Reference Free Cell Correction Comparison in Cord Blood

Meg Jones and Rachel Edgar

22/11/2017

Table 1: Cohort Description

	UBC data (n=24)	GenR (n=196)
Female	54.17%	
Mean Gestational Age	37.8	
FACS Cell Composition	**Mean % (sd)**	
Monocytes	6.68% (2.39)	
Granulocytes	49.35% (10.66)	
nRBCs	3.36% (2.73)	
NK Cells	4.15% (2.76)	
B Cells	6.51% (2.19)	
CD4 T Cells	25.92% (8.81)	
CD8 T Cells	4.03% (1.5)	

What point are we trying to make?

In cord blood we now have a reference based deconvolution method (**yay!**)
But in most other tissues we do not.
We can use the cord blood gold-standard correction to benchmark reference free methods to recommend a solution for tissues with no reference.

How will we benchmark?

- Which method gives components which are most similar to FACS cell counts?
- Which method produces corrected betas most like the gold-standard corrected betas?
- Which method's corrected betas are the most cell composition normalized (least variability in cell composition after correction)
- EWAS sensitivity specificity to gold-standard EWAS hits, and enrichment for cell type CpGs in EWAS hits.

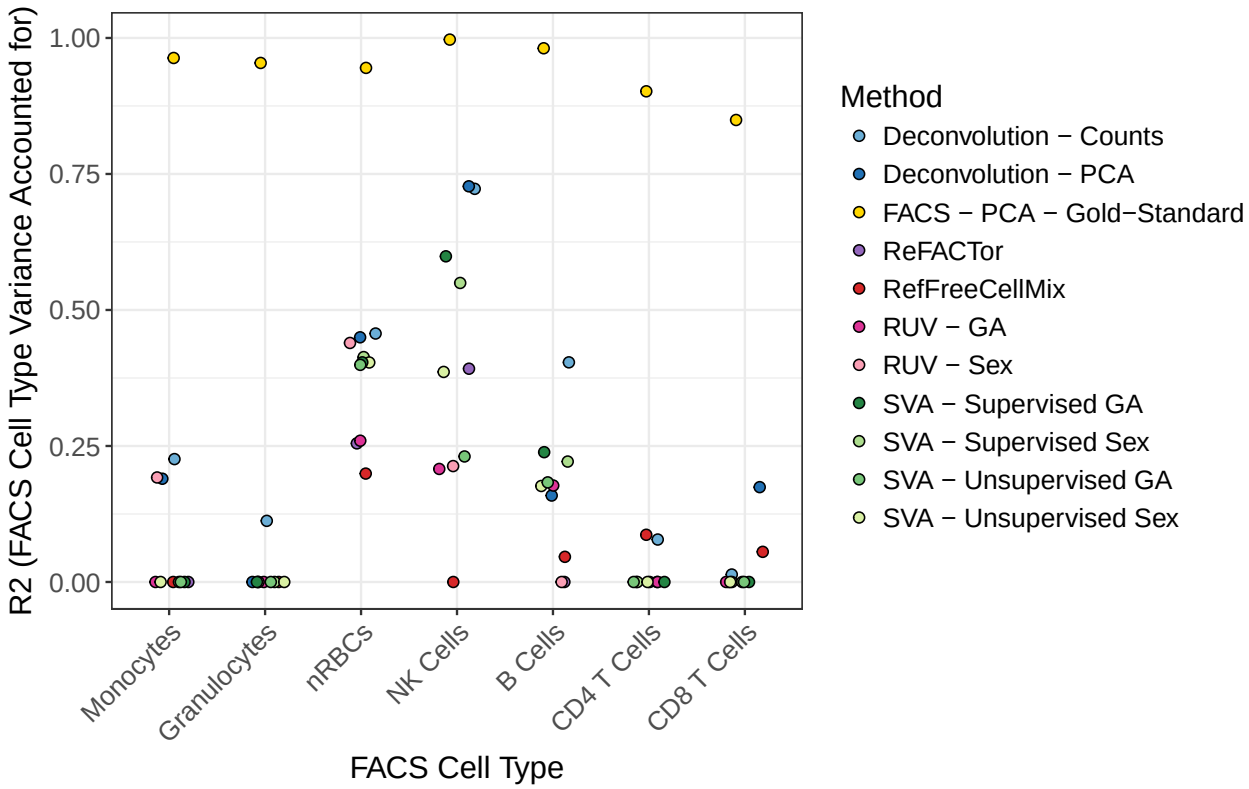

Figure 1: Variance explained in each FACS cell count by models using components from several cell type estimation methods. Models were fit used a nested-models likelihood ratio test. Components from each method were included if they significantly ($p<0.05$) improved model fit.

Table 2: Method name used in comparison, explanation of method, whether it is reference free, reference based or semi reference free, whether the method models with a given phenotype for EWAS, and the relevant citation for the method.

