nature COMMUNICATIONS

ARTICLE

Received 17 Sep 2015 | Accepted 11 Mar 2016 | Published 13 Apr 2016

DOI: 10.1038/ncomms11305

OPEN

High-dimensional genomic data bias correction and data integration using MANCIE

Chongzhi Zang[1,2,*], Tao Wang[3,4,*], Ke Deng[5], Bo Li[1,2,6], Sheng'en Hu[7], Qian Qin[7], Tengfei Xiao[1,2,8], Shihua Zhang[9], Clifford A. Meyer[1,2], Housheng Hansen He[1,2,8,10], Myles Brown[2,8], Jun S. Liu[6], Yang Xie[3,11,12] & X. Shirley Liu[1,2]

High-dimensional genomic data analysis is challenging due to noises and biases in high-throughput experiments. We present a computational method matrix analysis and normalization by concordant information enhancement (MANCIE) for bias correction and data integration of distinct genomic profiles on the same samples. MANCIE uses a Bayesian-supported principal component analysis-based approach to adjust the data so as to achieve better consistency between sample-wise distances in the different profiles. MANCIE can improve tissue-specific clustering in ENCODE data, prognostic prediction in Molecular Taxonomy of Breast Cancer International Consortium and The Cancer Genome Atlas data, copy number and expression agreement in Cancer Cell Line Encyclopedia data, and has broad applications in cross-platform, high-dimensional data integration.

[1] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02215, USA. [2] Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [3] Quantitative Biomedical Research Center, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [4] Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [5] Center for Statistical Science, Tsinghua University, Beijing 100084, China. [6] Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA. [7] Department of Bioinformatics, School of Life Sciences, Tongji University, Shanghai 200092, China. [8] Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts 02215, USA. [9] National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. [10] Department of Medical Biophysics, University of Toronto, Toronto, Ontatio M5G 1L7, Canada. [11] Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [12] Simons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Y.X. (email: yang.xie@utsouthwestern.edu) or to X.S.L. (email: xsliu@jimmy.harvard.edu).

High-throughput genomic technologies have made it possible to generate massive data for studying biological mechanisms or disease aetiology. Such high-dimensional genomic data usually can be presented as a matrix, with each column representing a sample (for example, a patient, a cell type, an experimental condition and so on), and each row representing a genomic feature (for example, a gene, a genomic locus and so on). By computational analyses of these high-dimensional data matrices using dimension reduction (for example, principal component analysis, PCA) or clustering approaches, one can learn characteristic information within samples and identify key features between samples to interrogate biological functions. In many cases, there can be multiple platforms of experiments on the same set of samples and they can generate more than one data matrices. For example, the ENCODE (Encyclopedia of DNA Elements) Consortium generated high-throughput data including ChIP-seq, DNase-seq, and exon array transcriptomes and so on. on a designated panel of human cell lines[1]; The Cancer Genome Atlas (TCGA) program[2] and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)[3] generated mutation and gene-expression profiles of patient tumours; and the Cancer Cell Line Encyclopedia (CCLE) project[4] provided copy number, gene expression for over a thousand cancer cell lines. Integrative analysis is critical for obtaining biological insights from these data sets, within which a common challenge exists in identifying and correcting hidden biases in such high-dimensional data matrices.

In high-throughput data with different experimental platforms, it is not uncommon for a subset of samples in a data matrix on one experimental platform to have technical biases[5,6]. For example, in a cohort of dozens of samples, the expression and ChIP-seq profiling were conducted under various batches, each with unique biases from sample collection and preparation, array hybridization, sequencing GC content[7] or coverage differences that are challenging to identify and remove. There have been methods developed to remove batch effect within one data matrix of the same platform. For example, PCA have been used to solve such problems. As an extension of PCA, Sparse PCA[5] uses the linear combination of a small subset of variables instead of all to generate the principal components and still explains most variances present in the data, while making the dimension reduction and bias removal clearer and easier to interpret[8]. Surrogate variable analysis (SVA)[9] models the gene-expression heterogeneity bias as 'surrogate variables' and separate them from primary variables that capture biologically meaningful information. These methods aim to normalize data within the same data matrix from the same platform. However, to our knowledge, methods that can normalize data from different matrices and borrow information between different platforms are still lacking.

Recently, Wang *et al.*[10] propose similar network fusion (SNF), a method that first generates sample networks from each data platform separately, then uses network fusion to merge the platform-specific networks together with confidence weighting. SNF demonstrated good performance on separating TCGA glioblastoma samples into subtypes using transcriptome and DNA methylome profiles. However, SNF does not provide the normalized data matrices that could be useful in the downstream analysis. In addition, SNF is based on network construction, which could be sensitive to strong biases in a subset of samples that result in 'high-weight' edges in the network and are difficult to remove in the fusion step. In other words, if the networks generated from each data matrices were too dissimilar, it is difficult to 'fuse'. A more general applicable method is needed to simultaneously provide better sample clustering and generate normalized data matrices.

To overcome the above challenges, we propose MANCIE (matrix analysis and normalization by concordant information enhancement), an integrative computational method that can conduct data normalization and bias correction by borrowing information from a column-matched associated data matrix. Applied to ENCODE, METABRIC, TCGA and CCLE data, MANCIE showed effectiveness in improved identification of biologically meaningful patterns.

