Bioinformatics, 37, 2021, i327–i333
doi: 10.1093/bioinformatics/btab306
ISMB/ECCB 2021

OXFORD

Long reads capture simultaneous enhancer–promoter methylation status for cell-type deconvolution

Sapir Margalit[1,2], Yotam Abramson[1,2], Hila Sharim[1,2], Zohar Manber[2,3], Surajit Bhattacharya[4], Yi-Wen Chen[4,5], Eric Vilain[4,5], Hayk Barseghyan[4,5], Ran Elkon[2,3], Roded Sharan[2,6,*] and Yuval Ebenstein[1,2*]

[1]Department of Physical Chemistry, Tel Aviv University, Tel Aviv 6997801, Israel, [2]Edmond J. Safra Center for Bioinformatics, Tel Aviv University, Tel Aviv 6997801, Israel, [3]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 6997801, Israel, [4]Center for Genetic Medicine Research, Children's National Hospital, Washington, DC 20010, USA, [5]Department of Genomics and Precision Medicine, George Washington University, Washington, DC 20052, USA and [6]School of Computer Science, Tel-Aviv University, Tel-Aviv 6997801, Israel

*To whom correspondence should be addressed.

Abstract

Motivation: While promoter methylation is associated with reinforcing fundamental tissue identities, the methylation status of distant enhancers was shown by genome-wide association studies to be a powerful determinant of cell-state and cancer. With recent availability of long reads that report on the methylation status of enhancer–promoter pairs on the same molecule, we hypothesized that probing these pairs on the single-molecule level may serve the basis for detection of rare cancerous transformations in a given cell population. We explore various analysis approaches for deconvolving cell-type mixtures based on their genome-wide enhancer–promoter methylation profiles.

Results: To evaluate our hypothesis we examine long-read optical methylome data for the GM12878 cell line and myoblast cell lines from two donors. We identified over 100 000 enhancer–promoter pairs that co-exist on at least 30 individual DNA molecules. We developed a detailed methodology for mixture deconvolution and applied it to estimate the proportional cell compositions in synthetic mixtures. Analysis of promoter methylation, as well as enhancer–promoter pairwise methylation, resulted in very accurate estimates. In addition, we show that pairwise methylation analysis can be generalized from deconvolving different cell types to subtle scenarios where one wishes to resolve different cell populations of the same cell-type.

Availability and implementation: The code used in this work to analyze single-molecule Bionano Genomics optical maps is available via the GitHub repository https://github.com/ebensteinLab/Single_molecule_methylation_in_EP.

Contact: uv@post.tau.ac.il or roded@tauex.tau.ac.il

1 Introduction

The accumulation of high-throughput genome-wide methylation data has enabled the analysis of human methylomes across distinct populations and medical cohorts via epigenome-wide association studies (EWAS) (Gorenjak *et al.*, 2020; Küpers *et al.*, 2019; Chu *et al.*, 2017). Such analyses have shown that predisposition to common human disease is frequently associated with specific methylation signatures in distal control regions also known as gene enhancers (Li *et al.*, 2013). While the contribution of DNA methylation in gene promoters to variation in intertumor gene expression was found to be low, enhancer methylation provides a much higher level of contribution to tumor heterogeneity and may further illuminate the mechanism of cancer predisposition (Li *et al.*, 2013). Furthermore, changes in DNA methylation patterns have been shown to correlate with early carcinogenesis, even prior to tumor formation, as well as with metastasis and response to therapy (Hentze *et al.*, 2019; Kurkjian *et al.*, 2008; Vrba and Futscher, 2019). Aran *et. al.* have shown that enhancer methylation is drastically altered in cancers and is closely related to altered expression profiles of cancer genes (Aran *et al.*, 2013; Aran and Hellman, 2013a,b). Hansen *et.al.* have shown that regions which are differentially methylated between cancer and normal tissue are more prone to variability in methylation levels, suggesting that stochastic epigenetic variation is a fundamental characteristic of the cancer phenotype (Hansen *et al.*, 2011). Nevertheless, available analyses only assess enhancer–promoter methylation on the population level, averaging out any differences between individual cells in the studied sample. We hypothesize that

© The Author(s) 2021. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article/37/Supplement_1/i327/6319710 by Tel Aviv University user on 19 July 2021

variability in methylation on the population level may be attributed to variability in the mixing ratios of cancer and benign phenotypes of the same cell type. In such cases, the detailed single-cell enhancer–promoter methylation profile may provide valuable information for studying the evolution of early carcinogenesis and tumor heterogeneity. Long reads present a unique opportunity to study the co-existence of methylation in a promoter and its enhancers along the same DNA molecule, in effect providing single-cell information for the studied locus. When examining many such molecules, the methylation pattern distribution of an enhancer–promoter pair may be directly recorded and used to enumerate cell subsets in a mixture, similar to what may be achieved with gene expression RNA mixtures (Newman *et al.*, 2015; Zaitsev *et al.*, 2019). Methylation profiles have already been utilized to infer cell mixture distribution with good accuracy (Houseman *et al.*, 2012). It remains to be tested whether pairwise analysis of enhancer–promoter pairs provides information on subtle transformations in cells with identical genetic backgrounds such as in early cancer. In order to establish the analytical framework for such data, we analyzed whole genome Bionano Genomics optical methylation maps (Sharim *et al.*, 2019). We identified ∼4 million long molecule reads encompassing enhancer–promoter pairs up to 200 kb apart at 30×–300× coverage and explored several analytical approaches to harness these data for cell mixture deconvolution.

