BIOINFORMATICS

Vol. 00 no. 00 2018
Pages 1–13

Evaluation of cross-platform and interlaboratory concordance via consensus modelling of genomic measurements

Timothy J. Peters [1,*], Hugh J. French [1,2], Stephen T. Bradford [1,3], Ruth Pidsley [1], Clare Stirzaker [1,4], Hilal Varinli [1,3,5,6], Shalima Nair [1], Wenjia Qu [1], Jenny Song [1], Katherine A. Giles [1], Aaron L. Statham [1], Helen Speirs [7], Terence P. Speed [8,9,†] and Susan J. Clark [1,4,†]

[1]Epigenetics Laboratory, Genomics and Epigenetics Division, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia,
[2]South Western Sydney Clinical School, Faculty of Medicine, University of New South Wales, Liverpool, NSW 2170, Australia,
[3]CSIRO Health and Biosecurity, PO Box 52, North Ryde, NSW 1670, Australia,
[4]St Vincent's Clinical School, Faculty of Medicine, UNSW, Darlinghurst, NSW 2010, Australia,
[5]Department of Biological Sciences, Macquarie University, North Ryde, New South Wales 2109, Australia,
[6]NSW Ministry of Health, LMB 961, North Sydney, NSW 2059, Australia,
[7]Ramaciotti Centre for Genomics, University of New South Wales, Randwick, NSW 2031, Australia,
[8]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia and
[9]Department of Mathematics & Statistics, University of Melbourne, Melbourne, VIC 3010, Australia.
[*]To whom correspondence should be addressed.
[†]Contributed equally.

ABSTRACT

Motivation: A synoptic view of the human genome benefits chiefly from the application of nucleic acid sequencing and microarray technologies. These platforms allow interrogation of patterns such as gene expression and DNA methylation at the vast majority of canonical loci, allowing granular insights and opportunities for validation of original findings. However, problems arise when validating against a "gold standard" measurement, since this immediately biases all subsequent measurements towards that particular technology or protocol.

Since all genomic measurements are estimates, in the absence of a "gold standard" we instead empirically assess the measurement precision and sensitivity of a large suite of genomic technologies via a consensus modelling method called the row-linear model. This method is an application of the American Society for Testing and Materials Standard E691 for assessing interlaboratory precision and sources of variability across multiple testing sites. Both cross-platform and cross-locus comparisons can be made across all common loci, allowing identification of technology- and locus-specific tendencies.

Results: We assess technologies including the Infinium MethylationEPIC BeadChip, whole genome bisulfite sequencing (WGBS), two different RNA-Seq protocols (PolyA+ and Ribo-Zero) and five different gene expression array platforms. Each technology thus is characterised herein, relative to the consensus.

We showcase a number of applications of the row-linear model, including correlation with known interfering traits. We demonstrate a clear effect of cross-hybridisation on the sensitivity of Infinium methylation arrays. Additionally, we perform a true interlaboratory test on a set of samples interrogated on the same platform across twenty-one separate testing laboratories.

Availability: A full implementation of the row-linear model, plus extra functions for visualisation, are found in the `R` package `consensus` at `https://github.com/timpeters82/consensus`.

Contact: t.peters@garvan.org.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The reproducibility of scientific results from multiple sources is critical to the establishment of scientific doctrine (Popper (2005); Fisher (1971)). The spread of these sources can occur over various domains, including temporal, geographical, and technological. From each of these domains, one can expect a degree of variation in results, attributable to the ambient laboratory environment or intrinsic qualities of the measurement device. In many cases, the degree of this variation exceeds that of what is expected, leading to a "crisis of reproducibility" (Baker (2016); Nosek and Errington (2017); Begley and Ellis (2012)) in science. These sources of variation are commonly known as the batch effect, particularly in

genomics, and ameliorating methods exist for when it confounds the biological effect of interest across samples (Johnson *et al.* (2007); Leek and Storey (2007); Oytam *et al.* (2016)). However, an additional and less studied source of variation is that which occurs when an attempt is made to recapitulate results from a given technological platform on a subsequent platform, on the same set of samples and target loci. Platform-specific effects exist, and cross-platform studies (SEQC/MAQC-III Consortium (2014); Irizarry *et al.* (2005); Wang *et al.* (2005); Li *et al.* (2014); Wang *et al.* (2014); Holik *et al.* (2017); Nazarov *et al.* (2017)) have been undertaken in order to characterise them. However, evaluation of concordance in these studies is mostly geared towards assessing relative performance of the technologies in question, and this is exemplified by associative metrics such as Venn diagrams of differentially expressed genes or pairwise coefficients of correlation. While informative, very little attention is given to evaluating the more fundamental trait of measurement quality. The reliability of a measurement not only influences the performance of the technology that occasioned it for scientific ends, but more practically it gives confidence to its operator of its fealty to the quantum of interest. A governing framework for the assessment of measurement robustness for a given suite of genomic technologies is absent from the current literature.

Since the turn of the 21^{st} century, genomic science has presented researchers with a choice of platforms on which to interrogate their biological samples, chief among them microarrays and sequencing technologies. The microarray revolutionised whole-genome science, especially in the field of transcriptomics, where the relative expression levels of most known genes could be profiled using high-throughput hybridisation techniques (Kevil *et al.* (1997)). Later, microarrays such as the Illumina Infinium HumanMethylation450 (Bibikova *et al.* (2011)) and MethylationEPIC (Pidsley *et al.* (2016)) were introduced to measure the levels of DNA methylation in human samples. The intensity of the fluorescent dye (indicating hybridisation levels) on RNA expression and DNA methylation microarray chips serves as the metric by which the genomic feature is measured. This is an analogue reading, and hence follows a continuous distribution.

In the last decade, high-throughput sequencing has replaced the microarray as the assay of choice for many researchers. Sequencing technologies such as RNA-Seq (Lister *et al.* (2008)) and whole genome bisulfite sequencing (WGBS - Lister *et al.* (2008, 2009)) provide a richer set of biological information than microarrays, especially on anomalous features such as single nucleotide variations (SNVs) and alternative splicing events. In contrast to the analogue measurements of microarrays, sequencing measurements are represented digitally as a pileup of molecule counts per feature of interest. However, these still will not be an exact measure of the true level of the quantum of interest, since the shearing and subsequent sequencing of nucleic acid molecules produces a non-exhaustive sample from the population. Nucleic acid fragments compete for amplification within the assay, thus there is a stochastic component of the measurement for a given genomic feature (such as transcript abundance or methylation level of a CpG site).

Following the maxim of scientific reproducibility, genomic studies nevertheless aim to recapitulate the results derived from one technology on those from a subsequent platform. But what if this doesn't happen? To what do we ascribe the discrepancy? When confronted with a discordant result such as a low coefficient of correlation or a Venn diagram with poor overlap, the first instinct is to define a gold standard to which each deviant measurement can be compared. However, this is a dubious strategy in a number of ways. If the gold standard is chosen from one of the available platforms or protocols, a bias is immediately incurred, since error will be present in every measurement, regardless of its source.

Targeted approaches such as quantitative PCR (qPCR) and amplicon sequencing are accepted as more reliable, yet are impractical to assess from a genome-wide perspective due to constraints on time, reagents and nucleic acid yield. Furthermore, qPCR is still susceptible to amplification biases relating to sequence composition (Warnecke *et al.* (1997); Aird *et al.* (2011)). Most alarmingly, thorough quality-control studies such as SEQC and MAQC (SEQC/MAQC-III Consortium (2014); Shi *et al.* (2006)) reveal systematic biases across assay sites for microarrays, qPCR and RNASeq, even with molecular-level quality control measures via ERCC spike-ins (Baker *et al.* (2005); Jiang *et al.* (2011)) taken for the latter. As such, the SEQC study concludes there is no single "gold standard" technology or strategy to which gene expression measurements ought to conform, and biases remain despite the best efforts made to standardise them.

