bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Removing unwanted variation between samples in Hi-C experiments

Kipper Fletez-Brant[1,2], Yunjiang Qiu[3,4], David U. Gorkin[4,5], Ming Hu[6], and Kasper D. Hansen[1,2,*]

[1]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine
[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[3]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego
[4]Ludwig Institute for Cancer Research
[5]Department of Cellular and Molecular Medicine, University of California at San Diego
[6]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation

Abstract

Hi-C data is commonly normalized using single sample processing methods, with focus on comparisons between regions within a given contact map. Here, we aim to compare contact maps across different samples. We demonstrate that unwanted variation is present in Hi-C data on biological replicates, and that this unwanted variation changes across the contact map. We present BNBC, a method for normalization and batch correction of Hi-C data and show that it substantially improves comparisons across samples.

*To whom correspondence should be addressed. Email: khansen@jhsph.edu

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

The Hi-C assay allows for genome-wide measurements of chromatin interactions between different genomic regions (Lieberman-Aiden et al., 2009; Wit and Laat, 2012; Dekker, Marti-Renom, and Mirny, 2013; Schmitt, Hu, and Ren, 2016; Davies et al., 2017). Hi-C has predominately been used to comprehensively study differences in 3D genome structure between loci within a cell type. Partly because of the high cost of the assay, the role of interpersonal variation in 3D genome structure is largely unexplored.

When comparing genomic data between samples, variation can arise from numerous sources that do not reflect the biology of interest including sample procurement, sample storage, library preparation, and sequencing. We refer to these sources of variation as "unwanted" here, because they obscure the underlying biology that is of interest when performing a between-sample comparison. It is critical to correct for this unwanted variation in analysis (Leek, Scharpf, et al., 2010). A number of tools and extensions have been successful at this, particularly for analysis of gene expression data (Leek and Storey, 2007; Leek and Storey, 2008; Gagnon-Bartsch and Speed, 2012; Johnson, Li, and Rabinovic, 2007; Stegle et al., 2010; Leek, 2014; Risso et al., 2014). Existing normalization methods for Hi-C data are single sample methods, focused on comparisons between different loci in the genome. To facilitate this, some methods explicitly model sources of unwanted variation, such as GC content of interaction loci, fragment length, mappability and copy number (Yaffe and Tanay, 2011; Hu et al., 2012; Vidal et al., 2017). Other methods are agnostic to sources of bias and attempts to balance the marginal distribution of contacts (Imakaev et al., 2012; Knight and Ruiz, 2013; Rao et al., 2014; Yan et al., 2017). A comparison of some of these methods found extremely high correlation between their correction factors (Rao et al., 2014); we will use HiCNorm as an exemplar of these within-sample normalization methods (Hu et al., 2012).

By contrast, there has been less work on between-sample normalization. Two existing methods have considered between-sample normalization in the context of a differential comparison, both based on the idea of loess normalization from gene expression microarrays (Yang, Dudoit, et al., 2002). In these methods, the estimated fold-change between conditions are modeled using a loess smoother as a function of either average contact strength (Lun and Smyth, 2015) or distance between loci (Stansfield and Dozmorov, 2017). Using the estimated model, the data are corrected so there is no effect of the covariate on the fold-change. These approaches require a specific comparison of interest and only stabilizes the mean fold-change.

To address the pressing need for cross-sample normalization methods for Hi-C, we developed a BNBC (Bandwise Normalization and Batch Correction), a method for normalization and batch correction of Hi-C data. The method is focused on making individual entries in the contact maps comparable across samples. Our approach is inspired by the observation that patterns of variation between replicates of Hi-C data are different depending on the distance between interacting loci. Therefore, our approach conditions

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on 1D genomic distance, aggregating all contacts between loci separated by a specific distance across all samples into a matrix where the columns are each sample's Hi-C interactions for a specific inter-locus interaction distance. Normalization and removal of unwanted variation proceeds on these separate matrices, which we term band matrices. We show that important biological and statistical features of the Hi-C contact maps are preserved using this approach, while stabilizing the marginal distributions across samples and substantially reducing unwanted variation.

Results

Unwanted variation in Hi-C data varies between distance stratum

It is well described that a Hi-C contact map exhibits an exponential decay in signal as the distance between loci increases (Lieberman-Aiden et al., 2009). When we quantify this behavior across biological replicates (lymphoblastoid cell lines generated from 3 individuals from each of 3 trios from the HapMap project, Table 1), each with 2 growth replicates, we observe substantial variation in the decay rate from sample to sample (Figure 1a). We use the term "biological replicate" here, as it is widely used in the context of a population-based study where each biological replicate is a sample from a different individual. Our samples are lymphoblastoid cell lines from the HapMap project (International HapMap Consortium, 2003), because these cell lines have been a widely used model system to study inter-individual variation and genetic mechanisms in numerous molecular phenotypes including gene expression, chromatin accessibility, histone modification, and DNA methylation (Stranger et al., 2007; Pickrell et al., 2010; Montgomery et al., 2010; Degner et al., 2012; Kasowski et al., 2013; McVicker et al., 2013; Kilpinen et al., 2013; Bell et al., 2011). The Hi-C data from 8 lymphoblastoid cell lines were normalized within growth replicate using HiCNorm (Hu et al., 2012). Following application of HiCNorm, contact maps were corrected for library size using the log counts per million transformation and smoothed using the HiCRep approach (Yang, Zhang, et al., 2017); a bandwidth of 5 was selected using this approach (Methods). Smoothing of the contact map has been found beneficial (Yang, Zhang, et al., 2017; Ursu et al., 2017; Yaffe and Tanay, 2011; Imakaev et al., 2012); recent work, which we confirm, has found that the correlation between technical replicates are increased by smoothing (Yang, Zhang, et al., 2017).

The library preparation of these samples were done at 3 different time points, and we use these different time points to define a batch factor (Table 1). This batch factor encompasses other potential differences between the samples (aside from library perpetration batch), because the trios sampled here come from 3 different human populations (Yoruba, Han Chinese and Puerto Rico), and all of the Yoruba libraries (from Ibadan, Nigeria) were prepared on the same date. The other two populations have one growth replicate in each of two batches. It has been established that phenotypic differences, which are unlikely to

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

be explained by genetics, exists between lymphoblastoid cell lines from different HapMap populations (Stark et al., 2010; Choy et al., 2008; Stranger et al., 2007). These differences might be related to cell line creation and division (Stark et al., 2010). For this reason, it is hard to separate out the effect of Hi-C experimental batch from cell line creation and division in our data. Nonetheless, both types of effects represents unwanted variation insofar as they confound attempts to study the inter-individual variation in 3D genome organization.

To assess unwanted variation beyond changes in the mean, we represented our data as a set of matrices indexed by genomic distance (Figure 1b). Each matrix contains all contacts between loci at a fixed genomic distance for all samples (Methods). We call this a band transformation, since these contacts form diagonal bands in the original Hi-C contact matrices. For each band, we observe substantial variation in the distribution of contacts between samples (Figure 1c-e). These marginal distributions suggests the presence of a unwanted variation (Leek, Scharpf, et al., 2010). We argue that this variation is unwanted across biological replicates, since our data reveal it is at least partly technical. Note that not all contact distances are treated equally when interpreting Hi-C data: one goal of Hi-C experiments is to identify enhancer-promoter contacts, which are thought to occur primarily with 1 Mb (Vernimmen and Bickmore, 2015).

To assess the impact of unwanted variation on our Hi-C data, we first asked, for each contact, how much variation is explained by the batch factor? We measured the amount of explained variation using R^2 from a linear mixed effects model with a random effect to model the increased correlation between growth replicates (Methods). We observe an association between explained variation and distance between loci (Figure 2a), with an average R^2 value of 0.667. This suggests that the effect of the batch factor is substantial and changes with distance. We note again that the "batch factor" here is not simply a Hi-C experiment batch, because the variation between these batches has several potential sources as explained above. To further explore the effect of batch, we performed PCA on each of the band matrices and computed Spearman correlation between each of the first four principal components and the batch indicator. Since our batch factor has 3 levels there are 3 possible orderings of the factor (Figure 2b for one ordering, Supplementary Figure S1 for the other two). This again shows substantial unwanted variation associated with the batch factor, and furthermore shows the dynamic nature of the unwanted variation as distance changes.