Results

Method overview. MANCIE takes in two data matrices and adjusts one (thereafter defined as the 'main matrix') using the other (thereafter defined as the 'associated matrix') by identifying and maintaining the concordant information and reducing the discordant information between them. The two data matrices contain profiles on the same set of samples generated using different experimental platforms (for example, copy number variation (CNV) and RNA-seq on the same collection of tumours), or generated independently (for example, expression profiles measured at different institutions on the same collection of cell lines). If the rows of the two matrices are unmatched (for example, genes versus ChIP-seq peaks), MANCIE first generates a summarized associated matrix that has matched rows with the main matrix using a biologically motivated matching process (Supplementary Fig. 1, see Methods for details). This matching step requires additional biological information to connect the rows between the two matrices, for example, each gene (as a row vector in the main matrix) will corresponds to a row vector summarized from a few nearby transcription factor (TF) -binding sites (as a few rows in the associated matrix). MANCIE assumes that pairwise sample distance as measured by different platforms should be similar, and discordance in the pairwise distance largely arise from technical biases and/or noises. Therefore, the second and key step of MANCIE adjusts the main matrix row by row by borrowing information from data in the associated matrix. In each row, depending on whether the correlation between the main and associated data is high, moderate or low, MANCIE takes the first principle component (scenario 3), a correlation-weighted sum (scenario 2) or the original data (scenario 1) as the adjusted data, respectively (Fig. 1, see Methods for details). The correlation cutoffs are determined empirically, by making roughly 1/3 of the rows be adjusted under scenario 3. Finally, the output of MANCIE is the normalized adjusted matrix that has the same dimension as the main matrix yet with the information from the associated matrix incorporated. It is worth noting that one can swap the main and associated matrices, so the quality of both data can be improved from each other. We show that this approach is an appropriate approximation of the full Bayesian inference for reducing noises from such data sets (Supplementary Notes). In the following sections, we applied MANCIE on a few data sets generated by large consortia including ENCODE[1], METABRIC[3], TCGA[2] and CCLE[4] to demonstrate its utility in genomic data integration.

ENCODE data. From ENCODE consortium[1], we obtained data from 61 cell lines where both DNase-seq data for chromatin accessibility profiling and Affymetrix exon array data for gene-expression profiling are available. These cell lines can be classified into seven groups by their tissues of origin (Fig. 2). DNase hypersensitive sites (DHS, measured by DNase-seq) mark open chromatin regions that can be considered as a repertoire of all putative *cis*-regulatory elements in the genome[11,12] that regulate gene expression as measured by exon arrays. Although active gene promoters are usually DHS, most DHS are located in introns or intergenic regions, marking distal enhancers[1,13] which are

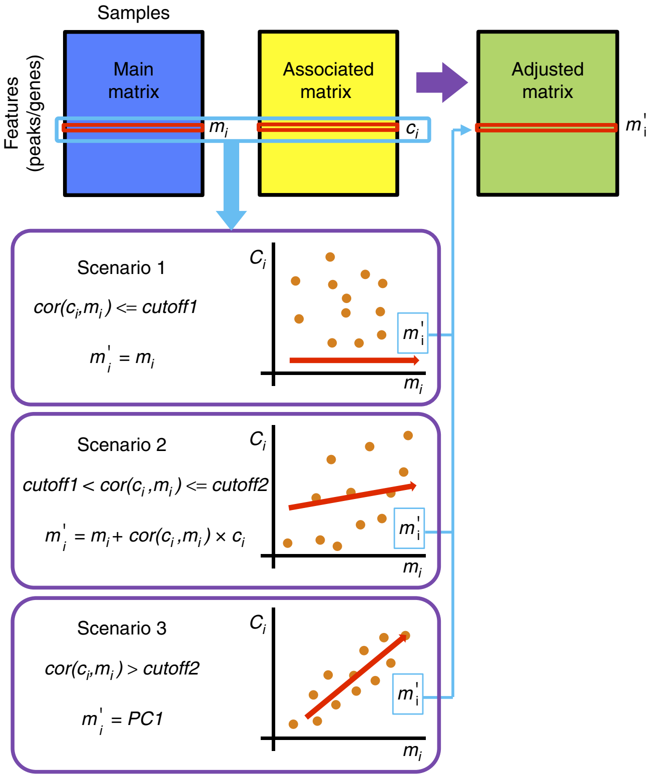

Figure 1 | Overview of MANCIE. Each row vector in the adjusted matrix is generated from the corresponding row vectors in the main matrix and the associated matrix. On the basis of the correlation between the main row vector m_i and the associated row vector c_i, one of three scenarios will be chosen. See more details in the online methods.

often more dynamic across different cell types or conditions[14–17]. Focusing on enhancers, we generated the DHS data matrix based on a union set of intronic and intergenic DHS peaks identified from all the DNase-seq data, and obtained the gene-expression data matrix based on the exon array data. We used MANCIE to adjust each data matrix using the other as the associated matrix. Since the rows between the two data matrices are not matched, we generated the summarized associated matrices using the genomic location information of genes (based on the transcription start site, TSS) and DHS (based on the DHS peak centre). For summarization of DHS data around genes, we used up to 50 nearby DHS located within 100 kb from the TSS of each gene as the local sub-matrix for that gene. For summarization of gene-expression data around DHS, we used a similar approach but up to 20 nearby genes, considering that there are much fewer genes than DHS. The adjusted main matrix was then generated by integrating the main matrix with the summarized associated matrix. We conducted multi-dimensional scaling on the adjusted data as well as the raw data, and plotted two principal components with each data point representing a cell line (Fig. 2a,b). We hypothesized that cell lines belonging to the same tissue type should be more similar to each other than cell lines from different tissue types, a pattern previously reported for these data types[16]. Indeed, although MANCIE only aims to make cell lines similar in one platform more similar in the other platform, the end result is that the adjusted data, both DHS and expression data, show better cell line clustering according to their tissue types (Fig. 2a,b). To assess the better clustering quantitatively, we performed K-means clustering with randomly sampled seeds for 1,000 times on each data set and calculated the adjusted Rand index[18] that measures the similarity between the K-means clustering and the actual tissue-type clustering for each random sample. The average adjusted Rand index is significantly higher for MANCIE-adjusted data than that for the raw data, indicating that MANCIE improves the tissue-type clustering (Fig. 2c,d). In contrast, cell line clustering using SVA-adjusted data is actually worse than using the raw data (Supplementary Fig. 2a,b).

