Computational and Systems Biology

Alleviating cell-free DNA sequencing biases with optimal transport

Reviewed Preprint
v1 • August 5, 2024
Not revised

Antoine Passemiers, Tatjana Jatsenko, Adriaan Vanderstichele, Pieter Busschaert, An Coosemans, Dirk Timmerman, Diether Lambrechts, Daniele Raimondi, Joris Robert Vermeesch, Yves Moreau

Dynamical Systems, Signal Processing and Data Analytics (STADIUS), KU Leuven, Leuven, Belgium • Laboratory for Cytogenetics and Genome Research, Department of Human Genetics, KU Leuven, Leuven, Belgium • Department of Gynaecology and Obstetrics, University Hospitals Leuven, Leuven, Belgium • Division of Gynaecological Oncology, Leuven Cancer Institute, KU Leuven, Leuven, Belgium • Department of Oncology, Laboratory of Tumor Immunology and Immunotherapy, Leuven Cancer Institute, Leuven, Belgium • Center for Cancer Biology, VIB, Leuven, Belgium • Laboratory for Translational Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium

Abstract

Cell-free DNA (cfDNA) is a rich source of biomarkers for various (patho)physiological conditions. Recent developments have used Machine Learning on large cfDNA data sets to enhance the detection of cancers and immunological diseases. Preanalytical variables, such as the library preparation protocol or sequencing platform, are major confounders that influence such data sets and lead to domain shifts (i.e., shifts in data distribution as those confounders vary across time or space). Here, we present a domain adaptation method that builds on the concept of optimal transport, and explicitly corrects for the effect of such preanalytical variables. Our approach can be used to merge cohorts representative of the same population but separated by technical biases. Moreover, we also demonstrate that it improves cancer detection via Machine Learning by alleviating the sources of variation that are not of biological origin. Our method also improves over the widely used GC-content bias correction, both in terms of bias removal and cancer signal isolation. These results open perspectives for the downstream analysis of larger data sets through the integration of cohorts produced by different sequencing pipelines or collected in different centers. Notably, the approach is rather general with the potential for application to many other genomic data analysis problems.

eLife assessment

This study presents a **useful** computational data preprocessing methodology for de-biasing/denoising high-throughput genomic signals using optimal transport techniques. The evidence supporting the claims of the authors is, however, in parts **incomplete**, with a partially insufficient experimental setup for validation. The method needs to be be compared with other algorithms, using datasets that demonstrate broad applicability of the algorithm presented. The work could be of interest to scientists in the field of computational genomics.

https://doi.org/10.7554/eLife.98116.1.sa2

1 Introduction

Cell-free DNA (cfDNA) has been identified as a promising source of biomarkers for the detection of fetal aneuploidy [1, 2], transplant rejection [3], incipient tumours [4], autoimmune disease [5] or inflammatory disease [6]. While cfDNA fragments in healthy individuals primarily originate from the apoptotic release of DNA from cells of hematopoietic origin [7], these fragments can also be of tumoural origin in cancer patients. While most clinical applications of cfDNA in oncology focus on finding tumour mutations (e.g., using a targeted panel of cancer driver variants) [8, 9], a lot of research has been carried out around the analysis of coverage and fragmentome profiles. Indeed, the copy number aberrations (CNAs) carried by the genome of cancerous cells are detectable by low-coverage whole-genome sequencing and downstream analysis of cfDNA from cancer patients [10, 11]. Because cfDNA can be collected in a non- or minimally-invasive manner (e.g., blood draw), and thanks to the cost-effectiveness of shallow whole-genome sequencing, liquid biopsies are a valuable candidate for population-wide cancer screening [7, 4] and diagnosis, and considerable research has been devoted to assessing their clinical utility [12].

Fragmentomic analysis of cfDNA offers the possibility to improve its use as a sensitive biomarker for cancer detection [13, 14], as cfDNA fragments mirror the chromatin accessibility, nucleosome positioning and degradation pattern of their tissue of origin [15, 16]. For this reason, CNA calling can be complemented with fragmentation profile analysis based on fragment length, as well as positional information [13, 17, 18, 19]. Indeed, the distribution of cfDNA fragment lengths is shifted downward in circulating tumour-derived DNA (ctDNA), supporting signatures of their cellular origin [15, 20].

Beyond fragmentomics, methylation patterns are indicative of the tissue of origin, and methylation signatures have been exploited for sensitive cancer detection and tissue-of-origin identification [21]. Finally, recent work has been devoted toward integrating multiple properties of cfDNA within a single multimodal analysis approach, including variant calling, CNAs, methylation and fragmentomic profiles, as well as other complementary sources of information such as nucleosome-depleted region (NDR) profiles [22] or fusion gene detection [23].

However, the development of reliable models that are predictive of relevant clinical outcomes (for example, diagnosis) remains challenging because of the limited number of available cases (especially for disorders with smaller incidence rate), the high dimensionality of cfDNA data and the various sources of biases related to preanalytical settings. These latter biases mainly arise when protocol changes are introduced over time or between different centers. For example, the choice of blood collection tube might affect cfDNA concentrations and the prominence of leukocyte DNA [24, 25], which could in turn affect the detection of low-frequency variations from cancerous cells. Other preanalytical factors include the delay before centrifugation and protocols for plasma separation, and plasma storage conditions [26]. For example, two-step centrifugation reduces contamination by genomic DNA thanks to reduced white blood cell lysis, compared to one-step centrifugation [27]. Moreover, some DNA extraction platforms, such as Maxwell and QIAsymphony, preferentially isolate short fragments over long ones [28]. The choice of library preparation kit directly affects the distribution of read counts, as the polymerase enzymes used in these kits have different levels of efficiency in amplifying fragments with low vs. high GC-content [29]. For instance, some library preparation kits (e.g., Nextera XT) introduce a bias toward low-GC regions [30]. Multiplexed sequencing without suitable dual indexing can result in barcode swapping, and the swapping rates are platform-dependent (e.g., higher on HiSeqX or 4000 compared to MiSeq) [31]. Index swapping mechanism is caused both by multiplex PCR and flow cell chemistry, and is responsible for cross-contamination within the same

pool [32]. Finally, the choice of sequencing instrument also plays a role. For example, different GC-content bias profiles have been reported for Illumina MiSeq and NextSeq platforms, compared to PacBio or HiSeq [33].

In this work, we focus on the bias correction of genome-wide copy-number (i.e., GIPseq [4]) profiles based on normalised read counts. The aforementioned preanalytical settings can affect the read counts, for example through differential coverage of regions differing by their GC content, thus invalidating direct statistical analysis (e.g., using *z*-scores) of CNA profiles. Moreover, these distributional shifts [34] are not properly handled by classical Machine Learning algorithms and are responsible for performance drops on test sets. Mitigating these biases is therefore of utmost importance in strengthening biological signals and guaranteeing performance on unseen data. Such a task typically falls in the category of domain adaptation (DA) [35] problems, where computational methods are needed to compensate for the fact that a given model is tested on data drawn from a different distribution than the ones on which it has been trained. In this work, we will refer to the samples being corrected as belonging to the source domain, while the fixed data lies in the target domain. We restricted ourselves to unsupervised DA, where the variable of interest (e.g., whether an individual is affected by a certain condition) is unknown. Such annotations are not necessarily available, especially for rarer diseases. Also, already-existing methods (GC correction) don't require such information and are widely applicable, as they can be applied in a sample-wise fashion. This is highly relevant due to GC and sequencing biases not only operating at the domain-level, but also at the individual level [36]. When multiple source domains coexist, the problem is referred to as a domain generalisation problem [37]. Since multiple domain shifts occurred in our data sets over time, our own method falls under this category.

Previous work on the bias correction of copy-number profiles has mostly been directed toward GC-content and mappability bias correction. Benjamini and Speed [38] originally categorised these methods as single position models, fragmentation models, read models, full-fragment models and global models. An example of global model is the widely-used LOESS GC-content bias correction [39, 40], which decorrelates the per-bin GC-content percentage from the normalised read counts based on local regressions. ichorCNA [41] is a tool for calling CNA from read counts, that internally performs mappability and GC-correction in a similar way. BEADS [36] falls into the category of read models, as it re-weights individual reads based on their GC-content before computing their per-bin counts. The single position model from Benjamini and Speed [38] relies on the computation of the mean fragment count for each GC stratum, by considering all the mappable positions along the genome having similar GC-content. Finally, the LIQUORICE algorithm [23] operates at the fragment-level, by computing a coverage weight for each position covered by each fragment. More recently, distance learning and k-nearest neighbours have been proposed [42] to correct coverage profiles. As opposed to previous work, the latter approach exploits information from the whole data set to correct each individual sample.

On the Machine Learning side, previous work on unsupervised DA includes the following. Discriminator-free domain adversarial learning [43] uses domain adversarial learning [44] to obtain a common representation space for all domains. Kernel mean matching [45] aims at matching the higher-order moments of the underlying distributions using kernel functions. Multilevel domain adaptive learning matches the distributions at each intermediate layer of the neural network in a hierarchical fashion [46]. Reconstruction-based methods, such as Cycle-Consistent Adversarial Domain Adaptation (CyCADA) [47], reconstruct samples from the target domain using the samples from the source domain as input. It should be noted that most existing methods use a latent space to represent the samples, which means that the debiased representation is not directly interpretable, which runs afoul of ubiquitous need for interpretability and explainability in human genetics [48]. A key motivation for our work is thus to design a domain adaptation method that adjusts cfDNA profiles in a transparently

interpretable manner, by operating at the read count level (i.e., without having recourse to a latent space as domain adversarial methods would) and preserving the *z*-scores produced by the original data.

In this article, we present an advanced data normalisation method for cell-free DNA sequencing data building on optimal transport (OT) theory [49, 50]. OT builds on strong mathematical bases and allows to define a patient-to-patient relationship across domains without the need to build a common latent representation space, as mostly done in the DA field. This enables high interpretability, as samples can be corrected in the original data space (e.g., read counts) directly. Because we originally designed this approach for the correction of normalised read counts within predefined bins, it falls under the category of "global models" according to the Benjamini/Speed classification [38]. In summary, we aim at correcting and mapping the data distribution from a source domain onto the data distribution obtained in a target domain, to enable more robust downstream analysis. As the ultimate goal is to go beyond the classical case–control setting and build models capable of accurately processing data from various sources, we hypothesised that bias removal is a good candidate to increase the effective size of available data sets through their fusion and thus benefit from the scalability of Machine Learning models and enhance their performance. This flexibility would, among other things, reduce the need for laboratories to consistently build new reference sets, as well as enable high reusability of older samples or data collected in unrelated studies. In this article, we report enhanced cancer detection with prior domain adaptation and show that cohorts can be corrected to match the same distribution while preserving the original biological signals (e.g., copy number aberrations) in each patient.

2 Results & discussion

In each of the following experiments, we compared our domain adaptation approach to the original data (no correction), as well as center-and-scale standardisation and LOWESS GC-content bias correction, when relevant. Center-and-scale standardisation consists in standardising the data points from each domain separately, by subtracting their median and dividing by the average squared deviation from the median. As this method is univariate, it has been performed on each bin and each data set separately.

2.1 Preanalytical biases can be accurately removed by optimal transport

In **Fig. 1A**, we performed a (Gaussian) kernel principal component analysis on the controls from the HEMA data set to illustrate the impact of the change in library preparation kit on the coverage profiles.

The two control sets belong to domains D_7 and D_8 from **Table 6**, respectively. It appears immediately that LOWESS GC-correction is not sufficient for superimposing the two panels of controls. Center-and-scale standardisation and optimal transport both succeed in that respect, which is expected since these two methods have been designed to explicitly correct data sets. Conversely, GC-correction alleviates GC-content biases at the level of individual samples only.

