Robust estimation of Hi-C contact matrices using fused lasso reveals preferential insulation of super-enhancers by strong TAD boundaries

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 have been shown to be largely conserved across cell types and even across species. Increasing evidence suggests that the seemingly invariant TADs may exhibit some plasticity in certain cases and their boundary strength can vary. However, a genome-wide characterization of TAD boundary strength in mammals is still lacking. 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 levels of CTCF and that TAD boundary insulation scores may differ across cell types. Intriguingly, we also found that super-enhancer elements are preferentially insulated by strong boundaries. Presumably, genetic or epigenetic inactivation of strong boundaries may lead to loss of insulation around super-enhancers, disrupt the physiological transcriptional program and cause disease.

INTRODUCTION

The advent of proximity-based ligation assays has allowed us to probe three-dimensional chromatin organization at unprecedented resolution [1, 2]. Hi-C, a high-throughput chromosome conformation variant has allowed genome-wide identification of chromatin-chromatin interactions [3]. Hi-C is prone to biases and multiple algorithms have been developed for Hi-C bias correction, including probabilistic modelling methods [4], Poisson or negative binomial normalization [5] and the widely popular Iterative Correction and Eigenvalue decomposition method (ICE) [6], which assumes “equal visibility” of genomic loci. A similar iterative method named Sequential Component Normalization was introduced by Cournac *et al.* [7]. Additional efficient correction methods have been developed to handle high-resolution Hi-C datasets [8]. Hi-C has revealed that the metazoan genome is organized in areas of active and inactive chromatin known as A and B compartment respectively [3]. These are further compartmentalized in super-TADs [9], topologically associating domains (TADs) [10–12] and sub-TADs [13], as well as gene neighbourhoods [14]. Some algorithms have been already developed to reveal this hierarchical chromatin organization, including Directionality Index (DI) [10], Armatus [15], TADtree [16], Insulation Index (Crane) [17], IC-Finder [18] and others. TADs are megabase-scale areas of highly interacting chromatin, demarcated by CTCF-enriched boundaries, and are highly-conserved across species and cell types [10, 19].

Genome compartmentalization in TADs confines enhancer-promoter interactions within the same domain [10, 12, 20] and during cell differentiation most changes have been shown to occur within TADs [21]. TAD boundaries have been found to be rich in tRNA genes, transposable elements, CCCTC-binding factor (CTCF), cohesin complex and other structural proteins [10–12]. More recently, proteins involved in chromatin remodelling such as BRG1 – an ATPase driving SWI/SNF activity – as well topoisomerase complexes have been implicated in boundary formation through regulation of chromatin compaction [22]. Whereas 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 [23]. Recent studies have shown that in *Drosophila*, exposure to heat-shock resulted in local changes in certain TAD boundaries resulting in TAD merging which is believed to have physiological consequences [24]. A recent study in mammals showed that during motor neuron (MN) differentiation in mammals, TAD and sub-TAD boundaries in *Hox* cluster are not rigid and their plasticity is linked to changes in the expression of genes of the *Hox* cluster during differentiation [25]. It has also been demonstrated that boundary strength is positively associated with the occupancy of certain structural proteins including CCCTC-binding factor (CTCF) [10]. Despite the fact that there is a handful of studies demonstrating that not all boundaries are equal and they 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 diseases like cancer. Here we introduce a new method based on fused two-dimensional lasso [26] in order to: (a) to improve the correlation of Hi-C contact matrices, (b) reveal the multiple levels of chromatin organization and (c) categorize TAD boundaries based on their corresponding strength.

MATERIALS AND METHODS

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 enzymes. We ensured that our datasets included samples 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 [21], K562, KBM7 and NHEK cells [27] and in-house generated CUTLL-1), or two enzymes (HindIII or MboI) (GM12878 [27], IMR90 [10, 28]), in order to examine the consistency of predicted Hi-C interactions across different enzymes.

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. The two types of correlation were: 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 [6] or scaled (see below) Hi-C contact matrices (Pearson) or the distance normalized ones (Pearson (*z*-score)).

Generation of scaled Hi-C contact matrices

In order to improve the cross-enzyme (and same-enzyme) correlation of Hi-C matrices we accounted for the total number of read pairs and the “effective length” [4]. More specifically, the scaled number of reads 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 normalization

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*, *μ(d)* to the mean (expected) number of interactions for distance *d*=|j-i| and *σ(d)* is the corresponding standard deviation of the mean. The higher the difference between the observed (y_{ij}) and expected number of interactions (*μ(d)*), the higher the corresponding *z*-score.

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 [26]. 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 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 applied one-dimensional lasso on the Hi-C contact matrices in order to estimate the matrices for high values of λ 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 (**Supplemental Figure 1**).

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 [29] 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 [30]**.** Total CTCF levels 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.

Association of boundary strength with super-enhancers and repeat elements

Super-enhancers were called using H3K27ac ChIP-seq data from GEO, ENCODE and in-house generated data. Reads were first aligned with Bowtie2 v2.3.1 [31] and then HOMER v4.6 [32] was used to call super-enhancers, all with standard parameters. For each super-enhancer in each sample, we identified the corresponding TAD and its TAD boundaries. We then counted (per sample) the percentage of super-enhancers that are surrounded by boundaries belonging in each boundary category, demonstrating that most super-enhancers are insulated by strong boundaries.

RESULTS

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

We identified publicly available human Hi-C datasets (described in Materials and Methods section) that fulfilled the following criteria: (i) two biological replicates and (ii) sufficient sequencing depth to robustly identify topologically-associating domains (TADs) as described in our TAD calling benchmark study [30]. All datasets were then comprehensively re-analysed using HiC-bench. 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 (**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) (**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 calling and thus none of them was excluded from downstream analysis. 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.

Assessment of same-enzyme and cross-enzyme reproducibility of Hi-C contact matrices

Although it has been demonstrated in the literature that Hi-C libraries are prone to enzyme biases (see Introduction), no systematic large-scale study has investigated in detail the reproducibility of Hi-C contact matrices. Here, we attempt to address this question using the most comprehensive Hi-C dataset that is currently available, as described in the previous section. More specifically, we will focus on multiple factors that may play an important role on reproducibility: first, we will separately consider biological replicates of Hi-C libraries generated with the same or different restriction enzymes; second, we will study the impact of Hi-C matrix resolution (i.e. bin size); third, we will assess 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 **Figure 1A**), (ii) iterative correction (ICE) which has already been demonstrated to improve cross-enzyme correlation, and (iii) our own simple 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. 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 1C**: 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 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 (**Supplemental Figure 2** and **Supplemental Figure 3** 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. 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 1C**).

