Qi and Teschendorff *Clinical Epigenetics* (2022) 14:31
https://doi.org/10.1186/s13148-022-01253-3

Clinical Epigenetics

PERSPECTIVE Open Access

Cell-type heterogeneity: Why we should adjust for it in epigenome and biomarker studies

Luo Qi[1] and Andrew E. Teschendorff[1,2*]

Abstract

Most studies aiming to identify epigenetic biomarkers do so from complex tissues that are composed of many different cell-types. By definition, these cell-types vary substantially in terms of their epigenetic profiles. This cell-type specific variation among healthy cells is completely independent of the variation associated with disease, yet it dominates the epigenetic variability landscape. While cell-type composition of tissues can change in disease and this may provide accurate and reproducible biomarkers, not adjusting for the underlying cell-type heterogeneity may seriously limit the sensitivity and precision to detect disease-relevant biomarkers or hamper our understanding of such biomarkers. Given that computational and experimental tools for tackling cell-type heterogeneity are available, we here stress that future epigenetic biomarker studies should aim to provide estimates of underlying cell-type fractions for all samples in the study, and to identify biomarkers before and after adjustment for cell-type heterogeneity, in order to obtain a more complete and unbiased picture of the biomarker-landscape. This is critical, not only to improve reproducibility and for the eventual clinical application of such biomarkers, but importantly, to also improve our molecular understanding of disease itself.

Keywords: Cell-type heterogeneity, Cell-type deconvolution, Epigenetic biomarkers, DNA methylation, Classification

Background

The cell-types that are present within a given tissue or organ are distinguished by a unique gene and protein expression profile [1]. This functional molecular profile is epigenetically determined via a complex interplay of histone modifications, chromatin accessibility and covalent DNA methylation marks [2]. Thus, epigenetic profiles are highly cell-type specific.

Clinical interest in measuring epigenetic profiles stems from the fact that such epigenetic profiles are often altered in disease and in association with disease risk factors [3–9], with some evidence also pointing to a potentially causal or causally-mediating role [3, 10–12]. DNA-based epigenetic marks like DNA methylation (DNAm) are also fairly stable and amenable to genome-wide measurement in large numbers of samples and in many types of clinical specimens [13, 14], including blood [15–17], cell-free DNA in serum [18, 19] and formalin-fixed paraffin embedded (FFPE) tissue [20–22], making it a very attractive substrate for biomarker studies. Indeed, the sensitivity of technologies like the Illumina DNAm beadarray is such that one can detect DNAm changes as small as 1–5% with over 90% sensitivity [23, 24]. This high sensitivity and precision has been confirmed by many biomarker studies: for instance, smoking-associated DNAm changes in blood tissue are characterized by such small effect sizes and are highly reproducible [25, 26].

*Correspondence: andrew@picb.ac.cn
[1] CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
Full list of author information is available at the end of the article

However, a key challenge for the biological interpretation of such epigenetic changes remains in that epigenetic measurements are generally performed on DNA extracted from heterogeneous sample specimens. This is because measuring epigenetic profiles, including DNA methylation, at cell-type resolution and for all cell-types in the tissue is currently very costly or technically challenging [27], and in the case of single-cells only generates very sparse and incomplete data [28, 29], which is therefore impractical for biomarker studies which aim to measure genome-wide profiles in hundreds if not thousands of clinical samples. Thus, the obtained measurements on bulk samples only reflects an average DNAm profile over all cells and cell-types within the specimen. It follows that this average DNAm profile will be affected by factors such as the cell-type composition of the sample, which could vary substantially between individuals. Thus, it might be difficult to ascertain in which cell-types a particular 5% change in DNAm is happening. Although similar considerations apply to other epigenetic marks such as histone modifications and chromatin accessibility, for reasons given above this perspective focuses on DNAm.

One of the first studies to demonstrate the big effect that cell-type heterogeneity (CTH) can have on subsequent statistical inference was an Epigenome-Wide Association Study (EWAS) performed in blood tissue from Rheumatoid Arthritis (RA) cases and controls [15]. In this study it was shown that there were a large number of CpGs differentially methylated between RA cases and controls, owing to a substantial shift in the granulocyte to lymphocyte proportions between cases and controls. While such alterations in blood composition could potentially be useful as a diagnostic marker (assuming they are specific to the disease), they don't reflect disease-associated DNAm alterations that occur in a given cell-type. It is now widely recognized that the large number of differentially methylated cytosines (DMCs) detected in this RA EWAS study is a reflection of the inflammatory response to the disease, which is therefore of limited interest for identifying disease risk markers. As shown by Liu et al. the great majority of these DMCs disappear once we adjust for the underlying changes in cell-type composition between RA cases and controls [15]. Another important and more recent application where adjustment for CTH is critical, is in the construction of diagnostic and pre-diagnostic disease predictors (e.g. cancer) from cell-free DNAm in serum, where such adjustment is necessary to remove the contaminating effect of lymphocyte DNA [18, 19, 30]. As shown in these studies, adjustment for CTH is critical to achieve the reported high prediction accuracies.

