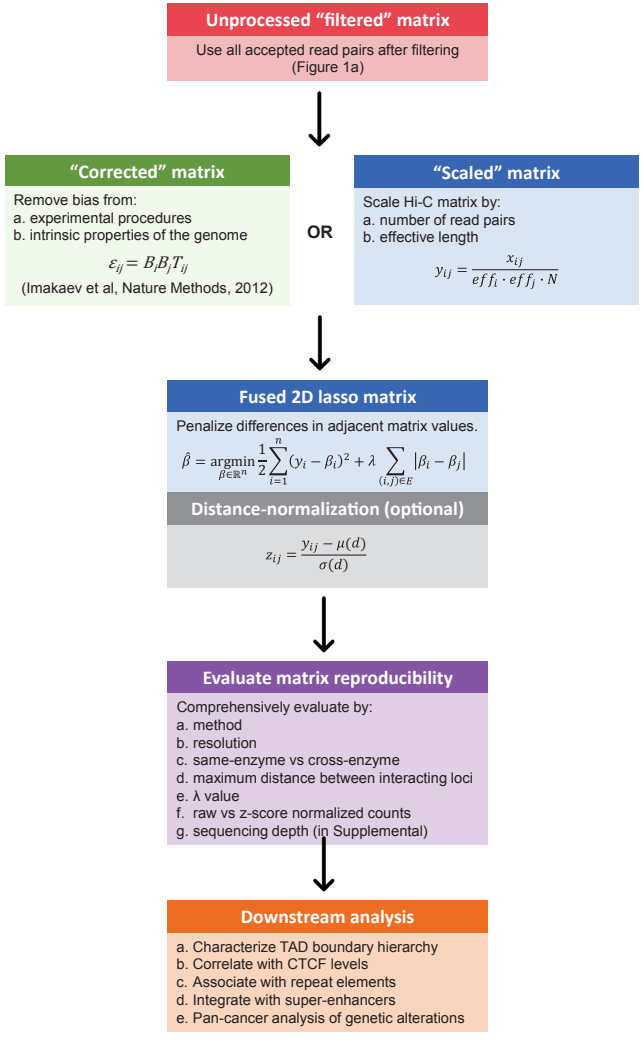
611 39. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR,
612 Fan ZP, Sigova AA, Reddy J, Borges-Rivera D, Lee TI, Jaenisch R, Porteus MH, Dekker J,
613 Young RA: **Activation of proto-oncogenes by disruption of chromosome**
614 **neighborhoods.** *Science* 2016, **351**:1454–1458.

615 40. Weischenfeldt J, Dubash T, Drainas AP, Mardin BR, Chen Y, Stütz AM, Waszak SM,
616 Bosco G, Halvorsen AR, Raeder B, Efthymiopoulos T, Erkek S, Siegl C, Brenner H,
617 Brustugun OT, Dieter SM, Northcott PA, Petersen I, Pfister SM, Schneider M, Solberg SK,
618 Thunissen E, Weichert W, Zichner T, Thomas R, Peifer M, Helland A, Ball CR, Jechlinger M,
619 Sotillo R, et al.: **Pan-cancer analysis of somatic copy-number alterations implicates**
620 **IRS4 and IGF2 in enhancer hijacking.** *Nat Genet* 2017, **49**:65–74.

