Estimating cell type composition in whole blood using differentially methylated regions

Stephanie Hicks

Bioconductor 2017

DANA-FARBER CANCER INSTITUTE

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

What is DNA Methylation?

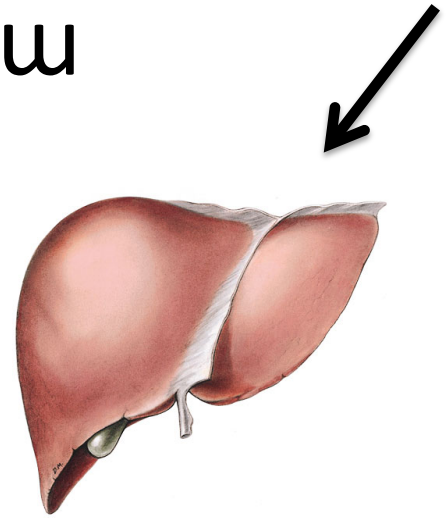
What is DNA Methylation?

m m

ATCGCGTTACTGCGGAA

TAGCGCAATGTCGCCTT

ɯ



What is DNA Methylation?

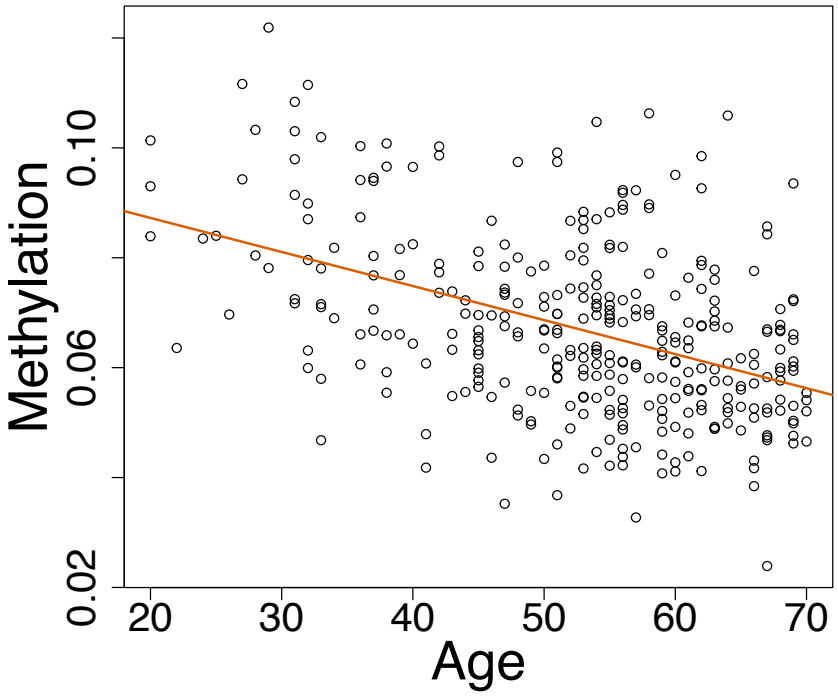
m

ATCGCGTTACTGCGGAA

TAGCGCAATGTCGCCTT

ɯ ɯ

DNA methylation in whole blood correlates with age at this one CpG



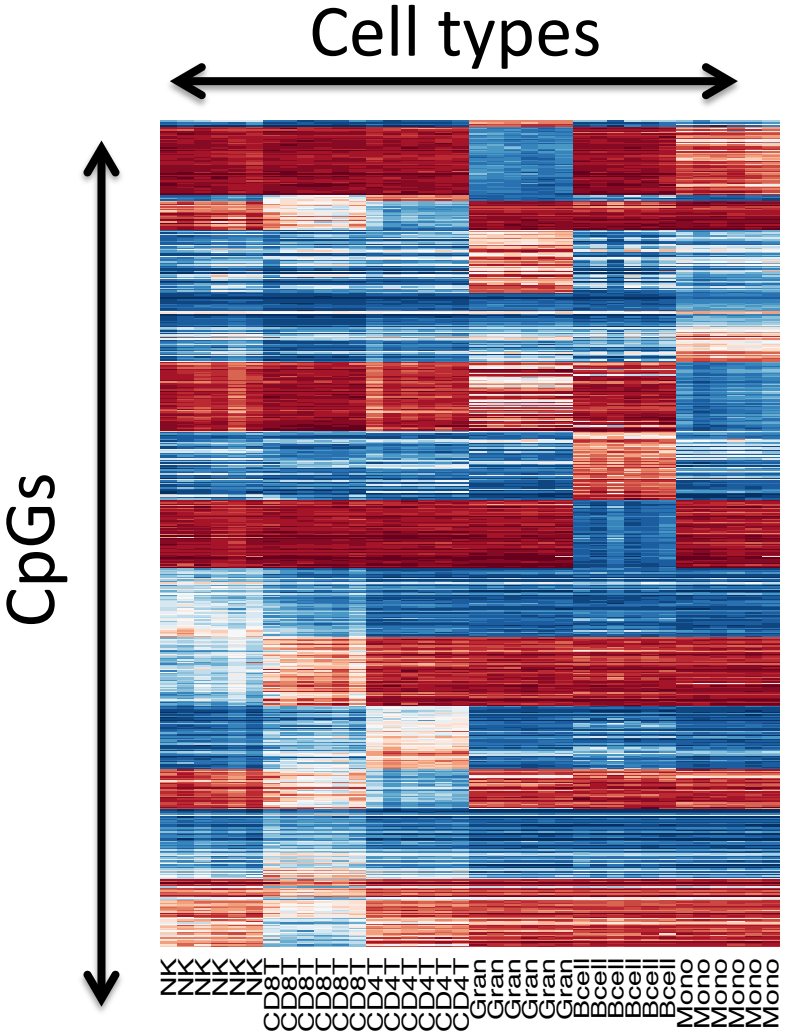
Data from GSE32148

Slide courtesy of A. Jaffe and R. Irizarry

Blood is a mixture of many cell types

Whole blood cell types:

- Tcells
  - CD8T
  - CD4T
  - Natural Killer
- Bcells
- Granulocytes
- Monocytes

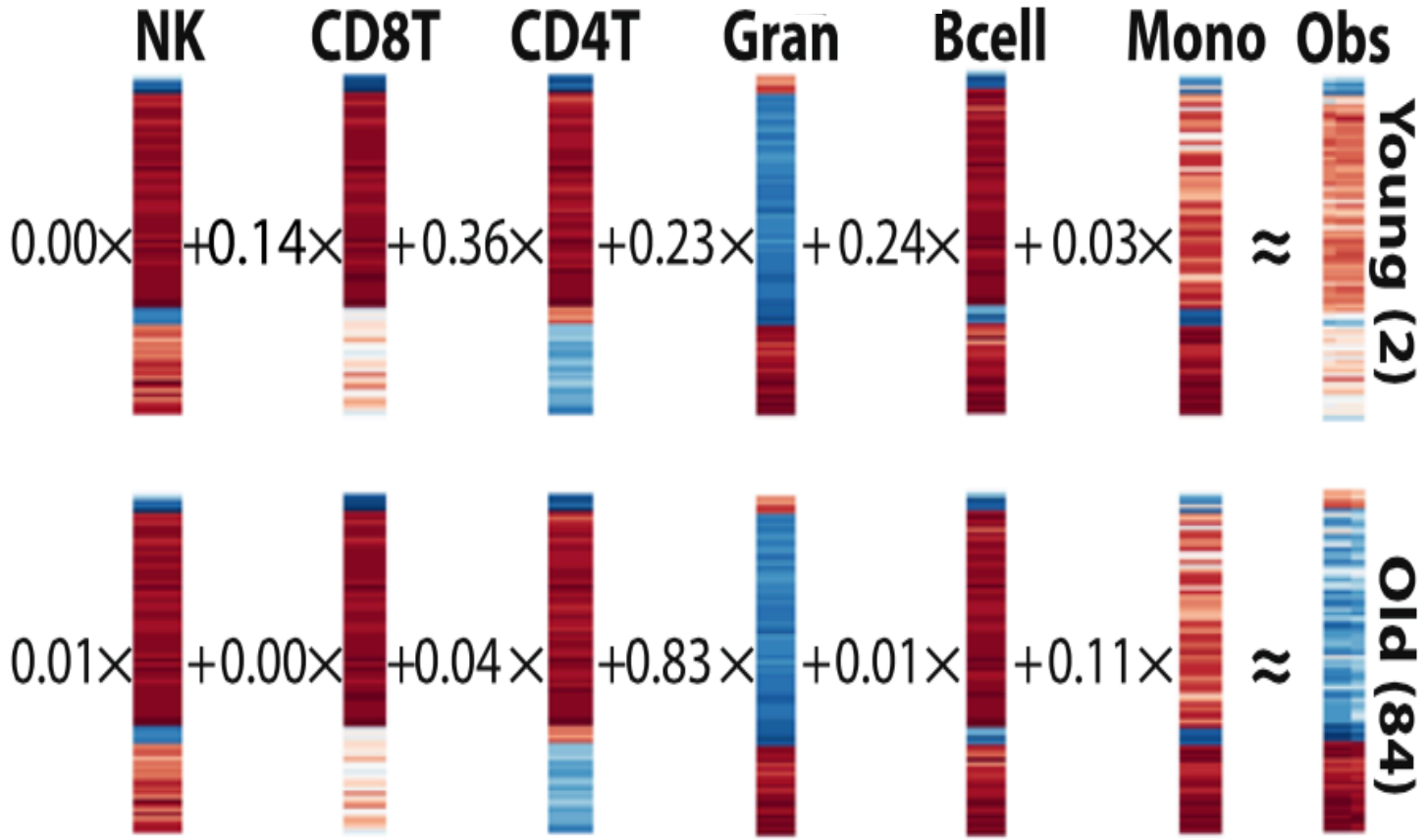


Bioconductor data package available:

- Data originally from Reinius et al. (2012)

```
> library(FlowSorted.Blood.450k)
```

Cell composition changes with age



- Different cell compositions in whole blood imply different observed whole blood DNA methylation profiles
- Important to estimate differences in cell composition