Although many other studies have re-emphasized the critical need to adjust for cell-type heterogeneity when analysing DNAm data [31–33], it is surprising that many epigenome studies continue to ignore this major confounder when identifying biomarkers [34–37], or when proposing novel disease classifications [38–40]. In our opinion, there are two reasons for this. First, adjusting for cell-type heterogeneity, specially in complex solid tissues, can be technically challenging and investigators may not even be aware that tools for such adjustments or partial adjustments exist. For instance, all Cancer Genome Atlas (TCGA) projects [41–46] do not adjust for CTH when proposing novel cancer taxonomies. Second, there is a common belief that adjustment for cell-type heterogeneity is not really necessary when searching for biomarkers, based on the premise that any biomarker with high prediction accuracy is valuable irrespective of the underlying biological process driving the association. While this 2nd argument is entirely valid, it does not justify not performing additional analyses that adjust for cell-type heterogeneity (CTH), as these additional analyses can lead to important novel biological insight or novel biomarkers. The purpose of this perspective is therefore to reinforce the argument and rationale for performing adjustments for CTH, as well as to make the community aware that appropriate computational tools for such corrections are available and that they often work better and cheaper than laborious experimental alternatives (e.g. generating epigenetic profiles in purified samples).

Why we should adjust for cell-type heterogeneity

One important argument in favor of performing a differential DNAm analysis that adjusts for CTH comes from consideration of the relative data variance that can be attributed to the various factors, including CTH and the phenotype or exposure of interest. In general, given a genome-wide DNA methylation dataset where samples are drawn from different genetic backgrounds (e.g. ethnicity) and sexes, these factors are likely to dominate the data variance alongside CTH. That these three factors would dominate the DNAm variation landscape is intuitively clear. First of all, it is well known that a substantial proportion of DNAm is under genetic influence [47] and in many instances the effect sizes are quite substantial, i.e. over 50% DNAm differences between the homozygote A/A and B/B genotypes is common. Thus, top principal components (PCs) are likely to correlate with ethnicity if the study contains roughly equal numbers of samples from each ethnic group. Variation associated with sex is also expected to contribute substantially to the data variance, assuming that probes on the X and Y chromosomes are retained and assuming the study contains balanced numbers of each gender. In the case of CTH, a substantial proportion of the DNA methylome differs between major cell-types, e.g. between neutrophils and CD4+ T cells, or between epithelial and fibroblast cells [48–50], with over 80% differences in DNAm at individual loci being very common [49]. Thus, if cell-type composition of a tissue also varies between individuals, then this can be a major source of data variation, and indeed in the great majority of studies and irrespective of tissue-type, the top PC is most often driven by such variations in CTH (Fig. 1a). Depending on the phenotype or exposure of interest, the data variance driven by CTH could be substantially higher than that associated with the phenotype/exposure (Fig. 1a). For instance, components of variation associated with age or smoking are generally ranked much lower than those associated with CTH, with the corresponding variance often a factor of 5 or 10 lower than that associated with CTH. On the other hand, a phenotypic comparison between cancer and normal tissue is generally associated with a substantial remodelling of the DNA methylation landscape and would generally account for the top PC alongside CTH (Fig. 1a).

Hence, the importance of CTH adjustment is primarily a function of the relative data variance that can be attributed to the phenotype/exposure relative to CTH. For biomarker studies performed in easily accessible tissues like blood, serum, saliva, buccal swabs, vaginal swabs and cervical smears, the effect sizes associated with the typical phenotypes or exposures of interest are generally