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 improving cross-enzyme at the expense of lower same-enzyme correlation when correcting for enzyme-related biases, we applied fused two-dimensional lasso (see Methods for details), a well-studied image denoising method, to generate Hi-C contact matrices with increased consistency between replicates. Briefly, two-dimensional fused lasso utilized a parameter λ which penalizes differences between neighboring values in the Hi-C contact matrix. 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, as done in the previous section, 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 and compared to the naïve filtering approach. The results are summarized in **Figure 2B**. Clearly, 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 3**.

Fused lasso reveals a TAD hierarchy linked to TAD boundary strength

After demonstrating that parameter *λ* helps improve reproducibility of Hi-C contact matrices independent of the bias-correction method, we further hypothesized that increased values of *λ* may define distinct classes of TADs with different properties. For this reason, we now allowed *λ* to range from 0 to the maximum possible value (after a finite value of *λ*, 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 **Supplemental Figure 1**). 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 4A**), suggesting that by increasing *λ*, we are effectively identifying larger TADs encompassing smaller TADs detected at smaller *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 as defined by the ratio score described in HiC-bench. Indeed, as hypothesized, TAD boundaries lost at higher values of parameter *λ* are associated with higher TAD insulation scores (**Figure 4B**). We then stratified TAD boundaries into six classes according to their strength, independently in each Hi-C dataset used in this study and generated a heatmap representation including all TAD boundaries and their associated class across all samples (**Figure 4C,D**). 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, 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 (**Figure 4E**). SINE elements have also been shown to be enriched at TAD boundaries [10], and apart from confirming this finding, we extended it and demonstrated that Alu elements (the most abundant type of SINE elements) are enriched at stronger TAD boundaries, whereas, interestingly, L1 elements (a subset of LINE elements) are enriched at weaker TAD boundaries (**Figure 4F**). A comprehensive analysis of all major repetitive element subtypes can be found in **Supplemental Figure 4**. Finally, we investigated the proximity of super-enhancers to TAD boundaries of different strength. Intriguingly, we found that super-enhancers are preferentially insulated by strong TAD boundaries (**Figure 4G**). Super-enhancers are thought to be cell specific and drive expression of key genes. Thus, a potential explanation of our finding is that super-enhancers should only target genes confined in the same TAD, while strongly insulated from genes in adjacent TADs. Genetic or epigenetic inactivation of strong boundaries may lead to loss of insulation around super-enhancers, disrupt the physiological transcriptional program and cause disease.

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. We also demonstrated that the boundaries of similar strength are largely conserved across the samples included in this study, however, a subset of TAD boundaries displays varying levels of insulation strength across samples. 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. Based on this observation, we believe that strong boundaries prevent the aberrant activation of genes residing in adjacent TADs, by consisting a physical barrier between the gene promoters and the super-enhancer elements. We predict that despite the fact that weak boundaries would be more prone to disruption, in many cancers, strong boundaries are actually disrupted by either genetic lesions or epigenetically, leading to aberrant activation of oncogenes by enhancers as recently demonstrated [33–36]. In future work, we will further characterize boundaries of variable strength, reveal their features and help with the identification of targets for pharmacological intervention, in order to restore disrupted boundaries.

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: Assessment of the reproducibility of Hi-C contact matrices across biological replicates. **(A)** Counts of Hi-C read pairs in various read categories: dark and light green indicate read pairs that were not designated as artifacts and can be used in downstream analyses, **(B)** Percentages of Hi-C reads in each category, **(C)** 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, **(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; correlations of Hi-C contact matrices generated by the naïve filtering method are marked by the red line in each panel.

Figure 3: Fused two-dimensional lasso improves reproducibility of distance-normalized Hi-C contact matrices. **(A)** Example of application of fused two-dimensional lasso on a distance-normalized Hi-C contact matrix focused on an 8Mb locus on chromosome 8 for different values of parameter λ, **(B)** 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.

Figure 4: Classification and characterization of TAD boundaries according to insulation score. **(A)** Number of TADs for λ values ranging from 0 to 5, **(B)** TAD boundaries lost at higher values of parameter λ are associated with higher TAD insulation scores, **(C)** heatmap representation of TAD boundary insulation strength across samples; hierarchical clustering correctly groups replicates and related cell types independent of enzyme biases or batch effects related to the lab that generated the Hi-C libraries, **(D)** Classification of boundaries according to boundary strength across samples, **(E)** TAD boundary strength is associated with CTCF levels, **(F)** Alu elements are enriched at stronger TAD boundaries whereas L1 elements are enriched at weaker TAD boundaries, **(G)** Super-enhancers are preferentially insulated by stronger TAD boundaries. The gradient of blue corresponds to λ values with darker blue denoting higher λ value.

Supplementary Figure 1: 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 2: 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 3: Fused one-dimensional lasso improves reproducibility of distance-normalized Hi-C contact matrices. **(A)** Hi-C contact matrix and **(B)** 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.

Supplementary Figure 4: Numbers of repeat elements in proximity to boundaries of certain boundary strength. Darker blue in the blue colour gradient denotes higher boundary strength.

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. doi:10.1038/nrg3454.

2. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol Cell Biol. 2016;17:743–55. doi:10.1038/nrm.2016.104.

3. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93. doi:10.1126/science.1181369.

4. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43:1059–65. doi:10.1038/ng.947.

5. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012;28:3131–3. doi:10.1093/bioinformatics/bts570.

6. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9:999–1003. doi:10.1038/nmeth.2148.

7. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. BMC Genomics. 2012;13:436. doi:10.1186/1471-2164-13-436.

8. Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis. 2013;33:1029–47. doi:10.1093/imanum/drs019.

9. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol. 2015;11:852. doi:10.15252/msb.20156492.

10. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80. doi:10.1038/nature11082.

11. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485:381–5. doi:10.1038/nature11049.

12. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148:458–72. doi:10.1016/j.cell.2012.01.010.

13. Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell. 2013;153:1281–95. doi:10.1016/j.cell.2013.04.053.

14. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, et al. Control of cell

identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell. 2014;159:374–87. doi:10.1016/j.cell.2014.09.030.

15. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithms Mol Biol. 2014;9:14. doi:10.1186/1748-7188-9-14.

16. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics. 2016;32:1601–9. doi:10.1093/bioinformatics/btv485.

17. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015;523:240–4. doi:10.1038/nature14450.

18. Haddad N, Vaillant C, Jost D. IC-Finder: inferring robustly the hierarchical organization of chromatin folding. Nucleic Acids Res. 2017. doi:10.1093/nar/gkx036.

19. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015;10:1297–309. doi:10.1016/j.celrep.2015.02.004.

20. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015;25:582–97. doi:10.1101/gr.185272.114.

21. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015;518:331–6. doi:10.1038/nature14222.

22. Barutcu AR, Lian JB, Stein JL, Stein GS, Imbalzano AN. The connection between BRG1, CTCF and topoisomerases at TAD boundaries. Nucleus. 2017;8:150–5. doi:10.1080/19491034.2016.1276145.

23. Cubeñas-Potts C, Corces VG. Topologically Associating Domains: An invariant framework or a dynamic scaffold? Nucleus. 2015;6:430–4. doi:10.1080/19491034.2015.1096467.

24. Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong CT, et al. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. Mol Cell. 2015;58:216–31. doi:10.1016/j.molcel.2015.02.023.

25. Narendra V, Bulajić M, Dekker J, Mazzoni EO, Reinberg D. CTCF-mediated topological boundaries during development foster appropriate gene regulation. Genes Dev. 2016;30:2657–62. doi:10.1101/gad.288324.116.

26. Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. Ann Appl Stat. 2007;1:302–32. doi:10.1214/07-AOAS131.

27. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80. doi:10.1016/j.cell.2014.11.021.

28. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-

dimensional chromatin interactome in human cells. Nature. 2013;503:290–4. doi:10.1038/nature12644.

29. Rocha PP, Raviram R, Bonneau R, Skok JA. Breaking TADs: insights into hierarchical genome organization. Epigenomics. 2015;7:523–6. doi:10.2217/epi.15.25.

30. 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. doi:10.1186/s12864-016-3387-6.

31. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. doi:10.1038/nmeth.1923.

32. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89. doi:10.1016/j.molcel.2010.05.004.

33. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25. doi:10.1016/j.cell.2015.04.004.

34. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature. 2016;529:110–4. doi:10.1038/nature16490.

35. Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351:1454–8. doi:10.1126/science.aad9024.

36. Weischenfeldt J, Dubash T, Drainas AP, Mardin BR, Chen Y, Stütz AM, et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat Genet. 2017;49:65–74. doi:10.1038/ng.3722.