2 Materials and methods

2.1 Data collection

Single-molecule methylation maps of three replicates of the B-lymphocyte cell line GM12878 were adapted from (Sharim *et al.*, 2019), and similar maps of immortalized myoblasts of two healthy human subjects (Wellstone center for FSHD, UMMS) were obtained by the same methods. Shortly, high molecular weight DNA was extracted to ensure long single molecules. DNA was then fluorescently barcoded at specific sequence motifs for alignment to an in-silico reference. Unmethylated cytosines in the recognition sequence TCGA were fluorescently labeled to perform reduced representation optical methylation mapping (ROM) on the Bionano Genomics Saphyr instrument. The genomic location of the labeled unmethylated cytosines on individual DNA molecules was inferred from direct alignments to the hg38 reference (Bionano Access and Solve). Label coordinates were extended to 1 kb to account for the optical measurement resolution, limiting localization accuracy to ∼1000 bp (Wang *et al.*, 2012). More details regarding ROM and optical mapping can be found in (Jeffet *et al.*, 2021; Sharim *et al.*, 2019).

2.2 Enhancer-promoter links and coordinates

Enhancers' target genes were predicted by the JEME method and adapted from (Cao *et al.*, 2017). Genomic coordinates of enhancers were converted from the human genome build hg19 to hg38 using UCSC liftOver (Haeussler *et al.*, 2019). Ambiguous genomic regions (Amemiya *et al.*, 2019) were subtracted from enhancer locations, and enhancers smaller than 200 bp were extended to 200 bp around their midpoint. Promoters were defined according to the transcription start sites (TSS) of the protein-coding genes [Gencode V.34 annotations, (Frankish *et al.*, 2019)], and taken as 2000 bp upstream and 500 bp downstream of the TSS. Enhancers and promoters not containing at least one potential site for methylation labeling were discarded. Enhancer–promoter (E–P) pairs in close proximity, less than 5000 bp, were also filtered out. All predicted E–P pairs under these conditions were used for our analysis, regardless of cell type and biological contexts used for their prediction.

For comparison against promoter-only analysis (see 2.4), we created a set of pairs that matches the size of the promoters set, by assigning a single enhancer to each promoter. We focused on enhancers that display the highest number of potential detectable methylation sites. In case of tie, enhancer size and proximity to the corresponding promoter were also considered. These criteria are unbiased toward specific cell types and select enhancers with the highest potential for reliable methylation calling by ROM. Nevertheless, these criteria are arbitrary and do not necessarily reflect biologically active enhancers.

2.3 Coexistence of methylation signals at the distal elements

We focused on DNA molecules that span entire E–P pairs and recorded the corresponding methylation states. The enhancers and promoters' states were reduced to binary methylation states: if the element showed any degree of fluorescence, it was identified as 'unmethylated'. Therefore, every enhancer and promoter pair coexisting on a DNA molecule displays one of four possible methylation combinations with the promoter and enhancer being methylated or unmethylated. For every pair, the number of molecules belonging to each class is counted in order to record the exact pairwise methylation distribution. Enhancer–promoter pairs that were covered by less than 30 molecules in an experiment were filtered out. Molecules were counted separately for each pair they covered.

2.4 Matched promoter methylation data

To benchmark our pairwise analysis, we extracted information of promoter-only methylation from the same molecules that span the E–P pairs. This was used to assess single-molecule level single-element-based deconvolution, as well as whether the coexistence of enhancer–promoter pairs holds similar or additional information beyond promoter methylation analysis.

2.5 Assembling datasets, training/test division of molecules and creating simulated mixtures

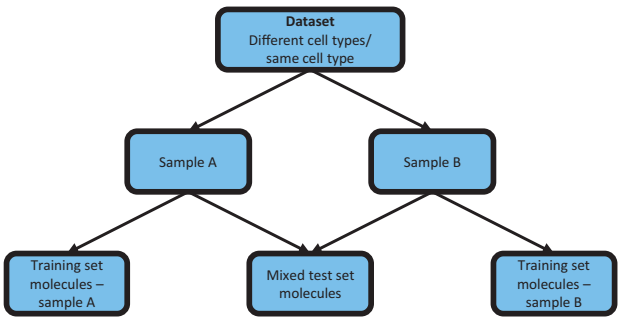
Two different datasets were assembled to test our deconvolution pipeline. Each dataset is composed of two samples. The two samples comprising the dataset will be referred to as sample 'A' and sample 'B'.

1. Dataset of two different cell types: contains B-lymphocyte cells (three GM12878 replicates were merged to a single sample) and myoblasts (two lines were merged).
2. Dataset of the same cell type from two different individuals: contains myoblast cells from two different human subjects.

Molecules of the two samples comprising each dataset ('A' and 'B') were randomly divided into a training set and a test set. Test sets of both samples in a dataset contain an equal number of molecules, accounting for 33% of molecules in the less-covered sample. The training sets were left pure while the test sets of both samples were randomly mixed at known ratios (0–100%, in 10% increments) to create a mixed test set (Scheme 1).

2.6 Deconvolution of mixtures

We explored several methods for deconvolving the mixtures to infer the mixing ratio: (i) local projection of vectors, (ii) global minimization of sum of squared errors (SSE) and Kullback-Leibler divergence (KLD) measures and (iii) global maximum



Scheme 1. Hierarchy of terms used: dataset, sample, training/test set.

Downloaded from https://academic.oup.com/bioinformatics/article/37/Supplement_1/i327/6319710 by Tel Aviv University user on 19 July 2021

likelihood estimation (MLE). We describe these methods in detail below.