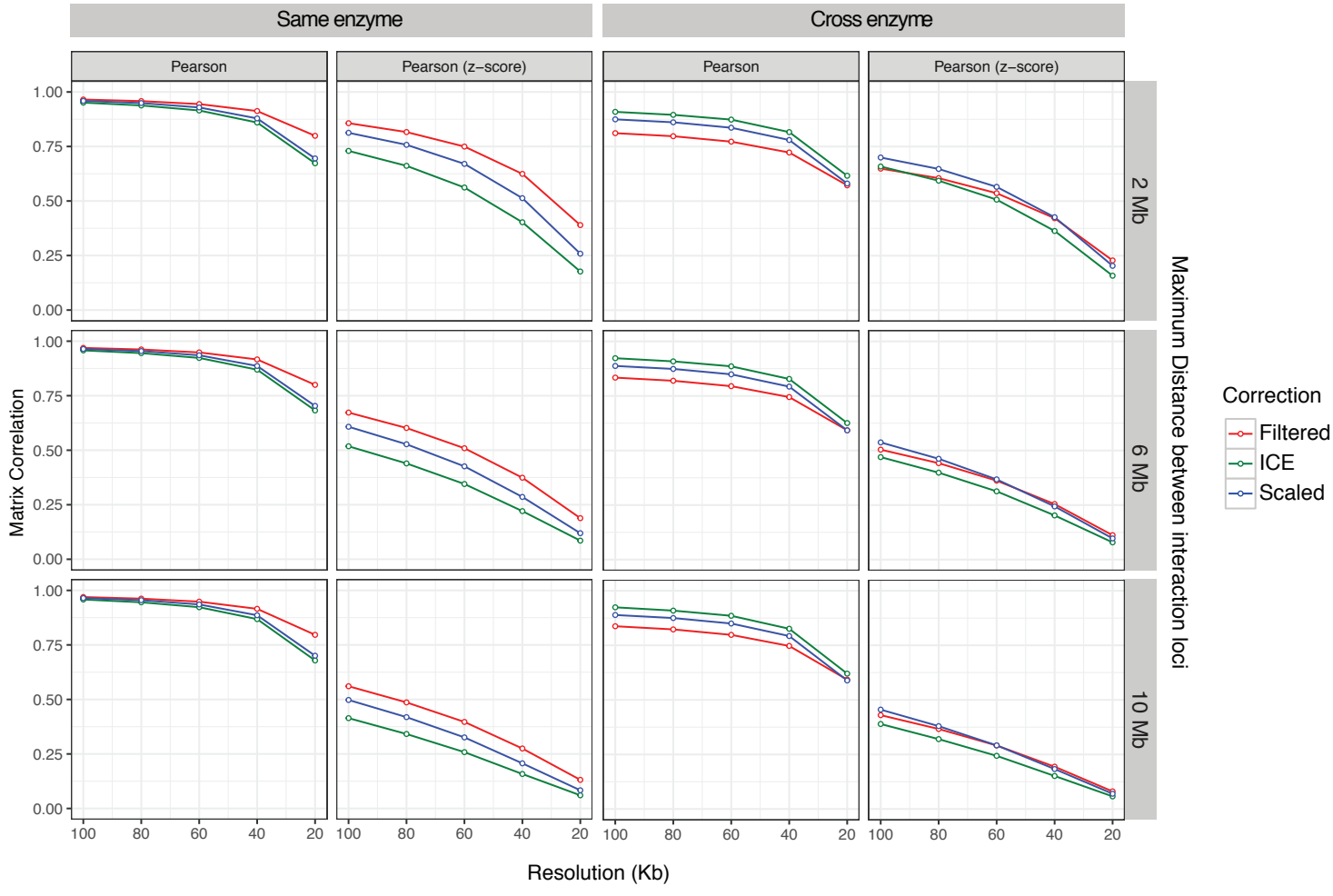
621

Figure 1

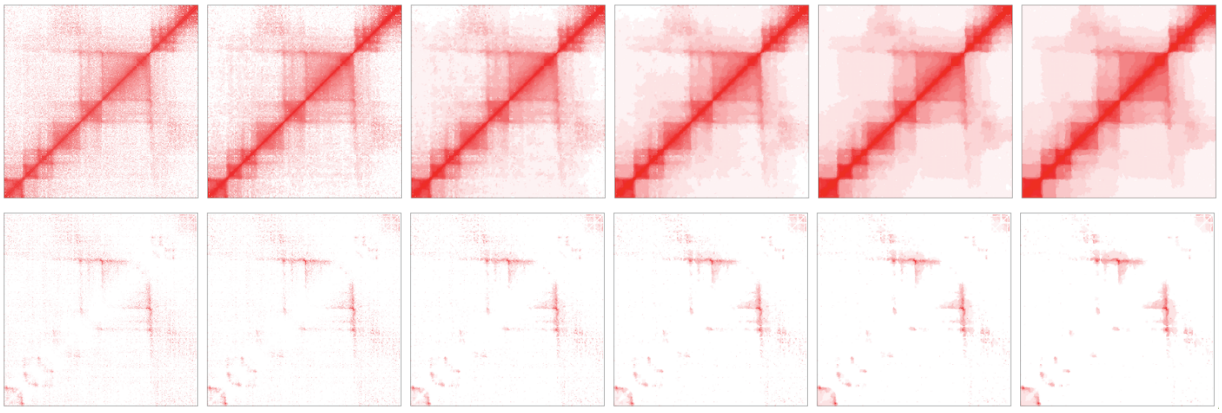
a


Unprocessed "filtered" matrix
Use all accepted read pairs after filtering (Figure 1a)
"Corrected" matrix
Remove bias from:
a. experimental procedures
b. intrinsic properties of the genome
$\varepsilon_{ij} = B_i B_j T_{ij}$
(Imakaev et al, Nature Methods, 2012)
OR
"Scaled" matrix
Scale Hi-C matrix by:
a. number of read pairs
b. effective length
$y_{ij} = \frac{x_{ij}}{eff_i \cdot eff_j \cdot N}$
Fused 2D lasso matrix
Penalize differences in adjacent matrix values.
$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_i)^2 + \lambda \sum_{(i,j)\in E} |\beta_i - \beta_j|$
Distance-normalization (optional)
$z_{ij} = \frac{y_{ij} - \mu(d)}{\sigma(d)}$
Evaluate matrix reproducibility
Comprehensively evaluate by:
a. method
b. resolution
c. same-enzyme vs cross-enzyme
d. maximum distance between interacting loci
e. λ value
f. raw vs z-score normalized counts
g. sequencing depth (in Supplemental)
Downstream analysis
a. Characterize TAD boundary hierarchy
b. Correlate with CTCF levels
c. Associate with repeat elements
d. Integrate with super-enhancers
e. Pan-cancer analysis of genetic alterations