For each 1 Mb bin, we ran a Two-sample Kolmogorov-Smirnov test (two-sided) to quantify the differences in normalised read counts between the two panels from domains D_7 and D_8, and reported the distribution of per-bin *p*-values in **Fig. 1B**, both in linear and log scales. While center-and-scale standardisation and optimal transport show similar distributions, the latter contains a median *p*-value close to 0.5 and a more uniform distribution. Such property is desirable since *p*-values are expected to be uniformly distributed under the null hypothesis and some other conditions [51]. Indeed, a distribution shifted leftward indicates the presence of confounders responsible for some discrepancy between the two distributions, as shown both in the absence of

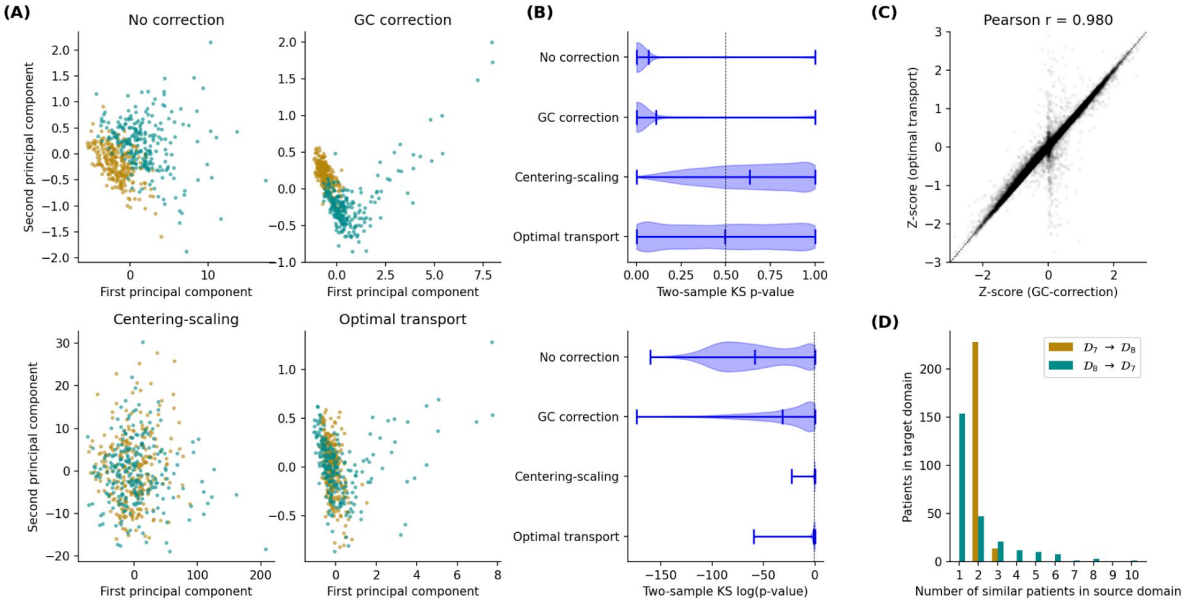

Figure 1

Correction of the healthy samples from the HEMA data set.

(A) Kernel principal component analysis of coverage profiles from the two control cohorts (haematological cancer data set). (B) Two-sample Kolmogorov-Smirnov testing, for each bin, of the difference between the two cohorts. *p*-values are shown in both linear and log scales. (C) *z*-scores of the coverage profiles before and after GC-correction and domain adaptation. (D) Histogram depicting, for each patient of the target domain, the number of patients in the source domain for which the transport plan shows a relationship.

correction or with GC-correction. Inversely, a rightward-shifted distribution illustrates overcorrection, as suggested for center-and-scale standardisation. This shift can however also occur for biological reasons, for example when the samples in the first domain are replicates of the samples from the second domain.

In **Fig. 1C**, we reported the *z*-scores of each sample and bin all together in a single scatter plot, before and after correction. We observe a good consistency between the GC-corrected normalised read counts and the OT-adapted ones, supported by a Pearson correlation coefficient of 0.98. This result suggest that our DA method does not appear to be overcorrecting the normalised read counts. We however observe a very small subset of the bins around 0 for which our method seems to overcorrect the normalised read counts. These bins are located in low-mappability regions, and we suggest that our method is correcting in these few regions relatively more due to the lack of information (the z-scores between these bins and the flanking bins are either not consistent, or mostly made of zeroes). These distortions are irrespective to the original sequencing depth, as the coverage profiles have been normalised. Given the lack of reliability of the original data in these bins (mostly zero counts), we suggest that the information loss is residual compared to high-mappability regions. Let's note that it is usually advised in the literature to disregard these bins before performing further analysis.

Finally, **Fig. 1D** reports the number of non-zero entries in the final transport plan Γ inferred by our model, for each patient of domains D_7 (green bars) and D_8 (golden bars). Without the use of entropic regularisation on Wasserstein distance, the model naturally assigns each control from the source domain to *multiple* controls from the target domain, thus reflecting the underlying complexity of the biological processes that generate the read counts. These peculiarities are implicitly acknowledged by our model, by not enforcing the patients to be assigned in pairs.

We conducted similar analyses on the OV and NIPT data sets and obtained slightly different results due to the smaller numbers of samples. In particular, visualisation based on kernel PCA show that the corrected cohorts are still not centered one onto the other. Indeed, since our convergence criterion builds on statistical tests, our algorithm is designed to halt earlier, due to *p*-values being higher when the number of samples is low. This mechanism prevents overcorrection when the available data is insufficient for accurate bias estimation. Also, histograms on the entries of the transport plans showed that each patient from the source was mapped on exactly one patient from the target domain, which met our expectations on these two data sets due to the way domains have been defined (samples were paired). Results have been reported in Suppl. Fig. 2-8.

2.2 Patient-to-patient mapping is accurate when the cohorts are representative of the exact same population

While our method is effectively capable of superimposing patients cohorts there is no *a priori* guarantee, besides theoretical considerations, that the coverage profiles are being corrected in the right direction. For this purpose, we considered 64 biological samples with ovarian carcinoma that have been processed with both Illumina HiSeq 2500 and HiSeq 4000 sequencing platforms. The two cohorts belong to domains D_{10} and D_9 from **Table 6**, respectively. In this section, we applied our domain adaptation technique on these two cohorts to see whether the bias removal is decreasing the distance between profiles originating from the same biological sample. Indeed, paired profiles are expected to overlap when the biological variation overcomes technical biases. By design, the cohorts consisted of the same patients, therefore we not only tested our algorithm with default hyper-parameters (noted as "default" in the table), but also without regularisation or early stopping criterion ("$\lambda = 0$"), and using the transport plan directly to assign pairs and compute accuracy. The purpose was mostly to test the assignment of patients and assess whether OT can map each sample to its correct counterpart.

In **Table 1**, we compared our correction method with the GC-correction approach, as well as center- and-scale standardisation. Because domain adaptation and estimation of accuracy can be done in two ways (as there are two domains), we reported both settings as separate columns in the table. While center-and-scale standardisation and GC-correction fails at pairing the samples with more than 50% accuracy, we observed a sharp improvement in accuracy with our domain adaptation approach when disabling regularisation. Without correction, only 17 and 14 profiles were correctly assigned to their counterpart in the source domain. GC-correction enabled the correct assignment of 23 and 30 patients, while our domain adaptation approach allowed the correct mapping of 47 and 48 patients when $\lambda = 0$.

In **Table 2**, we report the same metric on the NIPT data set, where 563 patients have been sequenced twice with different protocols. This data set has been divided in 6 validation groups and each group divided in 2 domains (see **table 6**). Each group was designed to control for exactly one preanalytical variable. As an example, the $D_{1,a}$ and $D_{1,b}$ domains differ by their library preparation kits, namely

TruSeq Nano and Kapa HyperPrep kits. We repeated the experiment done in previous section on each of these groups and reported accuracy in In **Table 2**. We can observe that our approach drastically improves over standard methods for all groups. In particular, the transport plan Γ inferred by our method perfectly identified the sample pairs for all 6 groups except the TruSeq Nano/Kapa HyperPrep and Kapa/IDT groups, while still improving accuracy by a large margin compared to GC-correction. These results suggest that OT is a suitable framework for estimating patient-to-patient similarities, even in the presence of a limited number of samples (i.e., 45).

2.3 Optimal transport disentangles cancer signals from non-biological sources of variation

We further tested the applicability of our method to the detection of haematological cancer and investigated whether data correction preserves the signals of interest (i.e., cancer). For this purpose, we trained simple Machine Learning models using the scikit-learn [52] Python library. The HEMA data set is composed of 179 cases of Hodgkin lymphoma, 22 of multiple myeloma and 37 of diffuse large B-cell lymphoma at different stages, as well as two control sets of size 242 and 257 respectively. The cancer samples as well as the healthy cohort from D_7 have been processed with the TruSeq ChIP kit (Illumina), while the second healthy set from D_8 has been prepared with the TruSeq Nano kit (Illumina). TruSeq Nano samples have been corrected to match the distribution of the TruSeq ChIP controls and validation was performed on the TruSeq ChIP cases and controls. As explained in the Methods section and as illustrated in **Fig. 6D**, corrected samples were used only for training, to avoid overoptimistic estimation of sensitivity and specificity resulting from controls being accidentally shifted away from the cancer cases.

In **Table 3**, we reported the performance of binary prediction of haematological cancers, namely Hodgkin lymphoma (HL), diffuse large B-cell lymphoma (DLBCL) and multiple myeloma (MM). Evaluation metrics have been estimated using fivefold cross-validation. Only samples from the training set have been corrected by our method, so as to avoid any data contamination between the training and validation set (see Methods). Sensitivity, specificity and MCC were determined based on the cutoff that produced the highest MCCs. As can be observed, data correction with our domain adaptation approach almost systematically improves cancer detection in terms of MCC, AUROC and AUPR, either through an increase in sensitivity, specificity, or both. In particular, it produced the best MCC in all of the nine settings (3 models × 3 pathologies), the best AUPR in 8 settings and the best AUROC in 7 settings. Strikingly, it outperformed GC-correction by 8.6% and 5.3% in MCC for DLBCL prediction with logistic regression and random forest, respectively.

Table 1

Performance assessment using paired samples from the OV data set.

Accuracy obtained after data correction, when assigning a sample from target domain to the closest sample in source domain, and using the Euclidean metric. In the first column, samples from domain D_9 have been corrected toward D_{10}, and vice versa.

	Correctly identified pairs	
Method	$\mathcal{D}_9 \to \mathcal{D}_{10}$	$\mathcal{D}_{10} \to \mathcal{D}_9$
No correction	17 / 64	14 / 64
Center-and-scale	20 / 64	27 / 64
GC-correction	23 / 64	30 / 64
Optimal transport (default)	24 / 64	31 / 64
Optimal transport ($\lambda = 0$)	47 / 64	48 / 64
Optimal transport (Γ)	51 / 64	50 / 64

Table 2

Performance assessment using paired samples from the NIPT data set.

Accuracy obtained after data correction, when assigning a sample from target domain to the closest sample in source domain and using the Euclidean metric, on each of the 6 validation groups. Each validation group was designed to control for one preanalytical variable at a time.

					Optimal transport		
Preanalytical variable	Comparison	No correction	Center-and-scale	GC-correction	$\lambda = 0.5$	$\lambda = 0$	Γ
Lib. prep. kit	TruSeq Nano/Kapa HyperPrep ($\mathcal{D}_{1,a}/\mathcal{D}_{1,b}$)	2 / 66	28 / 66	28 / 66	44 / 66	58 / 66	64 / 66
Adapters	Kapa/IDT ($\mathcal{D}_{2,a}/\mathcal{D}_{2,b}$)	15 / 179	32 / 179	96 / 179	98 / 179	151 / 179	157 / 179
Sequencer	HiSeq2000/NovaSeq ($\mathcal{D}_{3,a}/\mathcal{D}_{3,b}$)	1 / 45	38 / 45	19 / 45	29 / 45	45 / 45	45 / 45
Sequencer	HiSeq2500/NovaSeq ($\mathcal{D}_{4,a}/\mathcal{D}_{4,b}$)	2 / 45	43 / 45	40 / 45	43 / 45	45 / 45	45 / 45
Sequencer	HiSeq4000/NovaSeq ($\mathcal{D}_{5,a}/\mathcal{D}_{5,b}$)	5 / 93	86 / 93	75 / 93	78 / 93	93 / 93	93 / 93
NovaSeq Chemistry	V1/V1.5 ($\mathcal{D}_{6,a}/\mathcal{D}_{6,b}$)	12 / 135	60 / 135	82 / 135	89 / 135	135 / 135	135 / 135

Table 3

Haematological cancer detection using supervised approaches.

Sensitivity, specificity, Matthews correction coefficient (MCC), AUROC and AUPR obtained through validation of 3 supervised models. These models have been successively trained to distinguish Hodgkin lymphoma, DLBCL and multiple myeloma cases from healthy controls. Sensitivity, specificity and MCC were computed using the cutoff that maximises MCC.