2.6.1 Vector projections

In this local method, the mixing ratio is calculated separately for each enhancer–promoter pair. Specifically, the normalized proportions of each of the four possible methylation combinations define a 4-dimensional vector representing the pair in a given sample. Each E–P pair is characterized by three such vectors belonging to the mixed test set ($\overrightarrow{T_p}$) and the two pure training sets ($\overrightarrow{A_p}$, $\overrightarrow{B_p}$). In order to assess the proportion of molecules in the test set originating from each sample, the difference between the test set vector and vector $\overrightarrow{A_p}$, named vector $\overrightarrow{AT_p}$, was projected on the difference between vectors $\overrightarrow{B_p}$ and $\overrightarrow{A_p}$ of the training sets ($\overrightarrow{AB_p}$). The mixing ratio w.r.t. sample B is given by the ratio between the size of the difference between the projection vector and vector $\overrightarrow{A_p}$ (calculated as the dot product of $\overrightarrow{AT_p}$ and $\overrightarrow{AB_p}$ divided by the size of $\overrightarrow{AB_p}$), and the size of vector $\overrightarrow{AB_p}$. The mixing ratio w.r.t sample A completes this value to 1. The final selected mixing ratio is the average of mixing values calculated for all E–P pairs (Equation 1).

$$\alpha_A = \frac{1}{N_p}\sum_p \left(1 - \frac{\left(\overrightarrow{AT_p}\right)\cdot\left(\overrightarrow{AB_p}\right)}{\left|\left(\overrightarrow{AB_p}\right)\right|^2}\right) \quad (1)$$

In Equation 1, α_A is the mixing ratio relative to sample A; p represents the E–P pair; N_p is the total number of E–P pairs; $\overrightarrow{AT_p}$ represents the difference between the test set vector and the vector of the training set of sample A; $\overrightarrow{AB_p}$ represents the difference between the vectors of training sets B and A.

2.6.2 Minimizing the difference between the test set and linear combinations of the training sets via SSE and KLD computations

The counts of molecules with the different methylation combinations in each enhancer–promoter pair in the mixed test set and the pure training sets were normalized to obtain ratios, making sure all ratios are non-zeros by adding 0.01 to each ratio and renormalizing to 1. The ratios of the pure training sets were treated as the probability of a molecule spanning a certain E–P pair to have a specific methylation combination given the sample it came from, either 'A' or 'B'. Then, linear combinations of the two pure training sets were assembled per pair in all possible mixing ratios w.r.t. sample A (0–100% in 1% increments) (Equation 2).

$$\mathrm{LCT}_{\alpha,c,p} = \alpha \mathrm{A}_{c,p} + (1-\alpha)\mathrm{B}_{c,p} \quad (2)$$

$LCT_{\alpha,c,p}$ is the linear combination of the pure training sets per EP pair (p), methylation combination (c) and current parameter of the training sets' linear combination (α); $A_{c,p}$ and $B_{c,p}$ are the probabilities of molecules from training sets A and B respectively to have methylation combination c given the sample and the E–P pair p.

The test set is compared against all the 101 linear combinations of the training sets. The ratio α that minimizes these expressions is reported. We use two minimization criteria:

2.6.2.1 Sum of squared errors (SSE).

$$\sum_p \sum_c \left(\mathrm{test}_{cp} - \mathrm{LCT}_{\alpha,c,p}\right)^2 \quad (3)$$

In Equation 3, p represents E–P pairs; c represents methylation combinations; α represents the current parameter of the training sets' linear combination.

2.6.2.2 Kullback–Leibler divergence (KLD).

$$\mathrm{D_{KL}}\left(\mathrm{test}_{c,p}||\mathrm{LCT}_{c,p,\alpha}\right) = \sum_p \sum_c \mathrm{test}_{c,p}\log\left(\frac{\mathrm{test}_{c,p}}{\mathrm{LCT}_{c,p,\alpha}}\right) \quad (4)$$

$$\mathrm{D_{KL}}\left(\mathrm{LCT}_{c,\,p,\alpha}||\mathrm{test}_{c,\,p}\right) = \sum_p \sum_c \mathrm{LCT}_{c,\,p,\,\alpha}\log\left(\frac{\mathrm{LCT}_{c,\,p,\alpha}}{\mathrm{test}_{c,\,p}}\right) \quad (5)$$

$$\mathrm{D_{KL_{average}}} = \frac{\left[\mathrm{D_{KL}}(\mathrm{test}_{c,p}||\mathrm{LCT}_{c,p,\alpha}) + \mathrm{D_{KL}}(\mathrm{LCT}_{c,p,\alpha}||\mathrm{test}_{c,p})\right]}{2} \quad (6)$$

In Equations 4–6, p represents E–P pairs; c represents methylation combinations; α represents the current parameter of the training sets' linear combination.

2.6.3 Maximum likelihood estimation (MLE)

Last, we considered a method that is based on a probabilistic model of the data which asserts that each molecule is chosen from one of the samples based on the mixing ratio and then its methylation status is chosen based on the corresponding normalized counts vector. In detail, the probability of a molecule in the test set to originate from sample A, is the mixing ratio, α, whereas its probability to originate from sample B is 1-α (Equation 7).

$$\mathrm{P}(\mathrm{y_i} = \mathrm{A}) = \mathrm{P_A} = \alpha, \mathrm{P}(\mathrm{y_i} = \mathrm{B}) = \mathrm{P_B} = 1-\alpha \quad (7)$$

In Equation 7, i stands for a molecule in the test set; y is the hidden information about the sample it came from.