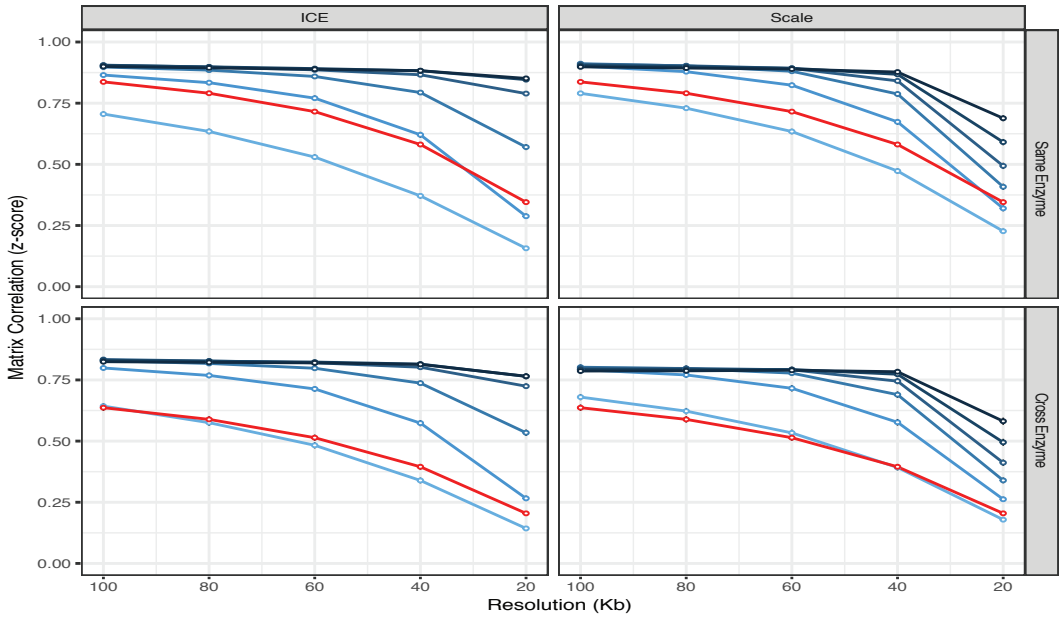
b


Same enzyme
Cross enzyme
Pearson
Pearson (z−score)
Pearson
Pearson (z−score)
2 Mb
6 Mb
10 Mb
Maximum Distance between interaction loci
Matrix Correlation
1.00
0.75
0.50
0.25
0.00
100
80
60
40
20
Resolution (Kb)
Correction
Filtered
ICE
Scaled

