

# Reliability and correlation of mixture cell correction in methylomic and transcriptomic blood data

**Boris Chaumette** (✉ [boris.chaumette@mcgill.ca](mailto:boris.chaumette@mcgill.ca))

McGill University <https://orcid.org/0000-0002-1313-2788>

**Oussama Kebir**

INSERM U1266

**Patrick A. Dion**

McGill University

**Guy A. Rouleau**

McGill University

**Marie-Odile Krebs**

INSERM U1266

---

## Research note

**Keywords:** DNA, RNA, leucocytes, chip, sequencing

**Posted Date:** January 3rd, 2020

**DOI:** <https://doi.org/10.21203/rs.2.12836/v2>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

---

**Version of Record:** A version of this preprint was published at BMC Research Notes on February 12th, 2020. See the published version at <https://doi.org/10.1186/s13104-020-4936-2>.

## Abstract

**Objectives:** DNA methylome and RNA transcriptome studies are growing, but investigators have to consider the cell type composition of tissues used. In blood samples, for instance, data reflect the picture of a mixture of cells. Specialized algorithms address the cell-type heterogeneity issue. We tested if the estimation of cell-type composition in the blood is correlated between two heterogeneous datasets.

**Results:** We used methylome and transcriptome datasets obtained from a cohort of ten individuals whose blood was sampled at two different timepoints. We examined how the cell composition derived from these omics correlated with each other using “CIBERSORT” for transcriptome and “EstimateCellCounts function” for the methylome. The correlation coefficients between the two omic datasets range from 0.45 to 0.81. There was a minimal correlation between the two timepoints using the same algorithm. Our results suggest that a posteriori correction of a mixture of cells present in blood samples is reliable. Using an omic dataset to correct a second dataset for relative fractions of cells seems to be appropriate, but only when the samples are simultaneously collected. Such approaches could be beneficial when there are difficulties to control the cell types in the second dataset, even when the sample size is limited.

## Introduction

*Omics* technologies are growing in many biomedical fields. In some of these fields, like psychiatry and neurology, access to tissues of interest is difficult in living humans. Consequently, a number of studies have come to rely on blood samples as an alternative source of material from patients (E. Andres Houseman et al. 2015). However, DNA methylation and gene expression profiles are relatively specific to a particular tissue and cell types, leading to frequent criticisms in regard to the reliability of results obtained from blood samples. Furthermore, one of the issues associated with the use of blood samples is that these are composed of various cell types. Consequently, the DNA methylation and RNA profiles that are obtained from blood samples are the results of a mixture of profiles. To detect statistically significant differences in methylation or gene expression data that are related to the experiment design, and not driven by the underlying variability and heterogeneity in cell-type composition, new algorithms were recently proposed to compute and address the issue. Yet, sometimes such correction cannot be applied because the cell counts composition can only be estimated from a whole-genome *omic* dataset (e.g. methylome and transcriptome) and not from a candidate gene study (e.g. Q-PCR or study of the methylation of one promoter). We want to test if the estimation of cell counts using one modality is reliable to correct the data obtained using a second modality. In a cohort of ten individuals, assessed at two different times (at baseline=T0 and one-year after=T1), we tested the correlation of two algorithms (EstimateCellCounts and CIBERSORT) to retrospectively estimate the cell counts of methylomic and transcriptomic datasets obtained from the same blood samples. We also tested for the longitudinal stability of the cell counts in the same individuals.

## Main Text

The participants were recruited through the ICAAR cohort (PHRC, AOM-07-118, see Chaumette et al. 2016 for a description of the cohort). For the methylomic analysis, genomic DNA (500 ng) was extracted from whole blood, treated with sodium bisulfite using the EZ-96DNA Methylation KIT (Catalog No D5004, Zymo Research, USA) following the manufacturer’s standard protocol. Then the DNA methylation was studied using the Illumina Infinium HumanMethylation450 BeadChip (Illumina, San Diego, CA, USA) which contains 485,000 probes across

the genome. The Illumina GenomeStudio software (Illumina, San Diego, CA, USA) was used to assess the signal intensities of each probe. The R Minfi package (Aryee *et al.*, 2014) enabled data quality checks and normalization. This *omic* data was previously described in Kebir *et al.* 2017. We used the EstimateCellCounts algorithm (Eugene Andres Houseman *et al.* 2012), which is implemented in Minfi package, to assess the abundances of various cell types in the methylomic dataset: "B cell", "CD4T", "CD8T", "Eosinophils", "Granulocytes", "Monocytes" and "NK cells". We decide to not consider Eosinophils as their estimates are effectively all zero (**Supplementary Table 1**).