Method	Explanation	Type	Phenotype Required	Citation
FACS – PCA	Top 5 PCs from PCA on FACS counts	Reference	N	Paper in prep
FACS - Drop One Cell Type	FACS cell types as covariates in correction model with one dropped	Reference	N	Paper in prep
Deconvolution - PCA	Top 5 PCs from PCA on deconvolution predicted cell counts	Reference	N	Paper in prep
Deconvolution - Drop One Cell Type	Decouvolution predicted cell counts as covariates in correction model with one dropped	Reference	N	Paper in prep
ReFACTor	Components as model covariates	Reference free	N	PMID:27018579
RefFreeCellMix	Components as model covariates	Reference free	N	PMID: 27358049
SVA - Unsupervised GA	SVA on all CpGs with gestational age (GA) in model	Reference free	Y	PMID: 17907809
SVA - Unsupervised Sex	SVA on all CpGs with sex in model	Reference free	Y	PMID: 17907809
SVA - Supervised GA	SVA on 700 cord blood cell type differentially methylated CpGs with gestational age (GA) in model	Semi Reference Free	Y	PMID: 17907809
SVA - Supervised Sex	SVA on 700 cord blood cell type differentially methylated CpGs with sex in model	Semi Reference Free	Y	PMID: 17907809
RUV - GA	RUV with 700 cord blood cell type differentially methylated CpGs as control probes and with gestational age (GA) in model	Semi Reference Free	Y	PMID: 25990733
RUV - Sex	RUV with 700 cord blood cell type differentially methylated CpGs as control probes and with sex in model	Semi Reference Free	Y	PMID: 25990733
Uncorrected	Not corrected for cell composition	No correction	N	NA