The probability of a molecule from the test set to exhibit a methylation combination c, given the sample it came from (A or B) and the E–P pair it spans can be evaluated as the proportion of combination c in that E–P pair in the training set of the corresponding sample (Equation 8).

$$\mathrm{P}(\mathrm{X_{i,p}} = \mathrm{c}|\mathrm{y_{i,p}} = \mathrm{j}) = \mathrm{P_{j,c,p}} \quad (8)$$

$X_{i,\,p}$ is the observed methylation combination of molecule i of pair p; j is the sample of origin, either A or B.

We estimate the mixing ratio for which the test data observations are most probable by maximizing the log-likelihood function (Equation 9):

$$\begin{aligned} log(L(\alpha)) &= \sum_i log\left(\alpha P_{A,p,c_i} + (1-\alpha)P_{B,p,c_i}\right) \\ &= \sum_p \sum_c N_{p,c} log\left(\alpha P_{A,p,c} + (1-\alpha)P_{B,p,c}\right) \end{aligned} \quad (9)$$

In Equation 9, i represents a test set molecule; p is the pair it belongs to; c is its methylation combination; A and B are the two samples; $N_{p,\,c}$ is the number of molecules in the test set that come from pair p and display methylation combination c.

As the log-likelihood is concave, we can find its maximum using gradient ascent. In our implementation of the algorithm (Equation 10), the difference in log-likelihood from previous iteration served as a stopping criterion (0.001), while confining the mixing ratio to the relevant range (0–1), as well as limiting the number of iterations (no more than 20 000). The mixing ratio that yielded the highest log-likelihood is reported.

$$\alpha^{new} = \alpha^{old} + \eta \frac{\partial log(L(\alpha))}{\partial \alpha} \quad (10)$$

α^{new} is the calculated mixing ratio in the current iteration; α^{old} is the mixing ratio obtained in the previous iteration (initialized to 0.5); η is the step size of the gradient ascent algorithm (fixed to 0.005).

2.7 Supervised selection of enhancer–promoter pairs

Several methods are proposed here to rank the different E–P pairs by their ability to discriminate between the training set samples. This ranking can serve to select a smaller subset of pairs, to ensure more accurate results and noise filtration. We explored three such methods, as detailed below.

2.7.1 Euclidean distances

$$\mathrm{Distance(A,\ B)_p} = \sqrt{\sum_c \left(\mathrm{P_{A,\,p,\,c}} - \mathrm{P_{B,\,p,\,c}}\right)^2} \quad (11)$$

In Equation 11, p is the current E–P pair; c represents methylation combinations; $P_{A,p,c}, P_{B,p,c}$ are the proportions of methylation combination c in pair p in samples A and B respectively.

Downloaded from https://academic.oup.com/bioinformatics/article/37/Supplement_1/i327/6319710 by Tel Aviv University user on 19 July 2021

2.7.2 KLD

$$D_{KL_p} = \frac{\sum_c P_{A,p,c}\log\left(\frac{P_{A,p,c}}{P_{B,p,c}}\right) + \sum_c P_{B,p,c}\log\left(\frac{P_{B,p,c}}{P_{A,p,c}}\right)}{2} \quad (12)$$

In Equation 12, p is the current E–P pair; c represents methylation combinations; $P_{A,p,c}, P_{B,p,c}$ are the proportions of methylation combination c in pair p in samples A and B respectively.

2.7.3 Weighted KLD (wKLD)

Molecule coverage of the different enhancer–promoter pairs in each training sample can vary. Multiplying the KLD score of each combination of every pair by the number of molecules displaying this combination in the corresponding sample and pair, puts more weight on the highly covered pairs, which presumably provide more reliable information.

$$wD_{KL_p} = \frac{\sum_c N_{A,p,c}P_{A,p,c}\log\left(\frac{P_{A,p,c}}{P_{B,p,c}}\right) + \sum_c N_{B,p,c}P_{B,p,c}\log\left(\frac{P_{B,p,c}}{P_{A,p,c}}\right)}{2} \quad (13)$$

In Equation 13, p is the current E–P pair; c represents methylation combinations; $P_{A,p,c}, P_{B,p,c}$ are the proportions of methylation combination c in pair p in samples A and B respectively; $N_{A,p,c}, N_{B,p,c}$ are the number of molecules in pair p that display methylation combination c in the training sets of samples A and B respectively.

3 Results

Our analytical exploration builds on the fact that new types of data that profile genomic methylation via long single-molecule reads have recently become available. Of special interest are datasets obtained by Bionano Genomics optical genome mapping. These data contain the largest fraction of molecules longer than 100 kbp in comparison to other long-read approaches, allowing us to explore distant enhancers in the context of their molecular promoter. The method is inherently poor in resolution but may effectively report on methylation status at the level of genomic elements such as gene bodies, promoters and enhancers. Figure 1 shows a stack of digitized DNA molecules mapped to a region in chromosome 17 containing the TP53 gene locus and three of its predicted enhancers, located ~50–100 kb away from the promoter. All individual molecules selected from these data span the gene promoter and at least one distant enhancer. Methylation labels shown in dark blue along the gray molecules contour denote unmethylated CpGs (non-methylation labels have been artificially enhanced inside the boxed promoter and enhancer regions). It can be clearly seen that the promoter and the two closest enhancers are highly labeled and thus unmethylated. The leftmost enhancer is almost void of methylation labels, indicating that it is highly methylated. The methylation status is well reflected in the average methylation profile shown on the bottom of the figure and reflecting the accumulation of methylation labels from all molecules along the region. We note that for many gene promoters and enhancers, the methylation status is variably distributed along individual molecules, with all four combinations of E–P methylation represented for some pairs.