Figure 1

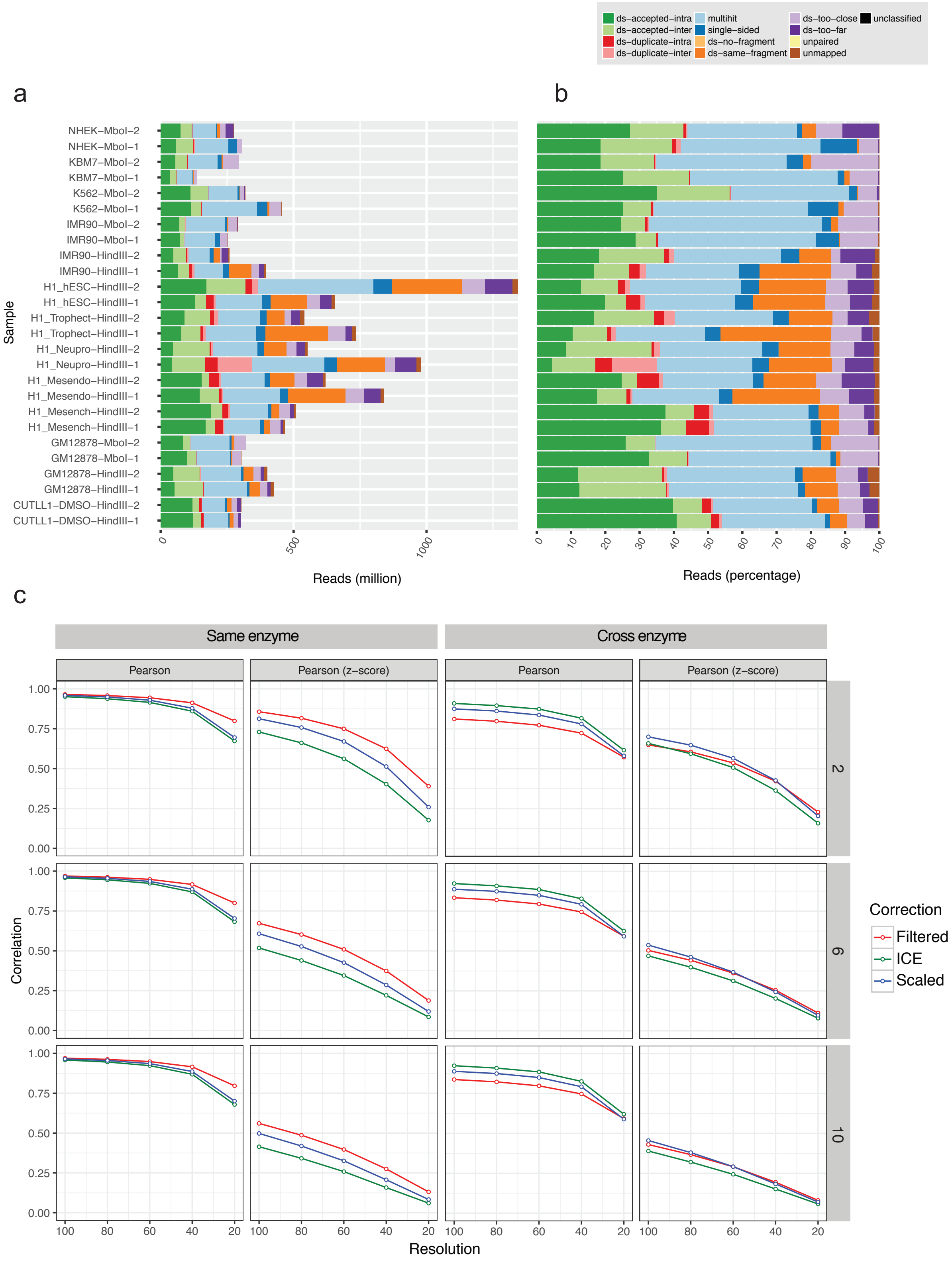

ds–accepted–intra
ds–accepted–inter
ds–duplicate–intra
ds–duplicate–inter
multihit
single–sided
ds–no–fragment
ds–same–fragment
ds–too–close
ds–too–far
unpaired
unmapped
unclassified
a
b
Sample
NHEK–MboI–2
NHEK–MboI–1
KBM7–MboI–2
KBM7–MboI–1
K562–MboI–2
K562–MboI–1
IMR90–MboI–2
IMR90–MboI–1
IMR90–HindIII–2
IMR90–HindIII–1
H1_hESC–HindIII–2
H1_hESC–HindIII–1
H1_Trophect–HindIII–2
H1_Trophect–HindIII–1
H1_Neupro–HindIII–2
H1_Neupro–HindIII–1
H1_Mesendo–HindIII–2
H1_Mesendo–HindIII–1
H1_Mesench–HindIII–2
H1_Mesench–HindIII–1
GM12878–MboI–2
GM12878–MboI–1
GM12878–HindIII–2
GM12878–HindIII–1
CUTLL1–DMSO–HindIII–2
CUTLL1–DMSO–HindIII–1
0
500
1000
Reads (million)
0
10
20
30
40
50
60
70
80
90
100
Reads (percentage)
c
Same enzyme
Cross enzyme
Pearson
Pearson (z–score)
Pearson
Pearson (z–score)
2
6
10
Correlation
1.00
0.75
0.50
0.25
0.00
100
80
60
40
20
Resolution
Correction
Filtered
ICE
Scaled

Figure 2

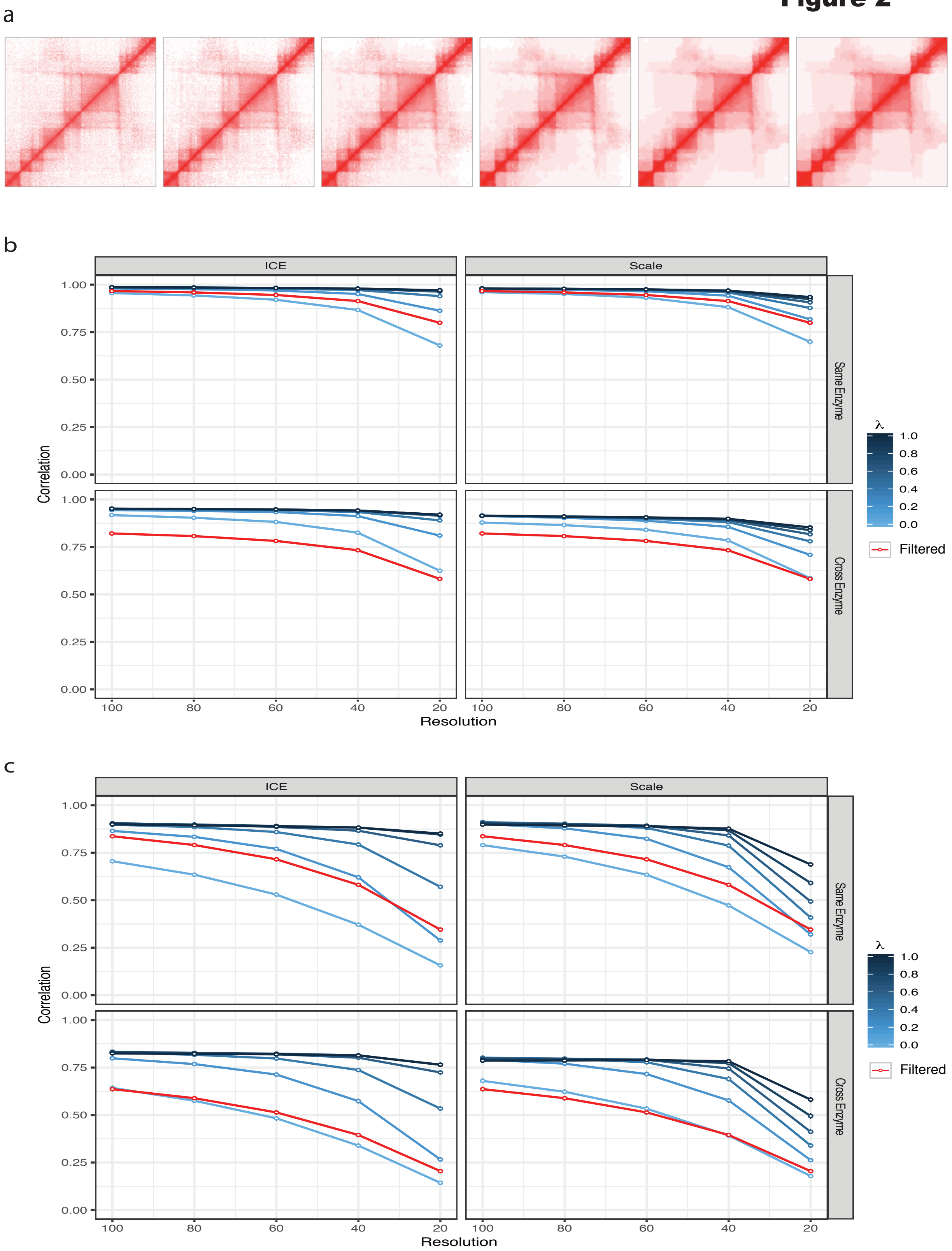

a
b
ICE
Scale
Same Enzyme
Cross Enzyme
Correlation
1.00
0.75
0.50
0.25
0.00
100
80
60
40
20
Resolution
λ
1.0
0.8
0.6
0.4
0.2
0.0
Filtered
c
ICE
Scale
Same Enzyme
Cross Enzyme
Correlation
Resolution
λ
Filtered