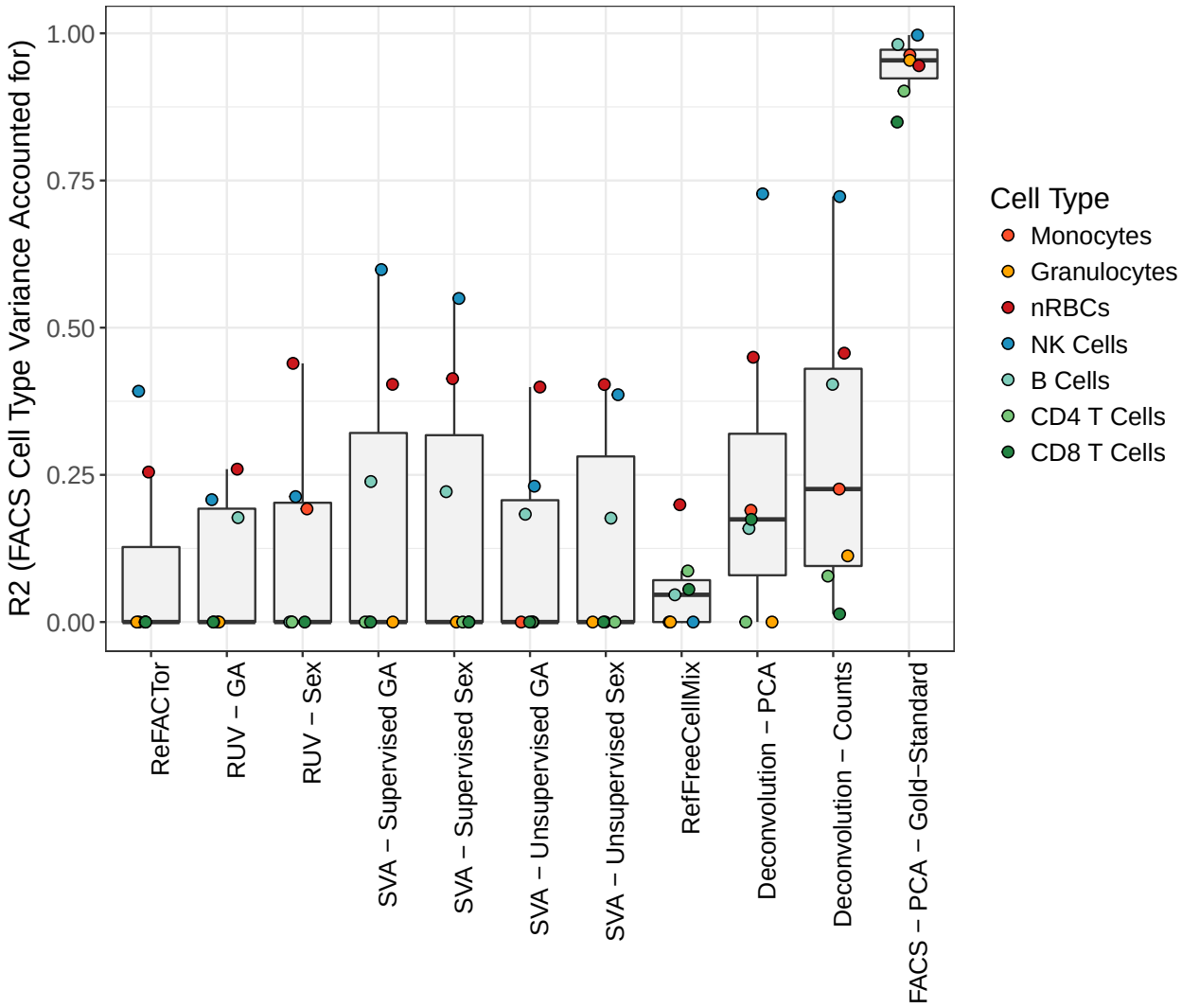

Figure 2: Variance explained in each FACS cell count by models using components from several cell type estimation methods. Models were fit used a nested-models likelihood ratio test. Models include components from each method if the component had significant likelihood ratio test p value ($p<0.05$). Expect for deconvolution - counts which gives measure of each cell type and R2 values are only for models with the deconvolution predicted cell type. Points represent each of 7 cell types, split by the method.