3.1 Comparison between deconvolution methods for promoters and E–P pairs

We first set out to compare the deconvolution efficacy of E–P pairwise methylation in comparison with promoter methylation. Given that each promoter has multiple predicted enhancers, which results in a much larger E–P dataset, we restricted this analysis to a single enhancer assigned to each promoter as described in Section 2.2. Deconvolution of simulated mixtures was performed by several methods: local projection of vectors, global minimization of the sum of squared errors (SSE), Kullback-Leibler divergence (KLD) and maximum likelihood estimation (MLE) (see Section 2.6). Mixtures containing two different cell types (B-lymphocytes and myoblasts) at 10% increments were subject to deconvolution by each of the methods for promoter-only methylation as well as in the context of E–P pairwise methylation (Fig. 2). The least accurate deconvolution was achieved by vector projections, yielding over 7% average error for the promoter-based analysis and over 4% error for the E–P analysis. The best deconvolution was achieved by MLE with 0.86% average error for promoter methylation and 0.69% error for E–P methylation. Since the entire range was assessed (mixing ratios from 0 to 1), the error rate of very small proportions can be obtained by interpolation (for median coverage per pair of ~50 molecules in the test set).

3.2 Deconvolution of myoblasts derived from two individuals using full E–P methylation

The abundance of enhancers and the interplay between their interaction with their gene promoters may hold important information on the precise state of a cell. While for comparison with promoter-based analysis we limited the number of enhancers to one per promoter, our E–P dataset is composed of over 100 000 different pairs with detailed pairwise distribution for each pair. B-lymphocytes and myoblasts, two distinct cell types, were successfully resolved with 1.01–1.36% accuracy by three of the four methods tested (Fig. 3). The incorporation of multiple enhancers per promoter slightly increases the overall deconvolution error relative to the more limited set used for comparison with promoters. This is likely due to the additional noise contributed by non-active predicted enhancers. Nevertheless, analyzing the full E–P methylation dataset is more biologically relevant as it does not make assumptions on the activity of enhancers and inherently contains more information (but also more noise). Using validated enhancer–promoter links as opposed to the predicted links used here, may further improve the pairwise estimates. Such validated links may be constructed by use of proximity ligation methods [such as Hi-C (Rao *et al.*, 2014)].

We next tested our deconvolution methods on mixtures of two myoblast cell lines from different donors. The mixtures were resolved with 5.81–6.82% accuracy by the same three methods (Fig. 4). The two myoblast samples are more similar to each other in genome-wide methylation profiles than they are to the unrelated B-lymphocytes in the previous mixtures. Hence, random alterations between the samples may occupy more weight, and can explain the decline in accuracy. We hypothesized that a smaller, educated subset of E–P pairs used for deconvolution could improve the analysis.

3.3 Supervised selection of enhancer–promoter pairs

With over 100 000 predicted enhancer–promoter links used for deconvolution, it is reasonable to assume that not all pairs contribute equally to the discrimination between sample types. Most probably the identity of the most contributing pairs is specific to the types of samples being resolved. Accordingly, relying only on a subset of most differentiating pairs, while filtering out the rest, has potential to improve deconvolution accuracy by filtering out invaluable and possibly noisy data. Additionally, if a small subset of pairs is sufficient for accurate deconvolution, it simplifies and shortens the required analysis. We tested three methods for ranking the pairs: Euclidean distances, KLD and Weighted KLD (wKLD) (see Section 2.7). The performance of the different deconvolution methods was compared for all ranking methods and with different numbers of highest-ranking pairs selected. The full set constituted 108 048 pairs in the mixture of B-lymphocytes and myoblasts, and 135 793 pairs for the two different myoblasts. The mean deconvolution error for several subsets of highest-ranking pairs are shown in $\log_{10}$ scale in Figure 5. The different ranking methods provide similar results and the differences in accuracy are mostly attributed to the deconvolution approach used. The lymphocytes and myoblasts mixtures were resolved with ~1% average deconvolution accuracy by MLE, using 75 000–108 048 pairs, and the mixture of the two myoblasts was resolved with ~1.4–1.8% average accuracy using only 100 pairs chosen by KLD or wKLD with MLE deconvolution.

Figure 5 reveals opposite trends in respect to the optimal size of pairs subset used for deconvolution. Whereas the mixture of the different cell types is monotonically resolved more accurately with

Downloaded from https://academic.oup.com/bioinformatics/article/37/Supplement_1/i327/6319710 by Tel Aviv University user on 19 July 2021