Jaffe and Irizarry (2014). *Genome Biology*

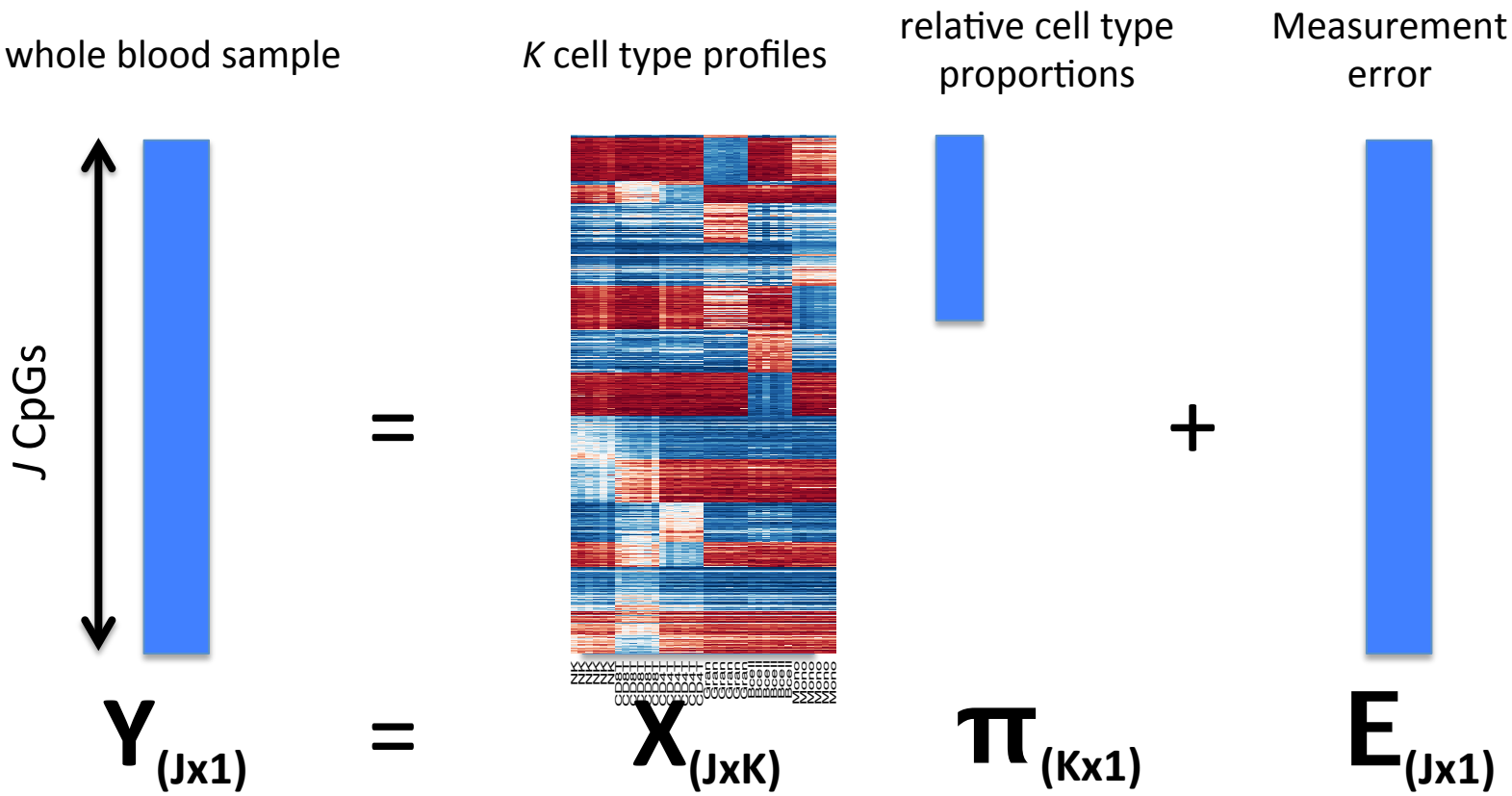
Statistical Model: Houseman et al. (2012)