We next investigated the implication of the MANCIE adjustment on the ENCODE data. As GC-content bias is one major sources of biases in next-generation sequencing data, we first checked whether MANCIE can reduce the GC-content biases in the DNase-seq data. For each cell line, we calculated the distribution of the GC-content of all sequence reads in the DNase-seq data set as well as the magnitude of MANCIE adjustment, measured by the Euclidean distance between the corresponding column vectors in the raw and the MANCIE-adjusted data matrices. Cell lines showing GC-content patterns that were farther away from average of all cell lines (Supplementary Fig. 2c) underwent a greater magnitude of MANCIE adjustment than the other cell lines (Fig. 2e and Supplementary Fig. 2d). This result indicates that MANCIE successfully corrected the GC-content biases in the DNase-seq data.

To further evaluate MANCIE performance in adjusting the DNase-seq data, we selected the top 2,000 DHSs with greatest increase after MANCIE adjustment in the cell lines with the biggest adjustment, and performed sequence motif analysis on these DHSs. We found that the sequence motifs enriched in these DHSs usually match cell-type-specific TFs (Fig. 2e). For example, ETS motif is enriched in both TH1 and TH2 cell lines, and the ETS-family TFs ERM and PU.1 are specific to TH1 and TH2 cell lines, respectively[19,20]. The motif of megakaryocyte-specific TF NF-E2[17] is enriched in the megakaryocyte cell line CMK. These results demonstrated that integrated with gene-expression data, MANCIE-adjusted DHSs show an increased pattern of cell-type specificity that is better correlated with cell-type-specific gene-expression pattern. Taken together, MANCIE is able to integrate the genomic DHS data with gene expression and to reduce potential biases, as well as to emphasize biologically meaningful signals.

METABRIC and TCGA data. We applied MANCIE on the METABRIC breast cancer data sets[3] to demonstrate its effectiveness for noise reduction on another data platform. METABRIC has two independent cohorts of breast cancer patients. Each cohort has around 1,000 patients with gene-expression values, CNV values and survival information. We used the first cohort as the training set and the second cohort as the independent set to predict survival information from gene-expression data. MANCIE was applied to adjust the gene-expression data based on CNV data. The underlying assumptions are: first, if gene-expression signatures can predict patient survival outcome, noise-reduced and bias-corrected gene-expression data should have better predictive accuracy in the patient survival; second, genes with concordant correlation between copy number and expression are more likely to be reliably measured and are more informative for outcome prediction. Indeed, we found that the MANCIE-adjusted data can better predict survival information than the original expression data, by better distinguishing patients with lower and higher risk of death, with an example shown in Fig. 3a. To assess the improvement quantitatively, we compared the logrank *P* values obtained using original training and original testing expression matrices with the logrank *P* values obtained using adjusted training and adjusted testing expression matrices. We limited our analysis to a subset of genes whose adjusted expression values are most different from