a



b



c

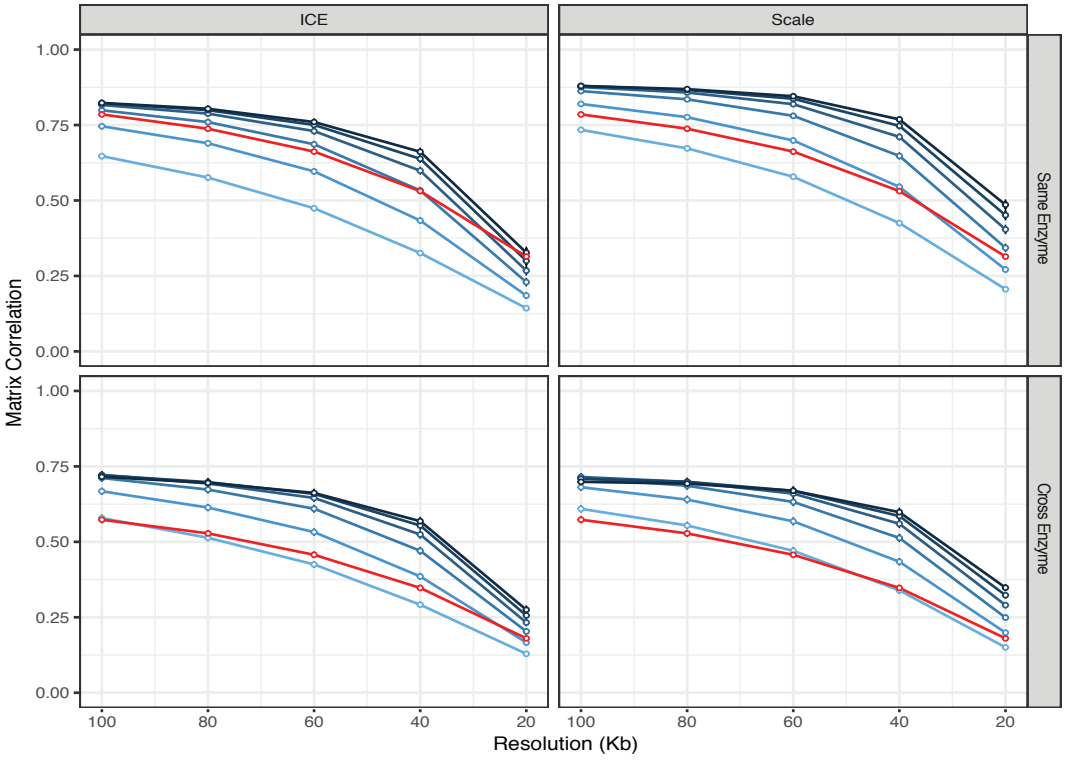
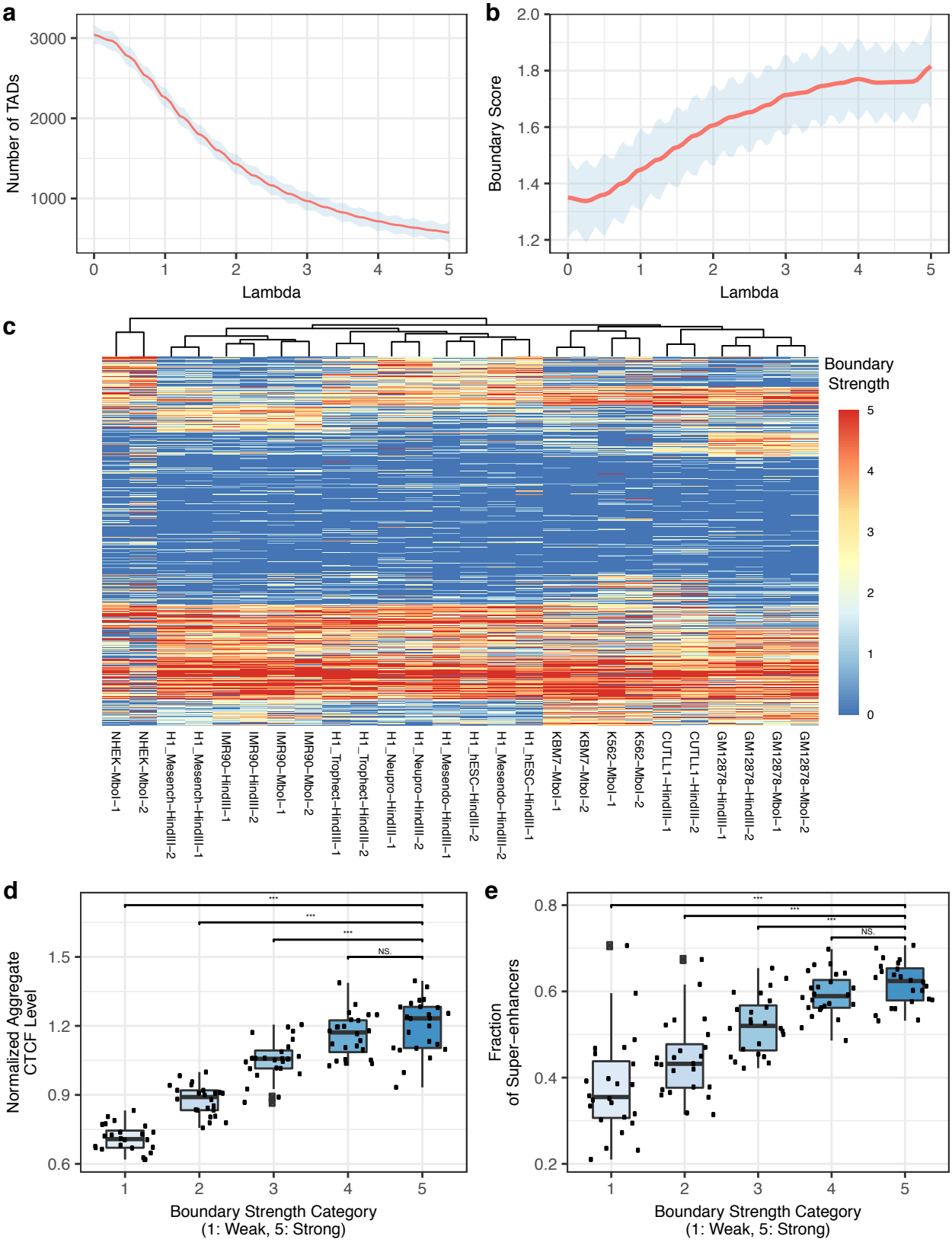