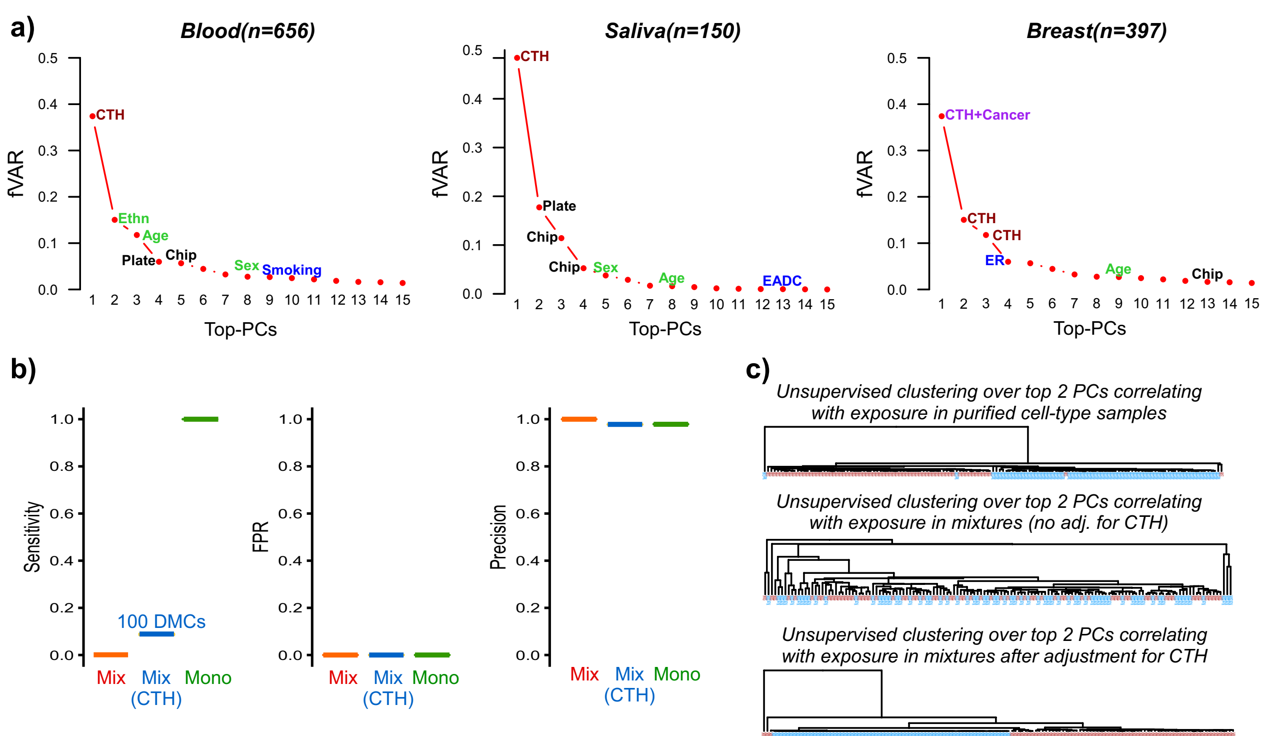

Fig. 1 The need to adjust for CTH in epigenome studies. **a** A comparison of the relative data variance, expressed as a fraction of the total variance accounted by the top 15 PCs (*y*-axis, fVAR), explained by each of the top-15 principal components (PCs) (*x*-axis) for 3 separate epigenome studies, with datapoints annotated to the main factor driving that PC. CTH = cell-type heterogeneity; Ethn = ethnicity; EADC = esophageal adenoma carcinoma; ER = estrogen receptor status. The tissue-type and number of samples in each study are given above plots. These plots derive from Illumina DNA methylation data from the following published works: Blood [51], Saliva [49] and Breast [52]. Briefly, the blood dataset is from healthy individuals, saliva samples are from EADC patients and matched healthy controls, and the breast tissue data is from breast cancers and normal-adjacent tissue. In the case of blood, the top-PC correlates with CTH, PC-2 correlates with ethnicity and PC-3 with age. **b** Sensitivity, false positive rate (FPR) and precision to detect 1000 simulated DMCs introduced in 139 monocyte samples from BLUEPRINT with an exposure distinguishing 69 cases from 70 controls. In each panel, we display the metrics when inferring DMCs from realistic mixtures of 3 cell-types (neutrophils, CD4+ T cells and monocytes) (Mix, red), when inferring DMCs from these same mixtures whilst adjusting for CTH (Mix CTH, blue) and when inferring DMCs from the purified monocyte samples (Mono, green). **c** For the same simulated data as in (**b**), the unsupervised hierarchical clustering obtained when clustering the 139 monocyte samples over the top 2 PCs correlating with the exposure (top panel), when clustering the 139 mixtures over the top 2 PCs correlating with the exposure without any adjustment for CTH (middle panel), and when clustering the 139 mixtures over the top 2 PCs correlating with the exposure after adjustment for CTH (lower panel). Note that in the second case, i.e. when clustering over the top 2 PCs derived from the mixtures without adjustment for CTH, that these PCs only exhibited very marginal associations with the exposure, hence why the samples do not segregate by exposure

quite small (usually less than 20% DNAm changes), which means that the majority of the DNAm variation would be associated with CTH, clearly justifying the need to adjust for that variation.