To ensure that these data characteristics were not introduced by HiCNorm combined with smoothing, we performed the same measurements on the raw Hi-C contact maps found similar characteristics (Supplementary Figure S2), with one exception. Specifically, we found that the smoothing with a larger bandwidth greatly increased the variation explained by the batch factor (Supplementary Figure S3).

Together, our results highlight the need for between-sample normalization and removal of unwanted variation for Hi-C data, and demonstrates that the effect of unwanted vari-

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ation depends on genomic distance between loci.

Band-wise normalization and batch correction

To normalize the data and remove unwanted variation we used the band transformation framework. Prior to band transformation we use a 2D smoother on the contact maps. Following smoothing we perform quantile normalization separately on each band matrix. Finally, we apply ComBat (Johnson, Li, and Rabinovic, 2007) separately on each band matrix using the batch factor variable. As said above, we refer to our method as band-wise normalization and batch correction (BNBC) (see Methods). This approach is not critically dependent on the choice of smoother or bandwidth nor on the usage of HiCNorm; we observe similar performance across these choices (Figure S3c).

To assess the effect of BNBC, we again measured the variation explained by the batch factor and observed a remarkable decrease of this quantity, which no longer changes as a function of distance (Figure 2c). Specifically mean R^2 decreases from 0.667 to 0.09. Comparing R^2 between HiCNorm and BNBC shows a decrease of essentially every individual contact (Figure 2e); this pattern depends on distance (Supplementary Figure S4). Likewise, the correlation between each of the first 4 principal components and the batch factor was close to zero (Figure 2d). In addition, the marginal distributions are stabilized, which is expected since we performed quantile normalization (Supplementary Figure S5). This shows that BNBC removes substantial unwanted variation associated with batch.

We next investigated the impact of BNBC on features of the contact map. The BNBC-corrected data exhibits the standard decay pattern of Hi-C data, without variation across replicates (Supplementary Figure S5a). More interestingly, we observe a contact map very similar to HiCNorm (Supplementary Figure S6). The same is true for its associated first eigenvector, which is commonly used to identify A/B compartments (Supplementary Figure S6). We conclude that the application of BNBC does not distort gross features of the contact map.

Above we show that increasing the bandwidth of the smoother increases the variation explained by the batch factor, and also increases the correlation between technical replicates. When we examine the impact of increased smoothing bandwidth following application of BNBC, we found little effect of bandwidth or put differently, that BNBC was able to correct for the increase. Since increasing the bandwidth does increase the correlation between technical replicates, we use the bandwidth recommended by the HiC-Rep approach (a bandwidth of 5). We note that the HiC-Rep criteria does not include consideration of biological signal and we caution that such signal could be diminished. For example, in work on normalization of DNA methylation arrays, we found that methods which performs best at reducing technical variation do not necessarily perform best when the assessment is replication of biological signal (Fortin, Labbe, et al., 2014).

Popular alternatives to ComBat include SVA (Leek and Storey, 2007; Leek and Storey,

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2008; Leek, 2014), RUV (Gagnon-Bartsch and Speed, 2012; Risso et al., 2014) and PEER (Stegle et al., 2010) which are all variation of factor models. These methods construct surrogate variables which represents unmeasured sources of unwanted variation. We experimented with the use of PEER instead of ComBat and observed dramatically reduced performance compared to ComBat (Figures S7, S1). Note that the choice of R^2 for evaluation metric can be considered unfair since ComBat uses the batch factor as input; the correlation plots should not be affected by this. We ran PEER using both 1 and 4 factors; results were very similar. The performance of PEER raises the question of how to best correct for unwanted variation when an explicit batch factor is unavailable. This is an important open question because (1) ComBat requires two samples for each level of the batch factor and (2) unwanted variation may be mediated through other factors than library preparation batch.

Discussion

To analyze Hi-C across samples, including biological replicates, it is clear that between-sample normalization methods are necessary. Here, we have characterized unwanted variation present in Hi-C contact maps and have developed a correction method named BNBC. We show unwanted variation exhibits a distance-dependent effect, in addition to known distance-based features of Hi-C contact maps. We present BNBC, a modular approach where we combine band transformation with existing tools for normalization and removal of unwanted variation. We show that BNBC performs well in reducing the impact of unwanted variation while still preserving important 3D features, such as the structure of the contact map and A/B compartments. Our focus in this work has been the normalization of individual entries in the contact map, but we note that proper normalization of such entries are not a requirement for normalization of higher-order structures. For example, we have previously observed that A/B compartments reproduce well between samples from the same cell type (Fortin and Hansen, 2015). Note that the batch factor we have analyzed here could be driven by either or both of cell line construction and Hi-C library preparation. Proper normalization and correction for unwanted variation will be critical for comparing Hi-C contact maps between different samples.

Methods

Data Generation

Hi-C experiments: Lymphoblast Hi-C data analyzed were generated by the dilution Hi-C method using HindIII (Lieberman-Aiden et al., 2009) on 9 lymphoblastoid cell lines derived from the 1000 Genomes project (Table 1). Data are publicly available through

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1000 genomes (Chaisson et al., 2017) as well as through the 4D Nucleome data portal (https://data.4dnucleome.org; accessions 4DNESYUYFD6H, 4DNESVKLYDOH, 4DNESHGL976U, 4DNESJ1VX52C, 4DNESI2UKI7P, 4DNESTAPSPUC, 4DNES4GSP9S4, 4DNESJIYRA44, 4DNESE3ICNE1). Hi-C contact matrices were generated by tiling the genome into 40kb bins and counting the number of interactions between bins. We refer to these as raw contact matrices.

Hi-C read alignment and contact matrices: Reads were aligned to hg19 reference genome using bwa-mem (Li, 2013). Read ends were aligned independently as paired-end model in BWA cannot handle the long insert size of Hi-C reads. Aligned reads were further filtered to keep only the 5′ alignment. Read pairs were then manually paired. Read pairs with low mapping quality (MAPQ¡10) were discarded, and PCR duplicates were removed using Picard tools 1.131 http://broadinstitute.github.io/picard. To construct the contact matrices, Hi-C read pairs were assigned to predefined 40Kb genomic bins. Bins with low mapping quality (< 0.8), low GC content (< 0.3), and low fragment length ($< 10\%$ of the bin size) were discarded.

Band Matrices

To make comparisons across individuals, we form band matrices, which are matrices whose columns are all matrix band i from each sample. A matrix band is a collection of entries in a contact matrix between two loci at a fixed distance. Formally, band i is the collection of j, k entries with $|j - k| + 1 = i$.

Log counts per million transformation

We use the logCPM (log counts per million) transformation previous described (Law et al., 2014). Specifically, for a contact matrix $\mathbb{X}$ we estimate library size L by the sum of the upper triangular matrix of each of the chromosome specific contact matrices. This discards inter-chromosomal contacts as well as the diagonal of the contact matrix. The logCPM matrix $\mathbb{Y}$ is defined as

$$Y_{ij} = \log\left(\frac{X_{ij} + 0.5}{L + 1} 10^6\right)$$

where X_{ij} refers to element i, j from the contact matrix $\mathbb{X}$ and L is the estimated library size for that matrix. For data normalized using HiCNorm both $\mathbb{X}$ and L are not integers.

HiCNorm

We normalized data using HiCNorm (Hu et al., 2012) with an updated implementation (https://github.com/ren-lab/HiCNorm). Following HiCNorm normalization, we

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

applied the log counts per million transformation (see above). We then smooth the contact matrices with a box smoother with a bandwidth of 5 bins; we use HiCRep to choose the bandwidth based on the correlation between technical replicates (Yang, Zhang, et al., 2017). The bandwidth we select is the same as the bandwidth selected for 40kb resolution Hi-C data in Yang, Zhang, et al. (2017). Smoothing was performed using the EBImage package (Pau et al., 2010); this is a separate but equivalent implementation to HiCRep.