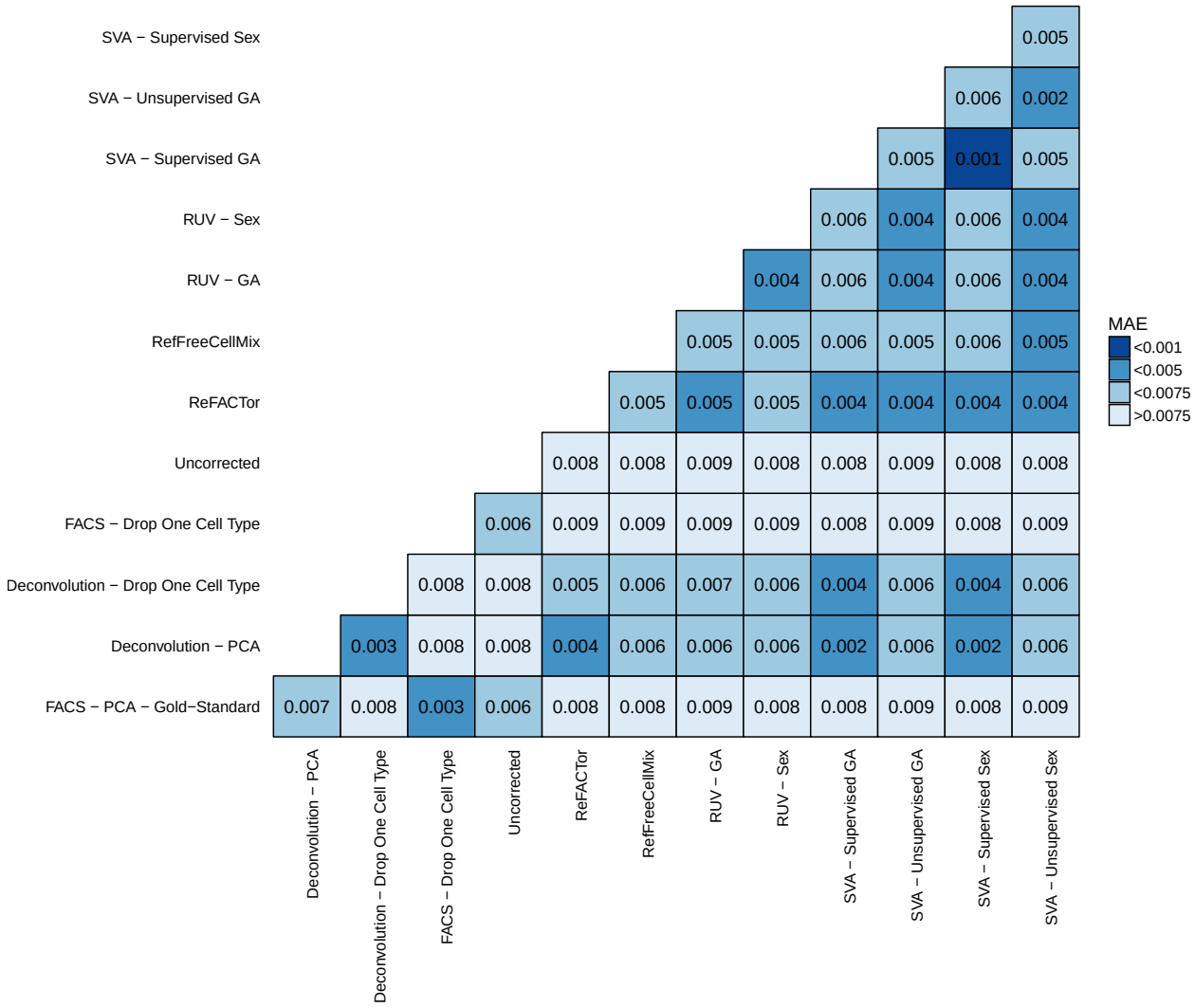

Figure 3: Mean absolute error(MAE) of CpGs compared to the gold-standard FACS-PCA corrected beta values. Errors shown are an average of all CpGs MAE. Boxes are coloured by the discretized MAE value to highlight method performance.