A second important reason for performing the adjustment for CTH is that ultimately we would like to know if an observed DNAm change is independent of a change in cell-type composition, and if so, in which specific cell-types that alteration is happening in. This is clearly important if the aim is to understand how an exposure affects a specific cell-function, or which particular signaling pathways and cell-types are affected. The power to retrieve such important knowledge would be much reduced if DMCs were swamped by those arising due to changes in cell-type composition. In other words, the list of DMCs, and subsequent Gene Set Enrichment Analysis (GSEA) results would be much more informative if derived from a CTH-adjusted analysis.

To illustrate the negative impact that not adjusting for CTH can have, we performed a simple simulation experiment: using Illumina 450k DNAm data from BLUEPRINT encompassing 139 matched neutrophil, CD4+ T cell and monocyte samples [50], we artificially generated 1000 DMCs in 69 of the 139 purified monocyte samples, with effect sizes drawn randomly between 0.01 and 0.2 (i.e. 1–20% DNAm changes). Effect sizes of 1–10% are typical of EWAS done in blood and in relation to factors such as smoking [25], obesity [53] or age [54], whilst larger effect sizes (> 10%) are often observed in solid tissues in relation to a disease like cancer [55]. We here did not consider effect sizes larger than 20% because although such larger DNAm changes are indeed observed in relation to cancer [56] or genotype [57], they are generally speaking not as frequent as those involving smaller effect sizes. We further required the DMCs to be cell-type specific markers. We generated mixtures of the 3 cell-types using the 139 neutrophil and 139 CD4+ T cell samples matched to the 139 monocyte specimens, with cell-type fractions drawn from empirical distributions derived from whole blood [58]. We then computed the sensitivity, false positive rate (FPR) and precision when identifying DMCs from the mixtures, adjusting and not adjusting for CTH, and benchmarked these measures against those obtained by identifying DMCs from the purified monocyte samples themselves. This simple analysis clearly shows that not adjusting for CTH would lack power to detect any DMCs in a cell-type that is not a major component of the tissue (e.g. monocytes in whole blood) (Fig. 1b). In contrast, adjusting for CTH, the power would be 10% at a precision of over 95%, which means that we would be able to detect 100 DMCs with 95% confidence (Fig. 1b). Although in theory one could in principle build a highly accurate predictor from just one true DMC (a scenario of low sensitivity but high precision as captured by our simulation model), it is clear that the ability to detect a larger number of true DMCs is important for building more robust predictors. A clear real-world example where adjustment for CTH is critical to achieve high prediction accuracies (AUC > 0.8) has been in the context of diagnostic and pre-diagnostic cancer classifiers built from cell-free DNAm markers in serum [18, 19, 30]. In this context, the need to adjust for CTH stems from the fact that most cell-free DNA in serum derives from circulating lymphocytes, hence candidate biomarkers are often pre-screened by comparing DNAm patterns between cancer-tissue and blood [18, 19, 30].

The need for CTH-adjustment would also be important in the context of unsupervised classification analyses where the aim would be to propose novel molecular subtypes of a particular disease. In Fig. 1c we provide a clear example of how an unadjusted unsupervised clustering analysis would fail to detect hidden DNAm variation associated with an exposure or factor of interest. Indeed, it is worth mentioning here again that all major Cancer Genome Atlas (TCGA) projects [41–46] have proposed molecular classifications which are largely confounded by underlying variations in cell-type composition. Thus, most proposed molecular classifications of disease do not necessarily reflect the specific patterns of DNAm change present in the individual cell-types of a tissue. Instead, they predominantly reflect variations in cell-type composition. For instance, mesenchymal and immune-cell enriched subtypes have been observed in many different solid cancer types, including brain [59] and breast [41], and that these subtypes reflect increased presence of fibroblasts and immune-cells is now well established [60].