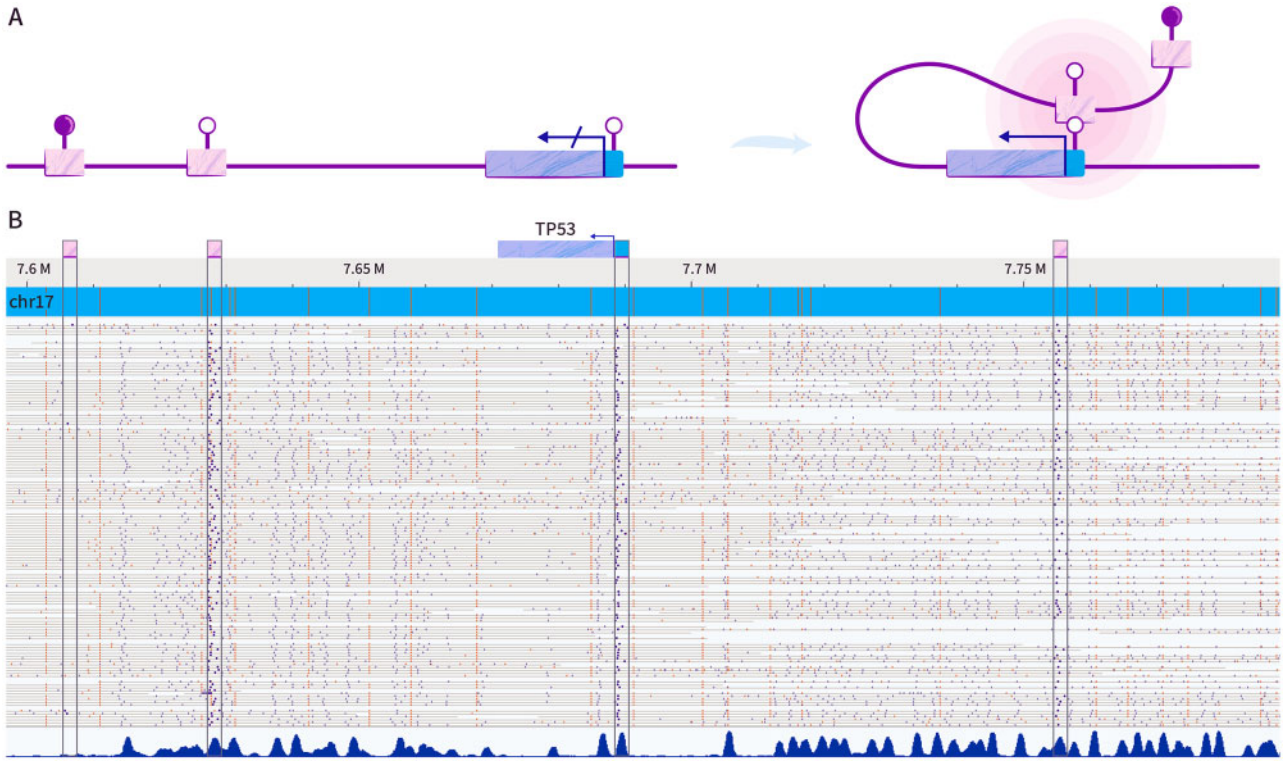


Fig. 1. Methylation states in predicted enhancer–promoter pairs. (**A**) schematic illustration of possible methylation states for a promoter and enhancers, and potential interaction between them. (**B**) Bionano Genomics optical methylation map of a region in chromosome 17 in GM12878 DNA. The region contains the gene TP53, its promoter (small blue box), and several predicted enhancers (pink boxes). Dark blue dots denote unmethylated sites and orange dots denote genetic tags used for alignment to the hg38 reference.

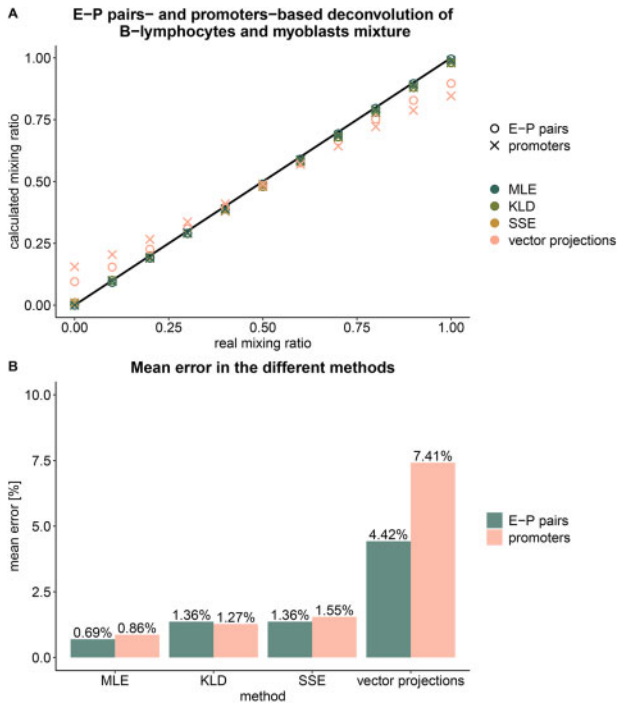


Fig. 2. Deconvolution of mixtures containing B-lymphocytes and myoblast cells by different methods using methylation states in promoters alone and enhancer–promoter pairs, accounting for one enhancer per promoter. (**A**) Calculated mixing ratio according to the different methods versus the known mixing ratio. (**B**) The mean error in calculated mixing ratio, calculated as the absolute distance from the known ratio, in the different methods.

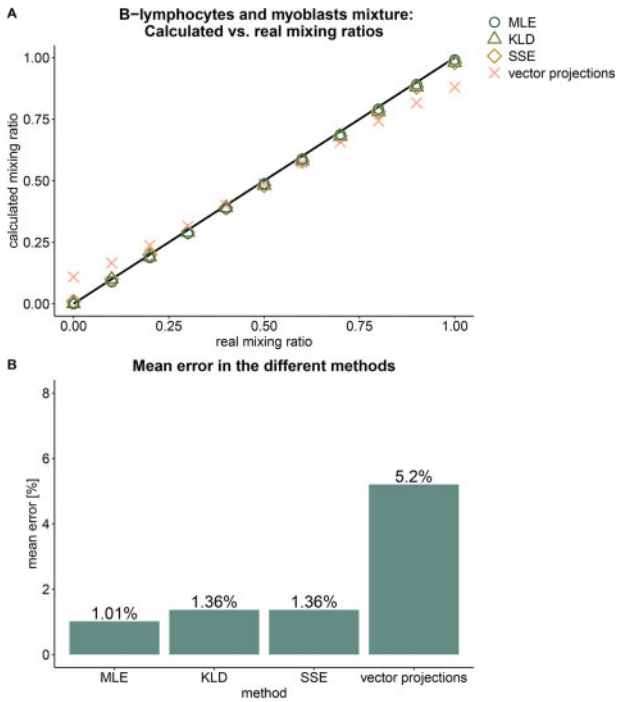


Fig. 3. Deconvolution of B-lymphocytes and myoblast cells mixtures by different methods using methylation states in all predicted enhancer–promoter pairs. (**A**) calculated mixing ratio according to the different methods versus the known mixing ratio. (**B**) the mean error in calculated mixing ratio, calculated as the absolute distance from the known ratio, in the different methods.