$$Y_{ij} = \sum_{k=1}^{K} \pi_{ik} X_{jk} + \varepsilon_{ij}$$

$i = (1,\ldots,N)$ = whole blood samples

$j = (1,\ldots,J)$ = CpGs

$k = (1,\ldots,K)$ = cell type profiles



New platform technologies emerging

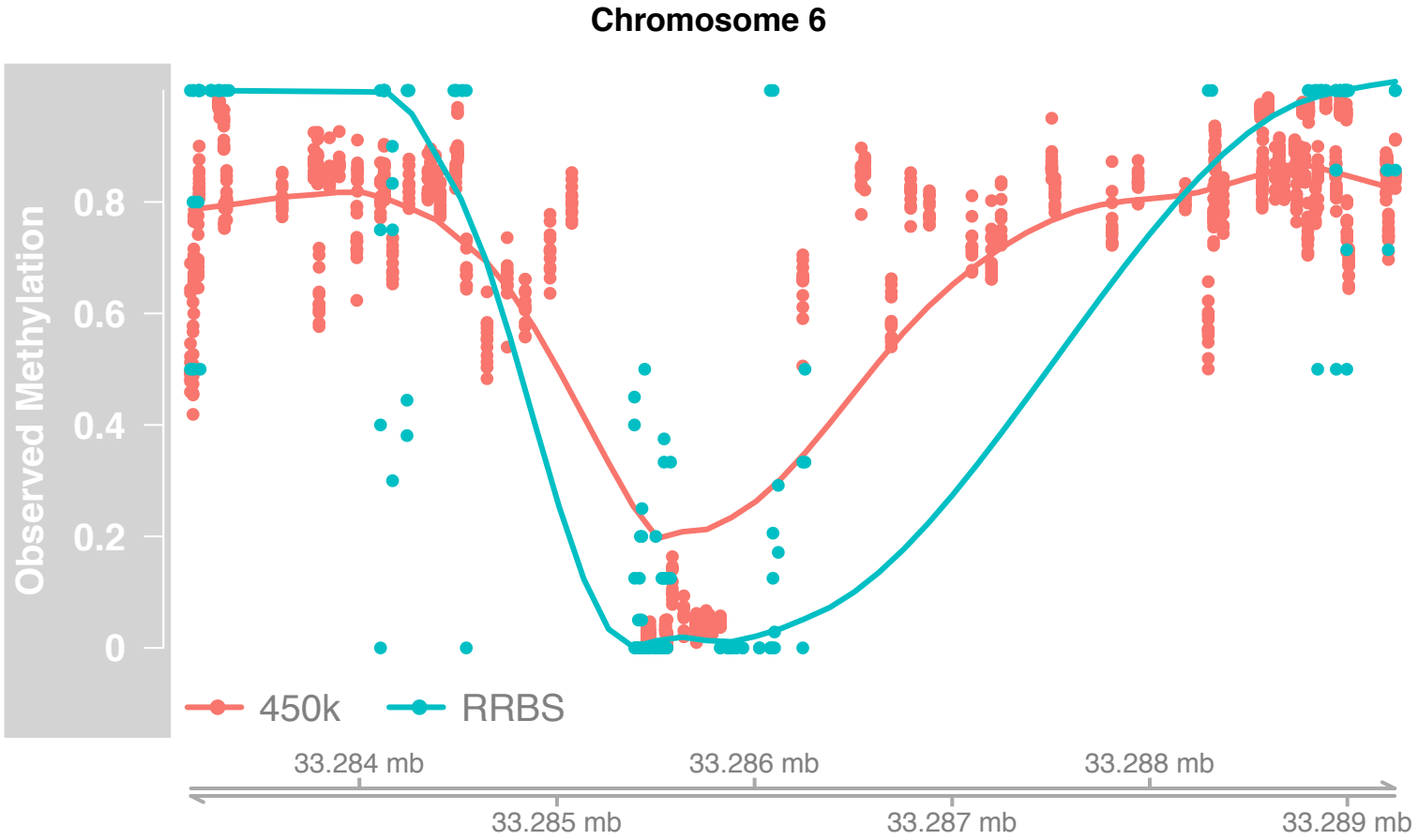
First approach

- Apply Houseman method using new platform technology

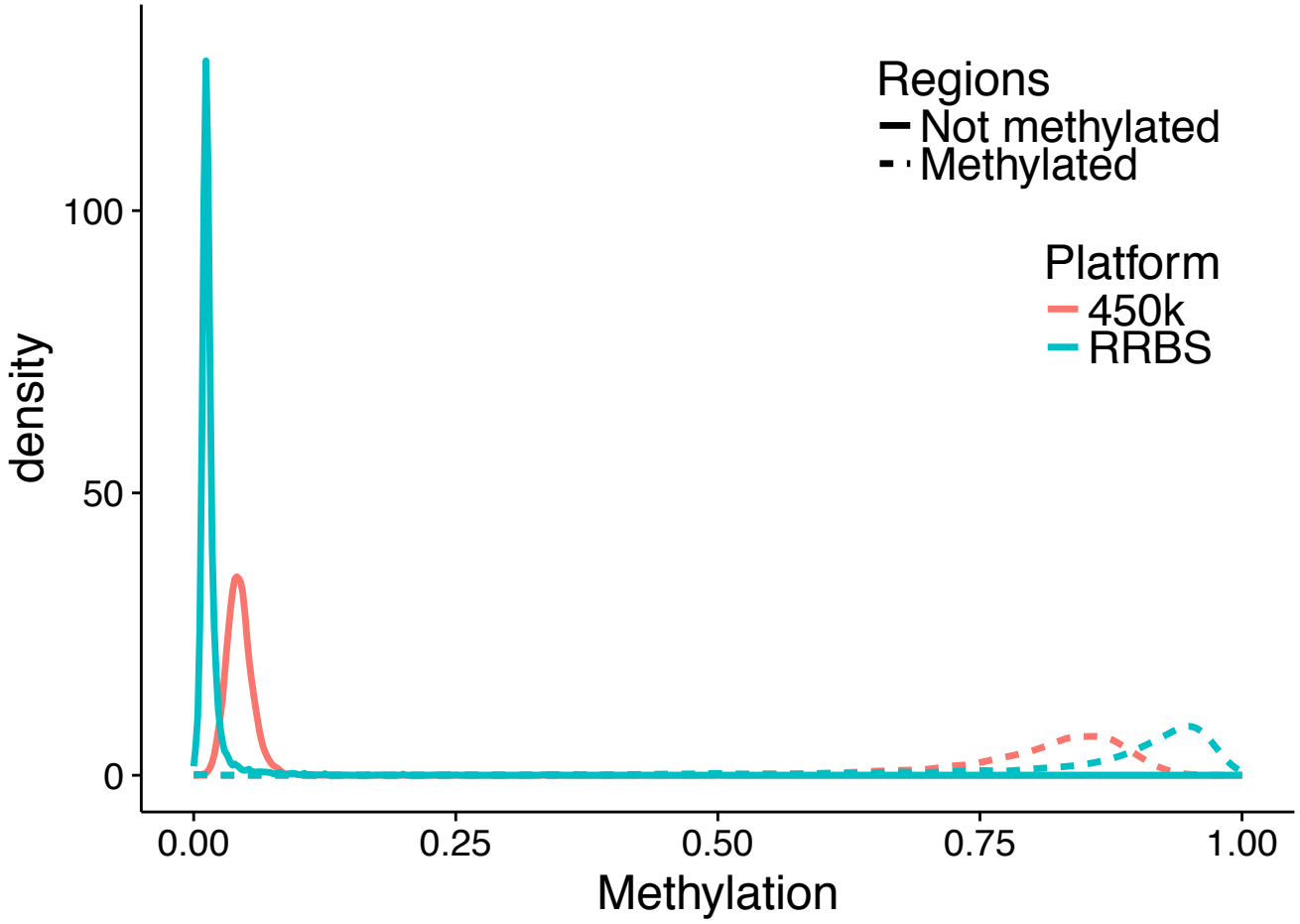
Problems with this approach

1. Observed methylation levels depend on platform used
2. Not all CpGs are included in new platforms