BNBC

BNBC has the following components: separate smoothing of each contact matrix, application of the band transformation, quantile normalization on each band matrix and finally application of ComBat on each band matrix.

Following the log counts per million transformation of the raw contact matrices, we smooth individual chromosome matrices using a box smoother with a bandwidth of 5, as selected by the HiCRep approach (Yang, Zhang, et al., 2017). Each contact matrix and each chromosome is smoothed separately. We next apply the band transformation (see above) and quantile normalize each band matrix separately (Bolstad et al., 2003).

Following quantile normalization we apply ComBat (Johnson, Li, and Rabinovic, 2007) to each band matrix separately. We apply the parametric prior described in Johnson, Li, and Rabinovic (2007). Prior to applying ComBat, we filter out matrix cells for which the intra-batch variance is zero for all batches. After applying ComBat we set filtered matrix cells to zero.

Our implementation of BNBC is available in the `bnbc` R package from the Bioconductor project(Gentleman et al., 2004; Huber et al., 2015) (`https://www.bioconductor.org/packages/bnbc`).

Explained variation and smoothed boxplot

To assess unwanted variation for each matrix cell in a contact matrix, we employ a linear mixed model approach. Specifically, we fit a mixed effect model regressing HiC contact strength on batch indicator, with a random effect at the subject level to capture the increased correlation between technical replicates. This model is fit using the R package *varComp* (Qu, Guennel, and Marshall, 2013) and R^2 for this model is calculated using the method of Edwards et al. (2008).

To display R^2 as a function of distance, we first compute a series of box plots of R^2, one for each band matrix. We extract the summary measures for the box plots (median, 1st and 3rd quantile and 1.5 times the inter-quartile range). We then display these 5 curves, with color fills. Medians are black, 1st and 3rd quartiles are pink and 1.5 times the inter-quartile range are blue.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A/B compartments from smoothed contact matrices

A/B compartments were originally proposed to be estimated using the first eigenvector of a suitable transform of the contact matrix Lieberman-Aiden et al., 2009. Specifically, the contact matrix was transformed using the observed-expected transformation where each matrix band was divided by its mean. Our contact matrices following application of the log counts per million transform and smoothing are on the log scale. To get A/B compartments from the output of BNBC (Supplementary Figure S6), we exponentiate every entry in the matrix, multiply by 10^6, apply the observed-expected transformation and compute the first eigenvector. Finally, we standardize the first eigenvectors to be in $(-1, 1)$ and then smooth the standardized eigenvectors using a moving-average as done by Fortin and Hansen (2015).

Acknowledgements

Funding: Research reported in this publication was supported by National Institute of Diabetes and Digestive and Kidney Diseases and the National Cancer Institute of the National Institutes of Health under award numbers 54DK107977 and U24CA180996. KFB was supported by the Maryland Genetics, Epidemiology and Medicine (MD-GEM) program. DUG was supported by funding from the A.P. Giannini Foundation and the San Diego Institutional Research and Academic Career Development Award (IRACDA) program.
Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Conflict of Interest: None declared.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Bibliography

Bell, J. T., A. A. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi, J. F. Degner, Y. Gilad, and J. K. Pritchard (2011). "DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines". *Genome Biology* 12, R10. DOI: 10.1186/gb-2011-12-1-r10.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". *Bioinformatics* 19, pp. 185–193. DOI: 10.1093/bioinformatics/19.2.185.

Chaisson, M. J. P., A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. Hastie, D. Antaki, P. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C.-S. Chin, Z. Chong, N. T. Chuang, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, G. David, M. Gujral, V. Guryev, W. Haynes-Heaton, J. Korlach, S. Kumar, J. Y. Kwon, J. E. Lee, J. Lee, W.-P. Lee, S. P. Lee, P. Marks, K. Valud-Martinez, S. Meiers, K. M. Munson, F. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stutz, D. C. J. Spierings, A. Ward, A. E. Welsch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P.-Y. Kwok, P. M. Lansdorp, G. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. Talkowski, R. E. Mills, T. Marschall, J. Korbel, E. E. Eichler, and C. Lee (2017). "Multi-platform discovery of haplotype-resolved structural variation in human genomes". *bioRxiv*, p. 193144. DOI: 10.1101/193144.

Choy, E., R. Yelensky, S. Bonakdar, R. M. Plenge, R. Saxena, P. L. De Jager, S. Y. Shaw, C. S. Wolfish, J. M. Slavik, C. Cotsapas, M. Rivas, E. T. Dermitzakis, E. Cahir-McFarland, E. Kieff, D. Hafler, M. J. Daly, and D. Altshuler (2008). "Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines". *PLOS Genetics* 4, e1000287. DOI: 10.1371/journal.pgen.1000287.

Davies, J. O. J., A. M. Oudelaar, D. R. Higgs, and J. R. Hughes (2017). "How best to identify chromosomal interactions: a comparison of approaches". *Nature Methods* 14, pp. 125–134. DOI: 10.1038/nmeth.4146.

Degner, J. F., A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard (2012). "DNase I sensitivity QTLs are a major determinant of human expression variation". *Nature* 482, pp. 390–394. DOI: 10.1038/nature10808.

Dekker, J., M. A. Marti-Renom, and L. A. Mirny (2013). "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data". *Nature Reviews Genetics* 14, pp. 390–403. DOI: 10.1038/nrg3454.

Edwards, L. J., K. E. Muller, R. D. Wolfinger, B. F. Qaqish, and O. Schabenberger (2008). "An R2 statistic for fixed effects in the linear mixed model". *Statistics in Medicine* 27, pp. 6137–6157. DOI: 10.1002/sim.3429.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fortin, J.-P. and K. D. Hansen (2015). "Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data". *Genome Biology* 16, p. 180. DOI: 10.1186/s13059-015-0741-y.

Fortin, J.-P., A. Labbe, M. Lemire, B. W. Zanke, T. J. Hudson, E. J. Fertig, C. M. Greenwood, and K. D. Hansen (2014). "Functional normalization of 450k methylation array data improves replication in large cancer studies". *Genome Biology* 15, p. 503. DOI: 10.1186/s13059-014-0503-2.

Gagnon-Bartsch, J. A. and T. P. Speed (2012). "Using control genes to correct for unwanted variation in microarray data". *Biostatistics* 13, pp. 539–552. DOI: 10.1093/biostatistics/kxr034.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang (2004). "Bioconductor: open software development for computational biology and bioinformatics". *Genome Biology* 5, R80. DOI: 10.1186/gb-2004-5-10-r80.

Hu, M., K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu (2012). "HiCNorm: removing biases in Hi-C data via Poisson regression". *Bioinformatics* 28, pp. 3131–3133. DOI: 10.1093/bioinformatics/bts570.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan (2015). "Orchestrating high-throughput genomic analysis with Bioconductor". *Nature Methods* 12, pp. 115–121. DOI: 10.1038/nmeth.3252.

Imakaev, M., G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny (2012). "Iterative correction of Hi-C data reveals hallmarks of chromosome organization". *Nature Methods* 9, pp. 999–1003. DOI: 10.1038/nmeth.2148.

International HapMap Consortium (2003). "The International HapMap Project". *Nature* 426, pp. 789–796. DOI: 10.1038/nature02168.

Johnson, W. E., C. Li, and A. Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". *Biostatistics* 8, pp. 118–127. DOI: 10.1093/biostatistics/kxj037.