Supervised model	Logistic regression					Random forest					Support vector machine				
Metric	Sen.	Spec.	MCC	AUROC	AUPR	Sen.	Spec.	MCC	AUROC	AUPR	Sen.	Spec.	MCC	AUROC	AUPR
Hodgkin lymphoma															
No correction	95.0 %	97.1 %	92.2 %	98.2 %	98.3 %	76.0 %	**97.5 %**	76.9 %	93.3 %	93.4 %	95.5 %	86.0 %	80.6 %	96.3 %	95.4 %
Center-and-scale	96.1 %	94.6 %	90.4 %	98.3 %	98.3 %	89.9 %	89.3 %	78.8 %	95.0 %	94.1 %	92.7 %	88.8 %	80.9 %	95.1 %	93.8 %
GC correction	92.7 %	**99.2 %**	92.8 %	98.7 %	98.8 %	87.7 %	**97.5 %**	86.5 %	97.0 %	97.1 %	**96.1 %**	95.9 %	91.8 %	**98.7 %**	**98.6 %**
Optimal transport	**96.6 %**	97.1 %	**93.7 %**	**98.9 %**	**99.0 %**	**92.2 %**	97.1 %	**89.8 %**	**97.8 %**	**97.9 %**	94.4 %	**97.5 %**	**92.2 %**	98.6 %	98.3 %
DLBCL															
No correction	75.7 %	**99.6 %**	83.6 %	92.8 %	86.1 %	32.4 %	**99.2 %**	49.1 %	84.2 %	56.2 %	78.4 %	97.5 %	77.7 %	95.5 %	83.9 %
Center-and-scale	86.5 %	98.8 %	87.3 %	96.6 %	92.8 %	64.9 %	97.5 %	68.3 %	94.1 %	79.0 %	**89.2 %**	95.9 %	79.9 %	96.8 %	84.2 %
GC correction	81.1 %	98.3 %	82.3 %	96.4 %	87.7 %	67.6 %	97.9 %	71.7 %	91.5 %	78.6 %	78.4 %	**99.2 %**	83.7 %	**97.4 %**	88.0 %
Optimal transport	**94.6 %**	98.3 %	**90.9 %**	**97.9 %**	**94.4 %**	**73.0 %**	98.3 %	**77.0 %**	**94.6 %**	**86.0 %**	**89.2 %**	97.5 %	**84.8 %**	96.7 %	**88.1 %**
Multiple myeloma															
No correction	90.9 %	99.6 %	92.4 %	98.0 %	92.1 %	63.6 %	98.8 %	70.3 %	91.0 %	70.6 %	63.6 %	96.7 %	60.3 %	93.7 %	62.4 %
Center-and-scale	72.7 %	99.6 %	81.4 %	97.3 %	89.4 %	63.6 %	98.3 %	68.0 %	93.3 %	64.6 %	59.1 %	97.5 %	60.5 %	93.6 %	63.1 %
GC correction	90.9 %	**100.0 %**	95.0 %	99.2 %	96.7 %	77.3 %	**100.0 %**	87.0 %	95.5 %	89.4 %	81.8 %	**97.9 %**	78.2 %	97.4 %	83.6 %
Optimal transport	**95.5 %**	**100.0 %**	**97.5 %**	**99.7 %**	**98.0 %**	**86.4 %**	99.6 %	**89.8 %**	**96.1 %**	**91.6 %**	**95.5 %**	95.9 %	**78.4 %**	**98.3 %**	**85.7 %**

Analogous results obtained on the OV data set have been reported in **Table 4**. Because the ovarian data set contains cases and controls from both domains, we could perform our validation in both directions (adapting samples from D_9 to D_{10} and assessing performance on remaining samples from D_{10}, and vice versa), corresponding to the top and bottom parts of **Table 4**. The proposed method systematically produced an improvement in MCC and AUPR in all three settings and improved AUROC in two out of the three settings. In particular, we noticed gains of 2.8%, 6.8% and 8.8% in MCC, respectively. In light of these results, we showed the ability of our method to disentangle sources of variation of biological and technical origins. Indeed, the supervised Machine Learning approaches for cancer detection benefited from the improvement in data quality resulting from domain adaptation, yielding better generalization and validation accuracy. Although these AUROC scores are far from being clinically relevant, they must be contextualized. First, we evaluated our predictive models in an artificially difficult setting where newly-collected samples have been processed with a different technology, while wet-lab protocols should be standardised in clinical settings. Second, cancer cohorts include many low-grade and borderline cases, which are most likely chromosomally stable and therefore may be overlooked by our CNA-based approach.

2.4 Optimal transport preserves copy number aberrations across domains

As we are fully aware of the overfitting risks associated with our model, we made sure the adapted samples were consistent with the original data by verifying whether the CNAs of each patient were conserved. For this purpose, clonal and subclonal CNAs were called using the ichorCNA v0.2.0 R package (details in Suppl. Mat. 2) and we benchmarked our method on the entire OV data set. Indeed, our domain adaptation method provides adjusted profiles that are transparently interpretable and are directly comparable across domains, allowing their comparison in terms of read counts before and after correction.

Let us denote by D_9 and D_{10} the wet labs from which the samples originate (see **Table 6**), respectively [53] and [54]. We first built a panel of controls (reference set) using the 79 controls from domain D_9 and called CNAs in cancer cases from D_9. Then, we built a panel of controls using the 39 controls from D_{10} and called CNAs in the same cancer cases from D_9, after adapting them with our proposed approach to match the distribution of cancer cases in D_{10}. Finally, we quantified the similarity of ichorCNA results using different metrics, as shown in **Table 5**. Because ichorCNA performs GC-correction “under the hood”, we did not include GC-correction in the benchmark, as it would produce results highly similar to the baseline. Also, in the case of center-and-scale standardisation we enforced positivity by clipping the corrected read counts, as negative values cannot be handled by ichorCNA.

Even in the absence of any correction, the per-bin copy numbers estimated for cancer CNAs are not consistent, as we see that the accuracy and SOV REFINE measures are far from being perfect. This can be attributed not only to (1) the difference in protocols used to produce the two panels of normals, but also (2) the limited number of controls, (3) the fact that the controls differ between the two domains, and (4) the experimental uncertainty (stochastic noise). Overall, both center-and-scale standardisation and optimal transport preserved the original normalised read counts sufficiently since they both provided results similar to the baseline (“no correction”). However, our method improves over center-and-scale standardisation regardless of the evaluation metric. After applying our DA method on the cancer cases, 73.4% of the bins were assigned the correct copy number and segmentation of CNAs produced an overlap score of 0.7552. As illustrated in **Fig. 2**, the copy numbers called by ichorCNA in D_9 (panel A) are mostly preserved without (B) or after (D) correction. The least consistent results were produced by center- and-scale standardisation, where some disruptions have been introduced at multiple locations (e.g., some higher copy numbers in chromosome 3).

Table 4

Ovarian carcinoma detection using supervised approaches.

Sensitivity, specificity, Matthews correction coefficient (MCC), AUROC and AUPR obtained through validation of 3 supervised models. These models have been trained to distinguish ovarian carcinoma cases from healthy individuals. Sensitivity, specificity and MCC were computed using the cutoff that maximises MCC.

Supervised model	Logistic regression					Random forest					Support vector machine				
Metric	Sen.	Spec.	MCC	AUROC	AUPR	Sen.	Spec.	MCC	AUROC	AUPR	Sen.	Spec.	MCC	AUROC	AUPR
$\mathcal{D}_9 \rightarrow \mathcal{D}_{10}$															
No correction	59.9 %	80.7 %	33.4 %	67.2 %	92.9 %	50.5 %	**81.3 %**	26.0 %	59.6 %	90.5 %	57.5 %	**87.9 %**	34.5 %	68.4 %	93.5 %
Centering-scaling	57.3 %	81.9 %	32.7 %	65.7 %	92.7 %	61.5 %	65.2 %	28.7 %	58.4 %	89.8 %	64.8 %	75.5 %	35.1 %	67.8 %	93.2 %
GC correction	**60.1 %**	79.4 %	33.0 %	68.0 %	93.2 %	64.7 %	72.0 %	32.1 %	67.0 %	92.7 %	62.1 %	80.4 %	33.3 %	68.8 %	93.5 %
Optimal transport	58.5 %	**86.6 %**	**35.4 %**	**70.5 %**	**94.0 %**	**75.5 %**	64.9 %	**38.0 %**	**70.8 %**	**93.7 %**	**71.4 %**	74.1 %	**40.0 %**	**73.7 %**	**94.5 %**
$\mathcal{D}_{10} \rightarrow \mathcal{D}_9$															
No correction	77.0 %	70.3 %	43.3 %	77.5 %	93.4 %	**78.6 %**	77.5 %	**51.0 %**	**83.0 %**	**95.5 %**	**75.8 %**	81.4 %	50.4 %	**82.5 %**	95.5 %
Centering-scaling	77.9 %	70.6 %	44.5 %	77.5 %	93.5 %	73.1 %	84.0 %	49.0 %	82.1 %	95.3 %	72.1 %	86.8 %	49.9 %	81.9 %	95.4 %
GC correction	**82.2 %**	68.0 %	47.3 %	80.3 %	94.4 %	67.4 %	90.6 %	48.4 %	80.0 %	95.0 %	70.7 %	86.7 %	48.4 %	81.0 %	95.2 %
Optimal transport	79.4 %	**78.5 %**	**52.0 %**	**83.3 %**	**95.7 %**	64.9 %	**93.4 %**	47.0 %	78.8 %	94.7 %	69.9 %	**91.6 %**	**50.7 %**	**82.5 %**	**95.7 %**

Table 5

Quantitative assessment of the consistency of CNA calling.

ichorCNA results on ovarian carcinoma cases from D_9 using the panel of controls from D_9, compared to the same cases corrected ($D_9 \rightarrow - D_{10}$) by each method and using the panel of controls from D_{10}. Metrics in the upper part of the table focus on per-bin metrics, namely the copy number in each bin, the presence of a CNA (copy number$\neq$ 2) in each bin, the SOV REFINE [55] score and the log-ratios. We used the SOV REFINE segmentation metric to measure the overlap between called CNAs. The metrics in the bottom section of the table are the average absolute errors on different model parameters estimated by ichorCNA.

Metric	No correction	center-and-scale standardisation	OT
Accuracy (copy number)	71.3 %	39.5 %	**73.4 %**
Accuracy (presence of CNA)	76.8 %	53.2 %	**78.7 %**
SOV_REFINE (copy number)	0.7401	0.5218	**0.7552**
Error on log-ratios	0.0083	0.1164	**0.0070**
Error on tumour fraction	0.0045	0.0197	**0.0042**
Error on tumour ploidy	0.0880	0.2293	**0.0836**
Error on tumour cellular prevalence	0.0888	0.1209	**0.0866**
Error on proportion of subclonal CNAs	0.1366	0.1414	**0.1268**

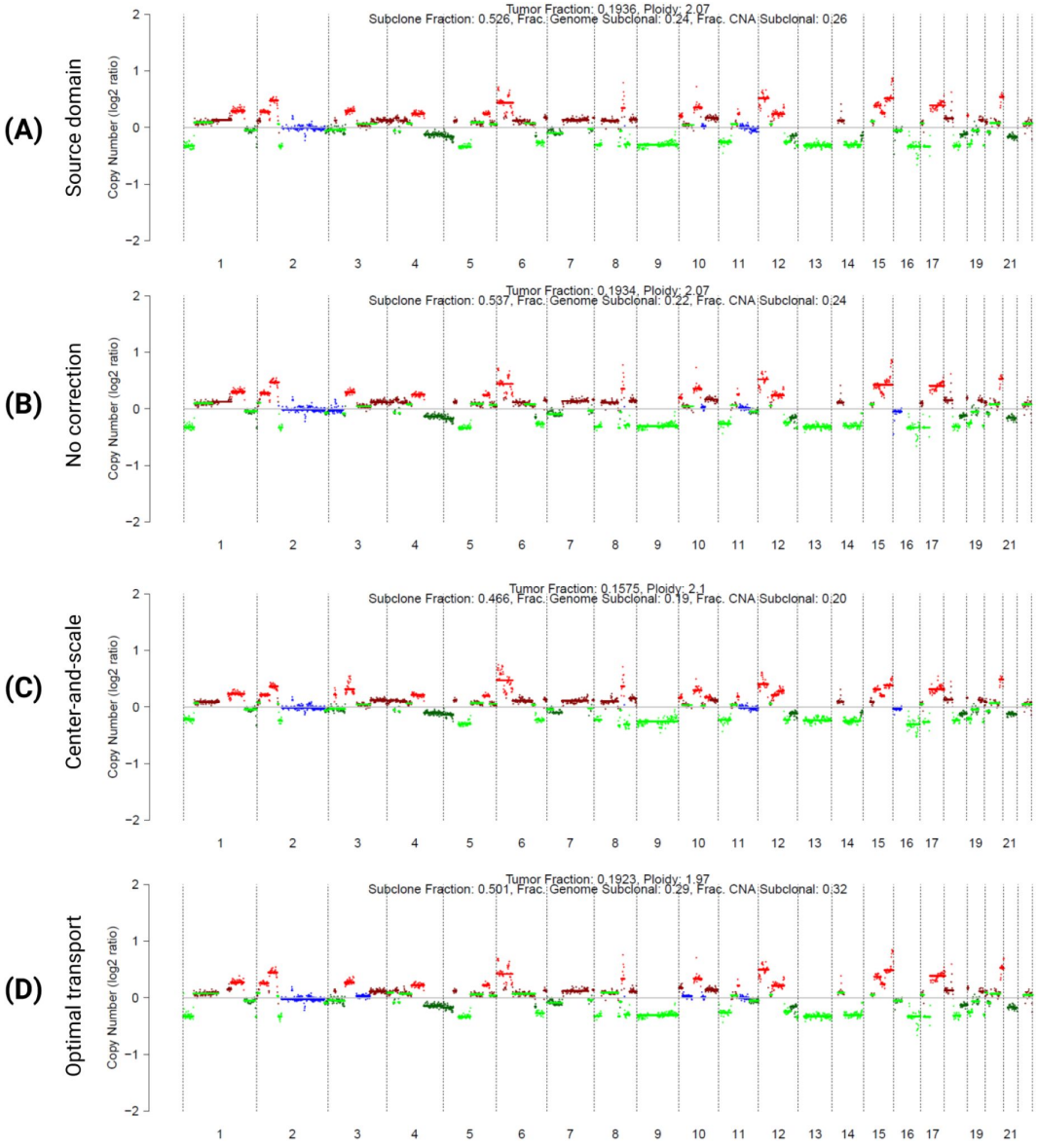

Figure 2

Qualitative assessment of the consistency of CNA calling.