Platform-dependent differences between 450k array and RRBS platforms



Platform-dependent differences between 450k array and RRBS platforms



New platform technologies emerging

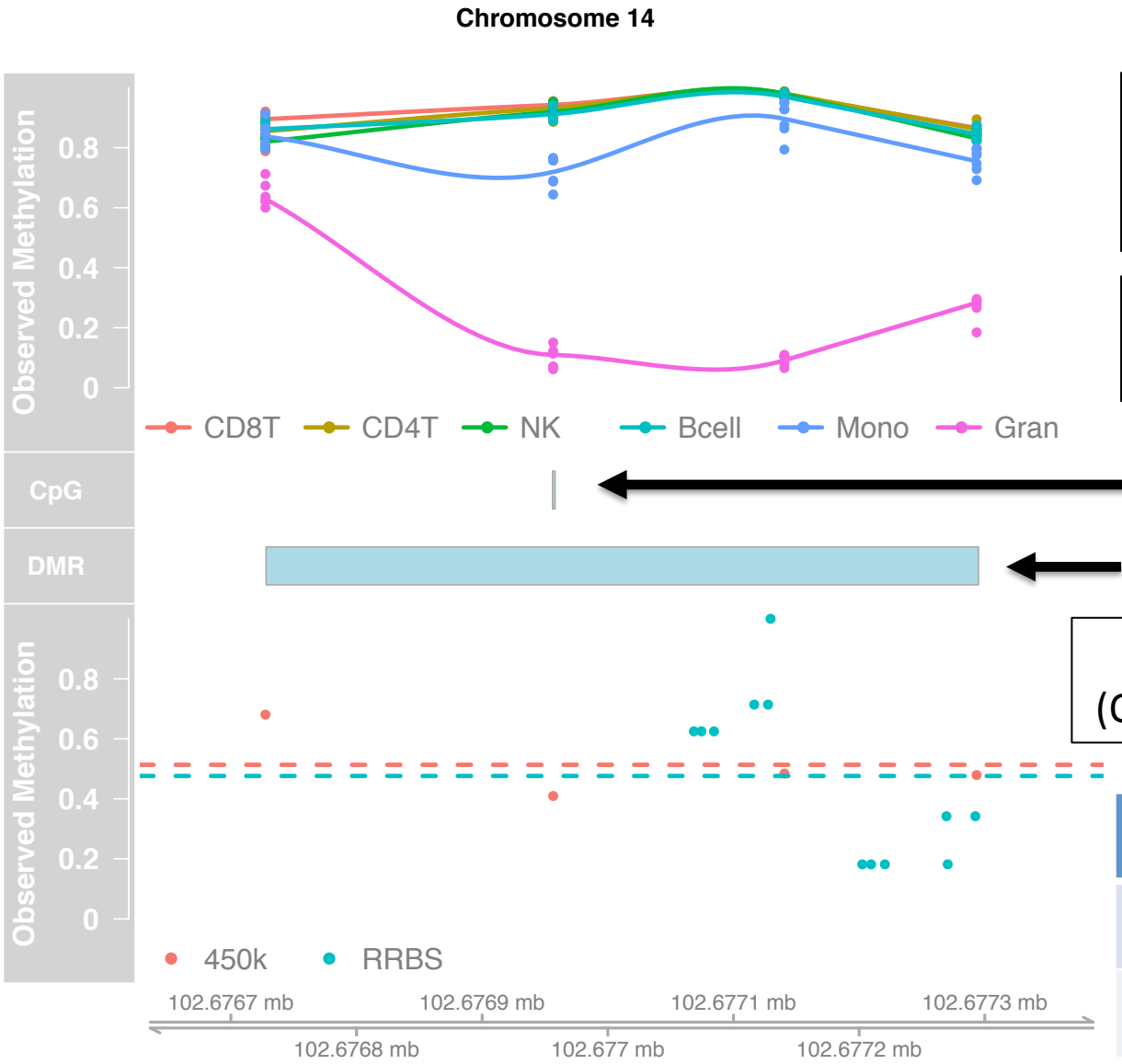
First approach

- Apply Houseman method using new platform technology

Problems with this approach

1. Observed methylation levels depend on platform
2. Not all CpGs are included in new platforms

Cell types preserve their methylation state across regions



Beta values
(Purified cell types on measured on 450k array)