For the transcriptomic analysis, total RNA was extracted from blood samples (PAXgene tubes) following the manufacturer protocol of PAXgene Blood RNA kit (QIAGEN) and a QIacube robot. Then the *omic* data was obtained from the sequencing of TruSeq libraries. The methods for RNA sequencing are detailed in Chaumette *et al.* 2019. Briefly, blood total RNA was processed using the mRNA-Seq Sample Prep Kit (Illumina) before poly(A) RNA was isolated, fragmented and purified by ethanol precipitation. The libraries were prepared using the TruSeq Stranded mRNA kit. Paired-end 75-bp sequencing runs were performed on an Illumina HiSeq 2000 instrument at over 80 million reads per sample. The Illumina software RTA1.12.4.2/HCS1.4.8 converted this fluorophore information to sequence data and FASTA files cell. Quality control was performed using ShortRead package for R (Morgan *et al.*, 2009). FASTA files were aligned to the reference genome (hg19) using TopHat2 to generate BAM files (Kim *et al.*, 2013). A matrix of read counts was then created using HTSeq (Anders *et al.*, 2015). Then we used the CIBERSORT algorithm (Newman *et al.*, 2015) to assess the abundances of 22 cell types in the transcriptomic data using the gene expression data and the LM22 signature gene file (default file). We only retained, for further analyses, the 6 cell types shared with the EstimateCellCounts algorithm (**Supplementary Table 1**).

Spearman's and Person's correlations between cell populations estimated using the methylomic data and the transcriptomic data were tested using SPSS software (IBM SPSS Statistics for Windows, Version 24, IBM Corp., Armonk, NY). The significance threshold was set to an alpha-risk of 0.05 and multiple testing corrections were done using the Benjamini-Hochberg method.

Fractions of cells derived from the methylomic and the transcriptomic datasets were computed and the results are presented in **Supplementary Table 1**. For the comparison between the two modalities, using all the samples, all Spearman's correlations were significant with a coefficient moderate to high (0.45 to 0.81). The weakest correlation was obtained for CD8T and the highest for CD4T (**see Table**). All correlations remained significant after multiple-testing correction by the Benjamini-Hochberg method. Pearson's correlations are reported in **Supplementary Table 2**. We then tested the correlation in the same individual between T0 and T1. As expected, due to the longitudinal variation of cells, there were fewer significant correlations with weaker coefficients (**see Table**).

These results are demonstrative that cell heterogeneity can be reliably computed using bioinformatic algorithms like CIBERSORT for transcriptomic data and EstimateCellCounts function in R for methylomic data. Moreover, such *a posteriori* corrections are easier to apply than *a priori* adjustment that would involve flow cytometry or MACS microbeads cell separations which are difficult to perform when the samples have been previously frozen.

As the correlation between transcriptomic and methylomic data is strong, it is reasonable to suggest using the first dataset to predict the cell composition of the second dataset if the reference-database is not provided for the later. Second, we can use the cell counts derived from a methylomic dataset to obtain cell mixture composition and then correct the targeted transcriptomic data (e.g. Q-PCR) where the cell counts could not be obtained. Inversely, the genome-wide transcriptome could be used to correct cell counts in a targeted methylation study (e.g.

pyrosequencing). Cell composition can be retrospectively estimated and correlated across the different sets of data, even in a dataset with limited sample size, but only when the samples are simultaneously collected.

## Limitations

The main limitation of our report is the sample size that is very limited with only 10 individuals. However, the strong correlations obtained demonstrated the high reliability of the algorithms even for small studies. We have to acknowledge that longitudinal correlations are based in a smaller dataset (10 samples with 2 timepoints) than the correlations between the two modalities (20 samples); this may have decreased the significance of the longitudinal correlations.