Figure 3

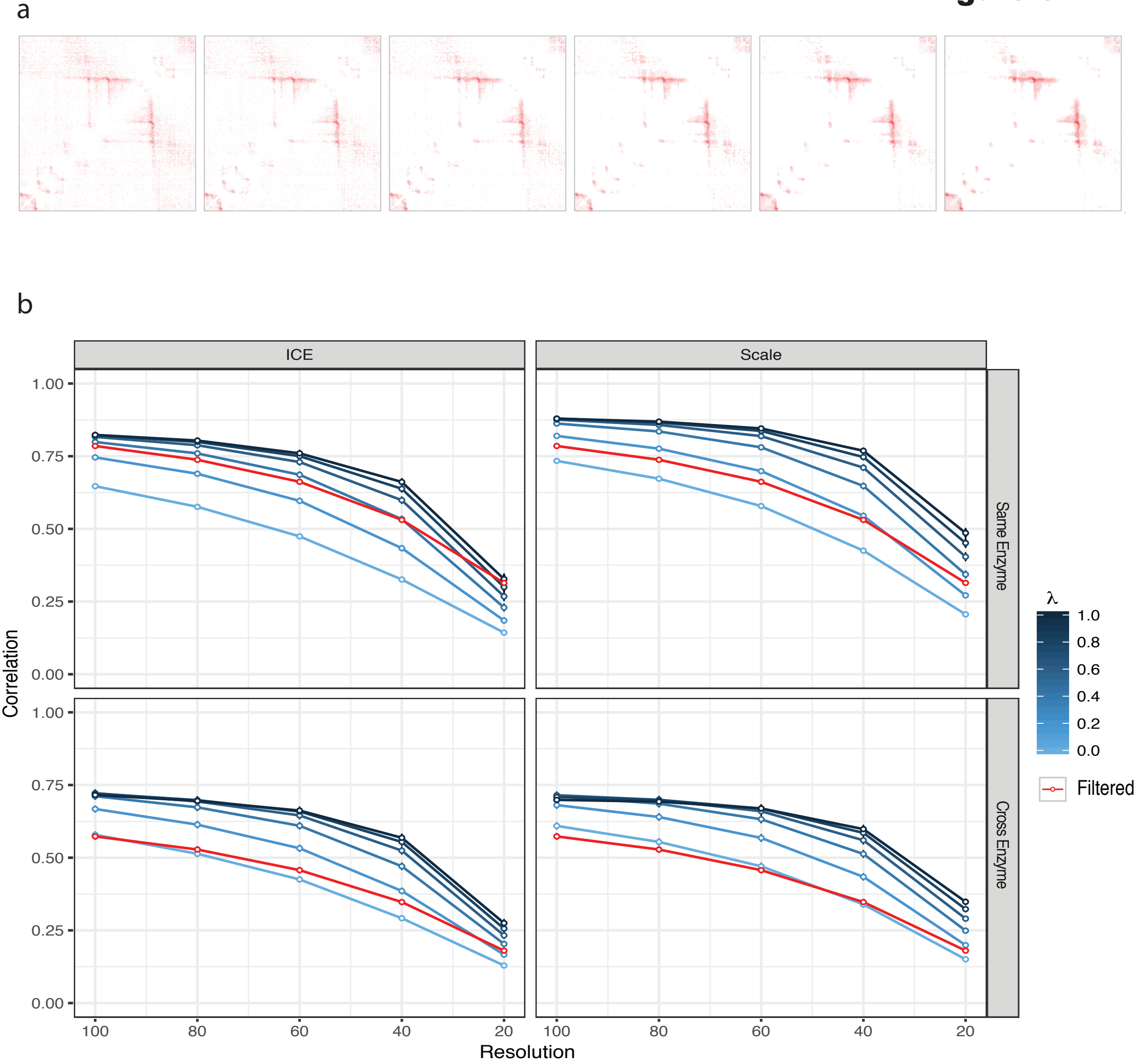