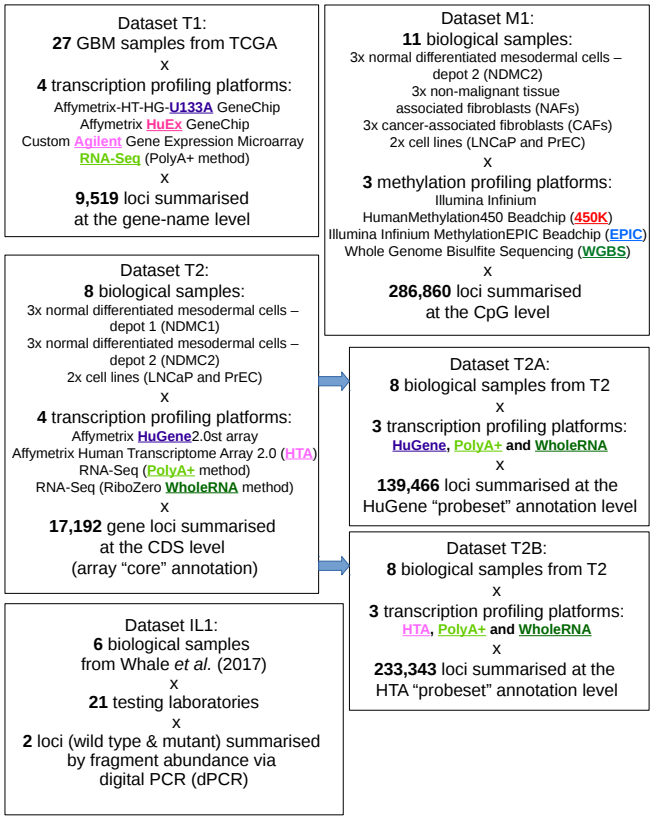


Fig. 1: Description of biological samples used in this study. Datasets are named T1 and T2(A,B) for transcription, M1 for methylation and IL1 for interlaboratory testing.

Instead of defining a gold standard from existing or further laboratory work, we present an alternative by building a consensus on existing data, based upon non-additive linear modelling. The method described forthwith was elucidated by the late American statistician John Mandel (Mandel (1984); Mandel and Lashof (1969)), in order to characterise both within-laboratory and cross-laboratory traits from a suite of manufactured products. It is now recognised as a standard by the American Society for Testing and Materials (ASTM International)(Mandel (1994)) for the identification of sources of variation and outliers from interlaboratory studies. Many variations on the interlaboratory test procedure exist (mostly summarised by Iyer *et al.* 2004), but we restrict the application to that described in the ASTM standard, in particular the final equation in that publication (see Methods section).

The method described here is applied to the cross-platform case, with measurements from each platform represented by those from separate "laboratories". By finding common loci across a minimum of three platforms, we can characterise per-locus, per-platform sensitivity and precision, and compare the resulting distributions. In addition, we perform a true interlaboratory study measuring DNA fragment abundance using digital PCR (dPCR)(Whale *et al.* (2017)), that evaluates measurement quality across testing sites. Given additional information from existing literature and platform annotation, we are then able to make valuable inferences about the susceptibility to bias of each platform that has contributed to the consensus.

2 METHODS

Approach

We applied the row-linear model to four distinct genomic data sets: two measuring gene expression levels, one measuring DNA methylation levels and one measuring DNA fragment abundance. An individual row-linear model was fitted to each locus $k = 1, \ldots, G$ for the G loci common to all platforms/conditions in each dataset. For some of these datasets G is in the order of hundreds of thousands, resulting in over half a million row-linear models being fit for this study. A complete summary of the samples and platforms used, and loci tested, can be found in Figure 1. In order for values to be made comparable, we use appropriate normalisations and transformations on each dataset (see Sample Sources and Preprocessing section in the Supplementary Material).

Dataset T1 allows us to assess the genewise measurement quality of four platforms (three microarray platforms and one RNA-Seq) using 27 samples from The Cancer Genome Atlas glioblastoma multiforme (GBM) study (Verhaak *et al.* (2010)).

Dataset T2 is an in-house dataset with 8 samples assessed over two more recently developed microarray platforms, and two separate RNA-Seq protocols. T2 was modelled at the gene locus (coding sequence) level, and then for T2A and T2B split into "probeset" levels according to annotation of the HuGene2.0st (HuGene) array and Human Transcriptome Array 2.0 (HTA) respectively. This was done for two reasons: firstly to fully interrogate the utility of these array platforms, and secondly to test whether a more granular summarisation of the datasets evinces subgenic trends. Both HuGene and HTA have targets summarising exonic or sub-exonic regions, over given genomic coordinates. These are characterised by the manufacturer as "probeset" level targets, whereas the gene-level summarisation is characterised as the "core" level. HuGene probeset targets are predominantly sub-exonic, whereas HTA targets are predominantly whole-exon.

Dataset M1 assesses the CpG-wise measurement quality of three DNA methylation platforms: two versions of Illumina's Infinium BeadChips (the discontinued HumanMethylation450 and its successor, the MethylationEPIC) and whole genome bisulfite sequencing (WGBS) across 11 samples, including those from Pidsley *et al.* (2016 & 2018).

Lastly, **Dataset IL1** represents a true interlaboratory study where the same 6 biological samples are tested for *KRAS* fragment abundance of wild type (WT) and mutant (G12D) loci at twenty-one separate laboratories (Whale *et al.* (2017)) using dPCR. For this dataset, the labwise sensitivity and precision is calculated.

Coloured text in Figure 1 sets the shorthand convention for technology platform/condition i for the remainder of this paper. For example, $b_{RNA-Seq}$ means the sensitivity b_i of RNA-Seq from dataset T1 when platform i is RNA-Seq, $d_{wholeRNA}$ means the estimate of precision d_i for RiboZero Whole RNA-Seq from dataset T2, and so on. Preprocessing and data accessibility for all data in this study can be found in the Sample Sources and Preprocessing section in the Supplementary Material.

The Row-Linear Model

We derive our method of cross-platform assessment from the description of ASTM standard E691 (Mandel (1994)), intended for the design and analysis of interlaboratory studies on a testing method.

Consider a matrix Z_{ij} of measurements at the same genomic locus, where the row index $i = 1, \ldots, p$ labels the whole-genome platforms (e.g. microarrays or sequencing assays) used and the column index $j = 1, \ldots, n$ labels the biological samples that are interrogated at that locus on each of the p platforms.

The row-linear model of the ASTM standard is

$$Z_{ij} = a_i + b_i(x_j - \bar{x}) + d_{ij} \tag{1}$$

where $x_j = p^{-1}\Sigma_i Z_{ij}$ is the average of the entries in the jth column, $a_i = n^{-1}\Sigma_j Z_{ij}$, the average of the entries in the ith row is the intercept, and b_i is the slope of the linear regression of $(Z_{ij}, j = 1, \ldots, n)$ on $(x_j - \bar{x}, j = 1, \ldots, n)$. Further, d_{ij} is the residual at x_j about the ith regression line, and the residual mean square about the ith fitted line is $d_i = (n-2)^{-1}\Sigma_j d_{ij}^2$.

Simply put, for each platform, the method regresses a set of measurements made on the n samples by that platform against the set of averages across the platforms of the p measurements on the samples.

We will refer to the parameters a_i, b_i and d_i as the *average*, the *sensitivity* and the *precision* of the ith platform at the locus in question, noting that higher precision corresponds to smaller values of d_i. Also, because the response variable is the sample mean of the predictor variables, we have the constraint $\Sigma_i b_i = p$, hence $\bar{b} = 1$. For most linear regressions, the residual term d_{ij} is informative about the suitability of the fit, but in the case of row-linear model it has more direct value in that it tells us about the precision of the group of measuring devices or strategies as a whole.