Only 6 cell types were shared between the two pipelines. However, these cell types are those mainly present in blood and worthy to be considered for cell mixture correction in methylomic or transcriptomic analyses.

Another limitation is that we have only performed our examinations from datasets that were derived from human blood. We cannot extrapolate on how significant such correlations would be in other tissues or species. Indeed, the two algorithms used here are anchored to reference-databases of methylomic or transcriptomic profiles obtained from major cell-types deemed to be present in the tissue of interest and uses this reference to infer sample-specific cell-type proportions. Sometimes however the reference of interest is not available for a particular tissue or species, but other algorithms can perform reference-free estimates (e.g. R package RefFreeEWAS for methylomic data) (Houseman *et al.*, 2014).

Then, a constraint to correct a dataset with the cell mixture composition estimated from another dataset is that the samples for the assessment should be taken simultaneously from the individuals used in a study. It appears to be not reliable to correct data from one modality if the sample for the second modality is taken at another time point.

Finally, it is important to remember that even if some bioinformatics corrections can be applied to the *omic* datasets of a particular tissue, the ideal tissue for a specific condition should ever be extensively discussed. Therefore, additional elements should also be considered when epigenetic and transcriptomic studies are being designed, including the batch effect.

## Declarations

### Ethics approval and consent to participate

The ICAAR study (PHRC, AOM-07-118) was approved by the institutional ethics committee “Comité de protection des personnes, Ile-de-France III, Paris, France” and written informed consent was obtained from all participants in accordance with the Declaration of Helsinki.

### Consent to publish

NA

### Availability of data and materials

Availability from the corresponding author on reasonable request.

## Competing interests

All the authors declare they have no competing interests related to this work.

## Funding

This work was supported by a grant by the French Government's Agence Nationale pour la Recherche (ANR, 08-MNP-007) and a grant by the French Ministry of Health's Programme Hospitalier de Recherche Clinique (PHRC, AOM-07-118). The Centre Hospitalier Sainte-Anne promoted the study. Additional financial support was obtained from the Institut National de la Santé et de la Recherche Médicale (INSERM), Université Paris Descartes (recurrent funding). Boris Chaumette receives a postdoctoral fellowship from the Healthy Brains for Healthy Lives project (Talent program) and a grant from the Bettencourt-Schueller foundation (CCA-INSERM-Bettencourt). The sponsors had no role in the design and conduct of the study, in the collection, management, analysis or interpretation of the data, in the preparation, review or approval of the manuscript, or in the decision to submit the manuscript for publication.

## Authors' Contributions

BC and OK designed the study. GR and MOK obtained funding and supervised the study. BC, OK and MOK collected the data. BC analyzed the data. BC, OK, PAD interpreted the data. BC, OK, PAD drafted the report. All authors contributed to and have approved the final manuscript.

## Acknowledgments

The authors thank all the participants to the ICAAR cohort as well as the practitioners from the C'JAAD team (Centre d'Evaluation pour Jeunes Adultes et Adolescents), at GHU Paris Psychiatrie et Neurosciences.

## References

- Anders, Simon, Paul Theodor Pyl, et Wolfgang Huber. 2015. « HTSeq—a Python Framework to Work with High-Throughput Sequencing Data ». *Bioinformatics (Oxford, England)* 31 (2): 166-69. <https://doi.org/10.1093/bioinformatics/btu638>.
- Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, et Rafael A. Irizarry. 2014. « Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays ». *Bioinformatics (Oxford, England)* 30 (10): 1363-69. <https://doi.org/10.1093/bioinformatics/btu049>.
- Chaumette, Boris, Oussama Kebir, Célia Mam-Lam-Fook, Yannick Morvan, Julie Bourgin, Bill P. Godsil, Marion Plaze, Raphaël Gaillard, Thérèse M. Jay, et Marie-Odile Krebs. 2016. « Salivary Cortisol in Early Psychosis: New Findings and Meta-Analysis ». *Psychoneuroendocrinology* 63: 262-70. <https://doi.org/10.1016/j.psyneuen.2015.10.007>.
- Chaumette, Boris, Oussama Kebir, Juliette Pouch, Bertrand Ducos, Fekrije Selimi, ICAAR study group, Raphael Gaillard, et Marie-Odile Krebs. 2019. « Longitudinal Analyses of Blood Transcriptome During Conversion to Psychosis ». *Schizophrenia Bulletin* 45 (1): 247-55. <https://doi.org/10.1093/schbul/sby009>.