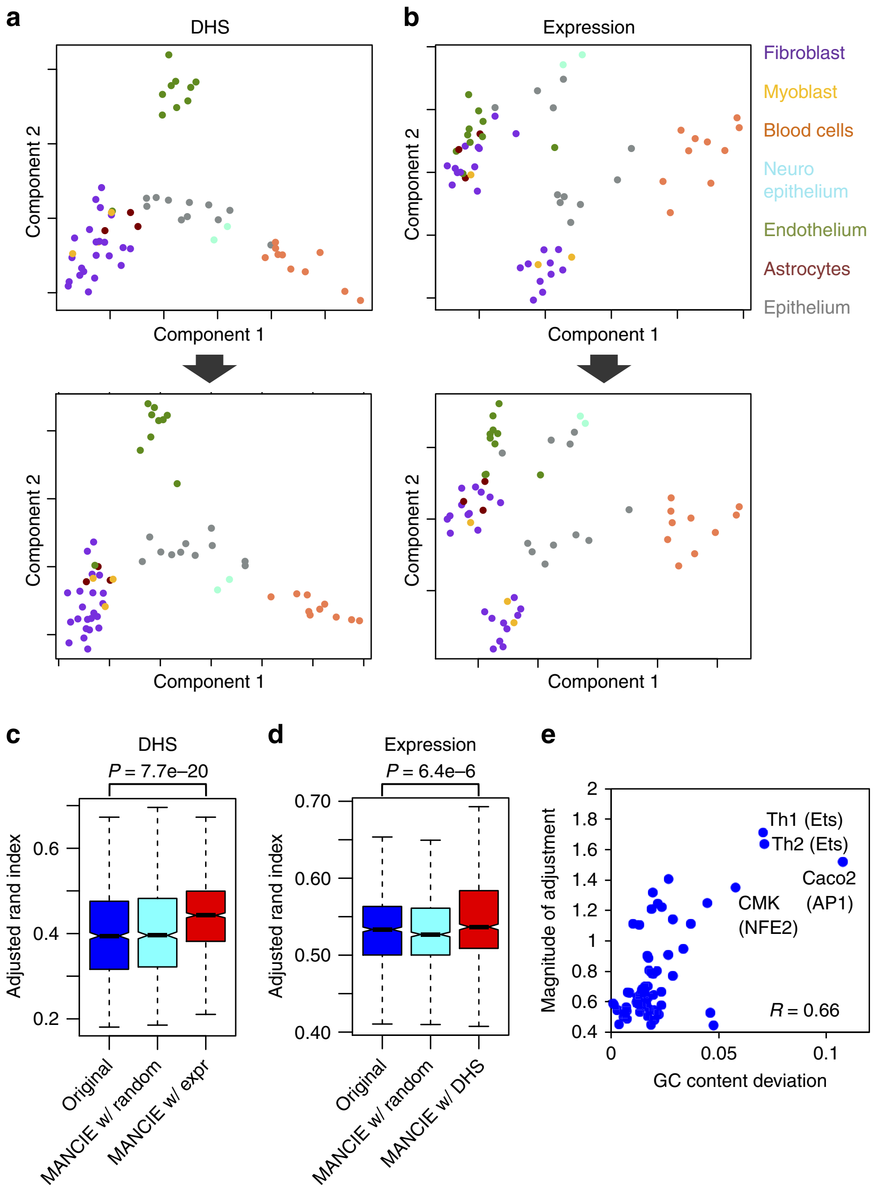

Figure 2 | Case study on ENCODE data. (**a**,**b**) Multi-dimensional scaling map representing genomic data from 61 cell lines. Each data point represents a cell line, with its tissue type labelled in the same colour as in the legend. (**a**, top) Raw DHS data; bottom, MANCIE-adjusted DHS data; (**b**, top) Raw expression data; bottom, MANCIE-adjusted expression data. (**c**,**d**) Adjusted Rand index comparing K-means clustering on the data with actual tissue-type clustering. K-means clustering was performed 1,000 times with random seeds. The three boxes represent original data (blue), MANCIE-adjusted with random data matrices (cyan) and MANCIE-adjusted with the other data type (red). (**c**) DHS data, (**d**) gene-expression data. *P* value was calculated using Wilcoxon rank sum test. (**e**) Relationship between the magnitude of MANCIE adjustment and the deviation of GC-content distribution of DNase-seq reads. The magnitude of MANCIE adjustment was calculated as the Euclidean distance between the sample data vectors before and after MANCIE adjustment. The deviation refers to the distance from each sample's data point to the centre of mass in the mean—coefficient of variation map of the GC-content distribution in Supplementary Fig 2c. Labels in the parentheses are the top sequence motif enriched in the most increased DHS in the corresponding cell line after MANCIE adjustment.

the original values, defined as correlation of the adjusted vector and original vector being smaller than a threshold for either the training or the testing data set. Under a series of threshold from 0.7–0.93, MANCIE consistently improved the prediction accuracy by generating smaller *P* values (Fig. 3b).

Although TCGA also has breast cancer profiles, the death events are too few to provide meaningful survival separation. Therefore, we applied MANCIE on TCGA lung adenocarcinoma data[2] for survival prediction. A total of 10,704 genes for 417 tumours with complete expression, CNV and clinical information were used, and MANCIE was applied to adjust the gene-expression data based on CNV data. For comparison, the gene-expression data matrix was also adjusted by the SVA method. To test the effectiveness of MANCIE adjustment, we selected six prognostic gene signatures for non-small cell lung cancer from previous publications[21–26] for survival prediction. The 417 tumours were sub-sampled for 1,000 times when each time 90% of these samples were randomly selected. For each sub-sample, supervised principal components for survival analysis (SuperPC)[27] was used to fit each gene signature and a continuous risk score is generated from the fitted model for the patients in the sub-sample. Then, similar to the METABRIC data

