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

Sample Sources and Preprocessing

Microarray intensities for dataset T1 were downloaded from [https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/](https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/) and RNA-Seq counts were downloaded from [https://portal.gdc.cancer.gov/](https://portal.gdc.cancer.gov/). U133A data points were natural log-transformed and RNA-Seq counts were normalised via limma voom. For all transcription datasets, any gene where a sample contained a zero count was filtered out, since zero counts artificially inflate the error term $d_i$ (Supplementary Figure 9).

For datasets T2, T2A and T2B, PolyA+ libraries were prepared using Illumina TruSeq Stranded Poly-A RNA Library Preparation Kit. Ribo-Zero libraries were prepared using Epicentre (Illumina) ScriptSeq v2 RNA-Seq Library Preparation Kit (LNCaP and PrEC) and Illumina TruSeq Stranded Total RNA Library Prep Kit (6x NDMCs). All sequencing was carried out on the Illumina HiSeq 2500 platform. HuGene array data for LNCaP and PrEC was pooled from GEO series accession GSE61790, and normal differentiated mesodermal cells (NDMCs, type 1 and 2) from GSE113506. HTA data for LNCaP and PrEC was pooled from GEO series GSE62497 and NDMCs from GSE69598. PolyA+ RNA-Seq for LNCaP and PrEC was pooled from GSE73784, and NDMCs were taken from GSE110820. Whole (Ribo-Zero) RNA-Seq for LNCaP and PrEC can be found at GSE113371 and NDMCs at GSE110915. Core and probeset loci across the HuGene array and HTA were matched by hg19 coordinate intervals by their respective annotations, and normalised separately using RMA to produce expression values. The RNA-Seq samples were aligned to hg19 using TopHat and read counts over the loci obtained from the array annotation were extracted from the alignment using bedtools multicov. Like the TCGA data, counts were then normalised (PolyA+ and WholeRNA separately) via limma voom. The library sizes (that is, the total counts in each sample for a given dataset) for all RNA-Seq samples can be found in Supplementary Table 2 (dataset T1) and Supplementary Table 3 (datasets T2, T2A and T2B).

For dataset M1, LNCaP and PrEC samples on all platforms, and CAF and NAF arrays were prepared as described in Pidsley and Zotenko et al. (2016) and can be found under GEO superseries GSE86833, namely individual series GSM2309154-63 (450K); 70-79 (EPIC) and 85-93 (Cell line WGBS). WGBS for CAF and NAF samples were taken from GSE86260, namely individual series GSM2279670-71, GSM2279673-75, and GSM2279677. NDMC type 2 data can be found at GSE110724 for HM450K (namely individual series GSM3015196.
For WGBS, adapter sequences and poor quality bases were removed using cutadapt version 1.9.1 in paired-end mode with default parameters. BWA-METH\(^7\) was then used to align reads to hg19 using default parameters. PCR duplicates were removed using GATK picard MarkDuplicates v1.91. Count tables of the number of methylated and unmethylated bases sequenced at each CpG site in the genome were constructed using the “tabulate” module of bwa-meth and BisSNP-0.82.2\(^8\) where reads mapped with a minimum mapping quality of 60. Average bisulfite-conversion of non-CpG sites was 0.68% per sample, with a maximum of 1.72%. Mean coverage of analysed CpG sites across the genome was 20.6x per sample, with a minimum of 5.91x.

450K and EPIC arrays were normalised separately using the preprocessFunnorm() function\(^9\) from the Bioconductor minfi package.\(^10\) To approximate normality and remove heteroscedasticity,\(^11\) methylation measurements were offset and logit-transformed to \(M\)-values via minfi for microarrays, and via the following formula for WGBS:

\[
M = \logit\left(\frac{C + 0.5}{C + T + 1}\right)
\]  

where \(C\) is the count of methylated reads, and \(T\) is the count of unmethylated reads for each CpG locus. Finally, to represent true variation in methylation signal from WGBS, CpG loci for which there were eight or more samples completely methylated or completely unmethylated were removed from the dataset since, similar to the zero-counts from RNA-Seq, these incurred a positive bias towards higher sensitivity and low precision (Supplementary Figures 10a and 10b), which is an artifact of the varying coverage at that locus, relative to the even “coverage” of the arrays. Loci were chosen on the availability of probes on both arrays, as well as each of the 11 samples having at least 1x coverage.