Comparison of the CNAs called by ichorCNA on a late stage ovarian carcinoma case from D_9, before and after domain adaptation. Green and red colouring correspond to deletions and gains, respectively. (A) Using D_9 controls. (B) Using D_{10} controls. (C) D_9 cancer cases (including the case shown) centered-and-scaled toward D_{10} and analysed with D_{10} controls. (D) D_9 cancer cases OT-corrected toward D_{10}, analysed with D_{10} controls.

The average absolute error on the estimation of tumour fraction is 0.0042, which is acceptable given the error of 0.0045 in the absence of any correction and the standard deviation of the tumour fraction estimates (0.0464). Despite the limited size of our reference sets and therefore the potentially overpessimistic assessment of the inconsistencies of ichorCNA's results, we conclude that most of the CNAs have been preserved and that the proposed method does not disrupt the original data, as a more straightforward standardisation approach would. Since ichorCNA offers the possibility to call CNAs without panels of normals, we ran similar analysis without controls and observed more consistent results between the two domains. We reported these results in Suppl. Tab. 1 and Suppl. Fig. 1.

Estimated tumour fractions before and after correction have been reported in **Fig. 3**, showing good consistency both in the presence (r=0.973, *p*-value=1.09e-205, two-sided test) or the absence (r=0.980, *p*-value=1.77e-224) of the D_{10} panel of controls.

2.5 Preanalytical variables are mostly discrete

A major limitation of our approach is its inherent restriction to discrete settings, where the technical counfounder is acting as a dummy variable and reflects whether some technology as been used to produce a sample or not. However, to the best of our knowledge there is no continuous preanalytical variable in whole-genome sequencing that induces gradual changes in the normalised read counts and in our data sets. A potential exception is the plasma separation delay, measured as the time elapsing between the blood draw and the separation of the plasma from the buffy coat. We tested Pearson and Spearman correlation for each 1 Mb genome bin on the HEMA data set, using a significance level of 0.01 and applying the Benjamini-Hochberg procedure to account for multiple testing. As shown in **Fig. 4B**, no bin was found to be significantly correlated with the plasma separation delay. The normalised reads counts of the NIPT samples in the first 1 Mb bin of chromosome 6 (bin showing highest correlation with plasma separation delay) are shown in **fig. 4A**.

2.6 GC-correction is not sufficient to decorrelate read counts from GC-content

In **Figure 1**, we showed that GC-correction did not help in reducing the dissimilarity between the two control sets from the HEMA data set. We also reported similar results on the two other data sets in Suppl. Mat. 3. While GC-correction succeeds at reducing the individual variability (experimental variance) of samples as shown by the improved accuracy in **Tables 1** and **2**, it fails at alleviating the biases introduced by changes in sequencer or library preparation method. Indeed, while this approach improves cancer detection on average by removing technical variations based on GC-content, it does not systematically produces performance gains, does not efficiently capture similarities between profiles originating from the same biological sample and does not completely remove the clustering effects introduced by the changes in the sequencing methodology. By contrast, our method showed that these expectations can be met through the modelling of patient-to-patient similarities and explicit constraining of the samples based on quantiles. Indeed, these latter constraints drastically lower the risks of overfitting and ensure that the mapping between the cohorts is performed in a biologically meaningful manner.

In **Figure 5**, we reported the two-sample Kolmogorov-Smirnov *p*-values from **Figure 1b** as a function of the GC-content, as well as the median *p*-value per 0.5% GC stratum. Normalised read counts (top left panel) exhibited strong relationship between median *p*-values and GC-content, demonstrating that regions with low and high GC-content are the most biased by the change from the KAPA HyperPrep to the

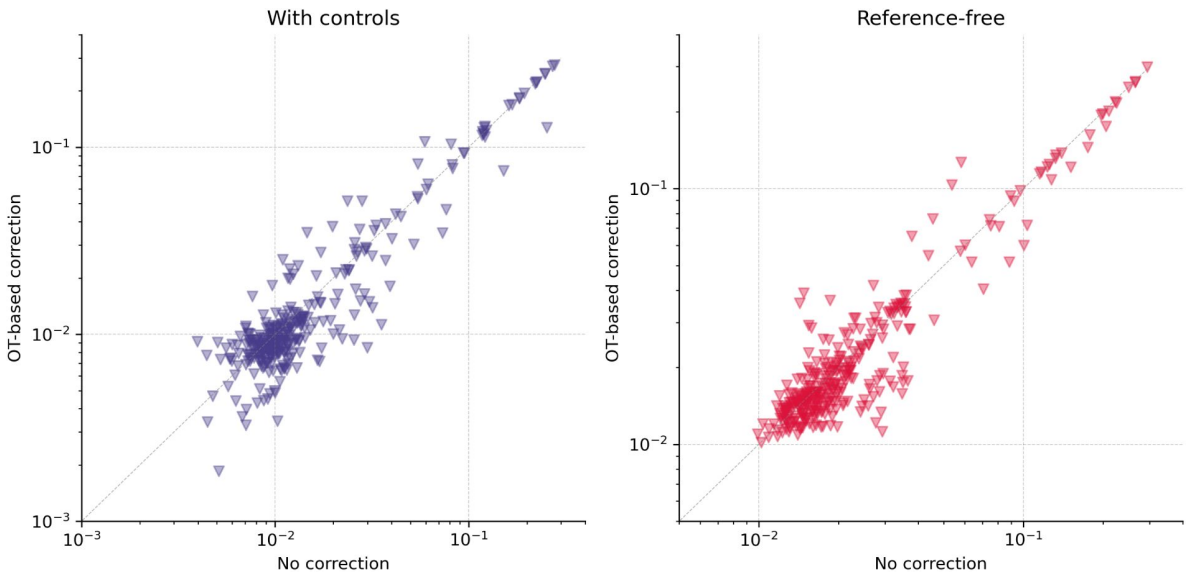

Figure 3

Tumour fractions before and after domain adaptation.

Fractions have been estimated with ichorCNA before and after adapting the D_9 ovarian carcinoma cases toward the D_{10} cases. Results have been produced both with (Left) and without (Right) the panel of controls from domain D_{10}. Fractions are shown in log-scale.

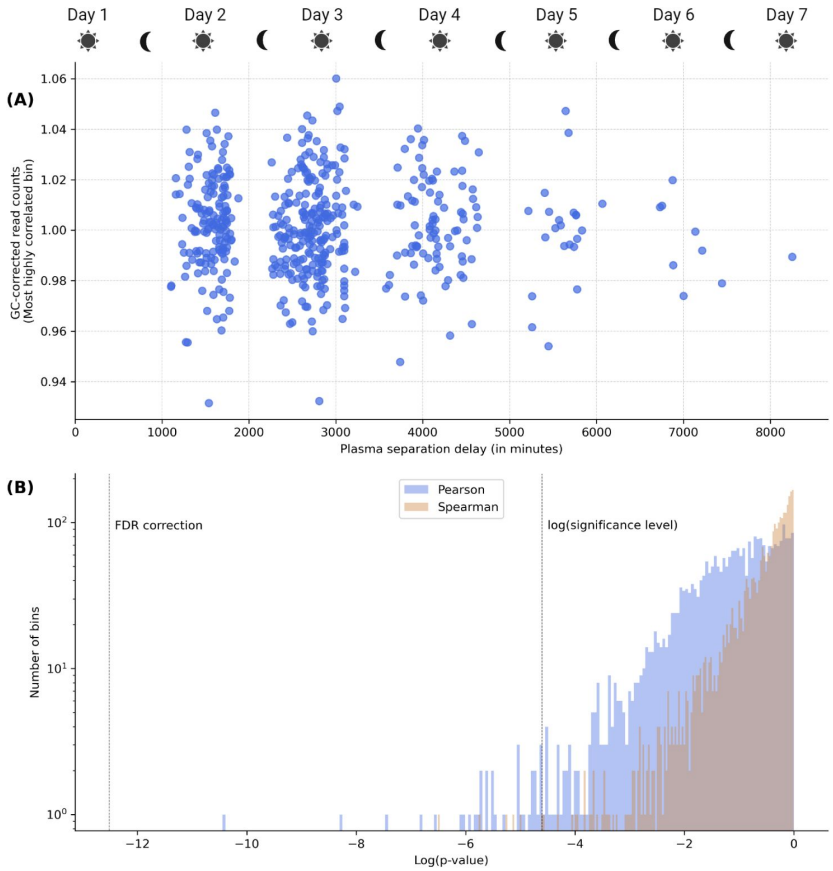

Figure 4

Effect of plasma separation delay on coverage profiles.

(A) GC-corrected normalised read counts of all samples from the NIPT data set for a specific bin (first 1 Mb bin from chromosome 6), namely the one giving the strongest linear correlation with the plasma separation delay (Pearson's *r*=0.1838, *p*-value=1.38e-5, two-sided test). (B) Distribution of the *p*-values computed likewise for each 1 Mb bin, and reported as a histogram. Histogram is shown in log-scale.

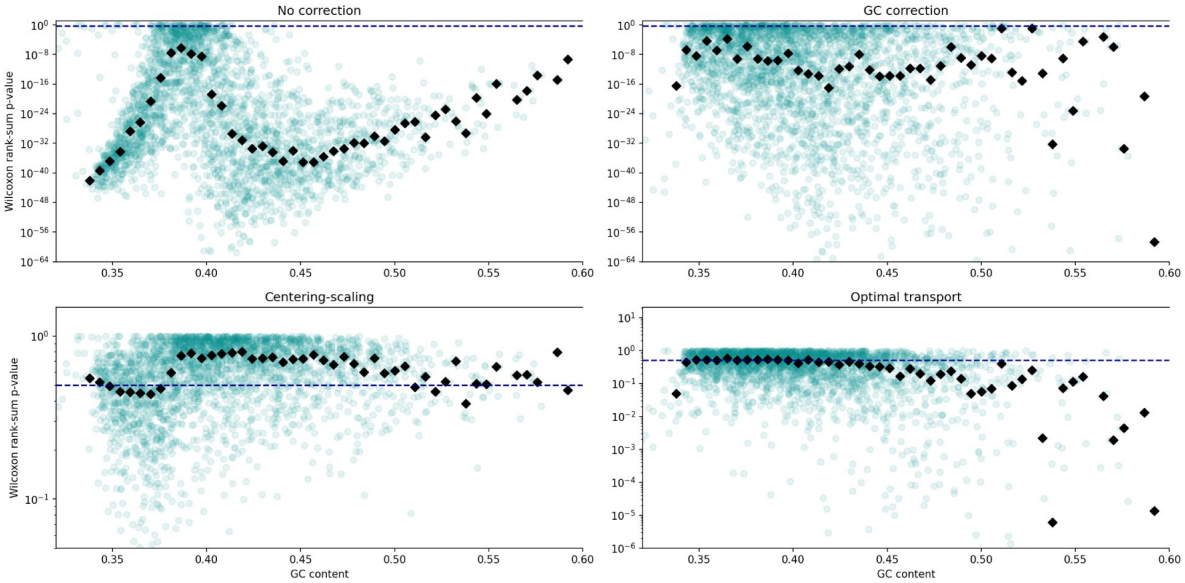

Figure 5

Coverage difference between domains as a function of GC-content.

Two-sample Kolmogorov-Smirnov testing, for each 1 Mb bin, of the difference between the two control sets from the HEMA data set processed with different library preparation kits. *p*-values are shown in log-scale, as a function of the GC-content of each bin. Dashed line corresponds to a 0.5 *p*-value and black markers to the median *p*-values per 0.5% GC stratum.

TruSeq ChIP library preparation kit. While center-and-scale standardisation is capable of centering these *p*-values around 0.5 (which is expected under the null hypothesis since *p*-values are uniformly distributed in the [0, 1] interval), we observed the same trend. GC-correction drastically improved in that respect, as no clear correlation can be observed. However, the median *p*-value still varies from stratum to stratum, suggesting that some subtle and nonlinear GC biases remained. Finally, our proposed approach produced the most consistent results across the GC-content values, showing that it more effectively alleviated these residual biases.

2.7 Domain expertise is the best regularisation

There are multiple mechanisms put in place within our model to constrain it to infer a matrix X that is as meaningful as possible, namely the quantile-based regularisation function, the positivity constraint on the read counts, the median normalisation and GC correction. While these constraints are not guarantees of performance *per se*, they restrict the size of the search space drastically, eliminating a large proportion of irrelevant solutions. While our method was originally designed for correcting normalised read counts with the detection of CNAs in mind, it remains sufficiently generic to be applied to any (whole-)genome sequencing or array-based data set. Indeed, the only requirement is the representativeness of the cohorts in all domains. However, the requirements imposed on the data are problem-dependent and heavily depend on the nature of the data sets. Therefore, from a general perspective, extra care should be given to the assumptions underlying the model. For instance, preserving the quantiles (e.g., *z*-scores) might not necessarily be desirable, as read counts are heavily subject to noisy fluctuations that should preferably not be transferred from domain to domain.