In relation to all previous arguments in favour of adjusting for CTH, we should clarify that this is always meant as an analysis to be done *in addition* to the standard unadjusted one. Indeed, an unadjusted analysis can detect shifts in cell-type composition that are associated with the phenotype or exposure of interest, and which could reflect important biological processes that have diagnostic or prognostic value. A case in point is the increased infiltration of CD8+ T cells in breast tumor tissue, which correlates with good outcome [61]. To emphasize this further, CTH-adjusted and unadjusted analyses largely yield complementary results and insights. For instance, by performing a CTH-adjusted analysis of buccal-swab DNAm profiles it has been shown that smoking-associated DNAm changes may differ between those occurring in the immune and squamous epithelial subsets of the tissue, with important implications for our understanding of how such DNAm alterations could mediate the risk of smoking-related diseases such as mouth cancer or cardiovascular disease [62, 63]. However, it is also worth pointing out that there could be other scenarios, for instance if the same DNAm changes are happening in all the underlying cell-types or in the dominant cell-type(s) of a tissue, where CTH-adjusted and unadjusted analyses would yield very similar results. For instance, there is evidence that DNAm changes associated with aging [54] and SNPs [50, 64] are largely independent of cell-type. In general, our proposed strategy to perform both adjusted and unadjusted analyses would lead to three classes of DMCs: (1) a set that is only significant in the unadjusted analysis, (2) a set that is only significant in the adjusted analysis, and (3) a set that is significant in both. As discussed above, their biological interpretations would be different: set (1) would correspond to biomarkers that are driven entirely by shifts in cell-type composition, set (2) would correspond to DMCs that are occurring in at least one of the cell-types and therefore independent of shifts in cell-type composition, whilst set (3) is more complex as it can include DMCs that are occurring in all underlying cell-types, or only in a predominant cell-type, or DMCs that are driven by both changes in cell-type composition as well as changes in individual cell-types. As to which set of biomarkers to take forward, the key priority would be to test their reproducibility using independent validation sets. Any biomarker that is highly reproducible and displays high sensitivity and specificity has great clinical potential, irrespective of whether it belongs to set (1), (2) or (3). Overall, our recommendation and guideline is to always perform both adjusted and unadjusted analyses, because only by doing both can we obtain a more complete picture and understanding of the observed DNAm changes.

Adjusting for cell-type heterogeneity: feasibility

Most of the community is now well aware that adjustment for CTH can be accomplished with relative ease in tissues like blood or cord blood using what is known as a reference-based cell-type deconvolution algorithm [65, 66]. There are two elements necessary for a reference-based approach: (1) a DNAm reference matrix defined over a selected set of cell-type specific marker CpGs and all main cell-types within a tissue, (2) a statistical algorithm which, given this reference matrix and an independent DNAm profile of a sample, yields cell-type fraction estimates for all main cell-types in the given sample.

The main limitation of a reference-based approach is the availability or construction of a DNAm reference matrix. However, for specific tissues like blood and cord blood such DNAm reference matrices have been built [58, 65, 66] and adjustment for CTH in these tissues can be easily performed, at least at the resolution of 6–7 cell-types. It is worth highlighting here that the observed correlation between DNAm-based cell-type fraction estimates and those obtained experimentally (e.g. FACS/MACS-sorting) are excellent (typical $R^2 \sim 0.8$), to the degree that the DNAm-based estimates could be viewed as providing the better gold-standard [67, 68]. For other tissues like saliva or buccal swabs, which contain squamous buccal epithelial cells in addition to the immune cells found in blood, reference-based cell-type deconvolution is possible using algorithms such as HEpiDISH [49], a method that infers cell-type fractions in a stepwise hierarchical fashion, first inferring fractions for the total epithelial and total immune cell fractions using an appropriately built DNAm reference matrix, and subsequently estimating fractions for all immune cell subsets using a separate carefully constructed DNAm reference matrix. These DNAm reference matrices are all available from the EpiDISH R-package and webserver [69]. For more complex solid tissues such as lung, which also contain stromal cells such as fibroblasts, the DNAm reference matrix within HEpiDISH can also be used to infer total epithelial, total fibroblast and total immune-cell fractions [49, 70]. A similar strategy is used by the MethylCIBERSORT algorithm [71]. More recently, the EpiSCORE algorithm [72], which leverages the high resolution nature of a tissue-specific scRNA-Seq atlas to impute a corresponding DNAm reference matrix, can be used to infer cell-type fractions in a much wider range of tissue types, including brain, liver, pancreas and skin. While it is clear that for solid tissue types the currently available DNAm reference matrices are not complete or may contain false positives, it is worth emphasizing that the inference of cell-type fractions is mathematically speaking a relatively robust procedure, i.e. it can tolerate missing minor cell-types or a number of false positives in the reference matrix [49]: similar to a voting algorithm, as long as the majority of the entries in the DNAm reference matrix are approximately correct, this will allow the algorithm to converge onto a relatively good proxy of the true cell-type fractions in the sample. Thus, while the quality of DNAm reference matrices for solid tissues will undoubtedly improve in the near future, either through improved imputation strategies, or because of improvements in single-cell methylomics that will enable direct construction of these reference DNAm matrices, the current ones are reasonably accurate to allow adjustment for CTH, as shown in the case of breast or lung tissue [72]. Indeed, CTH-adjustment when identifying DMCs should yield complementary and valuable insights to those of an unadjusted analysis. In particular, the ability to disentangle DMCs arising from a shift in cell-type composition from those present in the actual cell-types of the tissue is an important step towards a better understanding of the role of epigenetics in disease risk and disease progression. In addition, by estimating cell-type fractions in a given sample, this also opens up the possibility to infer cell-type specific differential DNAm using algorithms such as CellDMC [62], TOAST [73] and HIRE [74]. For instance, such an approach has led to the identification of a novel Endothelial-to-Mesenchymal transition (EndoMT) DNAm signature in lung cancer [72], or a novel smoking-associated DNA hypermethylation signature associated with acute myeloid leukemia [75].