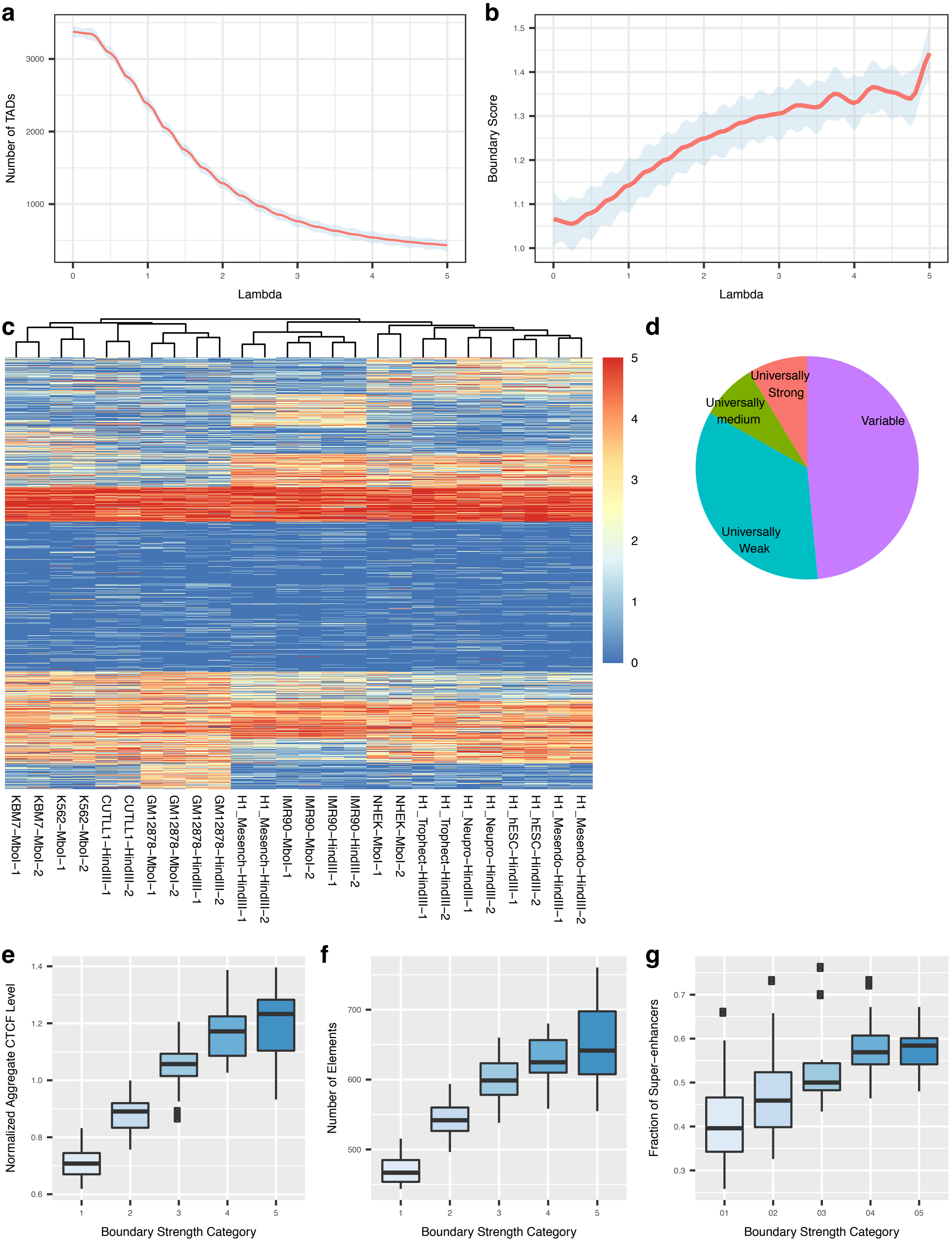
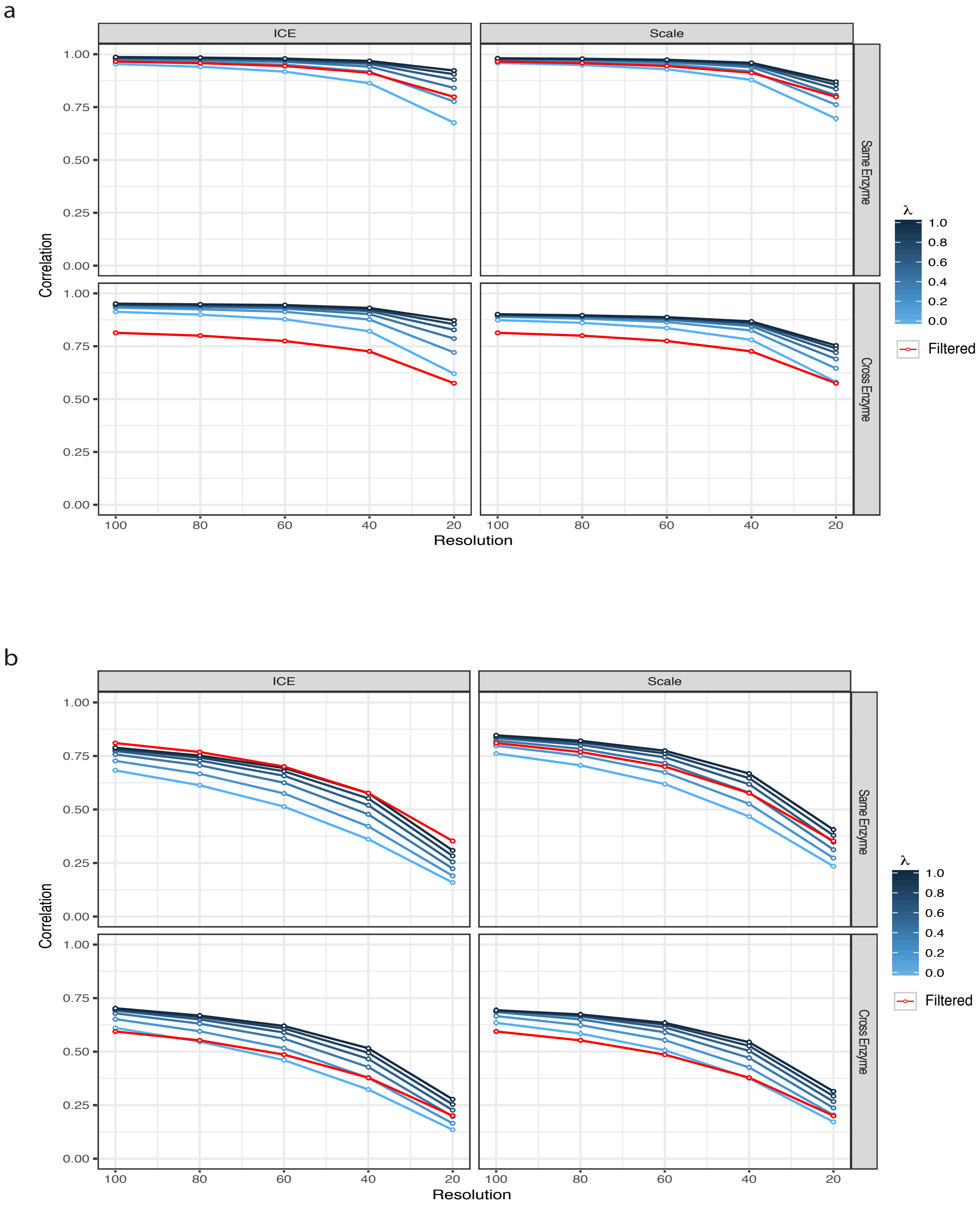