Beyond aforementioned limitations, our findings open new perspectives for the analysis of highdimensional whole-genome sequencing data and suggest that appropriate modelling of technical confounders enables the joint analysis of cohorts sequenced at different points in time (changes of sequencing platform, library preparation kit, DNA extraction method) and space (team, hospital, country). Finally, the analysis of larger data sets is expected to strengthen the detection power of statistical models based on cfDNA data and enable the presymptomatic detection of more subtle cancer signals.

3 Methods

Throughout the paper, we referred to *cohort* as a set of samples sharing similar high-level characteristics (e.g., a set of healthy controls, a set of pregnant women, a set of ovarian cancer patients) *and* processed using similar protocols. A *domain* is a set that can include multiple cohorts, with no regard for the biological state as only the protocol is considered. Finally, a *data set* can itself include multiple domains, as each data set has been used to assess our method's ability to correct for biases between the domains contained in this data set.

3.1 Clinical data

We benchmarked our method on three data sets produced in-house, each used for a different purpose. The peculiarities of each data set have been summarised in **Table 6**. All data sets have a median sequencing coverage between 0.1x and 0.2x.

The data sets used in the present work have been collected during studies previously approved by the ethical committee of the University Hospitals Leuven under S/57999, S/62285, S/62795, S/50623, S/56534, S/63240, S/51375, S/55904, S/57144, S/59207, S/64205 and S/64035. Blood samples were collected either into Streck cfDNA BCT or Roche Cell-Free DNA Collection Tubes. cfDNA was extracted using either the QIAamp Circulating Nucleic Acid Kit or the Maxwell automated

Data set	Condition/Setting	Domain	Size	Library preparation kit	Index type	Sequencer
NIPT	Pregnancy (kit validation)	$\mathcal{D}_{1,a}$	$2 \times 66^*$ { 66	TruSeq Nano	TruSeq Nano	HiSeq 4000
		$\mathcal{D}_{1,b}$	66	Kapa HyperPrep	Kapa dual	HiSeq 4000
	Pregnancy (adapter validation)	$\mathcal{D}_{2,a}$	$2 \times 179^*$ { 179	Kapa HyperPrep	IDT	HiSeq 4000
		$\mathcal{D}_{2,b}$	179	Kapa HyperPrep	Kapa dual	HiSeq 4000
	Pregnancy (sequencer validation)	$\mathcal{D}_{3,a}$	$2 \times 45^*$ { 45	Kapa HyperPrep	Kapa dual	HiSeq 2000
		$\mathcal{D}_{3,b}$	45	Kapa HyperPrep	Kapa dual	NovaSeq
		$\mathcal{D}_{4,a}$	$2 \times 45^*$ { 45	Kapa HyperPrep	Kapa dual	HiSeq 2500
		$\mathcal{D}_{4,b}$	45	Kapa HyperPrep	Kapa dual	NovaSeq
		$\mathcal{D}_{5,a}$	$2 \times 93^*$ { 93	Kapa HyperPrep	Kapa dual	HiSeq 4000
		$\mathcal{D}_{5,b}$	93	Kapa HyperPrep	IDT	NovaSeq
	Pregnancy (chemistry validation)	$\mathcal{D}_{6,a}$	$2 \times 135^*$ { 135	Kapa HyperPrep	IDT	NovaSeq (V1)
		$\mathcal{D}_{6,b}$	135	Kapa HyperPrep	IDT	NovaSeq (V1.5)
HEMA	Hodgkin lymphoma	$\mathcal{D}_7$	179	TruSeq ChIP	-	HiSeq 2000/2500
	Diffuse large B-cell lymphoma	$\mathcal{D}_7$	37	TruSeq ChIP	-	HiSeq 2000/2500
	Multiple myeloma	$\mathcal{D}_7$	22	TruSeq ChIP	-	HiSeq 2000/2500
	Healthy	$\mathcal{D}_7$	242	TruSeq ChIP	-	HiSeq 2000/2500
	Healthy	$\mathcal{D}_8$	257	TruSeq Nano	-	HiSeq 2000/2500
OV	Ovarian carcinoma	$\mathcal{D}_9$	223	KAPA HyperPrep	IDT	HiSeq 4000
			32	KAPA HyperPrep	-	HiSeq 4000
			1	KAPA HyperPrep	-	HiSeq 2000
			64^* { 61	KAPA HyperPrep	-	HiSeq 4000
			2	KAPA HyperPrep	-	NovaSeq V1
			1	KAPA HyperPrep	-	HiSeq 2000
	Ovarian carcinoma	$\mathcal{D}_{10}$	64^* { 64	KAPA DNA lib. prep.	-	HiSeq 2500
			156	KAPA DNA lib. prep.	-	HiSeq 2500
	Healthy	$\mathcal{D}_9$	79	KAPA HyperPrep	IDT	HiSeq 4000
	Healthy	$\mathcal{D}_{10}$	39	KAPA DNA lib. prep.	-	HiSeq 2500

Table 6

Summary of the data sets used in this study.

Samples in sets marked with a '$*$' have been processed twice, allowing quantitative assessment of the different biases caused by the changes of sequencing protocols. Domains have been defined based on our experiments, as well as the protocol differences shown in the table. For clarity purposes, we systematically refer to these domains in the results section.

protocol. Samples were pooled by batches of 20 for multiplex sequencing using all lanes of Illumina flow cells. Each pool was sequenced either on the Illumina HiSeq 2000, HiSeq 2500, HiSeq 4000 or NovaSeq 6000 platform, single-end 1×36bp, 1×50bp or paired-end 2×50bp.

The first data set consists of 563 validation samples collected in the context of Non-Invasive Prenatal Testing (NIPT) [56] and processed within the standard NIPT routine twice. These paired samples are divided in 6 validation groups (2 × 6 domains), each used to quantify the distributional shift introduced by the change of *one* preanalytical variable. The libraries of 66 biological samples have been prepared with either the TruSeq Nano DNA Sample Preparation Kit (Illumina) or the KAPA HyperPrep Kit (Roche) with Kapa Dual indexed adapters. 179 samples have been used with either IDT indexes or KAPA Dual indexed adapters. 45 samples have been processed either by the HiSeq 2000 or NovaSeq platform. 45 samples have been processed either by the HiSeq 2500 or NovaSeq platform. 93 samples have been processed either by the HiSeq 4000 or NovaSeq platform. Finally, 135 samples have been processed by a NovaSeq platform, with either V1 or V1.5 chemistry. In total, this results in 2 × 563 paired samples. We refer to this first data set as NIPT for short.

Our second data set (HEMA) focuses on haematological malignancies and is composed of 179 cases of Hodgkin lymphoma (HL), 37 of diffuse large B-cell lymphoma (DLBCL) and 22 of multiple myeloma, as well as 498 controls. Among those, 177 HL cases and 260 controls have been published in a previous study [57] and the entirety of the haematological cancer cases have been included in one of our studies (GipXplore [53]). The libraries of 242 out of the 499 controls have been prepared with the same kit as the haematological cancer cases, namely the TruSeq ChIP Library Preparation Kit (Illumina) [4]. The remaining ones have been prepared with the TruSeq Nano kit.

Finally, we further validated our OT-based bias removal approach with controls and ovarian carcinoma cases sequenced by two different teams [54] including ours. These samples were not derived from cancer patients with overt clinical disease, but rather the presence of a suspicious malignancy based on imaging. We refer to this last data set simply as OV. Protocols vary in multiple ways. As an example, all of the samples in D_9 (see **Table 6**) have been processed with HiSeq 2500, while all samples in domain D_{10} have been sequenced by an instrument that differed from HiSeq 2500. Samples from D_9 and D_{10} have been prepared with the KAPA HyperPrep and KAPA DNA library preparation kits, respectively. Ovarian carcinoma samples from D_{10} have been manually extracted with the QIAamp CAN kit. Let's note that D_9 does not strictly stick to our definition of domain. Despite the heterogeneity caused by the presence of multiple sequencers, we artificially grouped the samples in order to simplify the comparison between laboratories but also better reflect the heterogeneity expected to be encountered in Big Data settings.

3.2 Data preprocessing

Reads were first aligned to the reference genome hg38 using the Burrows-Wheeler aligner [58], only considering the 22 autosomes. Then, read duplicates were removed with Picard tools [59] and remaining ones were recalibrated with the Genome Analysis Toolkit [60]. Finally, reads were counted in predefined bins of size 1 Mb. Such size offers a good tradeoff between noise reduction and the granularity of achievable chromosomal aberration detection. Finally, counts were normalised by dividing by the median count per Mb of the whole profile to correct for the effective sequencing depth.

3.3 GC-content bias correction

We performed GC-correction by dividing normalised read counts by their estimate according to a Locally Weighted Scatterplot Smoothing (LOWESS) model [61], where the exogenous variable is the GC-content of the bin, and the endogenous variable is the corresponding normalised read count. 30% of the data points (bins) have been used to predict the endogenous variable. We used

the Python package statsmodels v0.12.2 [62] to implement the LOWESS correction. We also designed a differential version of GC-correction that is PyTorch-compliant and used by our DA method to ensure that the reads counts of adapted cohorts do not correlate with GC-content. More details are provided in Suppl. Mat. 1.

3.4 Center-and-scale standardisation

We benchmarked our method against a more straightforward approach, consisting in the standardisation of each cohort or data set separately. This approach ensures that the *z*-scores are centered in all domains and allows comparability as long as each cohort is representative of the population. For each 1Mb bin, we centered and scaled the normalised read counts as follows:

$$\tilde{X}_{ik} \leftarrow \frac{X_{ik} - \mu(X_{\cdot k})}{\sigma(X_{\cdot k})}, \qquad (1)$$

where $\mu(X_{\cdot k})$ is the median of normalised read counts within bin k across all samples from the same cohort and $\sigma(X_{\cdot k})$ the square root of the average squared deviation from this median. The median has been used in place of the mean for robustness against outliers, such as profiles with aberrant CNAs. We refer to this method as center-and-scale standardisation throughout the manuscript.

3.5 Domain adaptation using optimal transport

We defined the best correction function as the one that minimises statistical dissimilarity metric between two cohorts. Given the multivariate nature of the problem and the strong mathematical foundations behind optimal transport, we propose to use the Wasserstein distance to quantify the discrepancy between data sets.

The general principle of OT is to match two probability distributions by transporting the probability mass of one distribution into the other with minimal effort (hence the name *optimal transport*). Distributions that can be transported into each other at a low cost are considered highly similar. In the case of discrete samples, OT amounts to finding a discrete probabilistic mapping (called the *transport plan*) of the source samples onto the target samples where the mapping of a source sample to a target sample bears some associated cost. We consider thus two data matrices $X \in \mathbb{R}_+^{n \times q}$ and $Y \in \mathbb{R}_+^{m \times q}$, as illustrated by the normalised read counts matrices in **Fig. 6A**, where n and m are the sample sizes of the each domain and q is the number of predefined bins. As samples are all assumed to be of equal importance, we choose uniform probabilistic weights $\nu i = 1/n, \ \forall i$ and $\mu j = 1/m, \ \forall j$ to define a probability distribution on these discrete samples (assuming no two samples can be identical). We also consider the associated pairwise Euclidean distance matrix $C \in \mathrm{R}^{n \times m}$. For normalised read counts, the matrix Y is first GC-corrected. The matrix X will be GC-corrected during OT, as part of a joint optimisation process. The reason for not correcting X upfront is that OT might artificially introduce correlations with GC-content *a posteriori*. Instead, we implemented median-normalisation and GC-correction as a differentiable function f (see Suppl. Mat. 1).

The Wasserstein distance is defined, in its discrete form, by

$$
\begin{aligned}
W_p(C) = \min_{\Gamma} & \left(\sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij}^{p} \Gamma_{ij} \right)^{1/p} \\
\text{s.t.} \quad & \sum_{j=1}^{m} \Gamma_{ij} = \nu_i, && \forall i, \\
& \sum_{i=1}^{n} \Gamma_{ij} = \mu_j, && \forall j, \\
& \Gamma_{ij} \geq 0, && \forall i, j.
\end{aligned}
\tag{2}
$$

Matrix Γ, usually referred to as the transport plan and depicted in **Fig. 6B**, can be interpreted as the amount of probability mass transferred from points of the source domain to the target domain through optimal transport. In particular, Γ*ij* is the probability mass transferred from point *i* in the source domain to point *j* in the target domain. Such interpretation allows us to use Γ as a pairwise similarity matrix and express the domain adaptation problem as a multivariate regression problem. By choosing $p = 2$ and the Euclidean metric as function *d*, as well as by accounting for the reduction of variance caused by *Gamma*, and attaching equal importance to the bins (see details in Suppl. Mat. 1.2), the optimisation problem becomes

$$
\begin{aligned}
\min_{\Gamma, \mathcal{X}} \; & \frac{1}{nq} \sum_{i=1}^{n} \sum_{k=1}^{q} \frac{1}{\sigma(Y_{\cdot k})} \left(f(\mathcal{X})_{ik} - n \sum_{j=1}^{m} \Gamma_{ij} Y'_{jk} \right)^2 \\
& + \lambda R(\mathcal{X}), \\
\text{s.t.} \quad & \alpha(\Gamma) = \frac{\sum_{k=1}^{q} \sigma(Y_{\cdot k})}{\sum_{k=1}^{q} \sigma(\sum_{j=1}^{m} \Gamma_{\cdot j} Y_{jk})}, \\
& Y'_{jk} = \alpha(\Gamma)(Y_{jk} - \mu(Y_{\cdot k})) + \mu(Y_{\cdot k}), \\
& \Gamma \in \mathcal{F},
\end{aligned}
\tag{3}
$$

where $\alpha(\Gamma)$ is the variance correction factor, $R(X)$ is a regularisation term, and λ is the associated regularisation hyper-parameter.