- Identify regions using `bumphunter` BioC pkg

Cell type-specific **CpG**

Cell type-specific **region**

Beta values
(One whole blood sample)

	Microarray platform	Sequencing platform
Using CpGs	0.45	NA
Using Regions	0.55	0.50

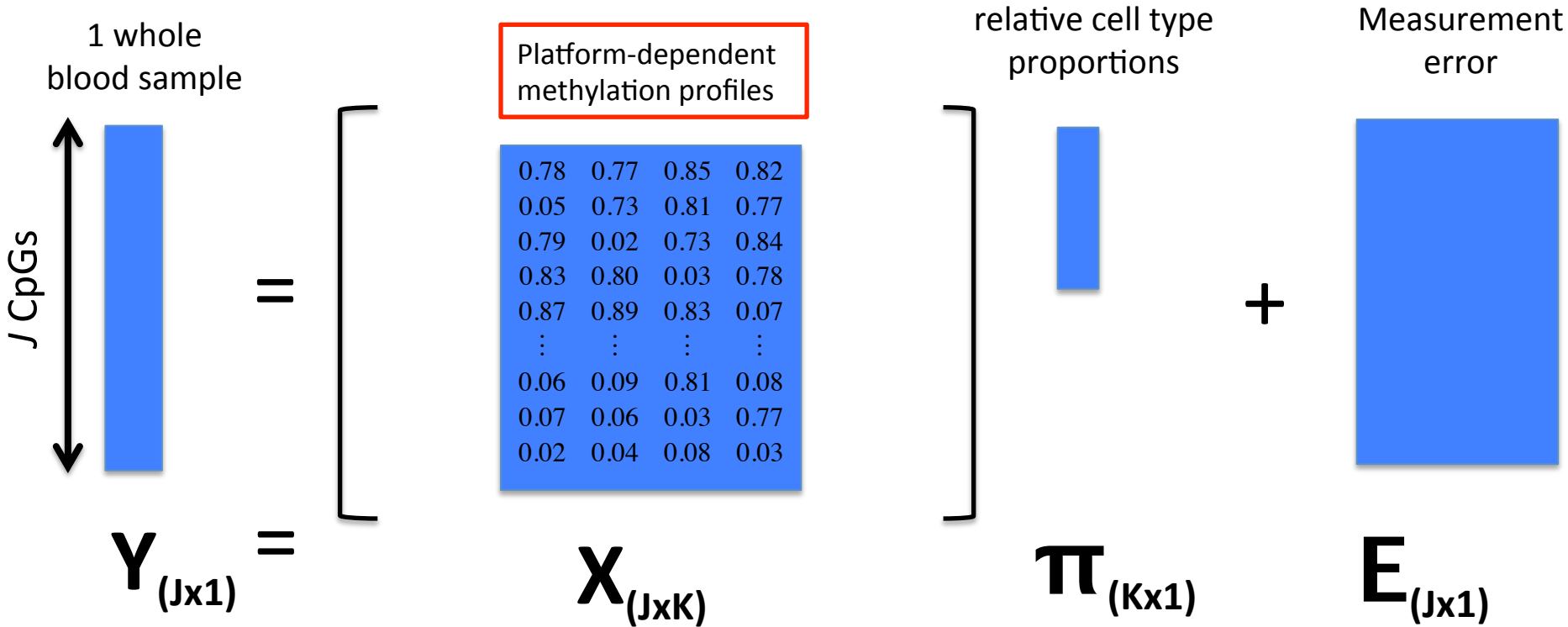
Recall Houseman Model:

$$Y_{ij} = \sum_{k=1}^{K} \pi_{ik} X_{jk} + \varepsilon_{ij}$$

$i = (1,\ldots,N) =$ whole blood samples

$j = (1,\ldots,J) =$ CpGs

$k = (1,\ldots,K) =$ cell type profiles



Our proposed model:

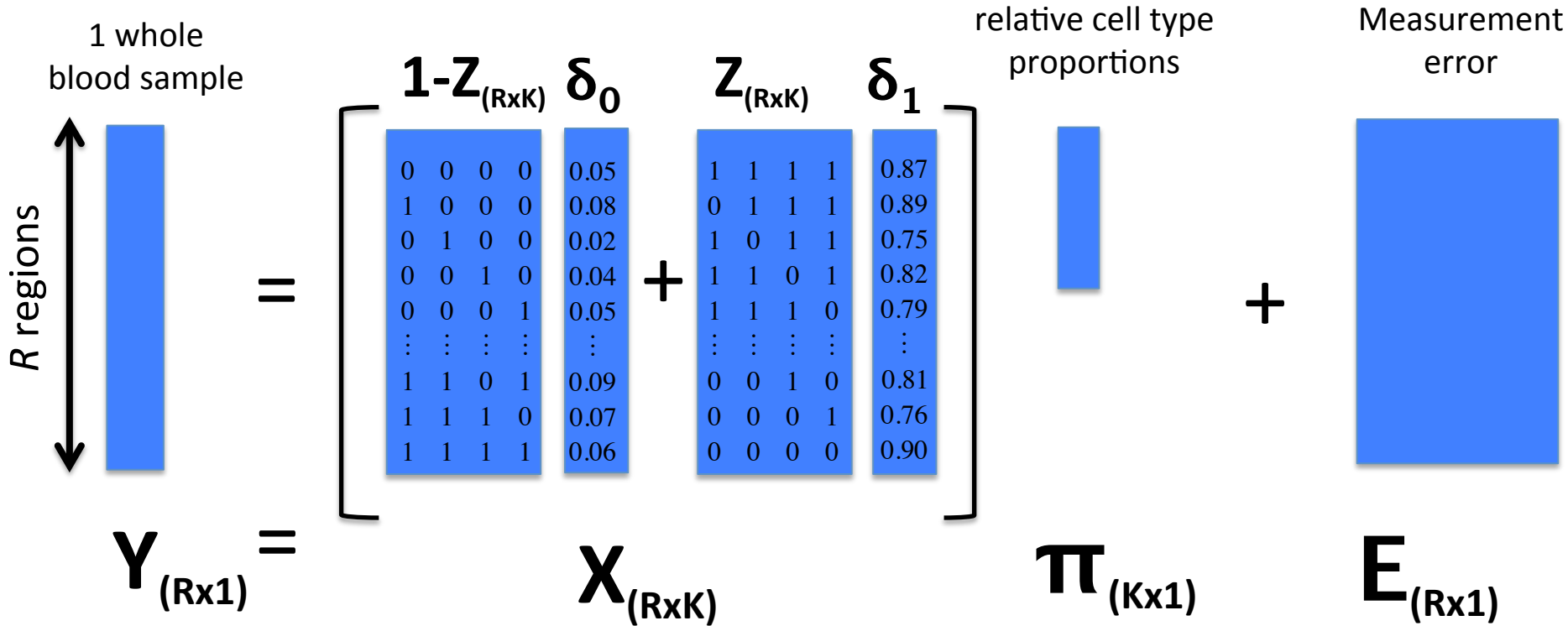
$$Y_r = \sum_{k=1}^{K} \pi_k \left[(1 - Z_{rk}) \delta_{o,r} + Z_{rk} \delta_{1,r} \right] + \varepsilon_r$$

$$\delta_{0,r} \sim N(\alpha_0,\ \sigma_0^2)$$
$$\delta_{1,r} \sim N(\alpha_1,\ \sigma_1^2)$$
$$\varepsilon_r \sim N(0,\ \sigma^2)$$

$$Z_{rk} = \begin{cases} 1 & \text{if region } r \text{ and cell type } k \text{ is methylated} \\ 0 & \text{otherwise} \end{cases}$$

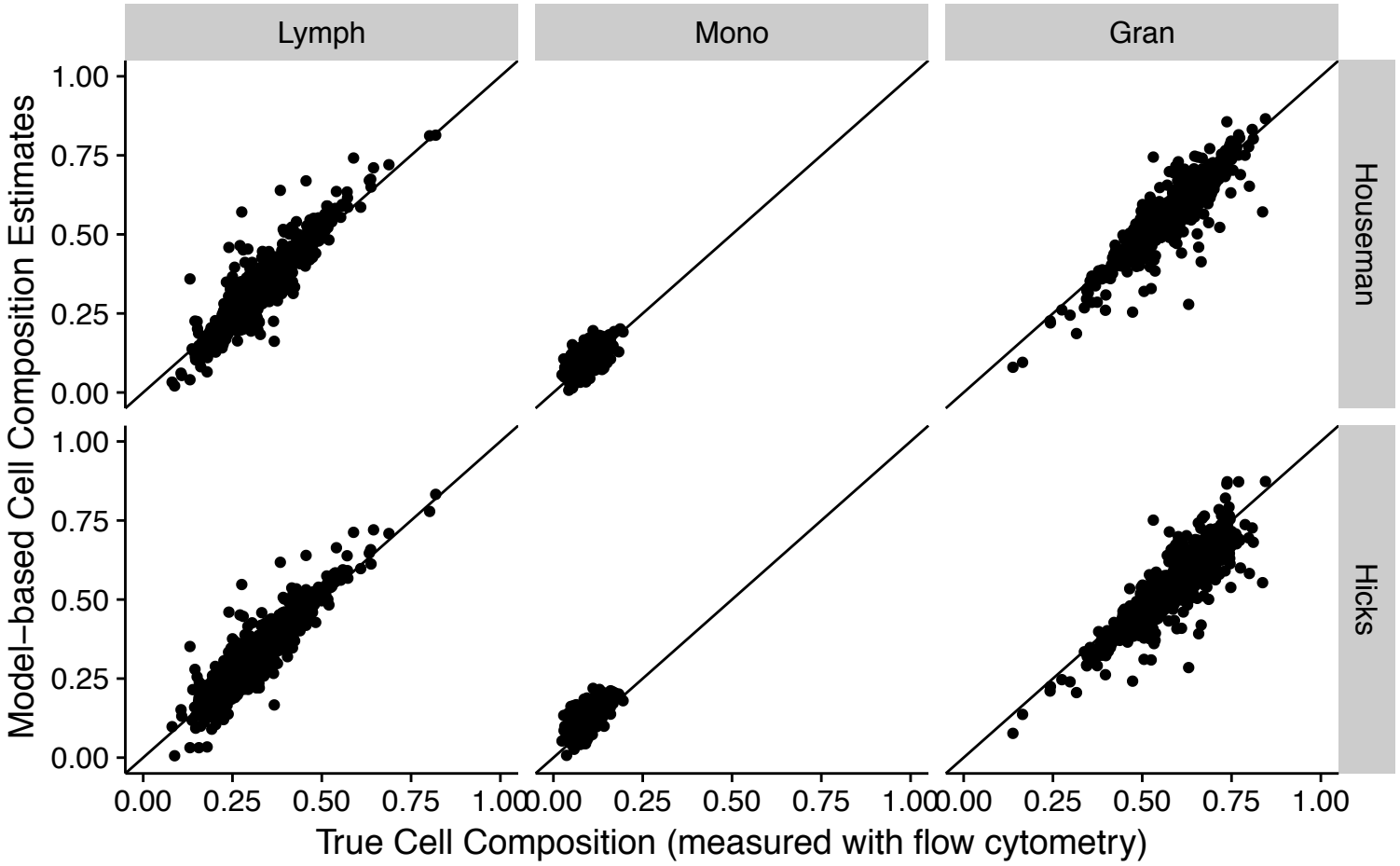
$r = (1, \ldots, R)$ = differentially methylated regions