A minimum of three platforms is needed to fit a row-linear model, since two degrees of freedom are needed to calculate the individual

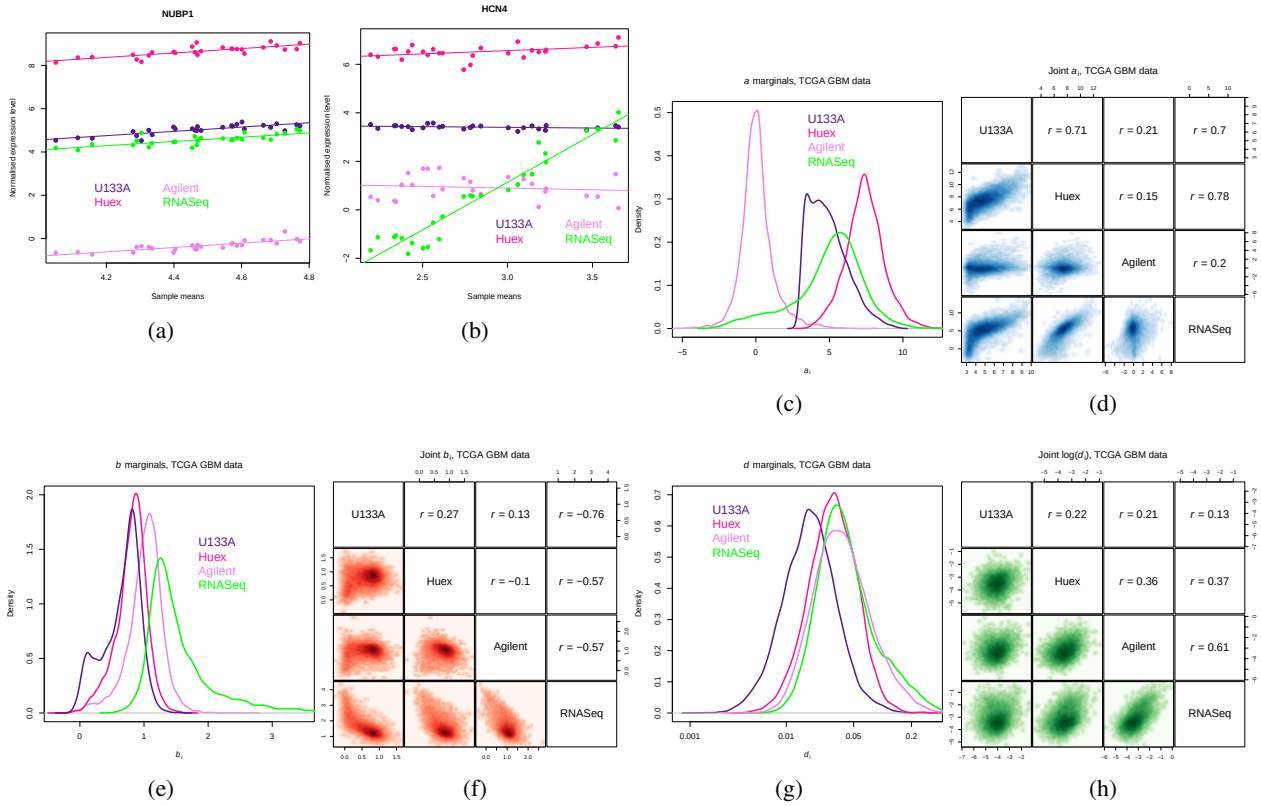


Fig. 2: Graphical depictions of row-linear fits for genes **(a)** NUBP1 and **(b)** HCN4 from Dataset T1. **(c)** Marginal and **(d)** joint distributions for parameter a_i, **(e)** marginal and **(f)** joint distributions for parameter b_i, **(g)** marginal and **(h)** joint distributions for parameter d_i, for the entirety of Dataset T1.

d_i from the residual sum of squares. Additionally, there must be true variation in each row i of Z_{ij}. Hence rows with, say, majority zero counts from RNA-Seq or fully methylated or unmethylated from whole genome bisulfite sequencing (WGBS) will produce trivial or artefactual results. Filtering steps for loci such as these are described in the Sample Sources and Preprocessing section in the Supplementary Material.

Informative extensions to the row-linear model (Mandel (1984)) include $V(a) = (p-1)^{-1}\Sigma_i(a_i - \bar{a})^2$ and $V(b) = (p-1)^{-1}\Sigma_i(b_i - 1)^2$. (In the 1984 Mandel monograph, the method is not referred to as the row-linear model, but instead an interlaboratory study of test methods. Also, the parameters a, b and d are instead represented by μ, β and η respectively). These give per-locus estimates of concordance (low values) or discordance (high values) over all technologies assessed. Given enough platforms, further sources of inter-platform variation could be explored such as the average precision $V(d) = (p-1)^{-1}\Sigma_i d_i$, the scatter about the regression line when b_i is plotted against a_i, or an estimate of z_0, a point along the dynamic range of the measurement at which (for some loci) the measurements converge. However, these extensions are beyond the scope of this study.

Independence between platforms is, in fact, not assumed for the row-linear model. Since the outputs are purely descriptive, any dependence between platforms will be reflected in the estimated parameters a_i, b_i and d_i. In this sense, the row-linear model should be seen as a *summarising*, rather than an *inferential* device.

All analyses performed in this paper, and the extensions mentioned above, are easily implemented using the R package *consensus* (`https://github.com/timpeters82/consensus`). The package performs multiple locus-wise row-linear fits per dataset. Plotting functions are available to visualise individual fits, marginal distributions of a_i, b_i and d_i, and heatmaps of the most discordant loci.

3 RESULTS

Assessment of the measurement quality of RNA expression platforms

Dataset T1 Varying levels of concordance between gene expression levels are found from the glioblastoma data. Graphical representations of row-linear fits for two gene loci can be found in Figures 2a and 2b, taken from dataset T1. Gene NUBP1 (Figure 2a) shows high concordance between samples, with the bundle of regression lines from the fit almost parallel, each with slope ≈ 1 (recalling the model constraint $\bar{b} = 1$), and the points falling

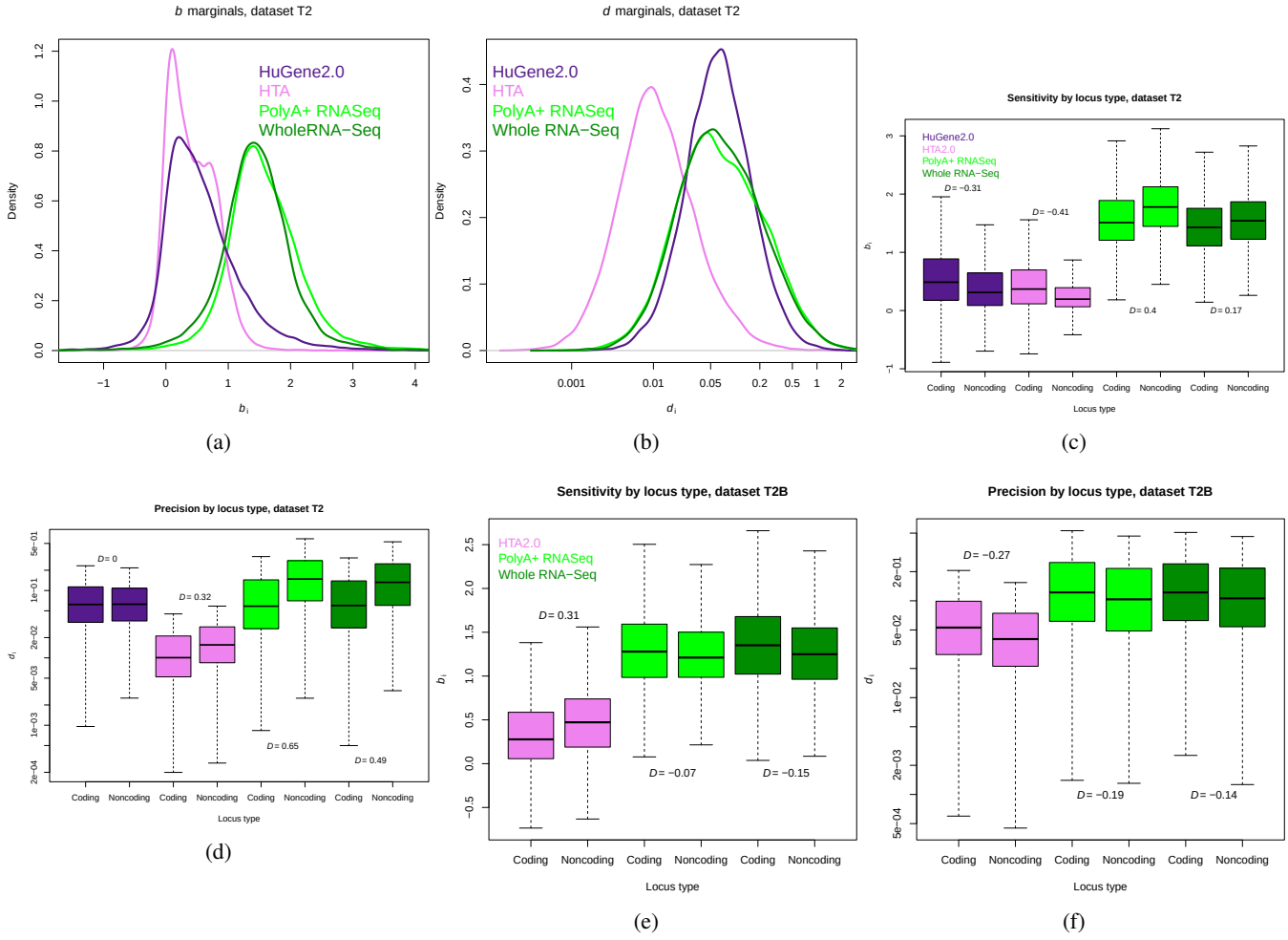


Fig. 3: Marginal distributions of **(a)** b_i and **(b)** d_i for dataset T2. Boxplots separating loci into coding and noncoding targets over all platforms for **(c)** dataset T2 sensitivity, **(d)** dataset T2 precision, **(e)** dataset T2B sensitivity and **(f)** dataset T2B precision.

close to their corresponding lines. By contrast, gene HCN4 (Figure 2b) shows much more discordance between the slopes b_i, with $b_{RNA-Seq}$ in particular explaining a disproportionate share of the change in expression, as well as noticeably larger residuals d_{ij} for three out of the four platforms represented, indicating decreased precision.