Houseman, E. Andres, Stephanie Kim, Karl T. Kelsey, et John K. Wiencke. 2015. « DNA Methylation in Whole Blood: Uses and Challenges ». *Current Environmental Health Reports* 2 (2): 145-54. <https://doi.org/10.1007/s40572-015-0050-3>.

Houseman, Eugene Andres, William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, et Karl T. Kelsey. 2012. « DNA Methylation Arrays as Surrogate Measures of Cell Mixture Distribution ». *BMC Bioinformatics* 13: 86. <https://doi.org/10.1186/1471-2105-13-86>.

Houseman, Eugene Andres, John Molitor, et Carmen J. Marsit. 2014. « Reference-Free Cell Mixture Adjustments in Analysis of DNA Methylation Data ». *Bioinformatics (Oxford, England)* 30 (10): 1431-39. <https://doi.org/10.1093/bioinformatics/btu029>.

Kebir, O., B. Chaumette, F. Rivollier, F. Miozzo, L. P. Lemieux Perreault, A. Barhdadi, S. Provost, et al. 2017. « Methyloomic Changes during Conversion to Psychosis ». *Molecular Psychiatry* 22 (4): 512-18. <https://doi.org/10.1038/mp.2016.53>.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, et Steven L. Salzberg. 2013. « TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions ». *Genome Biology* 14: R36. <https://doi.org/10.1186/gb-2013-14-4-r36>.

Morgan, Martin, Simon Anders, Michael Lawrence, Patrick Aboyoun, Hervé Pagès, et Robert Gentleman. 2009. « ShortRead: A Bioconductor Package for Input, Quality Assessment and Exploration of High-Throughput Sequence Data ». *Bioinformatics (Oxford, England)* 25 (19): 2607-8. <https://doi.org/10.1093/bioinformatics/btp450>.

Newman, Aaron M., Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn, et Ash A. Alizadeh. 2015. « Robust Enumeration of Cell Subsets from Tissue Expression Profiles ». *Nature Methods* 12 (5): 453-57. <https://doi.org/10.1038/nmeth.3337>.

## Table

**Table:** Spearman's correlation between cell counts computed using CIBERSORT and EstimateCellCounts

Cell type	Comparison between cell counts obtained from methylomic and transcriptomic dataset (n=20)			Longitudinal correlation in the methylomic dataset (n=10)		Longitudinal correlation in the transcriptomic dataset (n=10)	
	Coefficient	Significance	Significance after BH correction	Coefficient	Significance	Coefficient	Significance
CD8T	0.45	<b>0.044</b>	<b>0.044</b>	0.79	<b>0.007</b>	0.79	<b>0.006</b>
CD4T	0.81	<b>&lt;10<sup>-4</sup></b>	<b>&lt;10<sup>-4</sup></b>	0.72	<b>0.019</b>	0.81	<b>0.005</b>
NK	0.67	<b>1.10<sup>-3</sup></b>	<b>2.10<sup>-3</sup></b>	0.55	0.100	0.26	0.467
B cell	0.63	<b>3.10<sup>-3</sup></b>	<b>0.005</b>	0.68	<b>0.032</b>	0.52	0.128
Monocytes	0.55	<b>0.012</b>	<b>0.016</b>	0.82	<b>0.004</b>	0.79	<b>0.006</b>
Granulocytes	0.76	<b>0.017</b>	<b>0.019</b>	0.30	0.405	0.21	0.556
Lymphocytes (all)	0.76	<b>&lt;10<sup>-4</sup></b>	<b>3.10<sup>-4</sup></b>	0.38	0.276	0.66	<b>0.038</b>

NK: natural killer; CD8T: lymphocytes CD8+; CD4T: lymphocytes CD4+; B cell : lymphocytes B; BH: Benjamini-Hochberg

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ArticleCellcomparisonsupplementary.docx](#)