The output of our algorithm is matrix *X*, which we interpret as the surrogate of *X* in the target domain.

3.5.1 Regularisation function

While the Wasserstein distance is often supplemented with a regularisation term based on the entropy of Γ [63], we noticed that entropic regularisation tends to reduce the variance of the adapted samples, ultimately collapsing them onto their centroid. This is not a desirable property because in actual high-dimensional data the curse of dimensionality will naturally keep data points distant from each other in the presence of noise. This creates an obstacle to the idea of mapping a source sample to the "closest" target samples. Therefore, we do not regularise the Wasserstein distance based on entropy. Instead, we propose a more informative approach where the deviations (e.g., chromosome gains or deletions) of samples *X* from some reference should be preserved throughout the whole adaptation process.

The regularisation function is defined as a mean squared error function:

$$
\begin{aligned}
R(\mathcal{X}) &= \frac{1}{2nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{1}{\sigma_k^2}\left(S_{ik}-\hat{S}_{ik}\right)^2 \\
&\quad+\frac{1}{2nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{1}{\sigma_k^2}\left(T_{ik}-\hat{T}_{ik}\right)^2, \\
S_{ik} &= f(X)_{ik}-\mu(f(X)_{\cdot k}), \\
\hat{S}_{ik} &= f(\mathcal{X})_{ik}-\mu([Y;f(\mathcal{X})]_{\cdot k}), \\
T_{ik} &= Y_{ik}-\mu(Y_{\cdot k}), \\
\hat{T}_{ik} &= Y_{ik}-\mu([Y;f(\mathcal{X})]_{\cdot k}),
\end{aligned}
\tag{4}
$$

where $f(X)$ is our differentiable GC-correction function, applied independently on each row of X. For a more robust estimation, we computed $\mu(X_{\cdot k})$ as the median over bin k (rather than the mean). We used the MATLAB notation $[Y; f(X)]$ to denote the vertical concatenation of matrices Y and $f(X)$ (resulting in a matrix of dimension $(n + m) \times q$). This regularisation function is meant to preserve the quantiles (akin to the z-scores) across the two domains. In particular, the first term enforces the consistency of the samples from the source domain, while the second term operates similarly on the samples from the target domain.

3.5.2 Overfitting and stopping criterion

Since our inference process is iterative, a convergence criterion is required in order to end it and limit overcorrection risks. Since our goal is to merge cohorts in such a way that they appear to be drawn from the same distribution, we performed per-bin statistical tests. At each iteration, a p-value is computed based on the two-sample Kolmogorov-Smirnov test for each 1 Mb genome bin. Because p-values are randomly and possibly uniformly distributed under the null hypothesis [51] (that the two cohorts are representative of the same population), we assumed that the median of these p-values should be close to 0.5. In practice, we interrupt the optimisation process as soon as the median p-value exceeds this cutoff.

The convergence and its speed are to a large extent impacted by the regularisation rate λ, which needs to be picked carefully. When picking $p = 1$ and choosing the squared Euclidean distance as function d, then $Wp(C)$ and $R(X)$ are both average squared error functions and expected to have similar orders of magnitude after convergence. For this reason, a reasonable *a priori* choice for λ would be 1. However, we observed that this default value results in many situations where the model never satisfies the convergence criterion and ends up with a low median p-value (e.g., 0.25), despite a large number of iterations (> 1000). In many other situations, the model rapidly converges but introduces disruptions in the data, resulting in drastic information loss. Therefore, we chose to make λ adaptive and lower it every e iterations until the convergence criterion is met. Our motivation is to use the largest value for λ (to preserve the original data to the best extent possible) while ensuring that the two data distributions are no longer distinguishable. We arbitrarily chose an initial value $\lambda 0 = 1000$ and reduced λ by half every $h = 20$ iterations. Minimal value of λ was set to 1 in order to prevent overfitting risks in case where the model fails at reaching the convergence threshold.

3.6 Model validation and performance assessment

3.6.1 Downstream supervised learning

Domain adaptation aims at superimposing the data distributions originating from different domains. While this superimposition can be quantified through clustering metrics or visually assessed using kernel principal component analysis or *t*-SNE for example, additional validation is required to ensure that the predictive signal for the malignancies of interest has not been removed during the adaptation process. For this purpose, we trained widely used Machine Learning models for the detection of these malignancies before and after adaptation, using default hyperparameters. We trained logistic regressions, random forests and kernel support vector machines using the scikit-learn [52] Python package.

However, regular validation approaches, such as *k*-fold or leave-one-out cross-validation do not suit our setting, as they may show overoptimistic performance because of contamination between the training and the validation set. Indeed, the domain adaptation model should not be exposed to the validation set, since in the presence of overfitting some corrected samples from the training set will resemble samples from the validation set. Therefore, we propose a problem-specific validation method that excludes from the validation set any sample that does not belong to the target domain, as well as any sample that has been seen by the DA algorithm. During validation, adapted samples are only used to train the model and are not allowed to be left out, therefore not contributing to the estimation of performance. The aim is to prevent the supervised models to correctly assign a label (healthy/cancer) to the left-out sample just because the latter has been shifted arbitrarily far away from the real data distribution by the domain adaptation model. Such procedure can be repeated multiple times on random subsets, similarly to *k*-fold and leave-one-out cross-validation. The subsets generated during our validation procedure are illustrated in **Fig. 6D**.

Since the HEMA data set contains controls and cases in both domains, the validation has been performed in both directions: mapping samples from first domain to the second one before validating in the second one, and vice versa.

Performance of the supervised models for cancer detection was quantified using widely used metrics such as sensitivity, specificity, Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall curve (AUPR) and the Matthews Correlation Coefficient (MCC).

3.6.2 Evaluation of sample-to-sample mapping using paired samples

For the cohorts in which biological samples have been sequenced twice, we applied OT and assessed whether the distance between paired samples was indeed lowered by the adaptation process. More specifically we computed accuracy, measured as the percentage of profiles from the target domain correctly assigned to the corresponding profile in the source domain, using the closest profile (according to the Euclidean metric) as predictor. By construction, random counterpart assignment would result in a $1/n$ accuracy.

3.6.3 Downstream copy number aberration analysis

Finally, we assessed the ability of the different methods to preserve the copy number aberrations present in the original data. For this purpose, we built a reference set made of the 79 controls from domain D_9 (OV data set, see **Table 6**) and called CNAs in the cancer cohort from D_9, same data set. Next, we built a reference set based on the 39 controls from D_{10} and called CNAs in the same cancer cohort from D_9, after applying our tool to adapt them towards D_{10}. CNA calling was performed with ichorCNA, which would be theoretically expected to produce similar results in the two settings. Therefore, we quantified the agreement between ichorCNA results using accuracy and SOV REFINE score [55] between the estimated per-bin copy numbers. Also, we computed

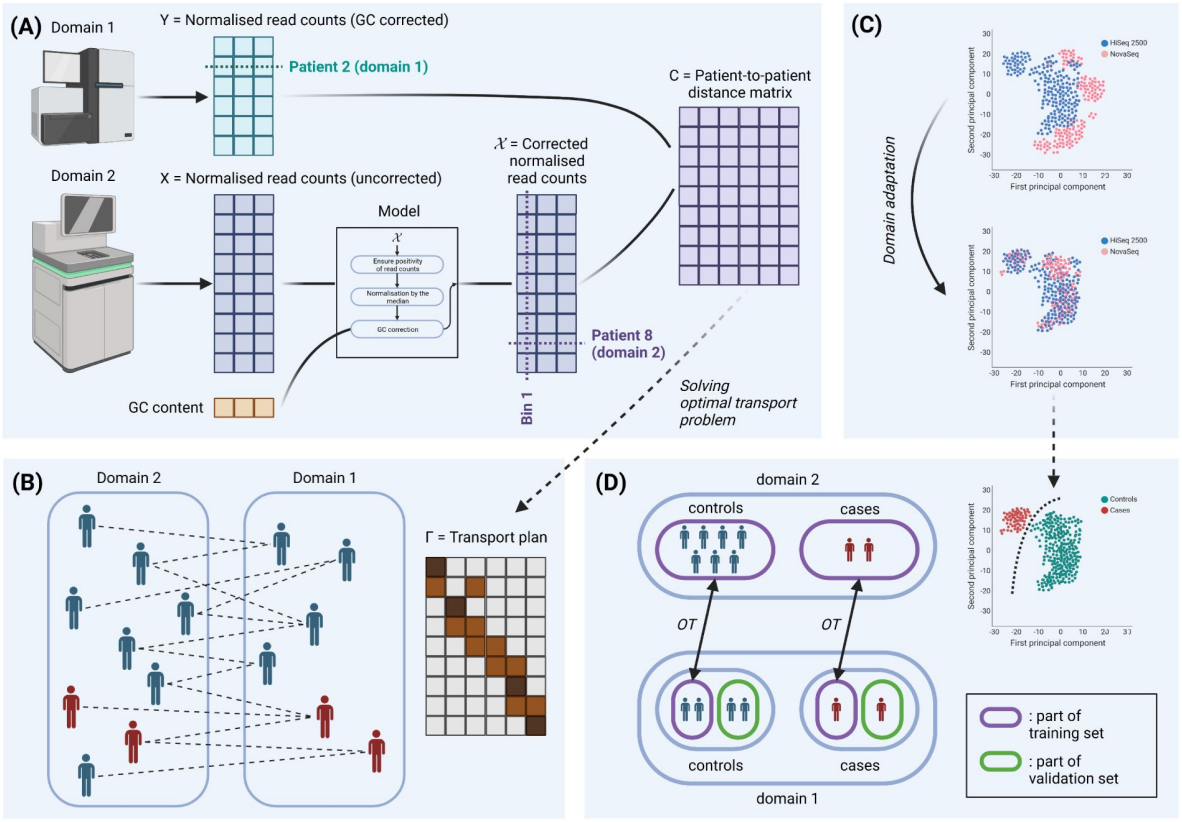

Figure 6

Illustrative summary of our methods.

(A) Given two cohorts of cfDNA samples differing by the sequencing pipeline that processed them, the model corrects the second cohort to match the distribution of the first one. After correction, the cost matrix for our OT problem is given by the pairwise Euclidean distances. (B) The solution of the OT problem, named transport plan, assigns patients from Domain 2 to similar patients in Domain 1. The model parameters are found by minimising the Wasserstein distance, as defined by the cost matrix and transport plan. (C) After inference, the two cohorts are merged and ready for downstream analysis. (D) Depiction of the validation procedure used for the purpose of this study.

the average absolute deviations in the parameters inferred by ichorCNA, including tumour fraction or tumour cellular prevalence. Since the tool also proposes to call CNAs in a reference-free fashion, we also conducted the same analysis without the two panels of normals.

Data availability

Haematological cancer cases and healthy controls constituting the HEMA data set are available from ArrayExpress (*https://www.ebi.ac.uk/arrayexpress/*) under accession number E-MTAB-10934 as part of the GIPXplore study.

Ovarian carcinoma and healthy controls (OV data set) from domain D_{10} of have been previously deposited at the European Genome-phenome Archive (EGA) under study no. EGAS00001005361.

The remaining samples (domain D_9 + NIPT dataset) are in-house cohorts.

Coverage profiles for all the samples have been compiled and uploaded to FigShare (DOI: 10.6084/m9.figshare.24459304).

Code availability

Our tool is available as an open source package at *https://github.com/AntoinePassemiers/DAGIP*.

Acknowledgements

Antoine Passemiers is funded by a Research Foundation – Flanders (FWO) doctoral fellowship (1SB2721N). Tatjana Jatsenko is funded by Agentschap Innoveren en Ondernemen (VLAIO; Flanders Innovation & Entreprenership grant HBC.2018.2108). Joris Robert Vermeesch is funded by FWO (G080217N, S003422N) and KU Leuven (no. C1/018). Daniele Raimondi is funded by a FWO post-doctoral fellowship (12Y5623N). Yves Moreau is funded by (1) Research Council KU Leuven: Symbiosis 4 (C14/22/125); CELSA Active Learning, (2) Innovative Medicines Initiative: MELLODY, (3) Flemish Government (ELIXIR Belgium, IWT, FWO 06260, FWO SBO MICADO S003422N, VLAIO ATHENA HBC.2019.2528) and (4) Impulsfonds AI: VR 2019 2203 DOC.0318/1QUATER Kenniscentrum Data en Maatschappij. An Coosemans is funded by the Flemish Cancer Society (2016/10728/2603). ctDNA samples within the Trans-IOTA study were financed by Kom Op Tegen Kanker (Stand Up to Cancer).