Broadening the perspective to all 9,519 genes assayed by these four platforms, we are able to see marginal and joint distributions for a_i, b_i, and d_i (Figures 2c-2h). The platform-wise intercepts from the row-linear model a_i are directly interpretable in the gene expression space, hence the distribution of a_i serves as a dynamic range for the ith platform (Figure 2c). Joint distributions (Figure 2d) show Huex and RNA-Seq are the most correlated, U133A shows little correlation with the other platforms at lower levels of its dynamic range, and the custom Agilent microarray has quite a different dynamic range, which is a consequence of it being a two-channel array. RNA-Seq is the platform with the greatest mean sensitivity to differences in gene expression (Figure 2e), and the insensitivity of U133A to expression change at lower values suggested in Figure 2d shows as a secondary mode $b_{U133A} = 0$ in Figure 2e. For the other three platforms, sensitivity to expression change b_i is centred around 1 (Figure 2f), indicating unimodal concordance, but with a substantial number of discordant loci. Correlations between the different b_is tend to be (but not always) negative, as the natural constraint $\bar{b} = 1$ renders sensitivity between platforms competitive. As the variation $V(b)$ increases—that is, as the cross-platform sensitivity becomes more discordant—the RNA-Seq begins to explain a disproportionate share of the cumulative b_i (Supplementary Figure (SF) 1). Cross-platform precision d shows similar marginal means for RNA-Seq, Agilent and Huex, but lower for that of U133A, indicating higher precision for this platform (Figure 2g). However, given that U133A also shows the lowest sensitivity (Figure 2e), from a platform design perspective there may have been a trade-off between risk and reward in detecting changes in gene expression that has now been surmounted by more recent

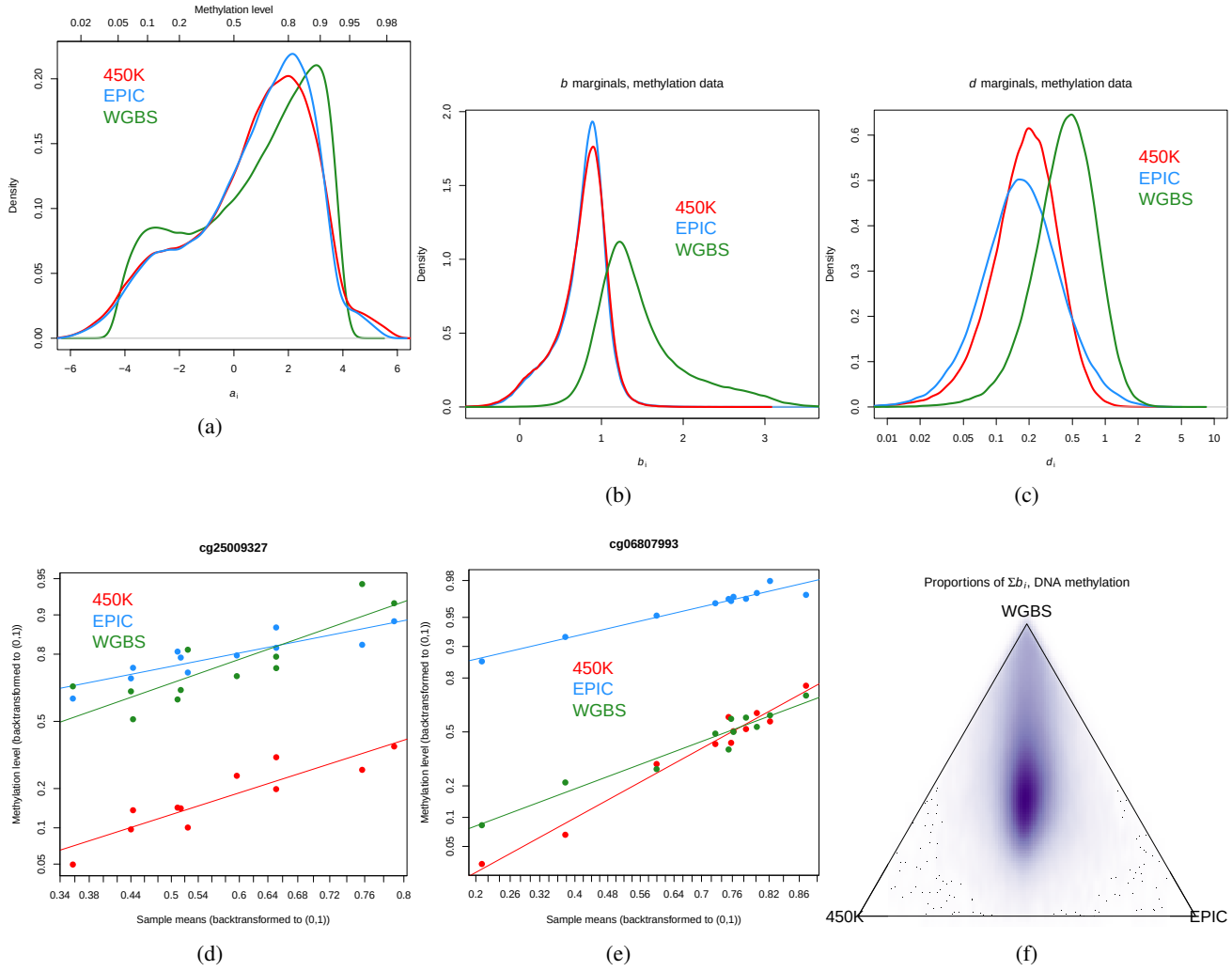


Fig. 4: Marginal distributions for **(a)** parameter a_i, **(b)** parameter b_i and **(c)** d_i, for the entirety of Dataset M1. Graphical depictions of row-linear fits for array-discordant CpG sites for which WGBS data favours **(d)** the EPIC array and **(e)** the 450K array. **(f)** DeFinetti diagram showing the proportions of b_i described by the three platforms in Dataset M1. We show backtransformed axes to the more interpretable methylation domain (0, 1) in **(a)**, **(d)** and **(e)**.

technologies. Given its greater overall sensitivity and competitive precision, RNA-Seq is the superior platform from dataset T1.

In addition to comparing distributions of platform sensitivity and precision to each other, we also examined whether these platforms showed any relative gene-specific biases by plotting CDS length (SFs 2a and 2b) and GC content (SFs 2c and 2d) against b_i and d_i. All effects of CDS length are mild to non-existent, including from the remaining transcription datasets (data not shown). There is a slight suggestion of relatively decreased precision of Huex and RNA-Seq at short CDSs (SF 2b). GC content has a greater effect, with mild to moderate relative decrease in sensitivity on the microarrays, offset by an increase for RNA-Seq (SF 2c). In addition, the three array platforms, especially Huex, struggle with maintaining relative precision assaying genes with low GC-content, with RNA-Seq proving to be the most robust (SF 2d).

Datasets T2, T2A and T2B Similar to dataset T1, RNA-Seq has greater sensitivity than microarrays, this time with both strategies (PolyA+ and WholeRNA) clustering together, and the two microarray platforms also doing so at a lower b_i (Figure 3a). While the HuGene array shows similar precision to the distributions from RNA-Seq, the Human Transcriptome Array shows a clearly superior precision (Figure 3b). Remaining marginal and joint distributions for a_i, b_i, and d_i for dataset T2 can also be found in SF 3, and those for datasets T2A and T2B can be found in SFs 4 and 5 respectively. When the HuGene array and HTA are broken down into their exonic and sub-exonic features in datasets T2A and T2B,

the superior precision of HTA to RNA-Seq is maintained (SF 5e), but the HuGene array shows both inferior sensitivity and precision to RNA-Seq (SFs 4c and 4e).