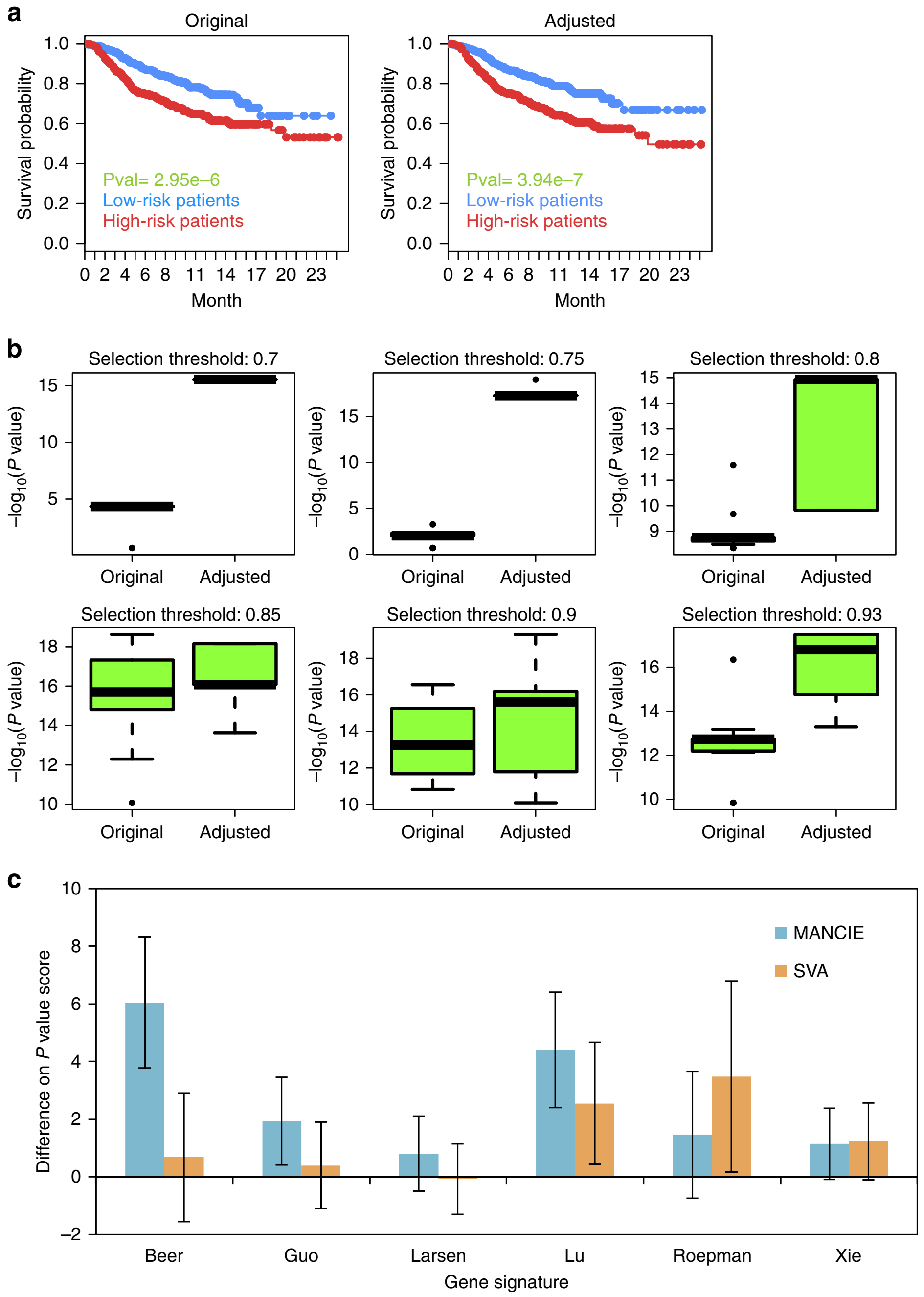

Figure 3 | Case studies on METABRIC and TCGA data. (**a**) The Kaplan–Meier plots for an example showing the dichotomized risk scores from the original matrices (left) and the adjusted matrices (right) under a correlation threshold of 0.93 using the METABRIC data. Patient samples were separated into two groups according to the predicted risk scores from the selected genes. High-risk group is labelled in red and low-risk group is labelled in blue. The high-risk group is better separated from the low-risk group by using the MANCIE-adjusted expression data (right), compared with using the original data (left). (**b**) *P* value scores ($-\log_{10}$Pvalue) in survival prediction using METABRIC gene-expression data comparing before or after MANCIE adjustment with CNV data. The gene selection thresholds are set as 0.7, 0.75, 0.8, 0.85, 0.9, 0.93, from left to right, from top to bottom, respectively. (**c**) Difference of *P* value scores ($-\log_{10}$Pvalue) in survival prediction with each gene signature using TCGA gene-expression data before or after adjustment by MANCIE or SVA. Gene signatures are labelled with the first author name of the publication. Error bar stands for s.d. of the results from 1,000 random samples.

analysis, a Cox proportional hazards model[28] was regressed on the risk score to test how well the trained risk score can explain the survival data, with smaller *P* values indicating better correlation between the risk score and the survival data (one example shown in Supplementary Fig. 3a). We then plotted the differences in negative log *P* values before and after either MANCIE or SVA adjustment for each gene signature over the 1,000 samplings (Fig. 3c and Supplementary Fig. 3b). MANCIE adjustment can improve the *P* values, indicating better association of the risk scores with survival data for all six gene signatures (Fig. 3c), and outperformed SVA in four out of the six gene signatures.

In addition, we showed the effect of different combinations of the two cutoff parameters on the performance of MANCIE. We used the differential log *P* values to evaluate the improvement of survival prediction of MANCIE-adjusted data over the original data. For both METABRIC (Supplementary Fig. 4a) and TCGA lung cancer data (Supplementary Fig. 4b), although MANCIE improvement varies with the cutoffs, overall it is robust and has good performance under the default (0, 0.5) parameters.

CCLE/GDSC data. We next demonstrated that MANCIE can reduce noise in a data set by leveraging information from a different source of the same data type. Two recent drug screen studies, CCLE[4] and Genomics of Drug Sensitivity in Cancer (GDSC) project[29] led by the Cancer Genome Project-profiled cancer cell line response to different drugs as well as the expression and CNV of each cell line. The two projects shared a total of 425 common cell lines and 10,461 genes measured. We adjusted the gene-expression data from CCLE project as the main matrix using the associated data matrix from the GDSC project, and compared the performance between MANCIE and SVA.

To evaluate the performance of MANCIE and SVA, the correlation between the expression level and CNV were examined for each gene using expression data with and without adjustment. The underlying assumption is that better noise reduction should give stronger correlation between expression and CNV as a general trend, given the fact that the expression of many genes is partially driven by its CNV. The difference in Spearman correlation after MANCIE adjustment is calculated for each gene (example of *NDUFC2* gene in Fig. 4a,b) and it is positive for 63% of the genes (Fig. 4c), compared with the 55% improved correlation with SVA (Fig. 4d). These results show that MANCIE is able to reduce noise in the expression data by using independent data from a different lab and to improve the overall data consistency.