Conclusions

In summary, the scientific rationale for not adjusting for CTH when inferring biomarkers from complex tissues, or when proposing novel molecular classifications of disease is weak. Given that single-cell epigenomics will remain costly and unscalable to large numbers of individuals in the near future, computational methods offer a cheap and equally accurate means to adjust for CTH. We strongly recommend the use of such methodology for an improved and more complete interpretation of epigenomic and biomarker data.

Abbreviations

DNAm: DNA methylation; scRNA-Seq: Single-cell RNA-seq; CTH: Cell-type heterogeneity; FFPE: Formalin-fixed paraffin embedded; TCGA: The cancer genome atlas; DMC: Differentially methylated cytosine; EWAS: Epigenome-wide association study; GSEA: Gene set enrichment analysis; FACS: Fluorescence-cytometric activated cell sorting; MACS: Magnetic activated cell sorting.

Acknowledgements

We thank the Chinese Academy of Sciences, the Shanghai Institute for Nutrition and Health, and the NSFC for their support.

Authors' contributions

Article was conceived and written by AET. LQ performed analyses and contributed data for the figure. Both authors read and approved the final manuscript.

Funding

This project was funded by the National Natural Science Foundation of China (NSFC) Grant Numbers 31571359 and 31771464.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China. [2]UCL Cancer Institute, University College London, London WC1E 8BT, UK.

Received: 19 October 2021 Accepted: 21 February 2022
Published online: 28 February 2022

References

1. Regev A, et al. The human cell atlas. Elife. 2017. https://doi.org/10.7554/eLife.27041.
2. Stunnenberg HG, International Human Epigenome C, Hirst M. The international human epigenome consortium: a blueprint for scientific collaboration and discovery. Cell. 2016;167:1145–9. https://doi.org/10.1016/j.cell.2016.11.007.
3. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature. 2010;465:721–7. https://doi.org/10.1038/nature09230.
4. Zheng SC, Widschwendter M, Teschendorff AE. Epigenetic drift, epigenetic clocks and cancer risk. Epigenomics. 2016;8:705–19. https://doi.org/10.2217/epi-2015-0017.
5. Teschendorff AE, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20:440–6. https://doi.org/10.1101/gr.103606.109.
6. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nat Rev Genet. 2017;18:441–51. https://doi.org/10.1038/nrg.2017.32.
7. Yang Z, Jones A, Widschwendter M, Teschendorff AE. An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. Genome Biol. 2015;16:140. https://doi.org/10.1186/s13059-015-0699-9.
8. Shenker NS, et al. Epigenome-wide association study in the European prospective investigation into cancer and nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet. 2013;22:843–51. https://doi.org/10.1093/hmg/dds488.
9. van Veldhoven K, et al. Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. Clin Epigenet. 2015;7:67. https://doi.org/10.1186/s13148-015-0104-2.
10. Fasanelli F, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun. 2015;6:10192. https://doi.org/10.1038/ncomms10192.
11. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet. 2018;19:371–84. https://doi.org/10.1038/s41576-018-0004-3.
12. Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet. 2013;22:R7–15. https://doi.org/10.1093/hmg/ddt375.
13. Beck S, Rakyan VK. The methylome: approaches for global DNA methylation profiling. Trends Genet TIG. 2008;24:231–7. https://doi.org/10.1016/j.tig.2008.01.006.
14. Beck S. Taking the measure of the methylome. Nat Biotechnol. 2010;28:1026–8. https://doi.org/10.1038/nbt1010-1026.
15. Liu Y, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31:142–7. https://doi.org/10.1038/nbt.2487.
16. Wojdacz TK. Biological and methodological aspects of assessment of locus specific de novo methylation in blood. Biomark Med. 2015;9:1291–9. https://doi.org/10.2217/bmm.15.83.
17. Taryma-Lesniak O, Sokolowska KE, Wojdacz TK. Current status of development of methylation biomarkers for in vitro diagnostic IVD applications. Clin Epigenet. 2020;12:100. https://doi.org/10.1186/s13148-020-00886-6.
18. Shen SY, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature. 2018;563:579–83. https://doi.org/10.1038/s41586-018-0703-0.
19. Chen X, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat Commun. 2020;11:3475. https://doi.org/10.1038/s41467-020-17316-z.
20. Thirlwell C, et al. Genome-wide DNA methylation analysis of archival formalin-fixed paraffin-embedded tissue using the Illumina Infinium HumanMethylation27 BeadChip. Methods. 2010;52:248–54. https://doi.org/10.1016/j.ymeth.2010.04.012.
21. Lechner M, et al. Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. Genome Med. 2013;5:15. https://doi.org/10.1186/gm419.