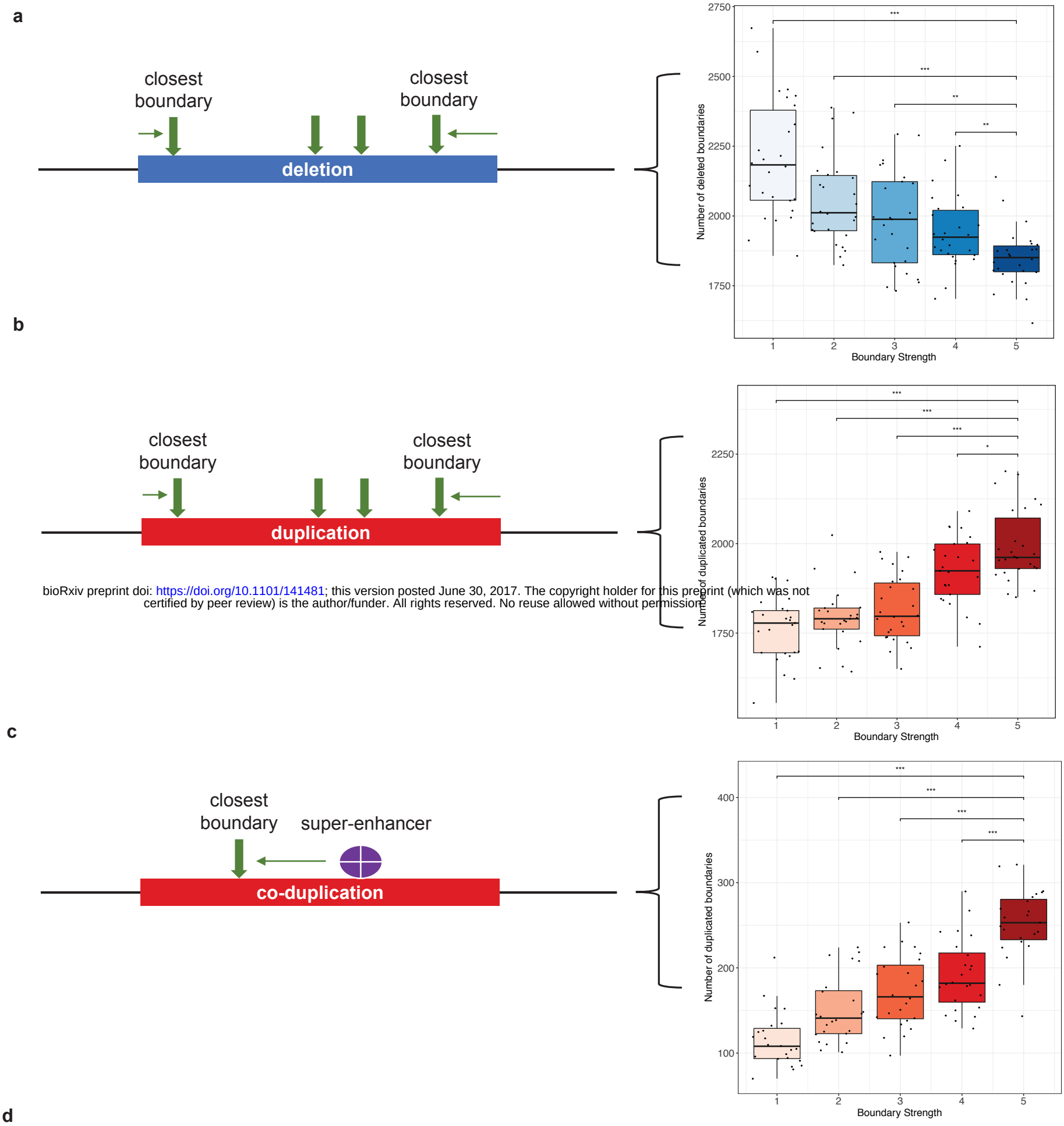
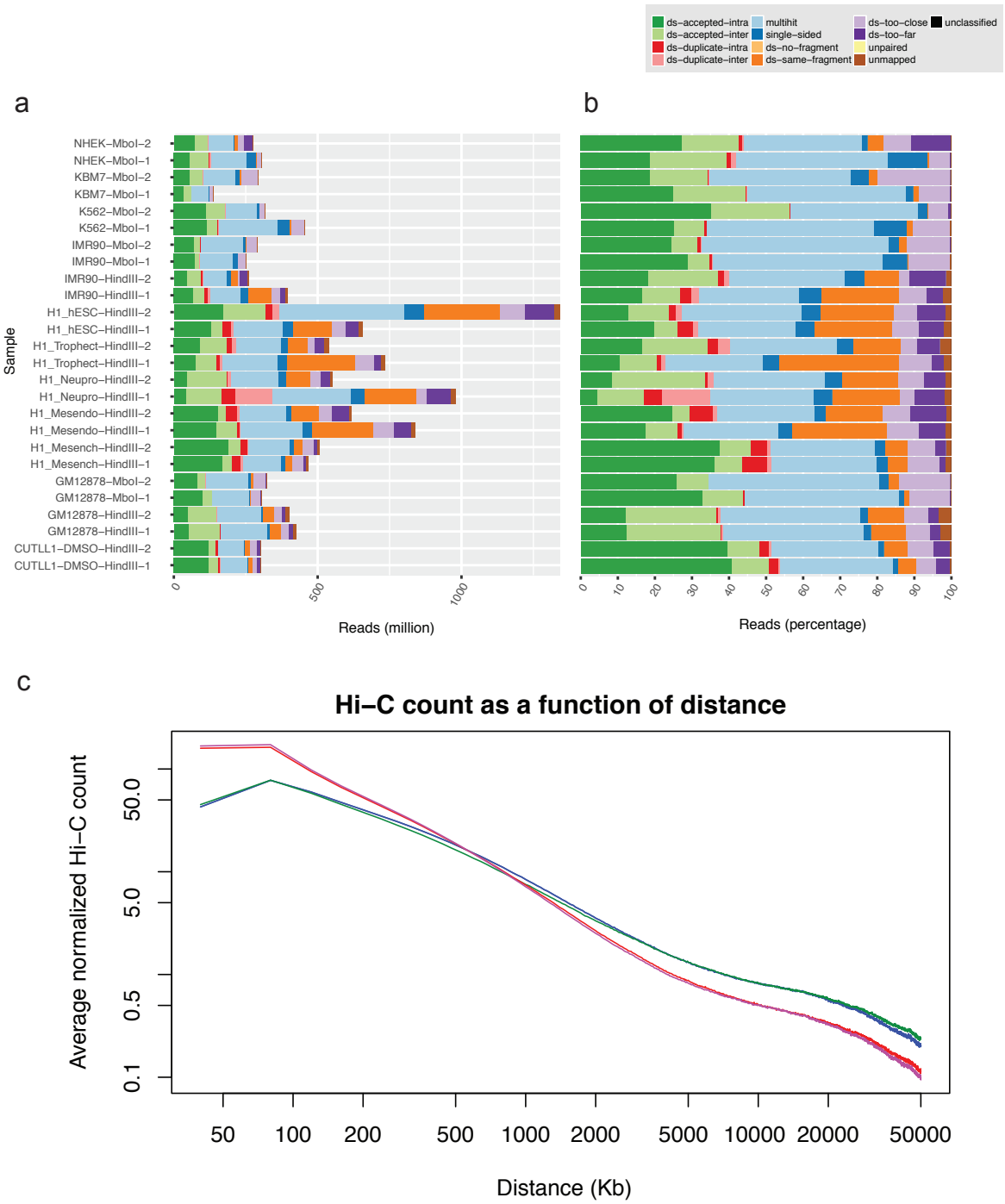