The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government. **Figure 6** has been created with BioRender.com.

Author contributions

Designed experiments: A.P, T.J, J.R.V, P.B, A.C, D.T, D.L. Designed computational methods: A.P, Y.M, D.R. Data collection and analysis: A.P, A.V, T.J. Performed computational experiments: A.P. Wrote the first draft of the manuscript: A.P, T.J, D.R, Y.M, J.R.V Revised and approved manuscript: all authors.

Competing interests

A.C is a contracted researcher for Oncoinvent AS and Novocure and a consultant for Sotio AS and Epics Therapeutics SA.

References

[1] Bianchi Diana W, Wilkins-Haug Louise (2014) **Integration of noninvasive dna testing for aneuploidy into prenatal care: what has happened since the rubber met the road?** *Clinical chemistry* **60**:78–87

[2] Leila Dardour *et al.* (2017) **Incidence of uncommon fetal aneuploidies detected by non-invasive prenatal testing** *17th annual Belgian Society of Human Genetics meeting: Human genetics goes somatic* :100–100

[3] De Vlaminck Iwijn *et al.* (2014) **Circulating cell-free dna enables noninvasive diagnosis of heart transplant rejection** *Science translational medicine* **6**:241–241

[4] Lenaerts L *et al.* (2019) **Genomewide copy number alteration screening of circulating plasma dna: potential for the detection of incipient tumors** *Annals of Oncology* **30**:85–95

[5] Tan EM, Schur PH, Carr RI (1966) **HG292857 Kunkel, et al. Deoxybonucleic acid (dna) and antibodies to dna in the serum of patients with systemic lupus erythematosus** *The Journal of clinical investigation* **45**:1732–1740

[6] Koffler David, Agnello Vincent, Winchester Robert, Kunkel Henry G, et al. (1973) **The occurrence of single-stranded dna in the serum of patients with systemic lupus erythematosus and other diseases** *The Journal of clinical investigation* **52**:198–204

[7] Jiang Peiyong, Dennis Lo YM (2016) **The long and short of circulating cell-free dna and the ins and outs of molecular diagnostics** *Trends in Genetics* **32**:360–371

[8] Newman Aaron M *et al.* (2014) **An ultrasensitive method for quantitating circulating tumor dna with broad patient coverage** *Nature medicine* **20**:548–554

[9] Phallen Jillian *et al.* (2017) **Direct detection of early-stage cancers using circulating tumor dna** *Science translational medicine* **9**

[10] Vandenberghe Peter *et al.* (2015) **Noninvasive detection of genomic imbalances in hodgkin/reed-sternberg cells in early and advanced stage hodgkin's lymphoma by sequencing of circulating cell-free dna: a technical proof-of-principle study** *The Lancet Haematology* **2**:e55–e65

[11] Lenaerts Liesbeth *et al.* (2020) **Breast cancer detection and treatment monitoring using a noninvasive prenatal testing platform: utility in pregnant and nonpregnant populations** *Clinical Chemistry* **66**:1414–1423

[12] Diaz Luis A, Bardelli Alberto (2014) **Liquid biopsies: genotyping circulating tumor dna** *Journal of clinical oncology* **32**

[13] Cristiano Stephen *et al.* (2019) **Genome-wide cell-free dna fragmentation in patients with cancer** *Nature* **570**:385–389

[14] Mouliere Florent *et al.* (2018) **Enhanced detection of circulating tumor dna by fragment size analysis** *Science translational medicine* **10**

[15] Mouliere Florent, Robert Bruno, Peyrotte Erika Arnau, Del Rio Maguy, Ychou Marc, Molina Franck, Gongora Celine, Thierry Alain R (2011) **High fragmentation characterizes tumour-derived circulating dna** *PloS one* **6**

[16] Snyder Matthew W, Kircher Martin, Hill Andrew J, Daza Riza M, Shendure Jay (2016) **Cell-free dna comprises an in vivo nucleosome footprint that informs its tissues-of-origin** *Cell* **164**:57–68

[17] Dennis Lo YM, Han Diana SC, Jiang Peiyong, Chiu Rossa WK (2021) **Epigenetics, fragmentomics, and topology of cell-free dna in liquid biopsies** *Science* **372**

[18] Jiang Peiyong *et al.* (2018) **Preferred end coordinates and somatic variants as signatures of circulating tumor dna associated with hepatocellular carcinoma** *Proceedings of the National Academy of Sciences* **115**:E10925–E10933

[19] Allen Chan KC *et al.* (2016) **Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred dna ends** *Proceedings of the National Academy of Sciences* **113**:E8159–E8168

[20] Jiang Peiyong *et al.* (2015) **Lengthening and shortening of plasma dna in hepatocellular carcinoma patients** *Proceedings of the National Academy of Sciences* **112**:E1317–E1325

[21] Liu Minetta C *et al.* (2020) **Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free dna** *Annals of Oncology* **31**:745–759

[22] Bae Mingyun *et al.* (2017) **Integrative modeling of tumor genomes and epigenomes for enhanced cancer diagnosis by cell-free dna** *Nature Communications* **14**

[23] Peneder Peter *et al.* (3230) **Multimodal analysis of cell-free dna whole-genome sequencing for pediatric cancers with low mutational burden** *Nature communications* **12**

[24] Messaoudi Safia El, Rolet Fanny, Mouliere Florent, Thierry Alain R (2013) **Circulating cell free dna: preanalytical considerations** *Clinica chimica acta* **424**:222–230

[25] Bronkhorst Abel Jacobus, Aucamp Janine, Pretorius Piet J (2015) **Cell-free dna: preanalytical variables** *Clinica Chimica Acta* **450**:243–253

[26] Meddeb Romain, Pisareva Ekaterina, Thierry Alain R (2019) **Guidelines for the preanalytical conditions for analyzing circulating cell-free dna** *Clinical chemistry* **65**:623–633

[27] Till Jacob E *et al.* (2021) **Optimization of sources of circulating cell-free dna variability for downstream molecular analysis** *The Journal of Molecular Diagnostics* **23**:1545–1552

[28] van Dessel Lisanne F, Vitale Silvia R, Helmijr Jean CA, Wilting Saskia M, van der Vlugt-Daane Michelle, Hoop Esther Oomen-de, Sleijfer Stefan, Martens John WM, Jansen Maurice PHM, Lolkema Martijn P (2019) **High-throughput isolation of circulating tumor dna: a comparison of automated platforms** *Molecular oncology* **13**:392–402

[29] Ross Michael G, Russ Carsten, Costello Maura, Hollinger Andrew, Lennon Niall J, Hegarty Ryan, Nusbaum Chad, Jaffe David B (2013) **Characterizing and measuring bias in sequence data** *Genome biology* **14**:1–20

[30] Sato Mitsuhiko P, Ogura Yoshitoshi, Nakamura Keiji, Nishida Ruriko, Gotoh Yasuhiro, Hayashi Masahiro, Hisatsune Junzo, Sugai Motoyuki, Takehiko Itoh, Hayashi Tetsuya (2019) **Comparison of the sequencing bias of currently available library preparation kits for illumina sequencing of bacterial genomes and metagenomes** *DNA Research* **26**:391–398

[31] Costello Maura *et al.* (2018) **Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms** *BMC genomics* **19**:1–10

[32] Sinha Rahul *et al.* (2017) **Index switching causes "spreading-of-signal" among multiplexed samples in illumina hiseq 4000 dna sequencing** *BioRxiv, page* **125724**

[33] Browne Patrick Denis, Nielsen Tue Kjærgaard, Kot Witold, Aggerholm Anni, Gilbert M Thomas P, Puetz Lara, Rasmussen Morten, Zervas Athanasios, Hansen Lars Hestbjerg (2020) **Gc bias affects genomic and metagenomic reconstructions, underrepresenting gc-poor organisms** *GigaScience* **9**

[34] Sun Baochen, Feng Jiashi, Saenko Kate (2016) **Return of frustratingly easy domain adaptation** *Proceedings of the AAAI Conference on Artificial Intelligence* **30**

[35] Ben-David Shai, Blitzer John, Crammer Koby, Kulesza Alex, Pereira Fernando (2010) **and Jennifer Wortman Vaughan. A theory of learning from different domains** *Machine learning* **79**:151–175

[36] Cheung Ming-Sin, Down Thomas A, Latorre Isabel, Ahringer Julie (2011) **Systematic bias in high-throughput sequencing data and its correction by beads** *Nucleic acids research* **39**:e103–e103

[37] Mancini Massimiliano, Bulo Samuel Rota, Caputo Barbara, Ricci Elisa (2018) **Best sources forward: domain generalization through source-specific nets** *2018 25th IEEE international conference on image processing (ICIP)* :1353–1357

[38] Benjamini Yuval, Speed Terence P (2012) **Summarizing and correcting the gc content bias in high-throughput sequencing** *Nucleic acids research* **40**:e72–e72

[39] Alkan Can *et al.* (2009) **Personalized copy number and segmental duplication maps using next-generation sequencing** *Nature genetics* **41**:1061–1067

[40] Chandrananda Dineika, Thorne Natalie P, Ganesamoorthy Devika, Bruno Damien L, Benjamini Yuval, Speed Terence P, Slater Howard R, Bahlo Melanie (2014) **Investigating and correcting plasma dna sequencing coverage bias to enhance aneuploidy discovery** *PLoS One* **9**

[41] Adalsteinsson Viktor A *et al.* (2017) **Scalable whole-exome sequencing of cell-free dna reveals high concordance with metastatic tumors** *Nature communications* **8**:1–13

[42] Larson Nicholas B, Larson Melissa C, Na Jie, Sosa Carlos P, Wang Chen, Kocher Jean-Pierre, Rowsey Ross (2020) **Coverage profile correction of shallow-depth circulating cell-free dna sequencing via multidistance learning** *PACIFIC SYMPOSIUM ON BIOCOMPUTING* :599–610

[43] Chen Lin, Chen Huaian, Wei Zhixiang, Jin Xin, Tan Xiao, Jin Yi, Chen Enhong (2022) **Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation** *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages* :7181–7190

[44] Ganin Yaroslav, Ustinova Evgeniya, Ajakan Hana, Germain Pascal, Larochelle Hugo, Laviolette François, Marchand Mario, Lempitsky Victor (2016) **Domain-adversarial training of neural networks** *The journal of machine learning research* **17**:2096–2030

[45] Huang Jiayuan, Gretton Arthur, Borgwardt Karsten, Schölkopf Bernhard, Smola Alex (2006) **Correcting sample selection bias by unlabeled data** *Advances in neural information processing systems* **19**

[46] Xie Rongchang, Yu Fei, Wang Jiachao, Wang Yizhou, Zhang Li (2019) **Multi-level domain adaptive learning for cross-domain detection** *In Proceedings of the IEEE/CVF international conference on computer vision workshops* :0–0

[47] Hoffman Judy, Tzeng Eric, Park Taesung, Zhu Jun-Yan, Isola Phillip, Saenko Kate, Efros Alexei, Darrell Trevor (2018) **Cycada: Cycle-consistent adversarial domain adaptation** *International conference on machine learning* :1989–1998

[48] Azodi Christina B, Tang Jiliang, Shiu Shin-Han (2020) **Opening the black box: interpretable machine learning for geneticists** *Trends in genetics* **36**:442–455

[49] Bonneel Nicolas, Van De Panne Michiel, Paris Sylvain, Heidrich Wolfgang (2011) **Displacement interpolation using lagrangian mass transport** *In Proceedings of the 2011 SIGGRAPH Asia conference* :1–12

[50] Courty Nicolas, Flamary Rémi, Habrard Amaury, Rakotomamonjy Alain (2017) **Joint distribution optimal transportation for domain adaptation** *Advances in Neural Information Processing Systems* **30**

[51] Murdoch Duncan J, Tsai Yu-Ling, Adcock James (2008) **P-values are random variables** *The American Statistician* **62**:242–245

[52] Pedregosa F. *et al.* (2011) **Scikit-learn: Machine learning in Python** *Journal of Machine Learning Research* **12**:2825–2830

[53] Che Huiwen *et al.* (2022) **Pan-cancer detection and typing by mining patterns in large genome-wide cell-free dna sequencing datasets** *Clinical Chemistry* **68**:1164–1176

[54] Vanderstichele Adriaan *et al.* (2017) **Chromosomal instability in cell-free dna as a highly specific biomarker for detection of ovarian cancer in women with adnexal masses** *Clinical Cancer Research* **23**:2223–2231

[55] Liu Tong, Wang Zheng (2018) **Sov refine: a further refined definition of segment overlap score and its significance for protein structure similarity** *Source code for biology and medicine* **13**:1–10

[56] Lenaerts Liesbeth *et al.* (2021) **Comprehensive genome-wide analysis of routine non-invasive test data allows cancer prediction: A single-center retrospective analysis of over 85,000 pregnancies** *EClinicalMedicine* **35**

[57] Buedts Lieselot, Wlodarska Iwona, Finalet-Ferreiro Julio, Gheysens Olivier, Dehaspe Luc, Tousseyn Thomas, Fornecker Luc-Matthieu, Lazarovici Julien, et al. (2021) **The landscape of copy number variations in classical hodgkin lymphoma: a joint ku leuven and lysa study on cell-free dna** *Blood advances* **5**:1991–2002

[58] Li Heng, Durbin Richard (2009) **Fast and accurate short read alignment with burrows–wheeler transform** *bioinformatics* **25**:1754–1760

[59] **Picard toolkit. https://broadinstitute.github.io/picard/, 2019.**

[60] McKenna Aaron *et al.* (2010) **The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data** *Genome research* **20**:1297–1303

[61] Cleveland William S (1979) **Robust locally weighted regression and smoothing scatterplots** *Journal of the American statistical association* **74**:829–836

[62] Seabold Skipper, Perktold Josef (2010) **Statsmodels: Econometric and statistical modeling with python** *In Proceedings of the 9th Python in Science Conference* **57**:10–25080

[63] Genevay Aude, Cuturi Marco, Peyré Gabriel, Bach Francis (2016) **Stochastic optimization for large-scale optimal transport** *Advances in neural information processing systems* **29**

Editors

Reviewing Editor
Yongliang Yang
Dalian University of Technology, Dalian, China

Senior Editor
Aleksandra Walczak
École Normale Supérieure - PSL, Paris, France

Reviewer #1 (Public Review):

Summary:

The authors applied a domain adaptation method using the principal of optimal transport (OT) to superimpose read count data onto each other. While the title suggests that the presented method is independent from and performs better than other methods of bias correction, the presented work uses a self-implemented version of GC bias correction apart of the OT domain adaptation. Performance comparisons were done both on normalized read counts as well as on copy number profiles which is already the complete set of presented use cases. Results involving copy number profiles from iChorCNA were also subjected to the bias correction measures implemented there. It is not clear at many points which correction method actually causes the observed performance.