a
b
ICE
Scale
Same Enzyme
Cross Enzyme
Correlation
1.00
0.75
0.50
0.25
0.00
Resolution
100
80
60
40
20
λ
1.0
0.8
0.6
0.4
0.2
0.0
Filtered

Figure 4

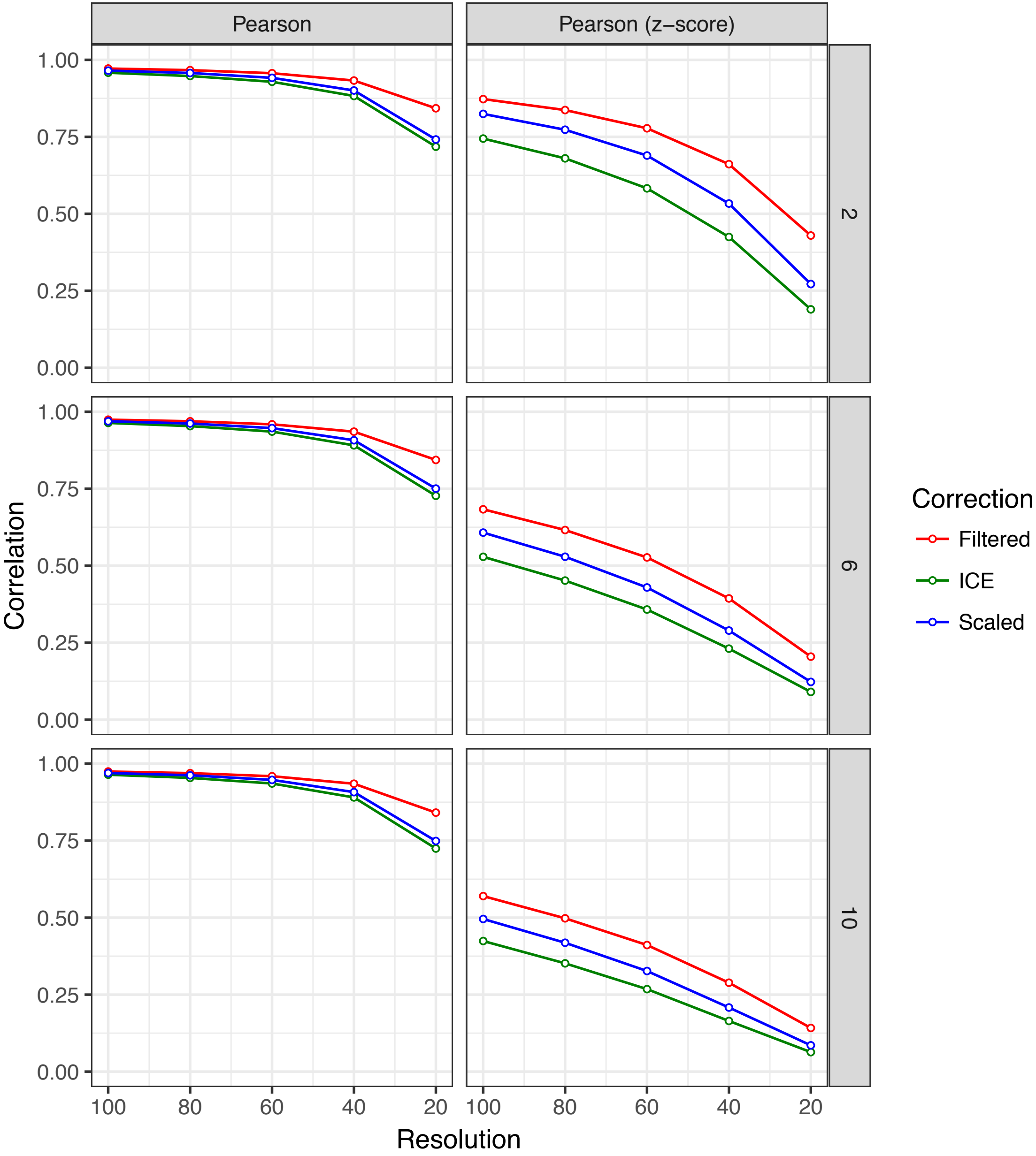


a
Number of TADs
3000
2000
1000
0
1
2
3
4
5
Lambda
b
Boundary Score
1.5
1.4
1.3
1.2
1.1
1.0
0
1
2
3
4
5
Lambda
c
5
4
3
2
1
0
KBM7−MboI−1
KBM7−MboI−2
K562−MboI−1
K562−MboI−2
CUTLL1−HindIII−1
CUTLL1−HindIII−2
GM12878−MboI−1
GM12878−MboI−2
GM12878−HindIII−1
GM12878−HindIII−2
H1_Mesench−HindIII−1
H1_Mesench−HindIII−2
IMR90−MboI−1
IMR90−MboI−2
IMR90−HindIII−1
IMR90−HindIII−2
NHEK−MboI−1
NHEK−MboI−2
H1_Tropect−HindIII−1
H1_Tropect−HindIII−2
H1_Neupro−HindIII−1
H1_Neupro−HindIII−2
H1_hESC−HindIII−1
H1_hESC−HindIII−2
H1_Mesendo−HindIII−1
H1_Mesendo−HindIII−2
d
Universally Strong
Universally medium
Variable
Universally Weak
e
Normalized Aggregate CTCF Level
1.4
1.2
1.0
0.8
0.6
1
2
3
4
5
Boundary Strength Category
f
Number of Elements
700
600
500
1
2
3
4
5
Boundary Strength Category
g
Fraction of Super−enhancers
0.7
0.6
0.5
0.4
0.3
01
02
03
04
05
Boundary Strength Category