Figure 3


a
Number of TADs
3000
2000
1000
0
1
2
3
4
5
Lambda
b
Boundary Score
2.0
1.8
1.6
1.4
1.2
0
1
2
3
4
5
Lambda
c
Boundary Strength
5
4
3
2
1
0
NHEK−MboI−1
NHEK−MboI−2
H1_Mesench−HindIII−2
H1_Mesench−HindIII−1
IMR90−HindIII−1
IMR90−HindIII−2
IMR90−MboI−1
IMR90−MboI−2
H1_Trophect−HindIII−1
H1_Trophect−HindIII−2
H1_Neupro−HindIII−1
H1_Neupro−HindIII−2
H1_Mesendo−HindIII−1
H1_hESC−HindIII−2
H1_Mesendo−HindIII−2
H1_hESC−HindIII−1
KBM7−MboI−1
KBM7−MboI−2
K562−MboI−1
K562−MboI−2
CUTLL1−HindIII−1
CUTLL1−HindIII−2
GM12878−HindIII−1
GM12878−HindIII−2
GM12878−MboI−1
GM12878−MboI−2
d
Normalized Aggregate CTCF Level
1.5
1.2
0.9
0.6
***
***
***
NS.
1
2
3
4
5
Boundary Strength Category
(1: Weak, 5: Strong)
e
Fraction of Super−enhancers
0.8
0.6
0.4
0.2
***
***
***
NS.
1
2
3
4
5
Boundary Strength Category
(1: Weak, 5: Strong)