Kasowski, M., S. Kyriazopoulou-Panagiotopoulou, F. Grubert, J. B. Zaugg, A. Kundaje, Y. Liu, A. P. Boyle, Q. C. Zhang, F. Zakharia, D. V. Spacek, J. Li, D. Xie, A. Olarerin-George, L. M. Steinmetz, J. B. Hogenesch, M. Kellis, S. Batzoglou, and M. Snyder (2013). "Extensive variation in chromatin states across humans". *Science* 342, pp. 750–752. DOI: 10.1126/science.1242510.

Kilpinen, H., S. M. Waszak, A. R. Gschwind, S. K. Raghav, R. M. Witwicki, A. Orioli, E. Migliavacca, M. Wiederkehr, M. Gutierrez-Arcelus, N. I. Panousis, A. Yurovsky, T. Lappalainen, L. Romano-Palumbo, A. Planchon, D. Bielser, J. Bryois, I. Padioleau, G.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Udin, S. Thurnheer, D. Hacker, L. J. Core, J. T. Lis, N. Hernandez, A. Reymond, B. Deplancke, and E. T. Dermitzakis (2013). "Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription". *Science* 342, pp. 744–747. DOI: 10.1126/science.1242463.

Knight, P. A. and D. Ruiz (2013). "A fast algorithm for matrix balancing". *IMA Journal of Numerical Analysis* 33, pp. 1029–1047. DOI: 10.1093/imanum/drs019.

Law, C. W., Y. Chen, W. Shi, and G. K. Smyth (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". *Genome Biology* 15, R29. DOI: 10.1186/gb-2014-15-2-r29.

Leek, J. T. (2014). "svaseq: removing batch effects and other unwanted noise from sequencing data". *Nucleic Acids Research* 42, gku864. DOI: 10.1093/nar/gku864.

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data". *Nature Reviews Genetics* 11, pp. 733–739. DOI: 10.1038/nrg2825.

Leek, J. T. and J. D. Storey (2007). "Capturing heterogeneity in gene expression studies by surrogate variable analysis". *PLOS Genetics* 3, pp. 1724–1735. DOI: 10.1371/journal.pgen.0030161.

– (2008). "A general framework for multiple testing dependence". *PNAS* 105, pp. 18718–18723. DOI: 10.1073/pnas.0808709105.

Li, H. (2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". *arXiv*, p. 1303.3997. URL: http://arxiv.org/abs/1303.3997.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". *Science* 326, pp. 289–293. DOI: 10.1126/science.1181369.

Lun, A. T. L. and G. K. Smyth (2015). "diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data". *BMC Bioinformatics* 16, p. 258. DOI: 10.1186/s12859-015-0683-0.

McVicker, G., B. van de Geijn, J. F. Degner, C. E. Cain, N. E. Banovich, A. Raj, N. Lewellen, M. Myrthil, Y. Gilad, and J. K. Pritchard (2013). "Identification of genetic variants that affect histone modifications in human cells". *Science* 342, pp. 747–749. DOI: 10.1126/science.1242429.

Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis (2010). "Transcriptome genetics using second generation sequencing in a Caucasian population". *Nature* 464, pp. 773–777. DOI: 10.1038/nature08903.

Pau, G., F. Fuchs, O. Sklyar, M. Boutros, and W. Huber (2010). "EBImage – an R package for image processing with applications to cellular phenotypes". *Bioinformatics* 26, pp. 979–981. DOI: 10.1093/bioinformatics/btq046.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard (2010). "Understanding mechanisms underlying human gene expression variation with RNA sequencing". *Nature* 464, pp. 768–772. DOI: 10.1038/nature08872.

Qu, L., T. Guennel, and S. L. Marshall (2013). "Linear score tests for variance components in linear mixed models and applications to genetic association studies". *Biometrics* 69, pp. 883–892. DOI: 10.1111/biom.12095.

Rao, S. S. P., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". *Cell* 159, pp. 1665–1680. DOI: 10.1016/j.cell.2014.11.021.

Risso, D., J. Ngai, T. P. Speed, and S. Dudoit (2014). "Normalization of RNA-seq data using factor analysis of control genes or samples". *Nature Biotechnology* 32, pp. 896–902. DOI: 10.1038/nbt.2931.

Schmitt, A. D., M. Hu, and B. Ren (2016). "Genome-wide mapping and analysis of chromosome architecture". *Nat. Rev. Mol. Cell Biol.* 17, pp. 743–755. DOI: 10.1038/nrm.2016.104.

Stansfield, J. and M. G. Dozmorov (2017). "HiCdiff: A method for joint normalization of Hi-C datasets and differential chromatin interaction detection". *bioRxiv*, p. 147850. DOI: 10.1101/147850.

Stark, A. L., W. Zhang, T. Zhou, P. H. O'Donnell, C. M. Beiswanger, R. S. Huang, N. J. Cox, and M. E. Dolan (2010). "Population differences in the rate of proliferation of international HapMap cell lines". *American Journal of Human Genetics* 87, pp. 829–833. DOI: 10.1016/j.ajhg.2010.10.018.

Stegle, O., L. Parts, R. Durbin, and J. Winn (2010). "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies". *PLoS Computational Biology* 6, e1000770. DOI: 10.1371/journal.pcbi.1000770.

Stranger, B. E., A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, S. Montgomery, S. Tavaré, P. Deloukas, and E. T. Dermitzakis (2007). "Population genomics of human gene expression". *Nature Genetics* 39, pp. 1217–1224. DOI: 10.1038/ng2142.

Ursu, O., N. Boley, M. Taranova, Y. X. Rachel Wang, G. G. Yardimci, W. S. Noble, and A. Kundaje (2017). "GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs". *bioRxiv*, p. 181842. DOI: 10.1101/181842.

Vernimmen, D. and W. A. Bickmore (2015). "The Hierarchy of Transcriptional Activation: From Enhancer to Promoter". *Trends in Genetics* 31, pp. 696–708. DOI: 10.1016/j.tig.2015.10.004.

Vidal, E., F. le Dily, J. Quilez, R. Stadhouders, Y. Cuartero, T. Graf, M. A. Marti-Renom, M. Beato, and G. Filion (2017). "OneD: increasing reproducibility of Hi-C Samples with abnormal karyotypes". *bioRxiv*, p. 148254. DOI: 10.1101/148254.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Wit, E. de and W. de Laat (2012). "A decade of 3C technologies: insights into nuclear organization". *Genes & Development* 26, pp. 11–24. DOI: 10.1101/gad.179804.111.

Yaffe, E. and A. Tanay (2011). "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture". *Nature Genetics* 43, pp. 1059–1065. DOI: 10.1038/ng.947.

Yan, K.-K., G. Gürkan Yardimci, C. Yan, W. S. Noble, and M. Gerstein (2017). "HiC-Spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps". *Bioinformatics* 33, pp. 2199–2201. DOI: 10.1093/bioinformatics/btx152.

Yang, T., F. Zhang, G. G. Yardimci, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li (2017). "HiCRep: assessing the reproducibility of Hi-C data using a stratum- adjusted correlation coefficient". *Genome Research*. DOI: 10.1101/gr.220640.117.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation". *Nucleic Acids Research* 30, e15.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tables

Table 1. Sample Information

Sample	Replicate	Ethnicity	Sex	Family	Role	Batch	Library preparation
GM19238	1	YRI	F	1	Mother	1	9/26/14
GM19238	2	YRI	F	1	Mother	1	9/26/14
GM19239	2	YRI	M	1	Father	1	9/26/14
HG00512	1	CHS	M	2	Father	2	3/4/15
HG00512	2	CHS	M	2	Father	3	5/28/15
HG00513	1	CHS	F	2	Mother	2	3/4/15
HG00513	2	CHS	F	2	Mother	3	5/28/15
HG00514	1	CHS	F	2	Child	2	3/4/15
HG00514	2	CHS	F	2	Child	3	5/28/15
HG00731	1	PUR	M	3	Father	2	3/4/15
HG00731	2	PUR	M	3	Father	3	5/28/15
HG00732	1	PUR	F	3	Mother	2	3/4/15
HG00732	2	PUR	F	3	Mother	3	5/28/15
HG00733	1	PUR	F	3	Child	2	3/4/15
HG00733	2	PUR	F	3	Child	3	5/28/15