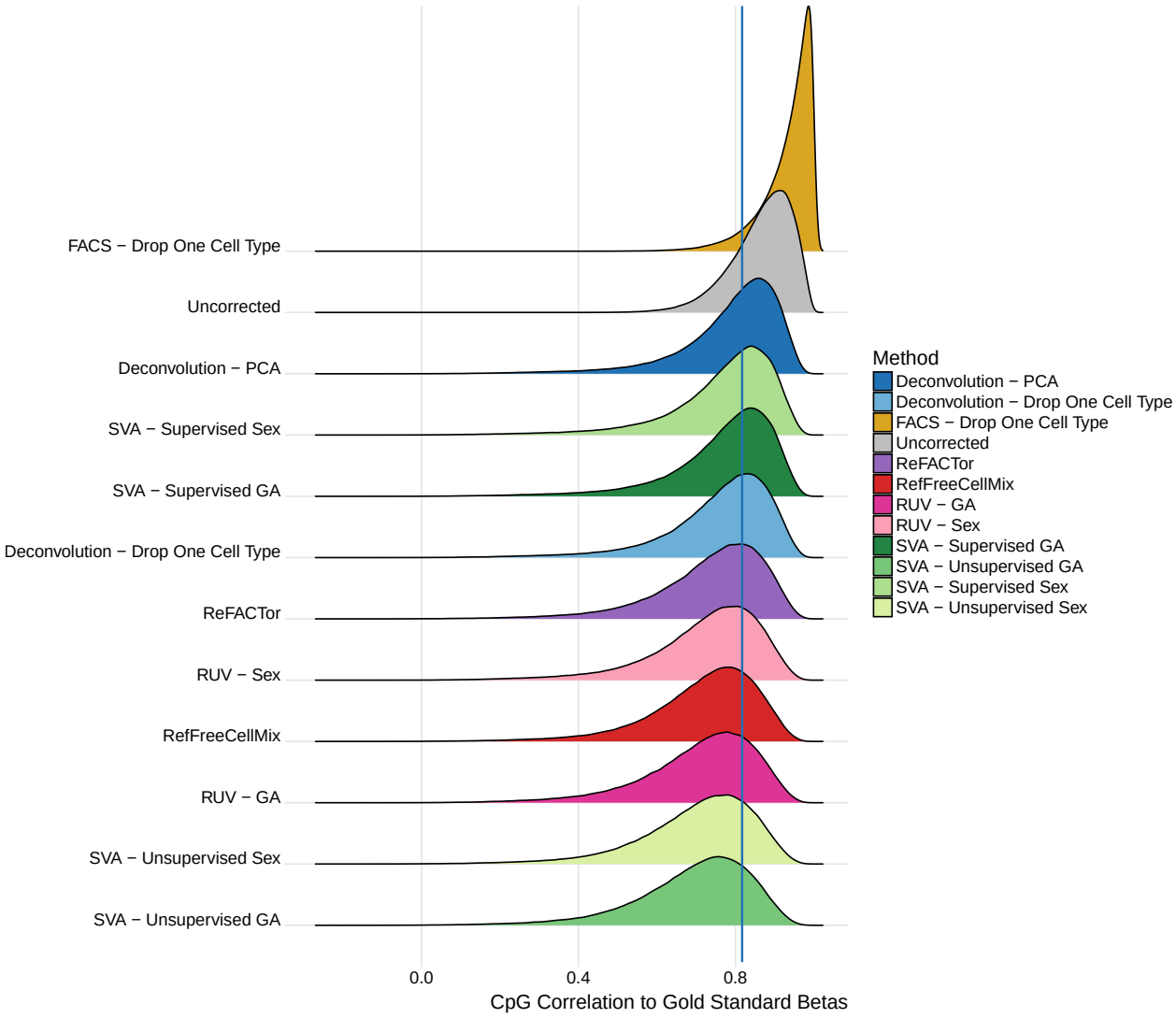

Figure 4: Distributions of correlation of beta values across samples at each CpG. Betas across samples were correlated between each method and the gold-standard FACS-PCA corrected betas. Vertical blue line shows the median correlation across all CpGs for the Deconvolution PCA corrected betas.

Table 3: Comparison of cell type correction methods through EWAS results. EWAS was performed for sex on each corrected dataset. True and false positives and negatives were cvalculated in comparison to the gold-standard FACS - PCA correction. Spearman is the correlation coefficient for all CpG p values, and Kendall is the rank correlation coefficient for the top 1000 CpG p values.

Method	Hits	True Positives	False Positives	False Negatives	Spearman	Kendall
FACS - PCA - Gold-Standard	5	-	-	-	-	-
FACS - Drop One Cell Type	0	0	0	5	0.636	0.42
Deconvolution - Drop One Cell Type	0	0	0	5	0.429	0.241
Deconvolution - PCA	0	0	0	5	0.458	0.264
ReFACTor	0	0	0	5	0.298	0.255
RefFreeCellMix	0	0	0	5	0.148	0.236
RUV	0	0	0	5	0.426	0.204
SVA - Supervised	4	0	4	5	0.56	0.279
SVA - Unsupervised	0	0	0	5	0.273	0.247
Uncorrected	631	4	627	1	0.595	0.283