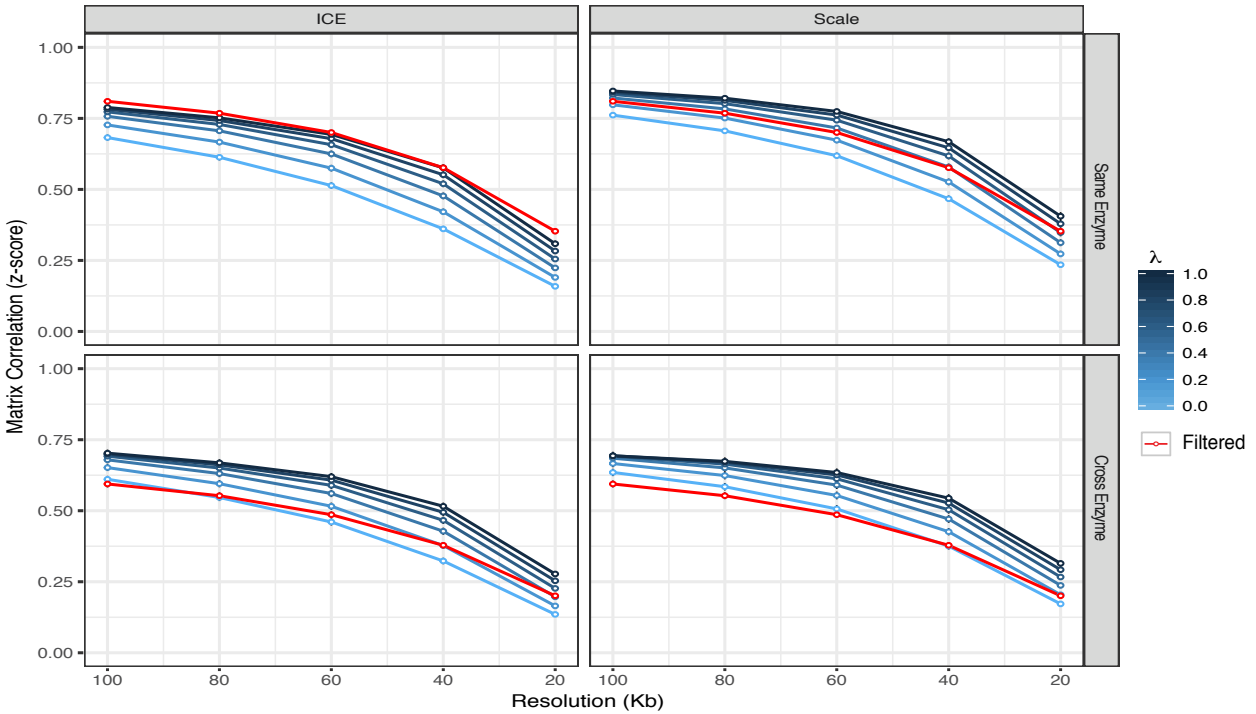


Supplementary Figure 1

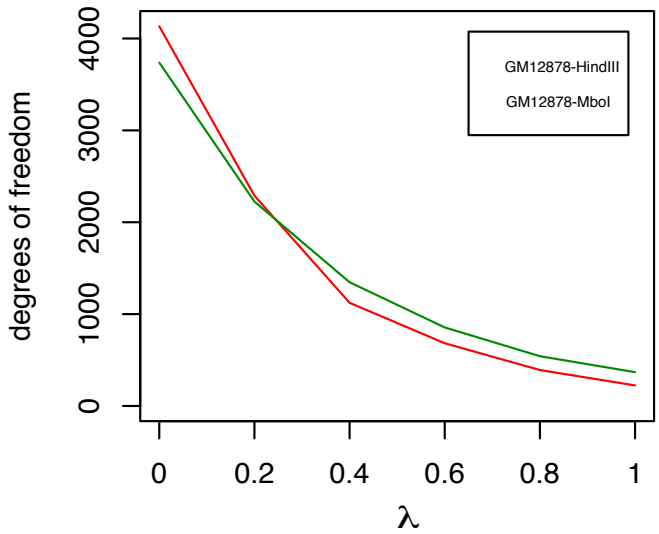


Supplementary Figure 2

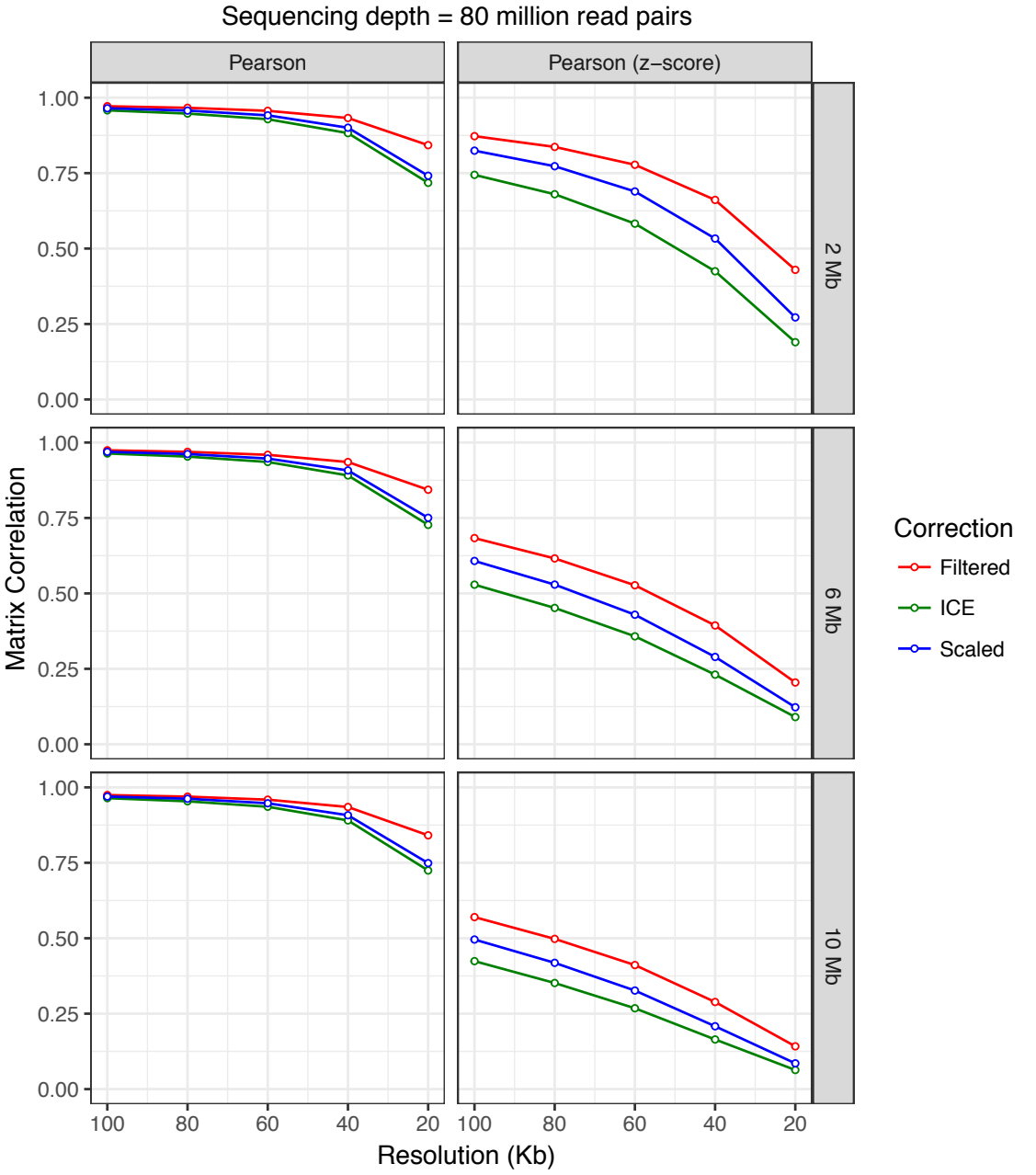