22. Daugaard I, Kjeldsen TE, Hager H, Hansen LL, Wojdacz TK. The influence of DNA degradation in formalin-fixed, paraffin-embedded (FFPE) tissue on locus-specific methylation assessment by MS-HRM. Exp Mol Pathol. 2015;99:632–40. https://doi.org/10.1016/j.yexmp.2015.11.007.
23. Sandoval J, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenet Off J DNA Methylation Soc. 2011;6:692–702.
24. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8:389–99. https://doi.org/10.2217/epi.15.114.
25. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clin Epigenet. 2015;7:113. https://doi.org/10.1186/s13148-015-0148-3.
26. Joehanes R, et al. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet. 2016;9:436–47. https://doi.org/10.1161/CIRCGENETICS.116.001506.
27. Das J, Idh N, Sikkeland LIB, Paues J, Lerm M. DNA methylome-based validation of induced sputum as an effective protocol to study lung immunity: construction of a classifier of pulmonary cell types. Epigenet Off J DNA Methylation Soc. 2021. https://doi.org/10.1080/15592294.2021.1969499.
28. Angermueller C, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016;13:229–32. https://doi.org/10.1038/nmeth.3728.
29. Smallwood SA, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11:817–20. https://doi.org/10.1038/nmeth.3035.
30. Kang S, et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biol. 2017;18:53. https://doi.org/10.1186/s13059-017-1191-5.
31. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15:R31. https://doi.org/10.1186/gb-2014-15-2-r31.
32. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics. 2017;9:757–68. https://doi.org/10.2217/epi-2016-0153.
33. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. Nat Rev Genet. 2018;19:129–47. https://doi.org/10.1038/nrg.2017.86.
34. Cheishvili D, et al. Identification of an epigenetic signature of osteoporosis in blood DNA of postmenopausal women. J Bone Miner Res: Off J Am Soc Bone Miner Res. 2018;33:1980–9. https://doi.org/10.1002/jbmr.3527.
35. Topham L, et al. The methyl donor S-adenosyl methionine reverses the DNA methylation signature of chronic neuropathic pain in mouse frontal cortex. Pain Rep. 2021;6:e944. https://doi.org/10.1097/PR9.0000000000000944.
36. Cao-Lei L, Elgbeili G, Szyf M, Laplante DP, King S. Differential genome-wide DNA methylation patterns in childhood obesity. BMC Res Notes. 2019;12:174. https://doi.org/10.1186/s13104-019-4189-0.
37. Li QS, et al. Association of peripheral blood DNA methylation level with Alzheimer's disease progression. Clin Epigenet. 2021;13:1–6.
38. Li J, et al. A genomic and epigenomic atlas of prostate cancer in Asian populations. Nature. 2020;580:93–9. https://doi.org/10.1038/s41586-020-2135-x.
39. Capper D, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018;555:469–74. https://doi.org/10.1038/nature26000.
40. Hoadley KA, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173:291–304. https://doi.org/10.1016/j.cell.2018.03.022.
41. Koboldt DC, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70. https://doi.org/10.1038/Nature11412.
42. Cancer Genome Atlas Research N, et al. Integrated genomic characterization of oesophageal carcinoma. Nature. 2017;541:169–75. https://doi.org/10.1038/nature20805.
43. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7. https://doi.org/10.1038/nature11252.
44. Cancer Genome Atlas Research Network. Integrated genomic characterization of pancreatic ductal adenocarcinoma. Cancer Cell. 2017;32:185–203. https://doi.org/10.1016/j.ccell.2017.07.007.
45. Alizadeh AA, et al. Toward understanding and exploiting tumor heterogeneity. Nat Med. 2015;21:846–53. https://doi.org/10.1038/nm.3915.
46. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25. https://doi.org/10.1038/nature11404.
47. van Dongen J, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat Commun. 2016;7:11115. https://doi.org/10.1038/ncomms11115.
48. Roadmap Epigenomics Consortium, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30. https://doi.org/10.1038/nature14248.
49. Zheng SC, et al. A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. Epigenomics. 2018;10:925–40. https://doi.org/10.2217/epi-2018-0037.
50. Chen L, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell. 2016;167:1398–414. https://doi.org/10.1016/j.cell.2016.10.026.
51. Hannum G, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359–67. https://doi.org/10.1016/j.molcel.2012.10.016.
52. Teschendorff AE, et al. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nat Commun. 2016;7:10478. https://doi.org/10.1038/ncomms10478.
53. Wahl S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541:81–6. https://doi.org/10.1038/nature20784.
54. Zhu T, Zheng SC, Paul DS, Horvath S, Teschendorff AE. Cell and tissue type independent age-associated DNA methylation changes are not rare but common. Aging. 2018;10:3541–57. https://doi.org/10.18632/aging.101666.
55. Irizarry RA, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41:178–86. https://doi.org/10.1038/ng.298.
56. Toyota M, et al. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci USA. 1999;96:8681–6.
57. Gaunt TR, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17:61. https://doi.org/10.1186/s13059-016-0926-z.
58. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinform. 2017;18:105. https://doi.org/10.1186/s12859-017-1511-5.
59. Parsons DW, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008;321:1807–12. https://doi.org/10.1126/science.1164382.
60. Moffitt RA, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet. 2015;47:1168–78. https://doi.org/10.1038/ng.3398.
61. Ali HR, et al. Association between CD8+ T-cell infiltration and breast cancer survival in 12,439 patients. Ann Oncol Off J Eur Soc Med Oncol ESMO. 2014;25:1536–43. https://doi.org/10.1093/annonc/mdu191.
62. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. Nat Methods. 2018;15:1059–66. https://doi.org/10.1038/s41592-018-0213-x.
63. Pan S, et al. DNA methylome analysis reveals distinct epigenetic patterns of ascending aortic dissection and bicuspid aortic valve. Cardiovasc Res. 2017;113:692–704. https://doi.org/10.1093/cvr/cvx050.
64. Hawe JS, et al. Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. Nat Genet. 2022;54:18–29. https://doi.org/10.1038/s41588-021-00969-x.
65. Houseman EA, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 2012;13:86. https://doi.org/10.1186/1471-2105-13-86.
66. Gervin K, et al. Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. Clin Epigenet. 2019;11:125. https://doi.org/10.1186/s13148-019-0717-y.