For dataset IL1, all data and metadata were downloaded from https://pubs.acs.org/doi/suppl/10.1021/acs.analchem.6b03980/suppl_file/ac6b03980_si_001.pdf.

**Supplementary Tables**

Supplementary Tables 1-3 can be found in Supplementary Table_[1-3].csv.

Table 1: Coefficients \(\hat{\beta}\) from LASSO fits from WGBS data for 450K and EPIC. \(\hat{\beta}_{\text{target}}\) is plotted on the x-axis in Figure 6c. Individual coefficients are given for target CpG sites, and coefficient sums for all off-target CpG sites.

Table 2: RNA-Seq library sizes for all samples from dataset T1.

Table 3: RNA-Seq library sizes for all samples from datasets T2, T2A and T2B.
Supplementary Figures

Supplementary Figure 1: Cumulative sensitivity for Dataset T1, where genes are ranked by $V(b) = (p - 1)^{-1} \sum_i (b_i - 1)^2$. Platform colours are matched as per Figure 2.

**Relative sensitivity to expression change, dataset T1**
Supplementary Figure 2: Boxplots of (a) sensitivity against CDS length, (b) precision against CDS length, (c) sensitivity against GC content, (d) precision against GC content for all platforms in Dataset T1. X-axes are binned at equal size for CDS length, but at equal intervals (0.05) for GC content.
Supplementary Figure 3: (a) Marginal and (b) joint distributions for parameter $a_i$, and joint distributions for parameters (c) $b_i$ and (d) $d_i$, for the entirety of Dataset T2.
Supplementary Figure 4: (a) Marginal and (b) joint distributions for parameter $a_i$, (c) marginal and (d) joint distributions for parameter $b_i$, (e) marginal and (f) joint distributions for parameter $d_i$, for the entirety of Dataset T2A.
Supplementary Figure 5: (a) Marginal and (b) joint distributions for parameter $a_i$, (c) marginal and (d) joint distributions for parameter $b_i$, (e) marginal and (f) joint distributions for parameter $d_i$, for the entirety of Dataset T2B.
Supplementary Figure 6: Boxplots of (a) sensitivity and (b) precision against GC content for all platforms in Dataset T2.
Supplementary Figure 7: Joint distributions for parameter estimates (a) $a_i$, (b) $b_i$ and (c) $d_i$, for the entirety of Dataset M1.
Supplementary Figure 8: Effect of cross-hybridising probes on HuGene2.0 sensitivity from (a) dataset T2 and (b) dataset T2A.
Supplementary Figure 9: Precision of RNA-Seq loci against the number of samples returning zero counts from (a) Dataset T1, (b) Dataset T2, PolyA+, (c) Dataset T2, wholeRNA, (d) Dataset T2A, PolyA+, (e) Dataset T2A, wholeRNA, (f) Dataset TB, PolyA+ and (g) Dataset T2B, wholeRNA. Only loci where all samples contained non-zero counts were retained in the analysis.
Effect of zero counts on PolyA+ RNA-Seq precision, dataset T2A

Effect of zero counts on Whole RNA-Seq precision, dataset T2A

Effect of zero counts on PolyA+ RNA-Seq precision, dataset T2B

Effect of zero counts on Whole RNA-Seq precision, dataset T2B
Supplementary Figure 10: Composition of WGBS data represented as number of fully (a) unmethylated (“zero-composition”) and (b) methylated (“one-composition”) samples per CpG site, and this effect on sensitivity and precision. CpG sites where 8 or more samples were completely methylated or unmethylated were not included in the analysis.
References


9. Jean-Philippe Fortin, Aurélie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig, Celia Mt Greenwood, and Kasper D