Strengths:

The quality of superimposing distributions of normalized read counts (and copy number profiles) was sufficiently shown using uniformly distributed p-values in the interval of 0 to 1 for healthy controls D7 and D8 which differed in the choice of library preparation kit.

The ability to select a sample from the source domain for samples in the target domain was demonstrated.

Weaknesses:

Experiment Design:

The chosen bias correction methods are not explicitly designed for nor aimed at domain adaptation. The benchmark against GC bias correction while doing GC bias correction during the OT procedure is probably the most striking flaw of the entire work. GC bias correction has the purpose of correction GC biases, wherever present, NOT correcting categorical pre-analytical variables of undefined character. A more thorough examination of the presented results should address why plain iChor CNA is the best performing "domain adaptation" in some cases. Also, the extent to which the implemented GC bias correction is contributing to the performance increase independent of the OT procedure should be assessed separately in each case.
Moreover, the center-and-scale standardization is probably not the most relevant contestant in domain adaptation that is out there.

Comparison of cohorts (domains) - especially healthy from D7 and D8 - it is not described which type of ChIP analysis was done for the healthy controls of the D7 domain. The utilized library preparation kit implies that D7 represents a subset of available cfDNA in a plasma sample by precipitating only certain cfDNA fragments to which undisclosed type of protein was bound. Even if the type of protein turns out to be histones, the extracted subset of cfDNA should not be regarded as coming from the same distribution of cfNDAs. For example, fragments with sub-mononucleosomal length would be depleted in the ChIP-seq data set while these could be extracted in an untargeted cfDNA sequencing data set. It needs to be clarified why the authors deem D7 and D8 healthy controls to be identical with regards to SCNA analysis. Best start with the protein targets of D7 ChIP-seq samples.

From the Illumina TruSeq ChIP product description page:
"TruSeq ChIP Libary Preparation Kits provide a simple, cost-effective solution for generating chromatin immunoprecipitation sequencing (ChIP-Seq) libraries from ChIP-derived DNA. ChIP-seq leverages next-generation sequencing (NGS) to quickly and efficiently determine the distribution and abundance of DNA-bound protein targets of interest across the genome."

Redundancy:

Some parts throughout the results and discussion part reappear in the methods. The description of the methodology should be concentrated in the method section and only reiterated in a summarizing fashion where absolutely necessary.
Unnecessary repetition inflate the presented work which is not appealing to the reader. Rather include more details of the utilized materials and methods in the corresponding section.

Transparency:

At the time point of review, the code was not available under the provided link.
A part of the healthy controls from D8 is not contained under the provided accession (367 healthy samples are available in the data base vs. sum of D7 and D8 healthy controls is 499)

Neither in the paper nor in reference 4 is an explanation of what was targeted with the ChIP-seq approach.

Consistency:

It is not evident why a ChIP-seq library prep kit was used (sample cohorts designated as D7). The DNA isolation procedure was not presented as having an immunoprecipitation step. Furthermore, it is not clear which DNA bound proteins were targeted during ChIP seq, if such an immunoprecipitation was actually carried out.The authors self-implemented a GC bias correction procedure although they already mentioned other procedures earlier like LIQUORICE. Also, there already exist tools that can be used to correct GC bias, like deepTools (github.com/deeptools/deepTools). Other GC bias correction algorithms designed specifically

for cfDNA would be Griffin (github.com/adoebley/Griffin) and GCparagon (github.com/BGSpiegl/GCparagon). When benchmarking against state-of-the-art cfDNA GC bias correction, these algorithms should appear in a relevant scientific work, somewhere other than the introduction, preferably in the results section. It should be shown that the chosen GC bias correction method is performing best under the given circumstances.

Accuracy:

Use clear labels for each group of samples. The domain number is not sufficient to effectively distinguish sample groups. Already the source name plus a simple enumeration would improve the clarity at some points.

The healthy controls of D7 and D8 are described but the numbers do not add up (257 healthy controls in line 227 vs. 260 healthy controls in line 389). Please double check this and use representative sample cohort labels in the materials description for improved clarity!

Avoid statements like "the rest" when talking about a mixed set of samples. It is not clear how many samples from which domain are addressed.

For optimal transport, knowledge about the destination is required ("where do I want to transport to?") and, thus, the proposed method can never be unsupervised. It is always necessary to know the label of both the source and target domains. In practice, this is not often the case and users might fall prey to the error of superimposing data that is actually separated by valid differences in some experimental variables.

Seemingly arbitrary cutoff values are mentioned. For example, it is not clear if choosing "the cutoff that produced the highest MCCs" is meant across methods or for each method separately (are the results for each method reported that also resulted in the highest MCC for that method?).

The Euclidean metric for assessing the similarity of (normalized) read counts is questionable for a high dimensional space: read counts are assessed for 1 Mb genomic intervals which yields around 3000 intervals (dimensions), depending on the number of excluded intervals (which was not described in more detail). There might be more appropriate measures in this high dimensional space.

It is sometimes not clear what data actually is presented. An example would be the caption of Figure 2, (C): it is suggested that all (320) ovarian cancer cases are shown in one copy number profile.

Furthermore, the authors do not make a distinction between male and female samples. A clarification is needed why the authors think SCNAs of ovarian cancer samples should be called against a reference set that contains male controls.
The procedure would likely benefit from a strict separation of male and female cases which would also allow for chrX (and chrY) being included in downstream analysis.

The GC bias and mappability correction implicitly done by iChorCNA for the SCNA profile comparison is presented as "no correction" which is highly misleading. (for clarification, this is also deemed inappropriate, not just inaccurate))

The majority of interpretations presented procedure does not give any significant improvement regarding the similarity of copy number profiles are off and in many instances favor the OT procedure in an unscientific and highly inappropriate manner.

Apart of duplicate marking (which is not specified any further - provide the command(s)!), there is no information on which read (pairs) were used (primary, secondary, supplementary, mapped in a proper pair, fragment length restrictions, clipping restrictions, etc.). The authors

should explain why base quality score re-calibration was done as this might be an unnecessary step if the base quality values are not used later on.

The adaptation method presented as "center-and-scale standardization" is inappropriate for unbalanced cancer profiles since it assumes the presence of identical SCNAs in all samples belonging to the same cancer entity.
Please explain why normalizing 1 Mb genomic intervals to the average copy number across different cancer samples should be valid or use another domain adaptation method for performance comparison.

Statements like in line 83 (unsupervised DA) are plain wrong because transport from one domain to another requires the selection of a target domain based on a label, e.g., based on health status, cancer entity, or similar.

Relevance and Appropriateness:

Many of the presented results are not relevant or details of the procedure were incomprehensible or incomplete: the results presented in table 2 - sample assignment. The Euclidean metric seems to be inappropriate for high dimensional data. Also the selection of the cutoff based on Euclidean distance seems to enable the optimization in favor of the OT procedure. It is hypothesized that there might exist other cutoff values for which the selection of samples form the source domain would also work for other correction methods but this is not further described. It could simply be the case that OT can assign a relationship between domains

The statement that there are no continuous pre-analytical variables is wrong (304). The effect of target depth-of-coverage (DoC) was not analyzed although this represents one of the most common (continuous) and difficult to control variables in NGS data analysis. The inclusion of multiple samples from a single patient in a cohort likely represents introduction of a confounding factor ["contamination"] to the model training procedure: the temporal difference that lies between the taken samples of that patient represents leakage of information. As far as can be told from the presented data, this potential bias has not been ruled out (e.g., exclusion of all samples beyond the first from each patient or alternatively: picking all samples of a patient either for the training set or the test set).

Conscientiousness:

Statements like "good"/"best" on their own should be avoided. A clear description of why a certain procedure/methodology/algorithm performs better should be preferred in scientific writing (e.g., "highest MCC values" instead of "best MCC values").
Otherwise, such statements represent mere opinions of the author rather than an unbiased evaluation of the results.
The domain D8 of healthy controls seems to contain samples from multiple sources (some published other in-house). Contrary to the data availability statement (533), not all healthy control samples of the HEMA data set are available from ArrayExpress

Other Major Concerns:

Potential Irrelevance:

The manuscript represents a mere performance assessment of the proposed sWGS per-bin-read-count fitting procedure and, thus, a verification in its character, not a validation (although the model training itself was "validated" - but this is to be viewed separately from the validity of the achieved correction in a biological context). A proper (biological) validation is missing.

It is of utmost importance that parameters of the adapted (transported) samples -that lie outside of what has been optimized to be highly similar- are checked to actually validate the procedure. Especially biological signals and genome-wide parameters (GC content distribution before/after transport) need to be addressed also in hindsight of the rampant criticism towards GC bias correction by the authors. At no point in the manuscript was GC bias addressed properly, i.e., how much of an improvement is expected from GC bias correction if there is no significant GC bias?

The (potential - not clear so far) ability of making ChIP-seq data look like cfDNA data (even if only the copy number profiles SCNAs appear highly similar) raises the concern of potential future users of the tool to superimpose domains that should not be superimposed form a biological point of view because the true domain the superimposed cohorts belong to are different. The ability to superimpose anything onto anything s troubling. There is no control mechanism that allows for failure in cases where the superposition is invalid.

Chromosome X was excluded which could be avoided if data sets were split according to biological sex.

The difference between the distributions was never attributed to GC bias, hence, the benchmark against GC bias correction tools might not be relevant in the first place.

Stability of OT data transformation:

The authors state that the straight forward choice of lambda resulted in many occasions where disruptions (of unspecified nature and amplitude) are introduced in the copy number profiles of transformed data. It is not evident from the proposed work to which extent this behavior was removed from the procedure and if it can occur and how the user could resolve such a problem on their own.

In summary, the presented work needs considerable adaptation and additions before it can actually be considered a valuable contribution to the liquid biopsy field.

https://doi.org/10.7554/eLife.98116.1.sa1

Reviewer #2 (Public Review):

The authors present a computational methodology for de-biasing/denoising high-throughput genomic signals using optimal transport techniques, thus allowing disparate datasets to be merged and jointly analysed. They apply this methodology on liquid biopsy data and they demonstrate improved performance (compared to simpler bias-correcting approaches) for cancer detection using common machine learning algorithms. This is a theoretically interesting and potentially useful approach for addressing a very common practical problem in computational genomics.

I have the following recommendations:

(1) When comparing performance metrics between different approaches (e.g., tables 3 and 4), 95% confidence intervals should also be provided and a pairwise statistical test should be applied to establish whether the observed difference in each performance metric between the proposed method and the alternatives is statistically significant, thus justifying the claim that the proposed method offers an improvement over existing methodologies.

(2) The commonly used center-and-scale and GC debias approaches presented by the authors are fairly simple. How does their methodology compare to more elaborate approaches, such as tangent normalisation (https://academic.oup.com/bioinformatics/article/38/20/4677/6678978) and robust PCA (https://github.com/mskilab-org/dryclean)?

(3) What is the computational cost of the proposed methodology and how does it compare to the alternatives?

(4) The proposed approach relies on a reference dataset, against which a given dataset is adapted. What are the implications for cross-validation experiments (which are essential for assessing the out-of-sample error of every methodology), particularly with regards to the requirement to avoid information leakage between training and validation/test data sets?

In conclusion, this is an interesting and potentially useful paper and I would like to encourage the authors to address the above points, which hopefully will strengthen their case.

https://doi.org/10.7554/eLife.98116.1.sa0