Downloaded from https://academic.oup.com/bioinformatics/article/37/Supplement_1/i327/6319710 by Tel Aviv University user on 19 July 2021

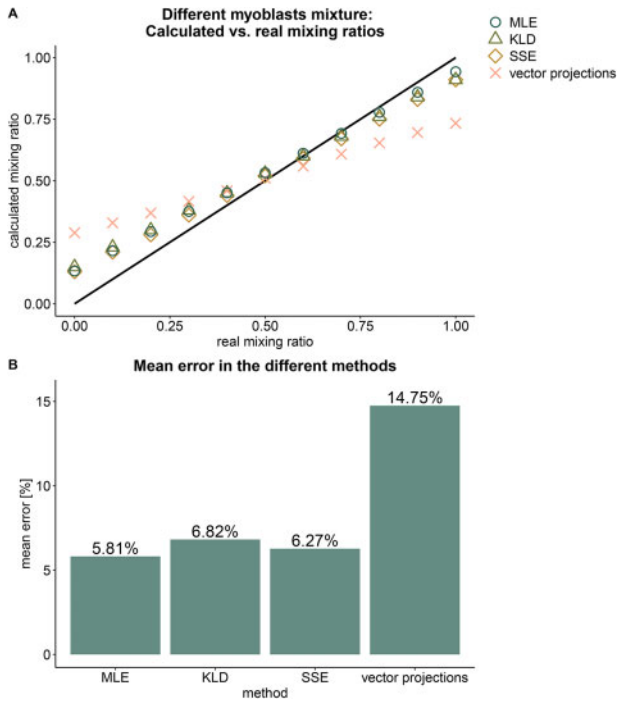


Fig. 4. Deconvolution of myoblast cells from two different donors by different methods. (**A**) calculated mixing ratio according to the different methods versus the known mixing ratio. (**B**) the mean error in calculated mixing ratio, calculated as the absolute distance from the known ratio, in the different methods.

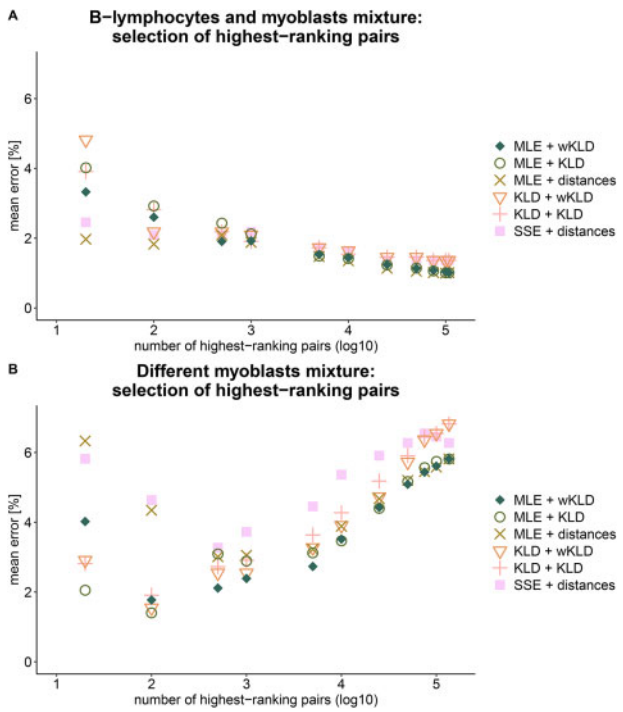


Fig. 5. Subsets of E–P pairs, selected by supervised selection methods. The mean error in calculated mixing ratio (distance from theoretical ratio) is displayed against the $\log_{10}$ of the number of best pairs selected in each combination of deconvolution method and ranking method. (**A**) A mixture of two cell types: B-lymphocytes and myoblasts. (**B**) A mixture of two myoblast cells derived from different individuals.

increasing number of pairs (>75 000 pairs), the mixture of myoblast cells derived from different individuals shows a distinct minimum in the mean error for 100–500 pairs. Since DNA methylation patterns are known to regulate the expression of cell-type specific genes (Dor and Cedar, 2018), a higher variance in methylation signatures is expected between different cell types such as lymphocytes and myoblasts. Different cell-specific methylation patterns imply that more regions along the genome are differential and may contribute to deconvolution, making their differentiation simple and accurate. Deconvolving mixtures of the same cell types such as the mixture of myoblasts from the different individuals is more challenging. We postulate that as may frequently happen in diseased tissue, the observed methylation differences are not related to the cell's identity, but factors as disease, age or exposure to environmental stimuli. In such cases, methylation variability at cell-type specific loci adds noise to the deconvolution analysis. Sorting the pairs by their information contribution provides a supervised educated approach for assembling the list of pairs that yields the most accurate differentiation.