Turning to gene-specific biases from dataset T2, we see the same loss of sensitivity with increasing GC content on the microarray platforms, with the compensatory increase from RNA-Seq (SF 6a), as we saw in dataset T1. We also see the same increased robustness to this domain of the precision estimate for RNA-Seq, whereas precision is slightly lower for low GC content loci in microarrays (SF 6b).

For both genewise (dataset T2) and exon/sub-exon (dataset T2B) target levels, HTA annotation lists targets that are either coding or noncoding, and we observe effects (calculated using Cohen's D: the difference of means divided by the pooled standard deviation) between these locus types. Here (and for the rest of the paper) we use Cohen's D to report effect size, capitalising to avoid confusion with precision d_i. The R package "effsize" (Torchiano (2017)) was used for all calculations. At the gene level, microarray sensitivity is greater for coding loci than for noncoding, and the opposite is true for RNA-Seq (Figure 3c). Precision is superior at coding loci on all platforms, except for HuGene, which evinces no effect (Figure 3d). However, for dataset T2B, these effects are reversed on sensitivity and precision across all three platforms, albeit to a smaller magnitude in all cases (Figures 3e and 3f). For example, noncoding loci show greater sensitivity than coding loci when assayed on HTA for dataset T2B (Figure 3e).

Assessment of the measurement quality of DNA methylation platforms

Dataset M1 As with the previous two datasets, we were able to apply a row-linear model to a suite of DNA methylation platforms containing both microarrays and sequencing assays. Perhaps unsurprisingly, the distributions of a_i, b_i, and d_i (Figure 4) for 450K and EPIC are extremely similar, owing to identical probe chemistry. (Joint distributions of these can be found in SF 7. The shrunk hypomethylated mode in Figure 4a is due to the filtering of CpG sites as outlined in the Sample Sources and Preprocessing section of the Supplementary Material, where CpG loci were discarded if they had at least 8 out of 11 samples completely unmethylated in WGBS.). However, there are a very small number of probes for which the signal is highly discrepant between the two platforms, and in the majority of these cases WGBS can provide clues as to which platforms measurements are more accurate. Two extreme examples, one for each array platform, are shown in Figures 4d and 4e. These were discovered by filtering for large values of $V(a)$ for this dataset. However, WGBS is unable to do this in a general way across all CpG sites with respect to sensitivity, showing little favour to either platform in the DeFinetti plot in Figure 4f. WGBS itself is superior to the arrays in sensitivity (Figure 4b and 4f), but has lower precision (Figure 4c). This is expected, however, given that WGBS is clearly the odd platform from a consensus highly influenced by very similar arrays.

To explore the effects of array normalisation, we also fit a separate set of row-linear models to the raw data, calculating methylation "beta" ratios from the raw array signal and transforming them to M-values as outlined for WGBS in the Sample Sources and Preprocessing section in the Supplementary Material. Illumina arrays have two separate probe types with different biochemistries, with Type I probes assuming a uniform methylated epitype across the length of the 50 base hybridisation locus, and Type II probes using a single-base extension at the 3' end to circumvent this assumption (Bibikova *et al.* (2011)). A clear decrease in precision (increase in d_i) is seen in the non-normalised Type I probes, compared to Type II on both 450K and EPIC (Figure 5a). Similarly, a considerable increase in precision is apparent with an increase of the total raw intensity (M+U) from Type II probes (Figure 5b). The normalisation procedure `preprocessFunnorm()` (Fortin *et al.* (2014)) does an excellent job at correcting these discrepancies in d_i, as shown in Figures 5c and 5d.

Elements particular to the human genome may interfere with the methylation measurement of CpG loci. For example, a small decrease in sensitivity is noted for when probe hybridisation overlaps a repeat region (Figure 6a), as per Naeem *et al.* (2014) . Hence it is likely that the methylation state of the repeated CpG locus is non-uniform. Additionally, cross-hybridisation is a known confounder of microarray technology, where probes designed to hybridise to an intended genomic site instead hybridise to an off-target site with high homology (Casneuf *et al.* (2007); Chen *et al.* (2013)). It is known that Illumina array signals are prone to potential cross-hybridising events, based on in silico alignment of probe sequences to the human reference (Pidsley *et al.* (2016); Chen *et al.* (2013)). Unlike the expression arrays in dataset T2, the hybridisation is intended to have one target only (as opposed to a composite set of targets over a gene locus), and so we are able to elucidate the off-target effects more precisely. A subset of these probes for which homology is 47/50 base pairs or higher is listed in Pidsley and Zotenko *et al.* (2016). We find that the sensitivity these of probes is blunted in comparison to probes for which no high-level homology was found (Figure 6b), and this effect is more severe on the EPIC array ($D = 0.73$) than it is on the 450K array ($D = 0.43$). To further characterise the degree of off-target hybridisation, we then posited that the signal from these genomically promiscuous probes could be expressed as a linear combination of the methylation levels from the off-target sites, as well as the target site. This would simulate the competitive homology between all potential binding sites for the given probe. The WGBS data is useful for this since its measurements do not have a cross-reactive bias, and that we have methylation measurements for the off-target CpG sites that are not assayed by the array technology. For each individual CpG site flagged as having a potentially cross-reactive probe on both 450K and EPIC arrays, we fit two individual LASSO (Tibshirani (1996)) sparse regression models with response either $y_i = Z_{450K}$ or $y_i = Z_{EPIC}$ from the measurement matrix Z_{ij}, and set the predictor matrix $x^i = (Z_{WGBS}, A_1, \ldots, A_m)^T$ where A is a mxn matrix of WGBS methylation measurements from the corresponding CpG loci of the off-target hybridisation sites. The corresponding estimated regression coefficients $\widehat{\beta} = (\widehat{\beta}_{target}, \widehat{\beta}_{off-target_1}, \ldots, \widehat{\beta}_{off-target_m})$ were calculated. Figure 6c plots array sensitivities b_{450K} and b_{EPIC} against $\widehat{\beta}_{target}$ for all known cross-reactive probes/CpG loci. Non-zero target coefficients correlate well with the sensitivity of the array, indicating that b_{450K} and b_{EPIC} serve as reasonable predictors of the degree of target homology. However, many coefficients are at or very close to zero. For these probes, this means that the entirety of their signals can be explained by a linear combination of the methylation values