a



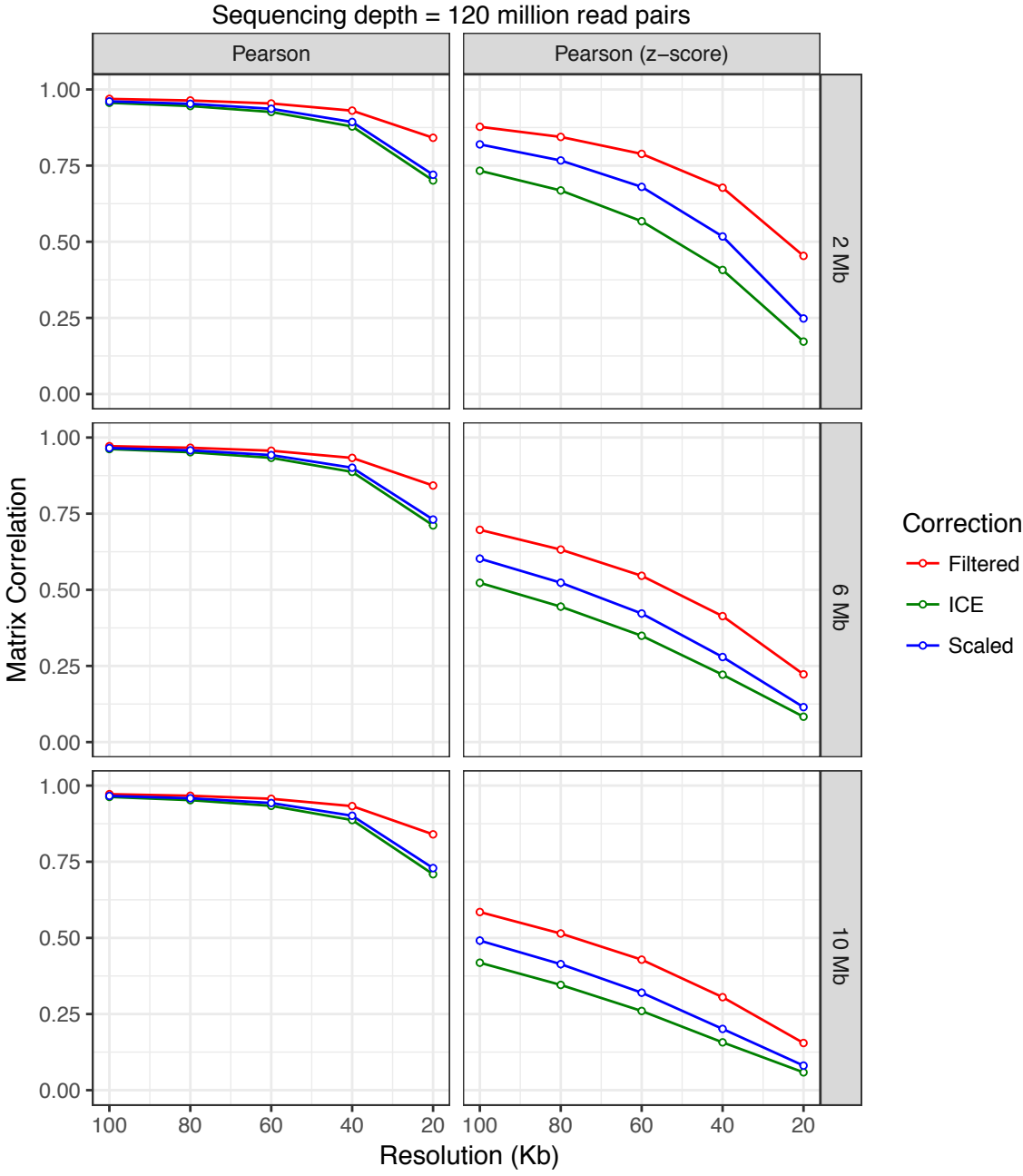
b



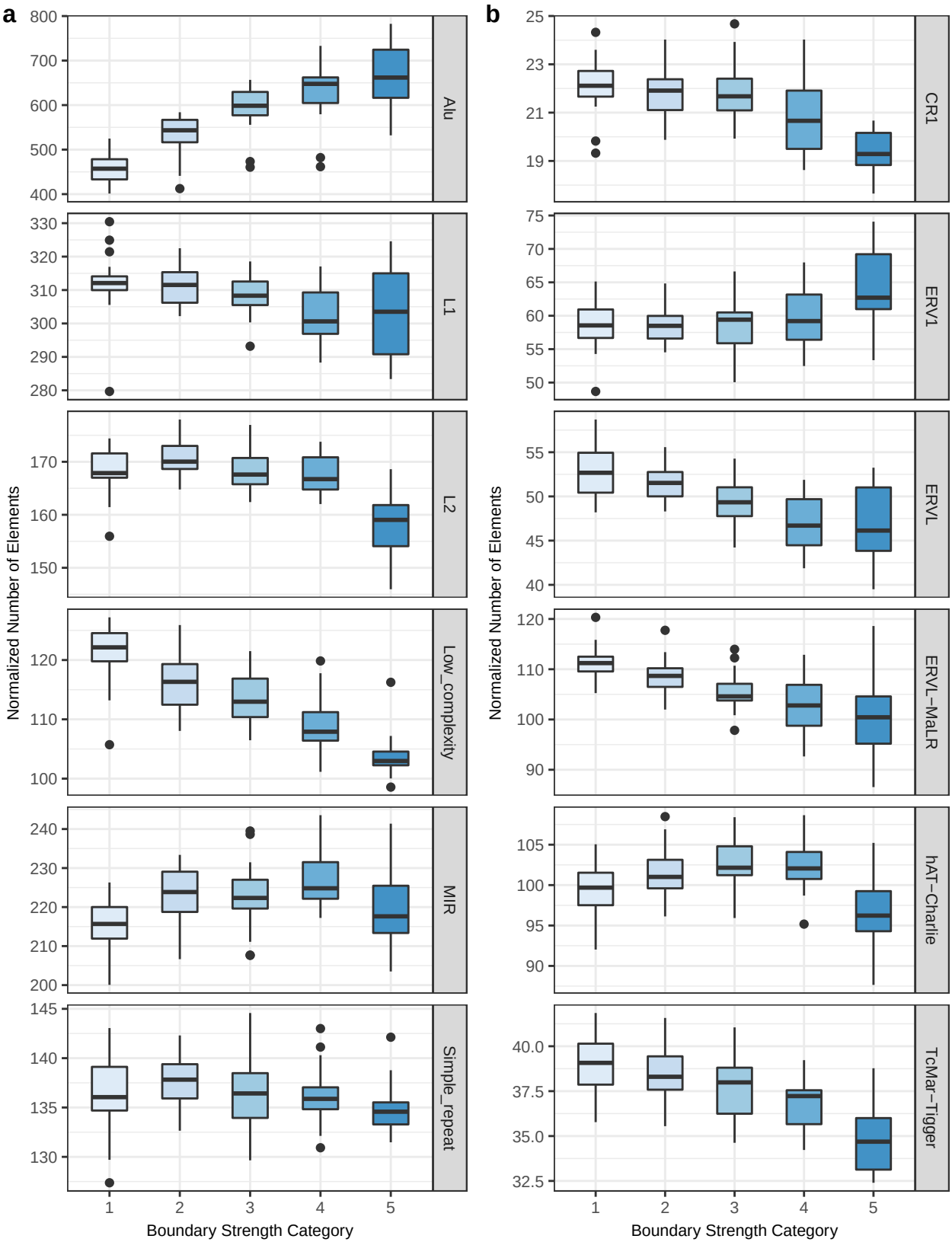
Supplementary Figure 3



Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6