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figures

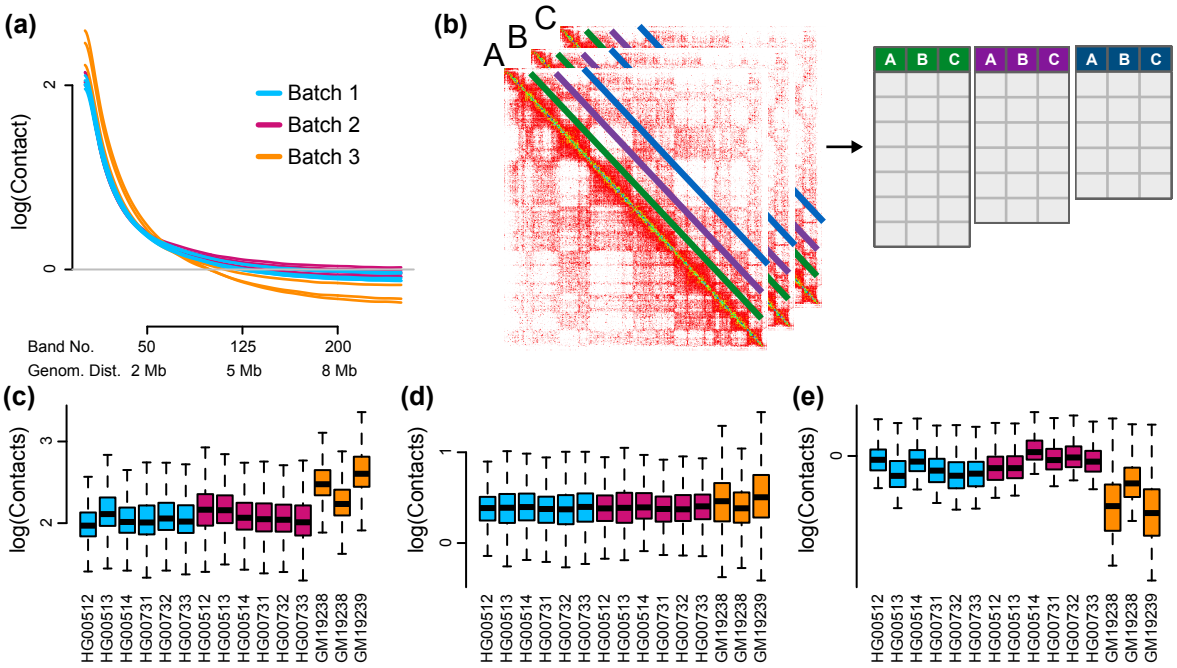


Figure 1. Unwanted variation in Hi-C data. We display Hi-C data from chromosome 14 from 8 different individuals, 7 of which have 2 technical replicates, processed in 3 batches. Each sample is normalized using HiCNorm followed by spatial smoothing using HiCRep; data is on a logarithmic scale. **(a)** Mean contact as a function of distance. Each sample is a separate curve. **(b)** Band transformation of a collection of Hi-C contact maps. **(c)-(e)** Boxplots of the marginal distribution of contacts across samples, for loci separated by **(c)** 40 kb (band 2), **(d)** 2 Mb (band 50) and **(e)** 8 Mb (band 200).

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

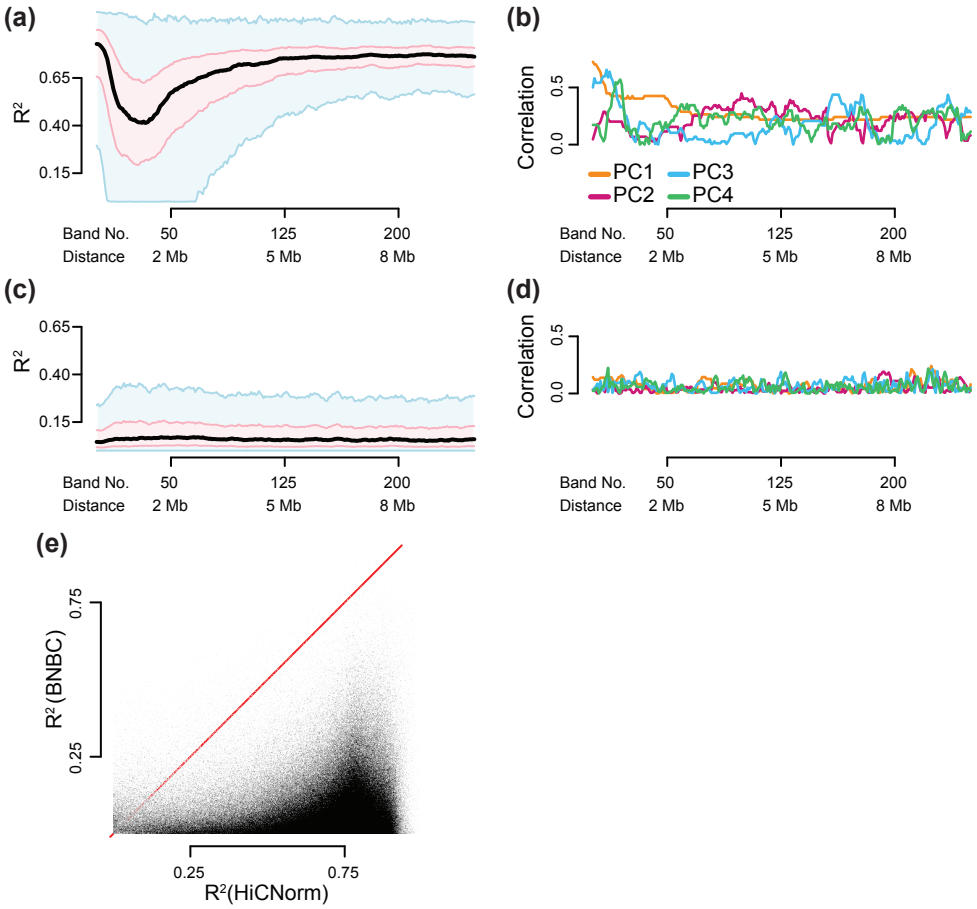


Figure 2. Substantial unwanted variation in Hi-C data. (a) The percentage of variance explained (R^2) by the batch factor for the HiCNorm processed data, as a function of distance (Methods). The distributions are displayed as a series of smoothed boxplots (black: median, pink: 1st and 3rd quartiles, blue: 1.5 times inter-quartile range, see Methods). **(b)** The Spearman correlation of the 1st-4th principal components of each band matrix with the batch factor, as a function of distance, for the HiCNorm processed data. Other permutations of the batch factor are shown in Supplementary Figure S1. **(c)** Like (a) but for data processed using BNBC. **(d)** Like (b) but for data processed using BNBC. **(e)** A scatterplot of R^2 for data processed using BNBC vs. data processed using HiCNorm, for entries in the contact map separated by less than 10 Mb (band 250).