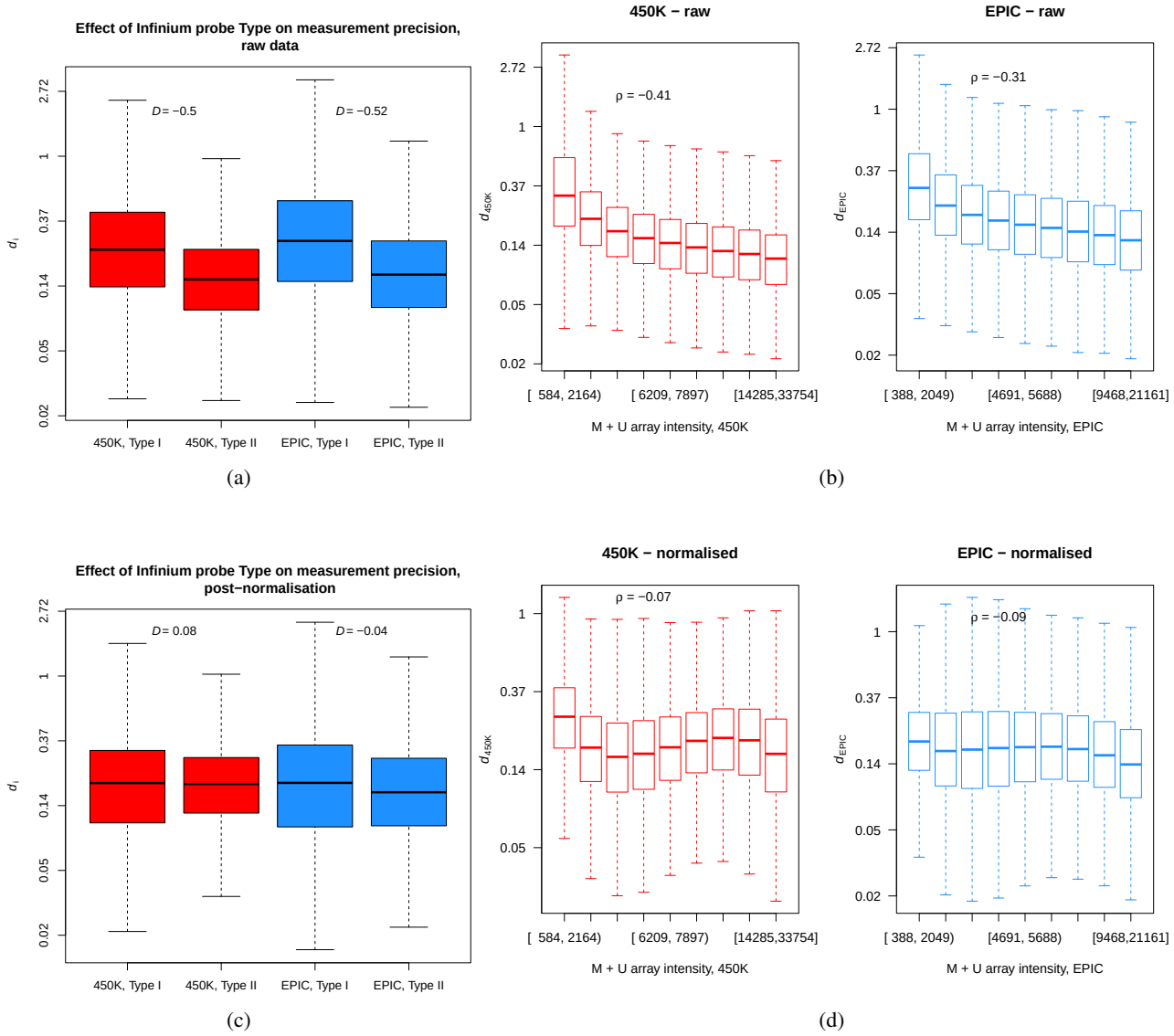


Fig. 5: Effect of array normalisation on Dataset M1. Precision of raw 450K and EPIC data **(a)** split by Type I and Type II probes and **(b)** total intensity (methylated + unmethylated channel) of Type II probes, and **(c and d)** the same values post-normalisation.

from their cross-hybridising genomic sites, and thus are unlikely to be measuring the target site at all. To check this was not an artefact of the LASSO fits themselves, we then randomly selected an identical number (n=11,646) of *non*-cross-hybridising probes from dataset M1 and also fit a LASSO model for each one, substituting their matching WGBS values for the original Z_{WGBS} but retaining the *same* list of A matrices used for the cross-hybridising LASSO fits. The resulting $\widehat{\beta}_{target}$s are higher than those from the cross-hybridising probes (Figure 6d), providing strong evidence for competitive homology having a detrimental effect on Infinium probe sensitivity. A complete list of the LASSO coefficients for the target site, and sum of coefficients from the off-target sites for cross-hybridising probes can be found in Supplementary Table 1, for both 450K and EPIC arrays.

Returning to datasets T2 and T2A, the HuGene annotation also describes targets that potentially cross-hybridise to off-target sites. However, there are negligible differences in sensitivity between these targets and the remainder (SFs 8a and 8b). One explanation for this is that the measurements on the HuGene arrays are composite, containing a mixture of probes that potentially cross-hybridise, and those that don't. It may be that the cross-hybridising component

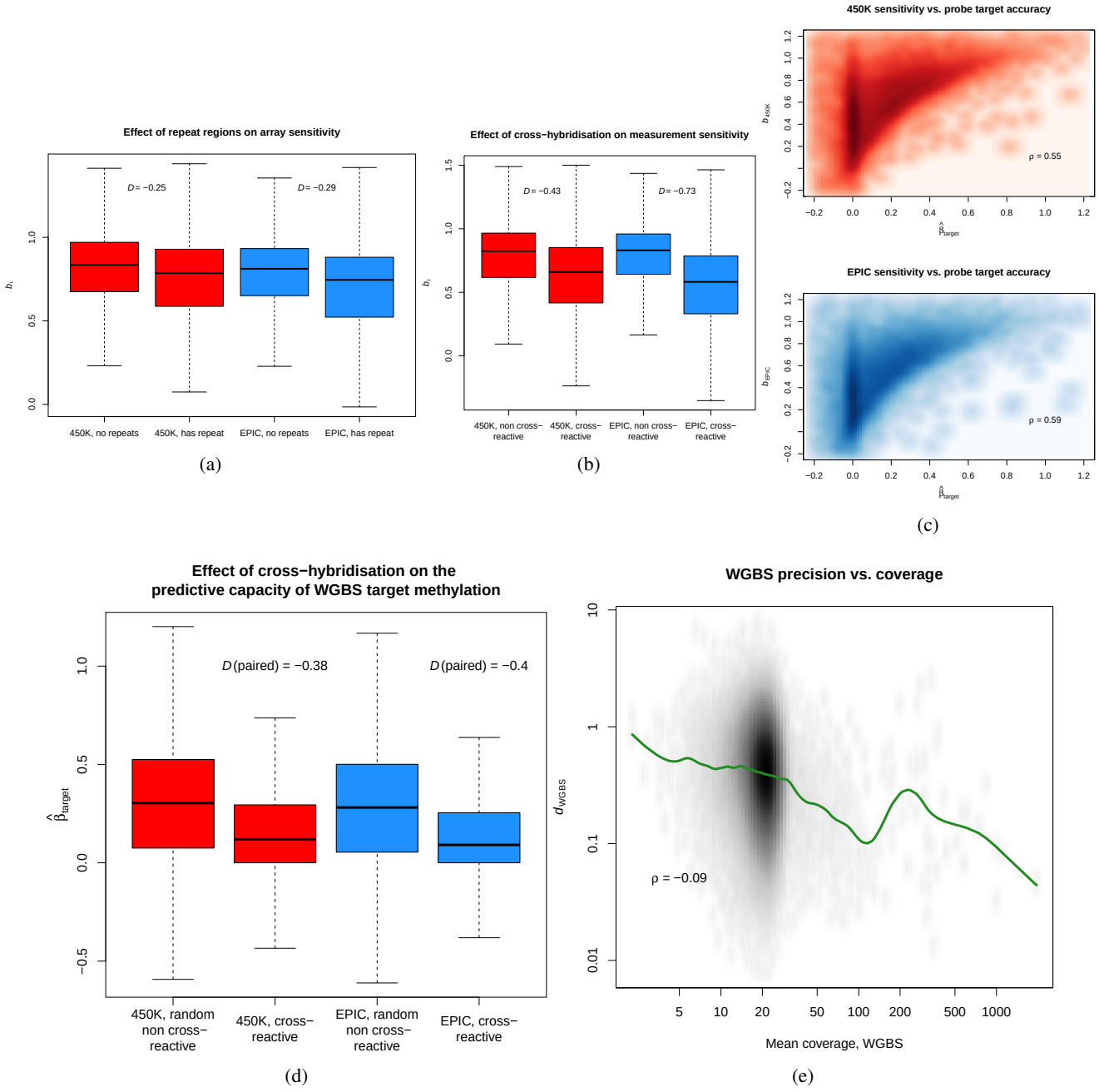


Fig. 6: Effect of **(a)** repeat regions and **(b)** cross-hybridisation on array sensitivity from dataset M1. **(c)** Sensitivity of cross-hybridising probes against the LASSO coefficient $\widehat{\beta}_{target}$ of target WGBS values from sparse linear modelling. **(d)** Effect of cross-hybridisation on the predictive capacity of WGBS measurements for their matched microarray measurements, via LASSO. **(e)** Precision of WGBS against mean coverage of the samples, for individual CpG loci.

of the signal is too weak to evince any systematic bias for this particular dataset.