Discussion

We present a general computational method for high-dimensional genomic data integration. By case studies using available large-scale genomics data cohorts from several consortium efforts, we demonstrated that MANCIE can reduce potential biases and noises without identifying them in prior and successfully derive biologically meaningful information from data integration. As a data-driven method, MANCIE uses the basic principles from PCA by identifying the largest concordant differences across samples. MANCIE has few parameters, therefore making it more generally applicable to different data types. To our knowledge, MANCIE is the best-performing computational method for bias correction and integration of two high-dimensional genomic data matrices. In general, MANCIE can be applied to integrate any two data matrices with matched columns, For example, it can integrate genetic mutation profiles with gene-expression profiles, RNA-level expression with protein expression, or chromatin profiles generated using similar but different techniques, such as DNase-seq and ATAC-seq and so on.

Like every computational method for high-dimensional data analysis, MANCIE has its specific scope and limitation. It holds a strong assumption that the concordant information between the two data matrices is biologically meaningful. However, if the biases or batch effects exist in both the main and associated matrices in a concordant manner, MANCIE cannot identify them, and might even enhance the consistent bias with detrimental effects. In addition, MANCIE requires additional information, such as physical distance information, for the associated matrix summarization step. It assumes that, for one row in the main matrix, the principal direction of the related features in the associated matrix should contain concordance

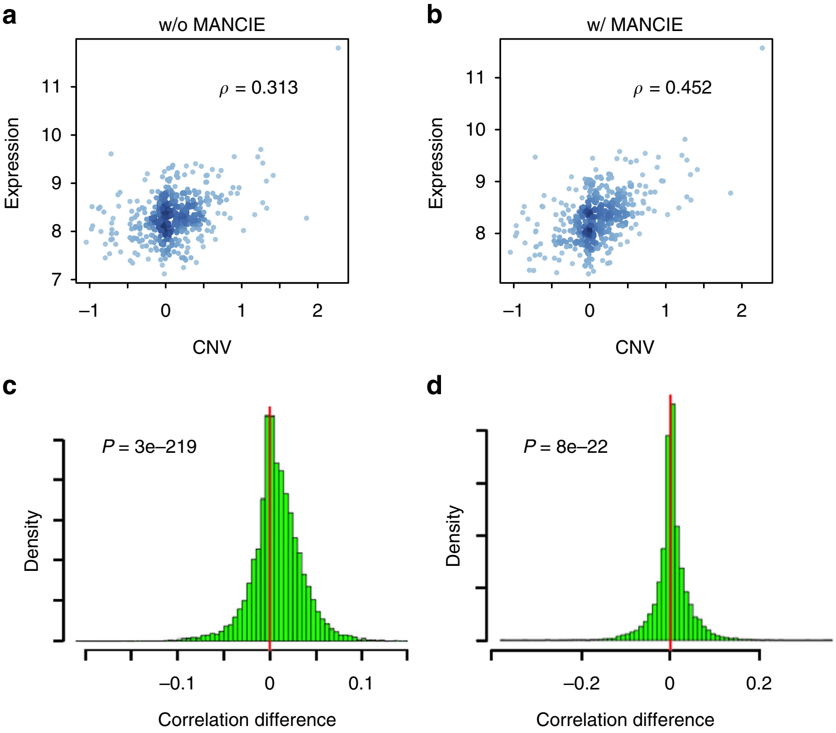

Figure 4 | Case Study on CCLE/GDSC data. (**a**,**b**) Correlation between the CNV and RNA expression for gene *NDUFC2*. The expression data were using raw CCLE data (**a**) or MANCIE-adjusted with GDSC data (**b**). ρ refers to Spearman correlation coefficient. (**c**) Distribution of the correlation difference comparing before and after MANCIE adjustment, for all genes. The Spearman correlation coefficient between CNV and RNA expression was calculated for each gene, and the correlation difference is calculated by subtracting with MANCIE adjustment by without MANCIE adjustment. *P* value was calculated using the one-tail paired *t*-test. (**d**) Distribution of the correlation difference comparing comparing the raw data with SVA-adjusted expression data. *P* value was calculated using the one-tail paired *t*-test.

information that can be borrowed. For computational convenience, we usually only consider the local effect, that is, only include nearby associated data rows in the summarization of the associated matrix into a row that matches the main data row. There are situations when this might be insufficient, for example, enhancer regulation at distal loci could be ignored in the summarization step of MANCIE. Furthermore, MANCIE carries out the adjustment through each corresponding row vectors independently. It does not use the information across rows, which could be valuable and helpful in more sophisticated statistical inference to be explored in the near future. Last but not least, MANCIE prefers many columns in the data matrices, because more measured samples could potentially increase the power of bias modelling and correction.

We tested different combinations of cutoff values in the METABRIC and TCGA data. In these cases, previous prognostic signatures were used to evaluate and guide the choice of parameters, but this type of orthogonal information is not always available to optimize parameter selection. According to our experience with the example data sets in this study and for simplicity's sake, we set the default cutoff1 to be 0. The underlying assumption is that only positive correlations between the main row vector and the corresponding associated row vector are biologically meaningful and can be integrated. Cutoff2 can be empirically set so that about 1/3 of the rows are adjusted under the third scenario, which has worked robustly in all the example data sets in our study.

Taking together, MANCIE can be used as a general method for high-dimensional data integration with the potential to reduce noise and correct biases. It serves as a proof of principle that bias can be removed by such a methodology and more advanced statistical algorithm may be developed. For example, MANCIE can be extended to include multiple associated matrices for correction of bias in the one main matrix. It provides a novel approach to genomic data integration, an area that will become increasingly important for mining big biological data in the near future.

Methods

MANCIE algorithm. *Data description*. The main matrix is denoted as $M=\begin{bmatrix} m_1 \\ m_2 \\ \cdots \\ m_{n_1} \end{bmatrix}$.