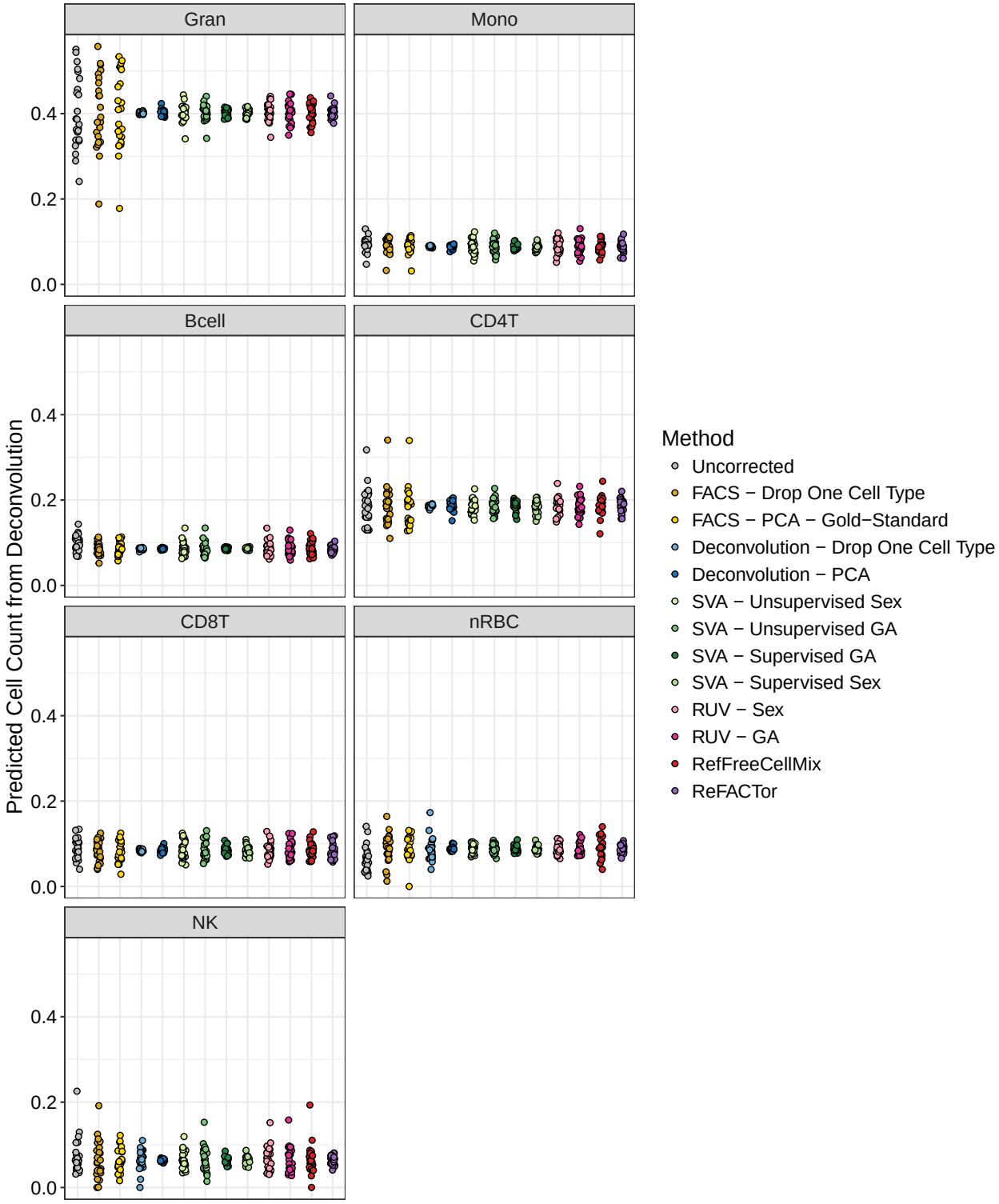

Figure 5: Varibility in the predicted cell counts on data already corrected for cell type. Across methods the deconvolution predicted cell counts for each cell type are shown.