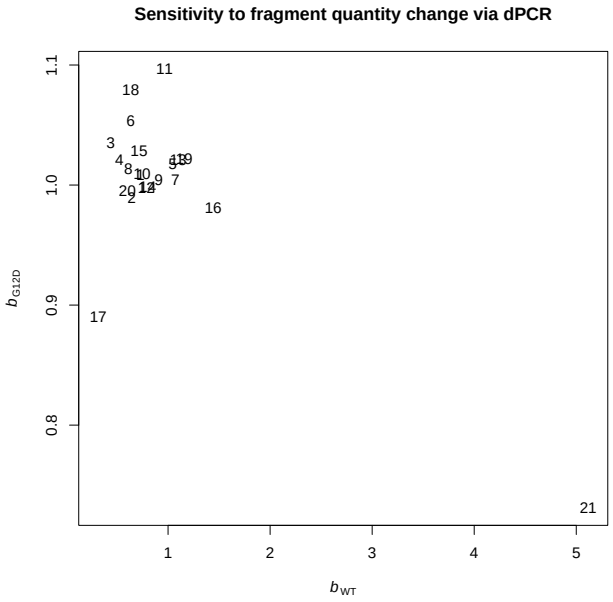
Lastly, the question of minimum WGBS sequencing coverage for detection of methylation differences is a practical consideration for many labs, and estimates at characterising this figure for a given methylation shift size have been made (Ziller *et al.* (2014)). Under binomial assumptions of simulated data, higher coverage results in a lower variability of the estimate when the methylation level

is known. In other words, the measurement is more precise. This is reflected in mildly decreasing d_i as the mean WGBS coverage increases (Figure 6e). Notably, the loess curve carries on decreasing at higher levels of coverage, even up to 1000x, evincing no "point of saturation", even though the data points are sparse.

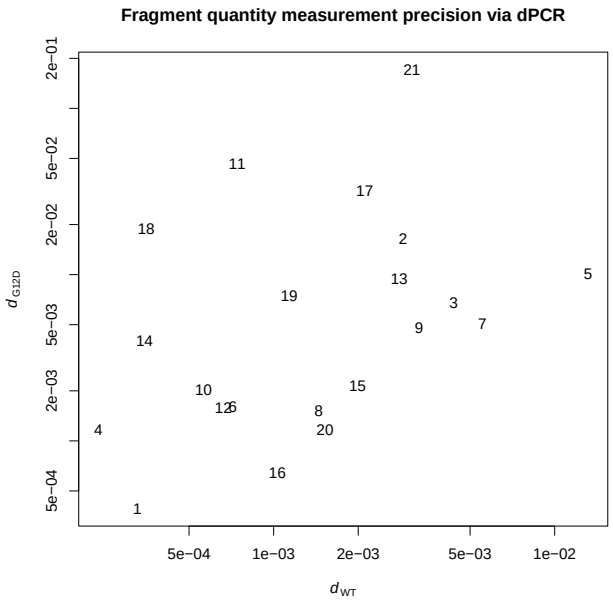
An interlaboratory test on DNA abundance measurements

Dataset IL1 Finally, to demonstrate the utility originally intended for the row-linear model, we perform a true interlaboratory test that assesses measurement quality not across technologies, but across geographical testing sites. The measurements used are log-transformed fragment concentrations (copies/μL in reaction) of 2 genotypic forms of human *KRAS* assayed using digital PCR (dPCR) on 6 biological samples from Whale *et al.* (2017) .

To test whether a site effect was indeed present, independent of biological variation, we constructed row-linear fits using the set of $i \in \{1, 2, ..., 21\}$ as numbered candidate laboratories to be compared. Deviant testing sites include laboratories 17 and 21 with respect to sensitivity (Figure 7a), and laboratories 5 and 21 with respect to lower precision (Figure 7b), with laboratories 1 and 4 showing the greatest measurement precision when compared to the consensus. These results are somewhat consonant with the reproducibility analysis in the original manuscript, in that laboratories 17 and 21 (as well as laboratory 2) are singled out for likely droplet misclassification. In addition, laboratory 21 has the greatest number of deviant analysis parameters of all testing sites (i.e. laboratory conditions, Table S3 from Whale *et al.* (2017)) including unique models of the PCR plate sealer and thermal cycler, potentially explaining its position as the most outlying laboratory.



(a)



(b)

Fig. 7: Scatterplots depicting **(a)** b_i and **(b)** d_i of twenty-one laboratories for both *KRAS* genotype abundances, from Dataset IL1. Number plotted denotes laboratory ID.

4 DISCUSSION

We have shown the row-linear model has the ability to empirically assess the sensitivity and precision of genomic platforms, given a sufficient corpus of variable samples. This model is highly versatile in that it not only allows direct comparisons of platforms with each other, but can be used to assess locus-specific biases and characteristics particular to both the platforms themselves and the genome they are measuring. Importantly, it assesses the measurement quality of each platform *independent of the biological variation in the data*. In addition, it can be used as a screening procedure to identify platforms with deviant measurements, and subsequently remove them prior to using applications that leveraging cross-platform information for biological purposes (Wang *et al.* (2014); Uziela and Honkela (2015); Thompson *et al.* (2016)).

Most effects we have shown are mild, or limited to a small subset of loci, so the results shown herein ought to be read with caution. Interpretation of trends as tendencies, rather than categorical biases, is advisable. The consensus improves as more technologies are included in the row-linear fit, since it is less susceptible to skewing due to measurement from a deviant technology. Positive correlations between the respective d_is (for example, in Figure 2h) are indicative that the ease with which quantification can be made varies from locus to locus—likely a result of stochastic "burstiness" (Raj *et al.* (2006)—and this variation is somewhat preserved across platforms. Platforms such as WGBS can be seen as higher-risk, higher-reward technologies, given their extreme ranks in both sensitivity and precision. Some trends are quite clear, such as an anticorrelation of

relative microarray sensitivity with GC content, which is likely due to differing amplification efficiencies (Arezi *et al.* (2003); Degrelle *et al.* (2008)). Plainly, RNA-Seq outperforms all microarrays in terms of sensitivity to changes in transcription and robustness to GC content. However, its measurement quality is less clear when it comes to precision. An anticorrelative effect can be seen between a_i and d_i on all RNA-Seq platforms (data not shown), though this is likely an artefact of the variance stabilisation via log transformation of counts, which may be suboptimal. Most of the library sizes used in this study (see Supplementary Tables 2 & 3) are below those recommended for analysis of differential expression (Liu *et al.* (2013)), speaking to the lack of stability of quantification estimates when overall count pileup is low. The HTA's superior precision at both the gene and exon level supports previous work, where stochastic variability of the signal was found to be higher in RNA-Seq than from this array (Nazarov *et al.* (2017)). No appreciable differences were observed between the two RNA-Seq strategies in datasets T2, T2A and T2B in terms of direct comparison. When comparing coding and noncoding loci, Whole RNA-Seq shows a smaller divergence (Figures 3c and 3d), which relates to earlier work showing this strategy shows less variability than PolyA+ for quantifying some noncoding RNAs (Holik *et al.* (2017)). The reasons for the reversal of effect between coding and noncoding RNAs from dataset T2 to T2B remain elusive, however this may be a consequence of the design of the HTA, giving more emphasis to quantifying canonical coding loci at the gene level, but removing this emphasis at the more granular exon level.

In terms of methylation, we have shown that a subset of potentially cross-hybridising probes on Illumina arrays show a lower sensitivity to change, as do probes known to hybridise to repeat regions. Both phenomena result in heterogeneous sources of signal intensity which, if hybridisation patterns are consistent, would not hinder the precision of the signal, but rather its sensitivity to methylation change compared to other loci, which is precisely what we see. Further evidence for this is gained from modelling the off-target methylation measurements from WGBS as predictors of the microarray measurements. The increased susceptibility of EPIC to blunted sensitivity via cross-hybridisation is without categorical explanation, but may be related to the fact that the overall intensities (M+U) of the EPIC arrays used in this study are, on average, only $\sim 70\%$ as strong as that of that of 450K, which may indicate fewer total hybridisation events for this technology. We have also shown that the row-linear model can be powerful tool in assessing current and future microarray normalisation procedures, as evidenced from the methylation results.

From the results taken from interlaboratory test on Dataset IL1, we can conclude that known deviations from consensus protocols can result in discordance of measurement between testing sites. A more standardised testing procedure, where intentional differences in protocol are implemented, replicated and blocked properly across all sites prior to row-linear modelling of their measurements would further elucidate the effect of such deviations.

To improve the characterisation of the biases outlined in this paper, a broader collection of platforms and samples will be needed. For example, the quality assessment of both transcription and methylation sequencing assays would be clarified by sequencing these (and other) samples across a range of library sizes and coverage depths. Data from the SEQC/MAQC-III consortium (SEQC/MAQC-III Consortium (2014)) would have been ideal to use for this paper, but the limited number of samples ($n = 4$) unfortunately precludes us from applying an informative row-linear fit.