The main matrix can be any high-dimensional genomic data such as gene expression or peak count data across many different samples/conditions (listed on columns). The main matrix can come from high-throughput experiments such as RNA-Seq, microarray, ChIP-Seq, DNase-Seq, ATAC-seq and so on. Each row vector $m_i, i=1, 2, \ldots n_1$ represents a feature such as one gene or one peak.

Similarly, an associated matrix is defined as $C=\begin{bmatrix} c_1 \\ c_2 \\ \cdots \\ c_{n_2} \end{bmatrix}$, where each row vector $c_i, i=1, 2, \ldots n_2$ represents a feature such as one gene or one peak. The associated matrix C could come from the same or different type of experiment as M, but must be conducted in exactly the same sample conditions as M. In other word, M and C must have matched columns. $n_1=n_2$ is satisfied when M and C are from the same experiment and may not be satisfied when M and C are from different experiments. In the latter case, one matrix could be DNAase-Seq and the other matrix could be RNA-Seq. For such scenarios, the annotation data relating both the main and associated matrix (for example, chromosome, start position of feature and ending position of feature) must be available.

Summarization of information in the associated matrix. This step is skipped when each row vector in C has a one-to-one matching row vector in M, which measures different features of the same entity. For example, when a gene's expression is measured in M and the same gene's CNV data is measured in C, the M and C matrices satisfy this condition. Otherwise, the C data need to be summarized to this format that is compatible with the M matrix in both dimensionality and biological meaning.

On the basis of the connection information, for each row in the main matrix, no more than p rows in the associated matrix associated with the row in the main matrix were selected and the first principal component of this sub-matrix is calculated as the summarized row in the summarized associated matrix (Supplementary Fig. 1).

One summarization function that works in the majority of all cases is described here. For each row vector in M, m_i, the nearest features defined by physical genomic distances calculated in the annotation data in the C matrix are extracted. The number of nearest features to extract can be set flexibly by users. Let $C_i=\begin{bmatrix} c_{i1} \\ c_{i2} \\ \cdots \\ c_{iS_i} \end{bmatrix}$ be the subset of C that are extracted whose features are closest to m_i. If not a single feature in C are found to be close to m_i in the main matrix, this main matrix row will be skipped for adjustment. Then the first principal component of C_i^T by PCA on correlation matrix is calculated as a row vectorc_i'. In the case of Pearson correlation $cor\left(c_i', m_i\right)$ smaller than 0, all elements in c_i' are converted to their additive inverse. These summarized vectors form a new matrix $C'=\begin{bmatrix} c_1' \\ c_2' \\ \cdots \\ c_{n_1}' \end{bmatrix}$.

The new matrix C' is compatible in dimension with M and each row represents summarized information carried by C that might be leveraged to remove the noise in the corresponding row in M.

Removing noise in M matrix by combination with C. With slight abuse of terms, if step (2) is carried out, still let C denote the new matrix C'. Now for each row m_i, a three-way strategy is employed to remove noise by borrowing information from the associated matrix C.

(a) If $cor(c_i, m_i) < = cutoff1$, the new row vector $m_i'=m_i$.

(b) If $cor(c_i, m_i) > cutoff1$ and $cor(c_i, m_i) < = cutoff2$, the new row vector $m_i'=scale(m_i)+cor(c_i, m_i)\times scale(c_i)$, where $cor(c_i, m_i)$ is the Pearson correlation and $scale(x)$ is a function that scales a row vector to an s.d. of 1.

(c) If $cor(c_i, m_i) > cutoff2$, the new row vector m_i' is the first principal component of the PCA of $\begin{bmatrix} m_i \\ c_i \end{bmatrix}^T$ on the 2×2 correlation matrix. In the case of Pearson correlation $cor\left(m_i', m_i\right)$ smaller than 0, all elements in m_i' are converted to their additive inverse.

The two cutoffs *cutoff*1 and *cutoff*2 can be set flexibly by the users, with default values of 0 and 0.5. In case summarization is involved and if the first cutoff is set to be equal or less than 0, the first option will not be used. Finally, m_i' calculated by either of the three options is scaled to have the mean and s.d. of the old row vector m_i. The adjusted main matrix is then $M'=\begin{bmatrix} m_1' \\ m_2' \\ m_{n_1}' \end{bmatrix}$.

ENCODE data analysis. DNase-Seq data sets from each individual cell line was downloaded from ENCODE[1]. DHS peaks were identified in each cell line using MACS2[30] with default parameters and only peaks whose summits are in introns or intergenic regions and have a fold enrichment of at least four were retained. Peaks wider than 150 bp were chopped to 150 bp centred at the summit location. Then overlapping DHS peaks were merged into a new region. The DHS level on each merged peak in each cell line were measured as RPKM. Exon array data were processed using JETTA[31]. The sequence motif scan analyses were performed using the MDSeqPos algorithm[32] on the Cistrome analysis pipeline[33].

Survival analysis for METABRIC data set. We adjusted both the training set expression matrix and the testing set expression matrix with the corresponding CNV data matrix using default parameters. Then we calculate the Pearson correlation of each row vector in the original training set expression matrix and the row vector in the adjusted training set expression matrix. We also calculated the Pearson correlation for the testing set matrix before and after adjustment. We focus our analysis on the row vectors (gene expression across patients) whose Pearson correlations are below a certain threshold for either the training set or the testing set. To be objective, we chose a series of threshold values in the downstream analysis. Then we used LASSO[34] with cox family to analyse selected genes from the previous step to further narrow down the genes whose expression levels are significantly correlated with survival in the training set. Then we fit a multivariate Coxph model[35] using these selected genes and predicted to the testing set after which Coxph model returned a risk score vector for the testing set patients. We dichotomized the risk score based on median risk and tested the Logrank P values of the overall survival difference between the low-risk group and high-risk group. Since the LASSO method is stochastic, generating slightly different results especially when the number of input features is large, we ran the same analysis 20 times and obtained a distribution of logrank P values.