4 Conclusions

This work lays the ground for cell-type deconvolution utilizing a new type of data structure now available via long single-molecule methylation maps. This data structure contains chromosome-level methylation profiles of gene bodies, promoters and one or more distant enhancers, all on the same molecule. We test several deconvolution methods and show that for differentiating two cell types, the pairwise analysis yields better deconvolution than promoter-based analysis, reaching an error rate of 0.7%. Since enhancer methylation is known to be a major contributor to methylation variability within a cell-type population, such as in cancer, we also analyzed mixtures of two myoblast cell-lines derived from two individuals. The full E–P pairs dataset yielded a deconvolution error of ∼6% for these highly similar samples. We reasoned that cells with similar methylomes will be differentially methylated only at a subset of loci while variability in common methylation loci will add noise to the deconvolution process. We tested several methods to rank the pairs according to their differentiation capacity. We assessed deconvolution fidelity for various numbers of highest-ranking pairs and found that for the two distinct cell types the deconvolution error monotonically declines with additional pairs. For the two myoblast samples on the other hand, a clear minimum was calculated at ∼100 pairs that reduced the error from ∼6% to ∼1.5%. These results constitute a first step toward harnessing enhancer–promoter linked methylation for deconvolution of cell populations with highly similar cell-type methylomes. Despite focusing on Bionano Genomics' optical methylation mapping (Gabrieli *et al.*, 2021; Sharim *et al.*, 2019), which currently provides the highest coverage of long reads, the principles are valid to other future datasets such as those produced by Oxford Nanopore ultralong-read sequencing protocol. Further exploration of these linkages, including the joint effects of multiple enhancers per promoter may shed light on insightful cellular transformations regulated by long-range epigenetic interactions.

Author Contributions

Y.E. and R.S. conceived and supervised the project. S.M., Y.A., H.S. and Z.M. performed data analysis. S.B., Y.W.C., E.V. and H.B. collected data. R.E. advised the project. Y.E., R.S. and S.M. wrote the manuscript. All authors read and edited the manuscript.

Funding

This study was funded by the European Research Council Consolidator grant [817811 to Y.E.]. Research reported in this manuscript was in part supported by Award Number UL1TR001876 from the NIH National Center for Advancing Translational Sciences (H.B.) and Award Number R21HG011424 from the NIH National Human Genome Research Institute (H.B.). R.S. was supported by a research grant from the Israel Science Foundation [715/18].

Downloaded from https://academic.oup.com/bioinformatics/article/37/Supplement_1/i327/6319710 by Tel Aviv University user on 19 July 2021

Conflict of Interest: H.B. and E.V. own a limited number of stock options of Bionano Genomics Inc. H.B. is also employed part-time by Bionano Genomics Inc.

Data availability

The raw data underlying this article are available on Zenodo (10.5281/zenodo.4679382).

References

Amemiya,H.M. *et al.* (2019) The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.*, **9**, 1–5.

Aran,D. *et al.* (2013) DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.*, **14**, R21.

Aran,D. and Hellman,A. (2013a) DNA methylation of transcriptional enhancers and cancer predisposition. *Cell*, **154**, 11–13.

Aran,D. and Hellman,A. (2013b) Unmasking risk loci: DNA methylation illuminates the biology of cancer predisposition. *Bioessays*, **36**, 184–190.

Cao,Q. *et al.* (2017) Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.*, **49**, 1428–1436.

Chu,A.Y. *et al.* (2017) Epigenome-wide association studies identify DNA methylation associated with kidney function. *Nat. Commun*, **8**, 1286.

Dor,Y. and Cedar,H. (2018) Principles of DNA methylation and their implications for biology and medicine. *Lancet*, **392**, 777–786.

Frankish,A. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.

Gabrieli,T. *et al.* (2021) Chemoenzymatic labeling of DNA methylation patterns for single-molecule epigenetic mapping. *BioRxiv*. doi: 10.1101/2021.02.24.432628.

Gorenjak,V. *et al.* (2020) Epigenome-wide association study in healthy individuals identifies significant associations with DNA methylation and PBMC extract VEGF-A concentration. *Clin. Epigenet.*, **12**, 79.

Haeussler,M. *et al.* (2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.

Hansen,K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.

Hentze,J. *et al.* (2019) Methylation and ovarian cancer: can DNA methylation be of diagnostic use? (Review). *Mol. Clin. Oncol.*, **10**, 323–330.

Houseman,E.A. *et al.* (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, **13**, 86.

Jeffet,J. *et al.* (2021) Single-molecule optical genome mapping in nanochannels: multidisciplinarity at the nanoscale. *Essays Biochem.*, **65**, 51–66.

Küpers,L.K. *et al.* (2019) Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. *Nat. Commun.*, **10**, 1893.

Kurkjian,C. *et al.* (2008) DNA methylation: its role in cancer development and therapy. *Curr. Probl. Cancer*, **32**, 187–235.

Li,Q. *et al.* (2013) Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*, **152**, 633–641.

Newman,A.M. *et al.* (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 453–457.

Rao,S.S.P. *et al.* (2014) A 3DMap of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Sharim,H. *et al.* (2019) Long-read single-molecule maps of the functional methylome. *Genome Res.*, **29**, 646–656.

Vrba,L. and Futscher,B.W. (2019) DNA methylation changes in biomarker loci occur early in cancer progression. *F1000Research*, **8**, 2106.

Wang,Y. *et al.* (2012) Resolution limit for DNA barcodes in the Odijk regime. *Biomicrofluidics*, **6**, 014101.

Zaitsev,K. *et al.* (2019) Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.*, **10**, 1–16.

Downloaded from https://academic.oup.com/bioinformatics/article/37/Supplement_1/i327/6319710 by Tel Aviv University user on 19 July 2021