As suggested from our discussion of Dataset IL1, applications of the row-linear model need not be restricted to different technologies, either. Potential biases from reagent concentrations, ambient laboratory temperature and humidity, and temporal and geographical variation are all able to be characterised by the row-linear model in the same fashion exemplified here, providing matched, homogenous aliquots of nucleic acid are distributed across all conditions. Given enough biological and technological replicates, a growing, dynamic repository of contributions sourced from throughout the world (in the vein of, for example, *recount2* (Collado-Torres *et al.* (2017))) would allow calculation of increasingly stable consensus estimates for sensitivity and precision. This is a conceivable empirical alternative to a static gold-standard benchmark of measurement. Through such an infrastructure, scientists will have increased confidence in the reproducibility of their results.

ACKNOWLEDGEMENTS

Thank you to Elena Zotenko, Fabian Buske, Phuc-Loi Luu and Cathryn M. Gould of the Garvan Institute of Medical Research and Firoz Anwar of CSIRO North Ryde for additional bioinformatic processing. Thank you also to Peter Molloy, Susan van Dijk, Brodie Sutcliffe, Julius von Martels, Rosanne Arnoldy, Michelle Peranec and Madhavi P. Maddugoda for patient acquisition, sample collection and processing of NDMC samples. Thank you to Caroline Janitz of UWS for performing RNA-Seq assays. We thank the Australian Genome Research Facility (Melbourne) for conducting the EPIC array experiments and Illumina for early and complimentary access to EPIC arrays. The results shown here are in whole or part based upon data generated by the TCGA Research Network: `http://cancergenome.nih.gov/`.

FUNDING

This work was supported by National Health and Medical Research Council (NHMRC) project grants 1088144 and 1106870; NHMRC Program Grant 1054618 to TPS; NHMRC Fellowship (SJC grant no. 1063559); Cancer Australia (grant no. 1044458); the Australian Prostate Cancer Research Centre NSW and the National Breast Cancer Foundation; the Science and Industry Endowment Fund (Australia) (grant no. RP03-064) and a CINSW Early Career Fellowship to RP (grant no. 14/ECF/1-23). The contents of the published material are solely the responsibility of the administering institution and individual authors and do not reflect the views of the NHMRC.

Conflict of interest statement: None declared.

REFERENCES

Aird,D., Ross,M.G., ... and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology,* **12** (2), R18.

Arezi,B., Xing,W., Sorge,J.A. and Hogrefe,H.H. (2003) Amplification efficiency of thermostable DNA polymerases. *Analytical Biochemistry,* **321** (2), 226–35.

Baker,M. (2016) 1,500 scientists lift the lid on reproducibility. *Nature,* **533** (7604), 452–454.

Baker,S.C., Bauer,S.R., ... and Zadro,R. (2005) The External RNA Controls Consortium: a progress report. *Nature Methods,* **2** (10), 731–734.

Begley,C.G. and Ellis,L.M. (2012) Drug development: Raise standards for preclinical cancer research. *Nature,* **483** (7391), 531–533.

Bibikova,M., Barnes,B., ... and Shen,R. (2011) High density DNA methylation array with single CpG site resolution. *Genomics,* **98** (4), 288–295.

Casneuf,T., Van de Peer,Y. and Huber,W. (2007) In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics,* **8** (1), 461.

Chen,Y.a., Lemire,M., ... and Weksberg,R. (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics : official journal of the DNA Methylation Society,* **8** (2), 203–9.

Collado-Torres,L., Nellore,A., ... and Leek,J.T. (2017) Reproducible RNA-seq analysis using recount2. *Nature Biotechnology,* **35** (4), 319–321.

Degrelle,S.A., Hennequet-Antier,C., ... and Hue,I. (2008) Amplification biases: possible differences among deviating gene expressions. *BMC Genomics,* **9**, 46.

Fisher,R.A. (1971) *The design of experiments.* 9th edition,, Hafner, New York.

Fortin,J.P., Labbe,A., ... and Hansen,K.D. (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology,* **15** (12), 503.

Holik,A.Z., Law,C.W., ... and Ritchie,M.E. (2017) RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. *Nucleic Acids Research,* **45** (5), e30–e30.

Irizarry,R.A., Warren,D., ... and Yu,W. (2005) Multiple-laboratory comparison of microarray platforms. *Nature Methods,* **2** (5), 345–350.

Iyer,H.K., Wang,C.M.J. and Mathew,T. (2004) Models and Confidence Intervals for True Values in Interlaboratory Trials. *Journal of the American Statistical Association,* **99**, 1060–1071.

Jiang,L., Schlesinger,F., ... and Oliver,B. (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Research,* **21** (9), 1543–51.

Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England),* **8** (1), 118–27.

Kevil,C.G., Walsh,L., ... and Alexander,J.S. (1997) An improved, rapid Northern protocol. *Biochemical and Biophysical Research Communications,* **238** (2), 277–9.

Leek,J.T. and Storey,J.D. (2007) Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics,* **3** (9), e161.

Li,S., Tighe,S.W., ... and Mason,C.E. (2014) Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nature Biotechnology,* **32** (9), 915–925.

Lister,R., O'Malley,R.C., ... and Ecker,J.R. (2008) Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell,* **133** (3), 523–536.

Lister,R., Pelizzola,M., ... and Ecker,J.R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature,* **462** (7271), 315–322.

Liu,Y., Ferguson,J.F., ... and Li,M. (2013) Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose. *PLoS ONE,* **8** (6), e66883.

Mandel,J. (1984) *The statistical analysis of experimental data*. Courier Corporation, New York.

Mandel,J. (1994) Analyzing Interlaboratory Data According to ASTM Standard E691. In *Quality and Statistics: Total Quality Management*. ASTM International 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959 pp. 59–70.

Mandel,J. and Lashof,T. (1969) The Interlaboratory Evaluation of Testing Methods. In *Precision measurement and calibration: selected NBS papers on statistical concepts and procedures*, (Ku,H.H., ed.),. US Department of Commerce pp. 170–178.

Naeem,H., Wong,N.C., ... and Macintyre,G. (2014) Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics,* **15** (1), 51.

Nazarov,P.V., Muller,A., ... and Vallar,L. (2017) RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples. *BMC Genomics,* **18** (1), 443.

Nosek,B.A. and Errington,T.M. (2017) Making sense of replications. *eLife,* **6**.

Oytam,Y., Sobhanmanesh,F., ... and Ross,J. (2016) Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets. *BMC Bioinformatics,* **17** (1), 332.

Pidsley,R., Lawrence,M.G., ... and Clark,S.J. (2018) Enduring epigenetic landmarks define the cancer microenvironment. *Genome Research,* .

Pidsley,R., Zotenko,E., ... and Clark,S.J. (2016) Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology,* **17** (1), 208.

Popper,K. (2005) *The logic of scientific discovery*. Routledge, London/New York.

Raj,A., Peskin,C.S., ... and Tyagi,S. (2006) Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology,* **4** (10), e309.

SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology,* **32** (9), 903–914.

Shi,L., Reid,L.H., ... and Slikker,W. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology,* **24** (9), 1151–1161.

Thompson,J.A., Tan,J. and Greene,C.S. (2016) Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ,* **4**, e1621.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B Methodological,* **58** (1), 267–288.

Torchiano,M. (2017) *effsize: Efficient Effect Size Computation*.

Uziela,K. and Honkela,A. (2015) Probe Region Expression Estimation for RNA-Seq Data for Improved Microarray Comparability. *PLOS ONE,* **10** (5), e0126545.

Verhaak,R.G.W., Hoadley,K.A., ... and Cancer Genome Atlas Research Network (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell,* **17** (1), 98–110.

Wang,C., Gong,B., ... and Tong,W. (2014) The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature Biotechnology,* **32** (9), 926–932.

Wang,H., He,X., Band,M., Wilson,C. and Liu,L. (2005) A study of inter-lab and inter-platform agreement of DNA microarray data. *BMC Genomics,* **6** (1), 71.

Warnecke,P.M., Stirzaker,C., ... and Clark,S.J. (1997) Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Research,* **25** (21), 4422–4426.

Whale,A.S., Devonshire,A.S., ... and Huggett,J.F. (2017) International Interlaboratory Digital PCR Study Demonstrating High Reproducibility for the Measurement of a Rare Sequence Variant. *Analytical Chemistry,* **89** (3), 1724–1733.

Ziller,M.J., Hansen,K.D., Meissner,A. and Aryee,M.J. (2014) Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature Methods,* **12** (3), 230–232.