Data analysis by SVA. For TCGA data, we used cancer stage and smoking status as the primary variables, no variable of known noise source and 1 surrogate variable to be estimated. For CCLE/GDSC data, we used no variable of known noise source and 1 surrogate variable to be estimated. For Encode data, we used tissue type as known noise source and 1 surrogate variable to be estimated. Each row vector is regressed on the estimated surrogate variable and replaced by the regression residuals.

Choice of parameters. For METABRIC and TCGA data studied in this paper, Cutoff1 was set as 0 and the Cutoff2 being 0.5. For CCLE data, the two parameters are 0 and 0.7, respectively. For Encode data, the two parameters are 0 and 0.5, respectively. We used the differences in negative log rank *P* values as a metric to evaluate the effect of different combinations of higher cutoff and lower cutoff on the efficiency of MANCIE for the METABRIC data set.

Availability. MANCIE is available as an R-package (http://cran.r-project.org/web/packages/MANCIE/).

References

1. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **488,** 75–82 (2013).
2. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511,** 543–550 (2014).
3. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (2012).
4. Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–307 (2013).
5. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11,** 733–739 (2010).
6. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* **15,** 1–13 (2014).
7. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic. Acids. Res.* **36,** e105–e105 (2008).
8. Zou, H., HASTIE, T. & Tibshirani, R. Sparse principal component analysis. *J. Comput. Graph. Stat.* **15,** 265–286 (2006).
9. Leek, J. T. & Storey, J. D. capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3,** e161–12 (2007).
10. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11,** 333–337 (2014).
11. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57,** 159–197 (1988).
12. Felsenfeld, G. & Groudine, M. Controlling the double helix. *Nature* **421,** 448–453 (2003).
13. Sabo, P. J. *et al.* Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nat. Methods* **3,** 511–518 (2006).
14. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39,** 311–318 (2007).
15. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
16. Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA Landscapes. *Cell* **154,** 888–903 (2013).
17. Luyten, A., Zang, C., Liu, X. S. & Shivdasani, R. A. Active enhancers are delineated de novo during hematopoiesis, with limited lineage fidelity among specified primary blood cells. *Genes Dev.* **28,** 1827–1839 (2014).
18. Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66,** 846–850 (1971).
19. Ouyang, W. *et al.* The Ets transcription factor ERM is Th1-specific and induced by IL-12 through a Stat4-dependent pathway. *Proc. Natl Acad. Sci. USA* **96,** 3888–3893 (1999).
20. Chang, H.-C. *et al.* The transcription factor PU.1 is required for the development of IL-9-producing T cells and allergic inflammation. *Nat. Immunol.* **11,** 527–534 (2010).
21. Beer, D. G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* **8,** 816–824 (2002).
22. Guo, L. Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. *Clin. Cancer Res.* **12,** 3344–3354 (2006).
23. Larsen, J. E. *et al.* Gene expression signature predicts recurrence in lung adenocarcinoma. *Clin. Cancer Res.* **13,** 2946–2954 (2007).
24. Roepman, P. *et al.* An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. *Clin. Cancer Res.* **15,** 284–290 (2009).
25. Xie, Y. *et al.* Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. *Clin. Cancer Res.* **17,** 5705–5714 (2011).
26. Lu, Y. *et al.* A gene expression signature predicts survival of patients with stage i non-small cell lung cancer. *PLoS Med.* **3,** 2229–2243 (2006).
27. Bair, E. & Tibshirani, R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.* **2,** e108 (2004).
28. Andersen, P. K. & Gill, R. D. Cox's regression model for counting processes: a large sample study. *Ann. Stat.* **10,** 1100–1120 (1982).
29. Yang, W. *et al.* Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic. Acids Res.* **41,** D955–D961 (2012).
30. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137–R139 (2008).
31. Seok, J., Xu, W., Gao, H., Davis, R. W. & Xiao, W. JETTA: junction and exon toolkits for transcriptome analysis. *Bioinformatics* **28,** 1274–1275 (2012).
32. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132,** 958–970 (2008).
33. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12,** R83 (2011).
34. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58,** 267–288 (1996).
35. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **34,** 187–220 (1972).

Acknowledgements

This work was partially supported by the US National Institutes of Health (NIH) grants U41HG7000 (X.S.L.), 1R01GM099409 (X.S.L.), 5R01CA172211 (Y.X.), 1R01CA152301 (Y.X.), Leukemia and Lymphoma Society (LLS) fellow award (C.Z.), and the National Natural Science Foundation of China (NSFC) grant 11401338 (K.D.) and 61422309 (S.Z.).

Author contributions

C.Z., T.W., C.A.M. and X.S.L. conceived the idea. C.Z. and T.W. developed the method with the help of K.D. and J.S.L. on the theoretical support. C.Z. and T.W. analysed the data with the help and inputs of B.L., S.H., Q.Q., T.X., S.Z., H.H.H. and M.B. C.Z., T.W., K.D., and X.S.L. wrote the paper. X.S.L. and Y.X. oversaw the research. All authors read and approved the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Zang, C. *et al.* High-dimensional genomic data bias correction and data integration using MANCIE. *Nat. Commun.* 7:11305 doi: 10.1038/ncomms11305 (2016).