67. Salas LA, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19:64. https://doi.org/10.1186/s13059-018-1448-7.
68. Koestler DC, et al. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinform. 2016;17:120. https://doi.org/10.1186/s12859-016-0943-7.
69. Zheng SC, et al. EpiDISH web server: epigenetic dissection of intra-sample-heterogeneity with online GUI. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz833.
70. Zheng SC, et al. Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. Nat Methods. 2017;14:216–7. https://doi.org/10.1038/nmeth.4187.
71. Chakravarthy A, et al. Pan-cancer deconvolution of tumour composition using DNA methylation. Nat Commun. 2018;9:3220. https://doi.org/10.1038/s41467-018-05570-1.
72. Teschendorff AE, Zhu T, Breeze CE, Beck S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. Genome Biol. 2020;21:221. https://doi.org/10.1186/s13059-020-02126-9.
73. Li Z, Wu Z, Jin P, Wu H. Dissecting differential signals in high-throughput data from complex tissues. Bioinformatics. 2019;35:3898–905. https://doi.org/10.1093/bioinformatics/btz196.
74. Luo X, Yang C, Wei Y. Detection of cell-type-specific risk-CpG sites in epigenome-wide association studies. Nat Commun. 2019;10:3113. https://doi.org/10.1038/s41467-019-10864-z.
75. You C, et al. A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. Nat Commun. 2020;11:4779. https://doi.org/10.1038/s41467-020-18618-y.

Publisher's Note