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Materials

for

Distance-dependent between-sample normalization for Hi-C experiments

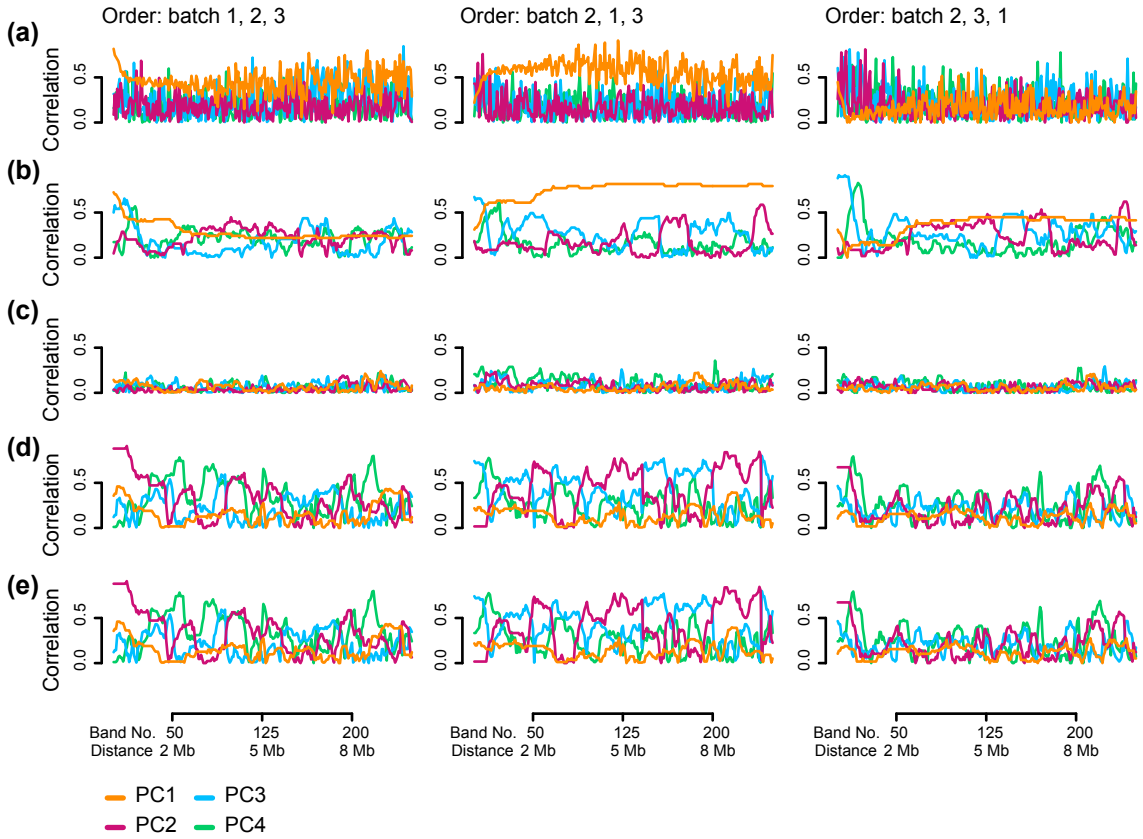
Kipper Fletez-Brant, Yunjiang Qiu, David U. Gorkin, Ming Hu, Kasper D. Hansen*

Contains Supplementary Figures S1-S7.

*To whom correspondence should be addressed. Email: khansen@jhsph.edu

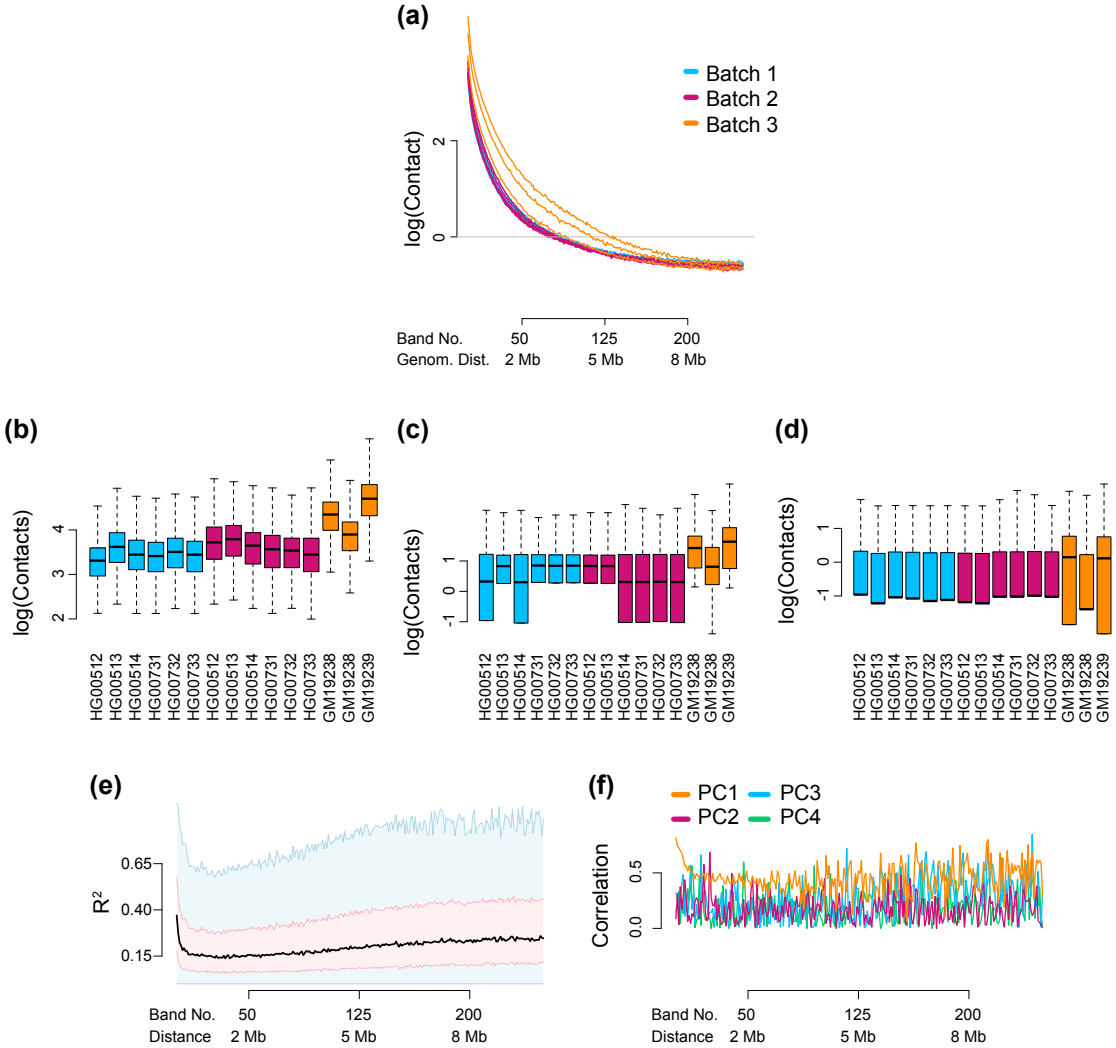
bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figures