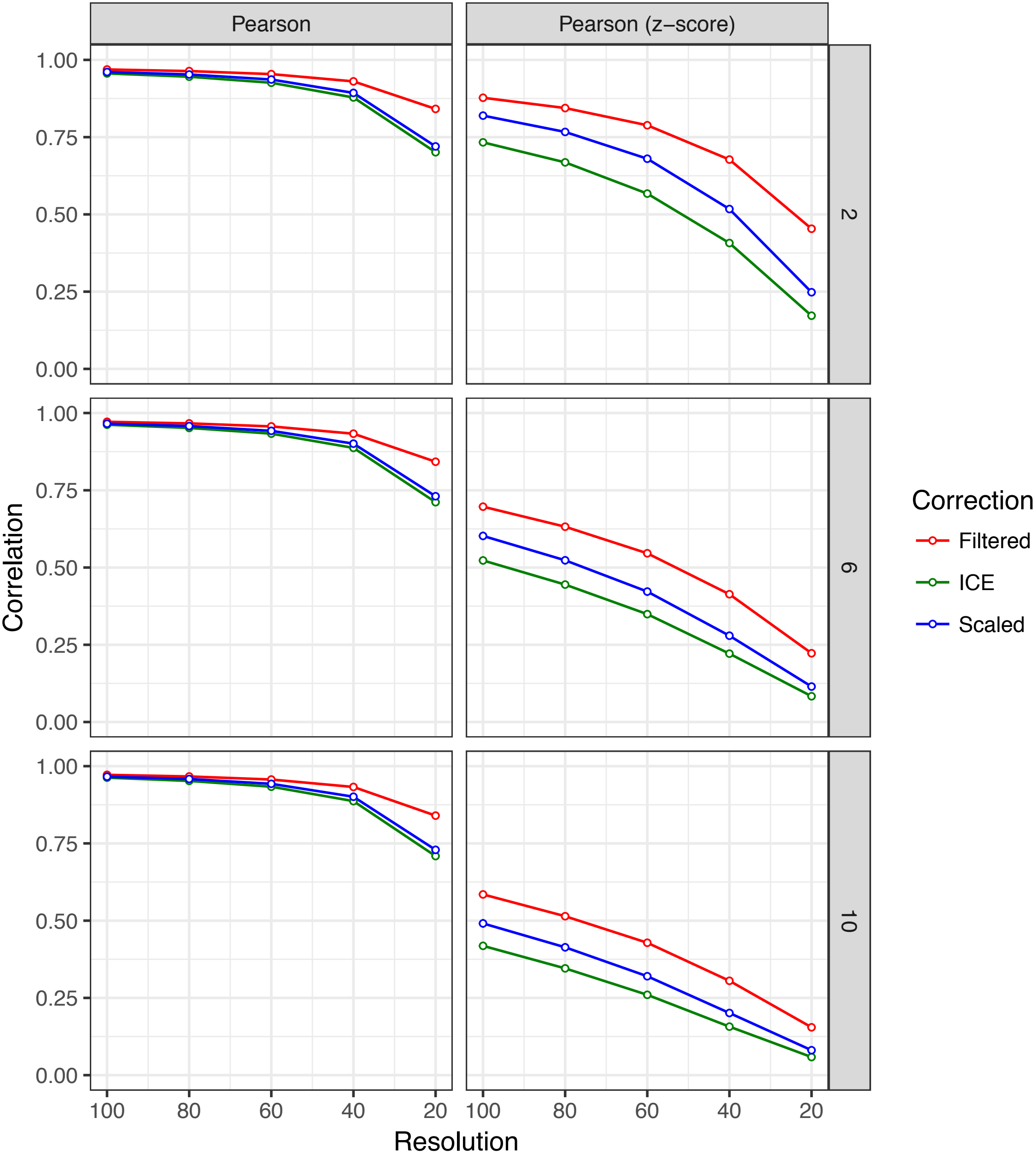
Supplementary Figure 1



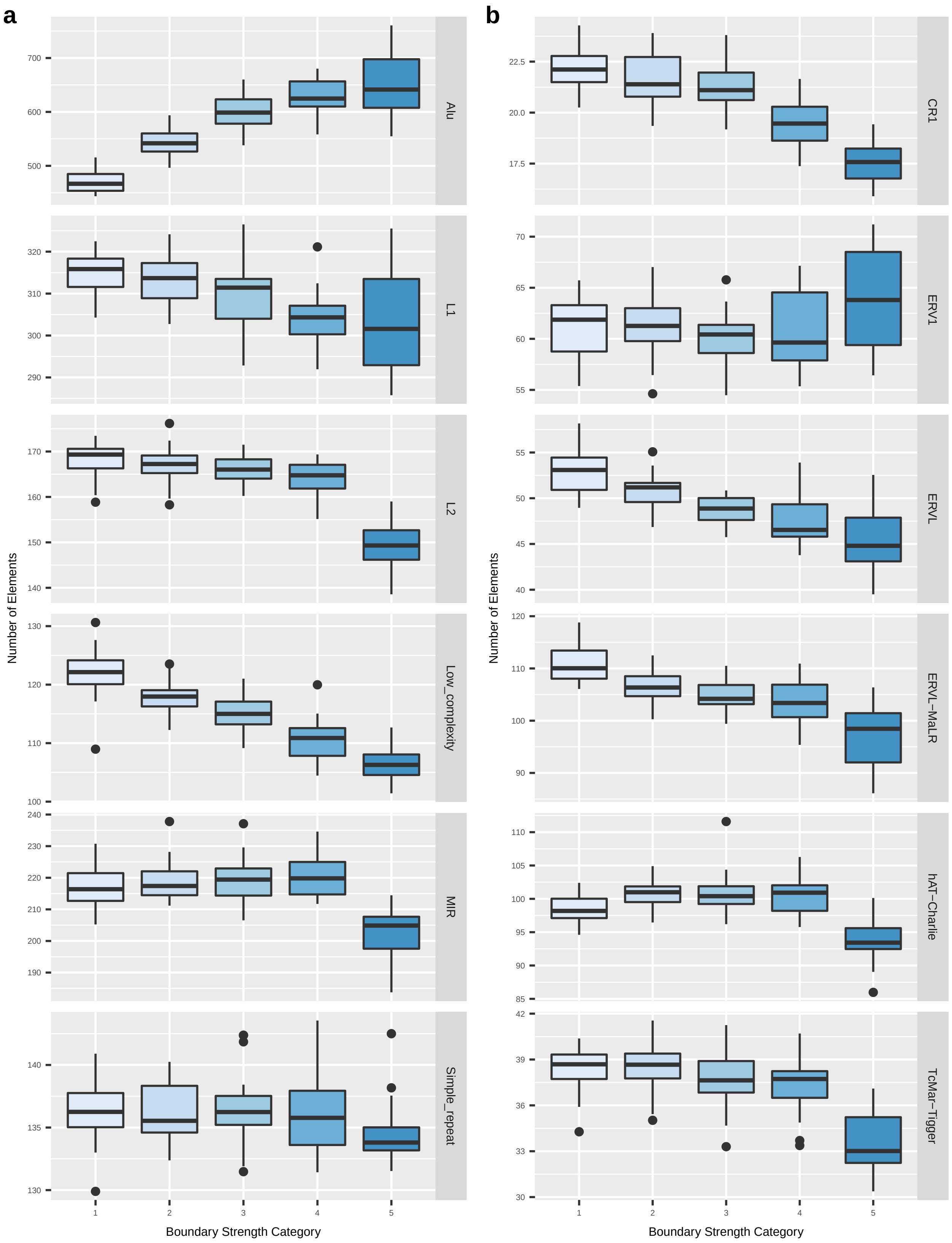
Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4


a
b
Number of Elements
Boundary Strength Category
Alu
L1
L2
Low_complexity
MIR
Simple_repeat
CR1
ERV1
ERVL
ERVL−MaLR
hAT−Charlie
TcMar−Tigger