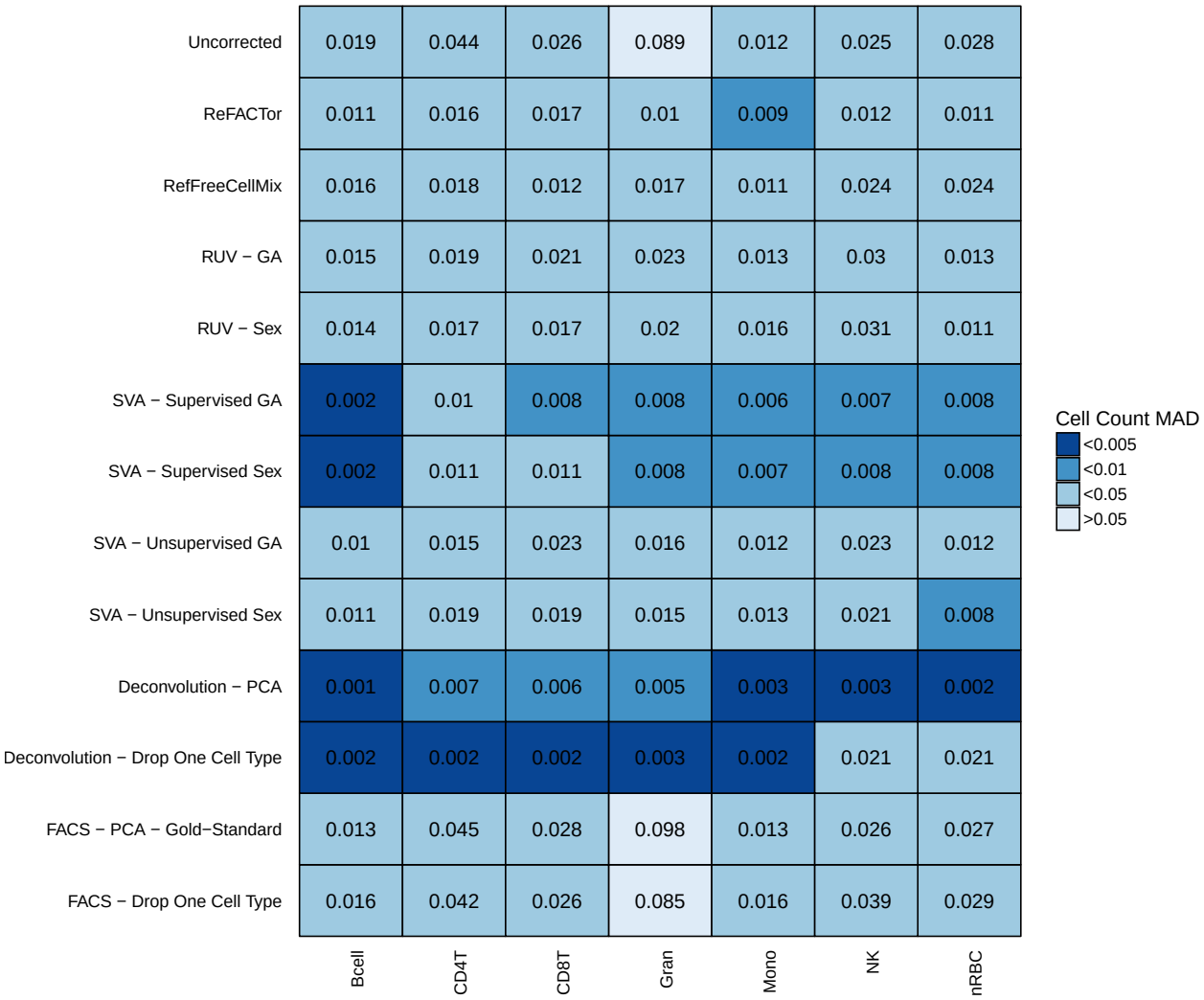

Figure 6: Median absolute deviation (MAD) of predicted cell type counts from deconvolution. Boxes are coloured by the discretized MAD value to highlight method performance.