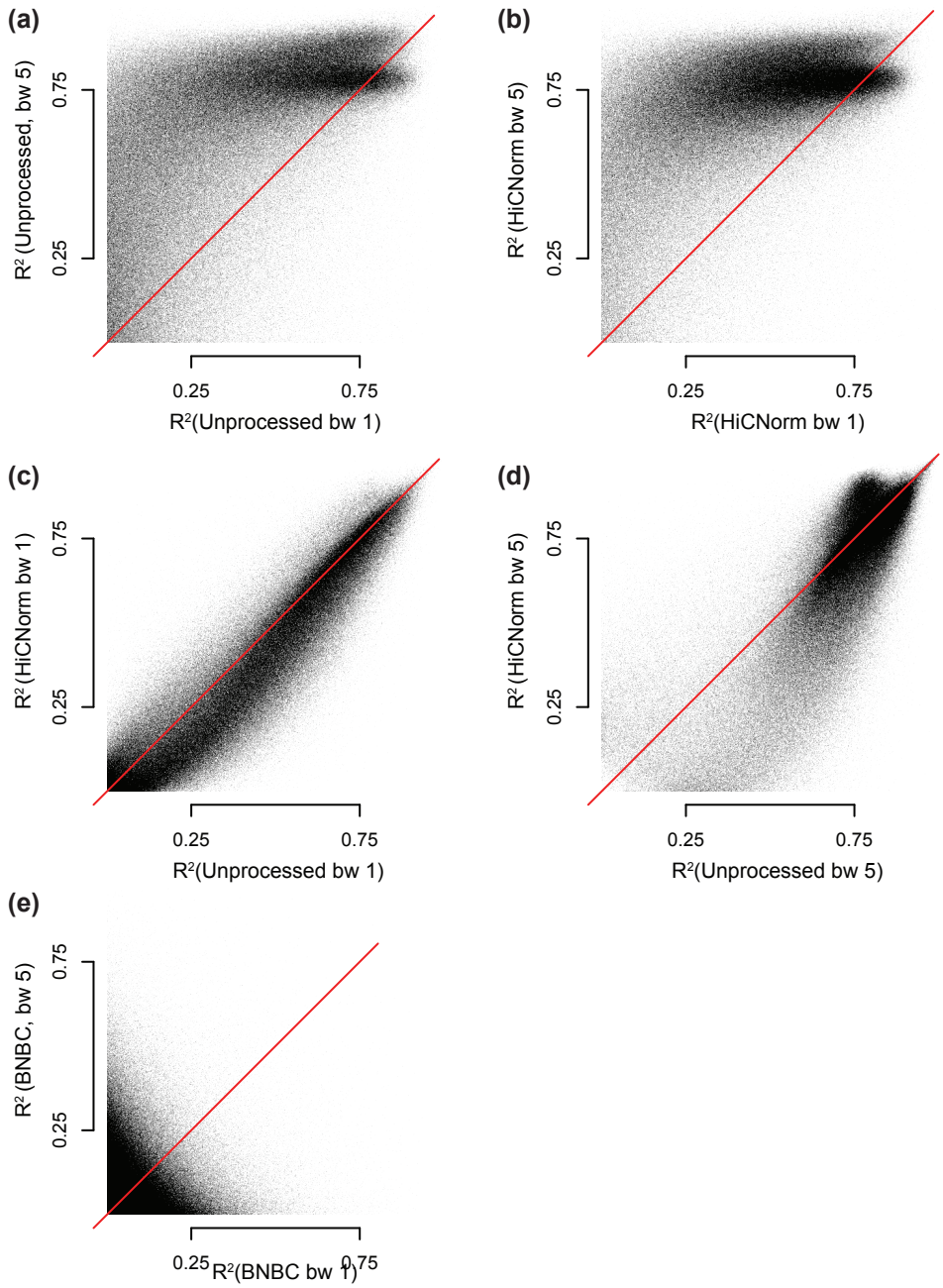
Supplementary Figure S1. The performance assessment of all methods using correlation of batch with principal components. As in Figure 2b we assess the influence of batch using Spearman correlation between the batch factor and the 1st-4th principal components of each band matrix, for various methods. Column 1 uses the ordering batch 1, batch 2, batch 3. Column 2 uses the ordering batch 2, batch 1, batch 3. Column 3 uses the ordering batch 2, batch 3, batch 1. **(a)** Unprocessed data **(b)** HiCNorm **(c)** BNBC. **(d)** BNBC using PEER with 1 hidden factor. **(e)** BNBC using PEER with 4 hidden factors. The first column and the first 3 rows reproduces Supplementary Figure S2f, Figure 2b, and Figure 2d.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



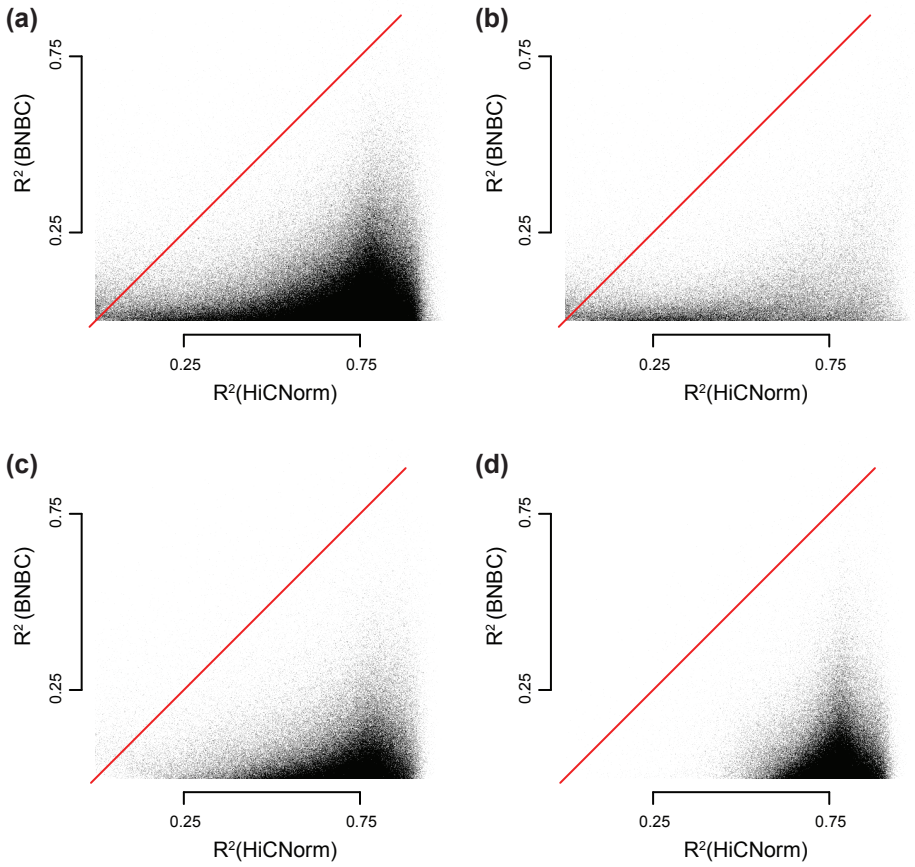
Supplementary Figure S2. Unprocessed data. As Figures 1 and 2, but using data prior to normalization by HiCNorm and smoothing by HiCRep. Data has been corrected for library size using the log counts per million transformation. **(a)** Mean contact as a function of distance. Each sample is a separate curve. **(b)-(d)** Boxplots of the marginal distribution of contacts across samples, for loci separated by **(b)** 40 kb (band 2), **(c)** 2 Mb (band 50) and **(d)** 8 Mb (band 200). **(e)-(f)** As Figure 2a,b. The correlations with the first four principal components are jagged due to the lack of smoothing.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



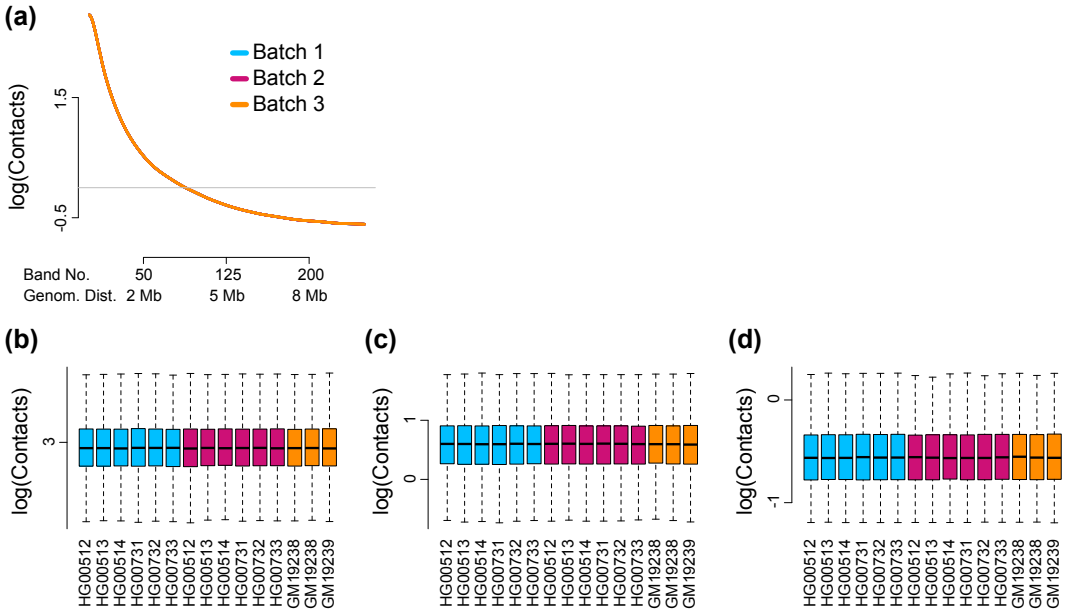
Supplementary Figure S3. Use of HiCNorm and choice of width impact unwanted variation. Pairwise scatterplots of explained variation by batch (R^2), comparing various methods. As Figure 2e. **(a)** Unprocessed data, smoothed with a bandwidth of 1 and 5. **(b)** HiCNorm data, smoothed with a bandwidth of 1 and 5. **(c)** HiCNorm data vs. Unprocessed data, both smoothed with a bandwidth of 1. **(d)** HiCNorm data vs. Unprocessed data, both smoothed with a bandwidth of 5. **(e)** Data processed using BNBC, smoothed with a bandwidth of 1 and 5.

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



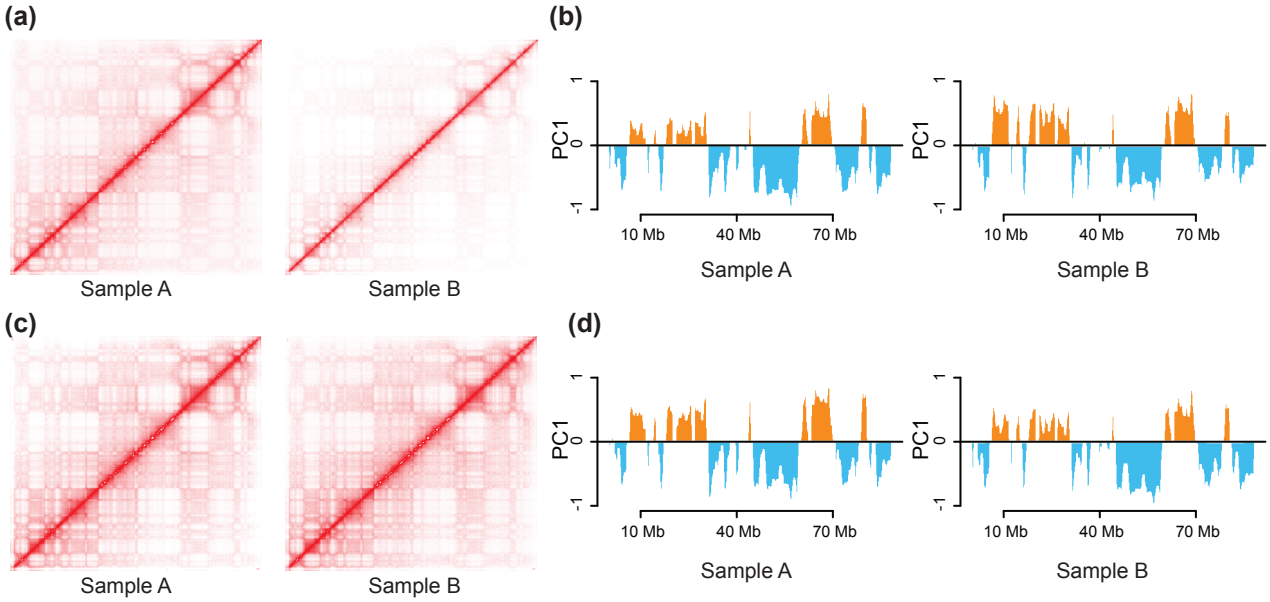
Supplementary Figure S4. The performance of BNBC by distance. We show a comparison between R^2 for data processed using HiCNorm and BNBC. **(a)** Loci separated by 10 Mb or less (bands 2-251) (Figure 2e reproduced). **(b)** Loci separated by 0-2Mb (bands 2-51). **(c)** Loci separated by 2-6Mb (bands 52-151). **(d)** Loci separated by 6-10Mb (bands 152-251).

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



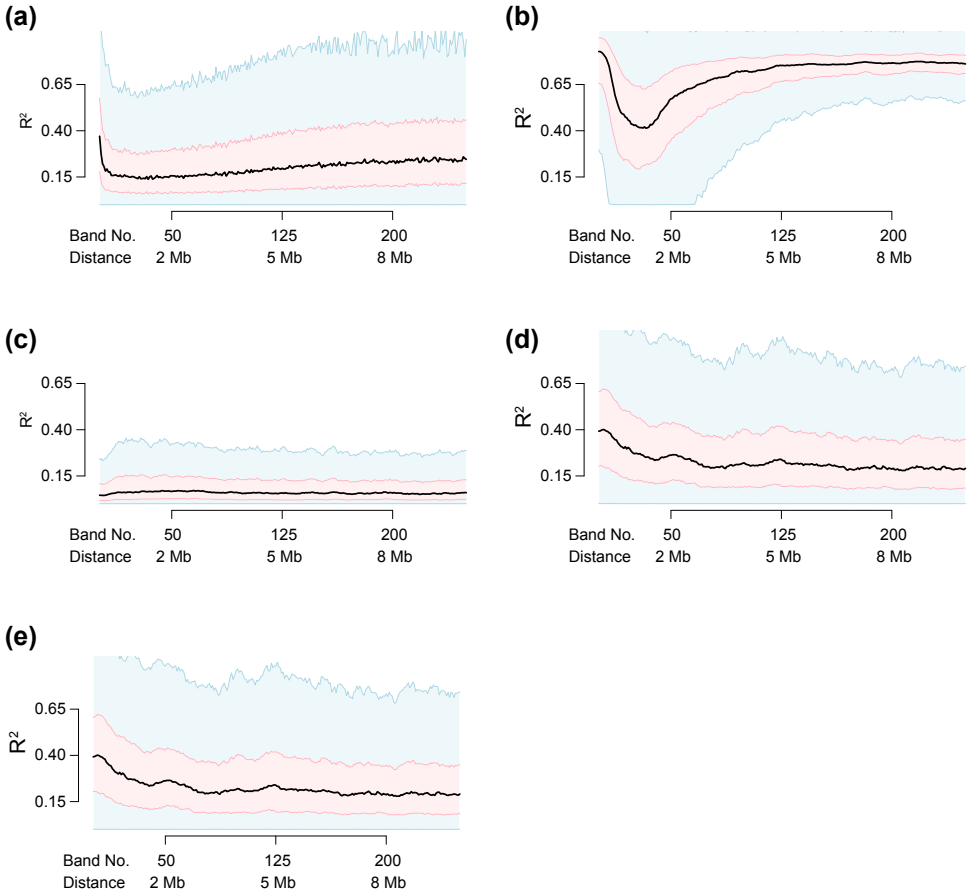
Supplementary Figure S5. Marginal distributions after BCBN. As Figure 1 but for data processed using BNBC. **(a)** Mean contact as a function of distance. Each sample is a separate curve. **(b)-(d)** Boxplots of the marginal distribution of contacts across samples, for loci separated by **(b)** 40 kb (band 2), **(c)** 2 Mb (band 50) and **(d)** 8 Mb (band 200).

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure S6. BNBC preserves structural features of Hi-C contact maps. Data from two different biological replicates (samples A, B) on chromosome 14. **(a)** Contact maps for data processed using HiCNorm. **(b)** First eigenvector of the contact maps in (a); this is used to estimate A/B compartments. **(c)** Contact maps for data processed using BNBC. **(b)** First eigenvector of the contact maps in (c).

bioRxiv preprint doi: https://doi.org/10.1101/214361; this version posted November 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure S7. The performance assessment of all methods using R^2. As in Figure 2a we assess the influence of batch using the percent variation explained by the batch factor (R^2), as a function of distance, for various methods. **(a)** Unprocessed data (Supplementary Figure S2e reproduced). **(b)** HiCNorm (Figure 2a reproduced). **(c)** BNBC (Figure 2c reproduced). **(d)** BNBC using PEER with 1 hidden factor. **(e)** BNBC using PEER with 4 hidden factors.