$k = (1, \ldots, K)$ = cell types



How does our model perform?

N = 800 whole blood samples run on 450k microarray platform

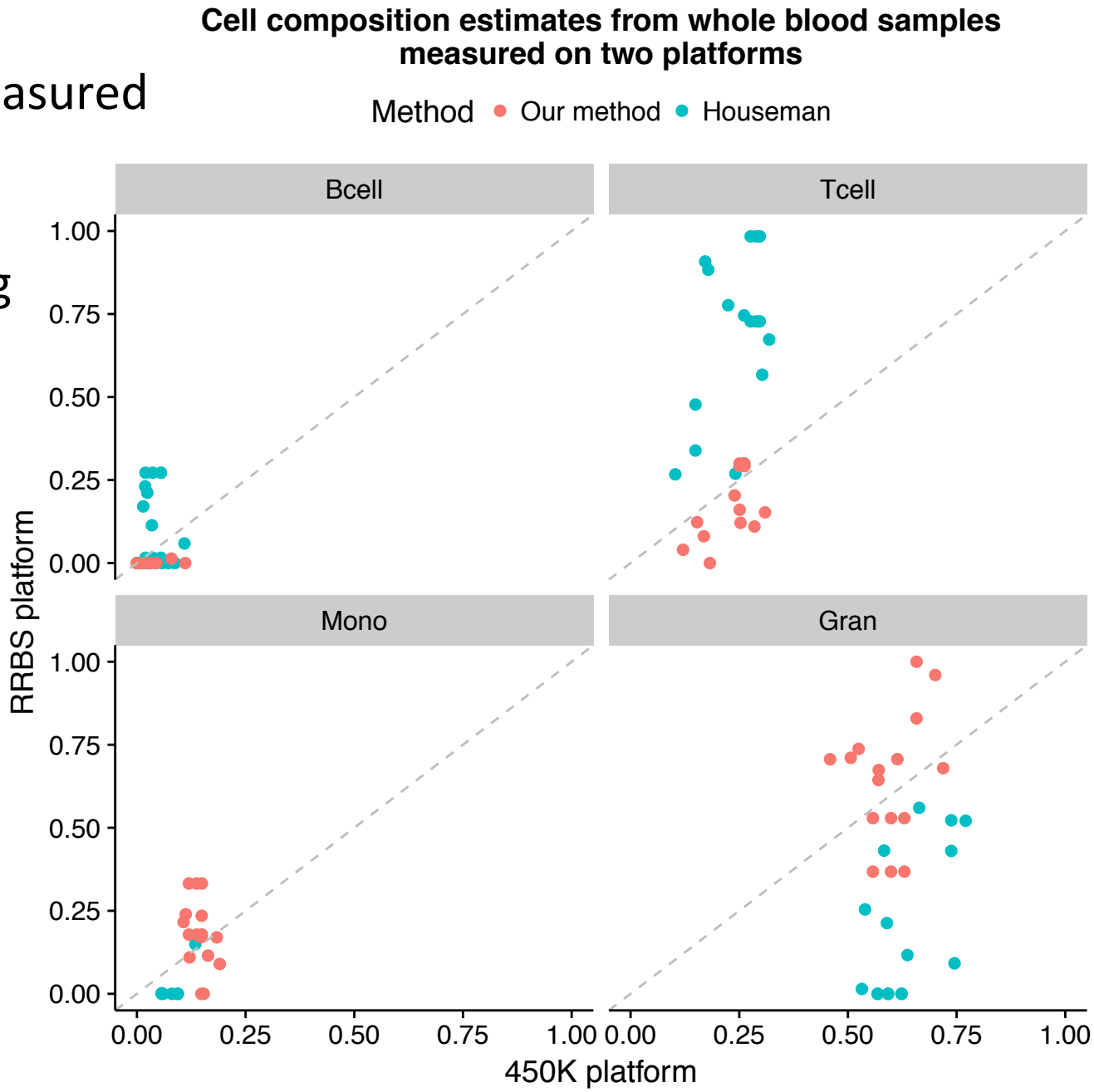


RMSE: 0.0385

RMSE: 0.0531

N = 12 samples measured on two platforms:

- 450k microarray
- RRBS sequencing



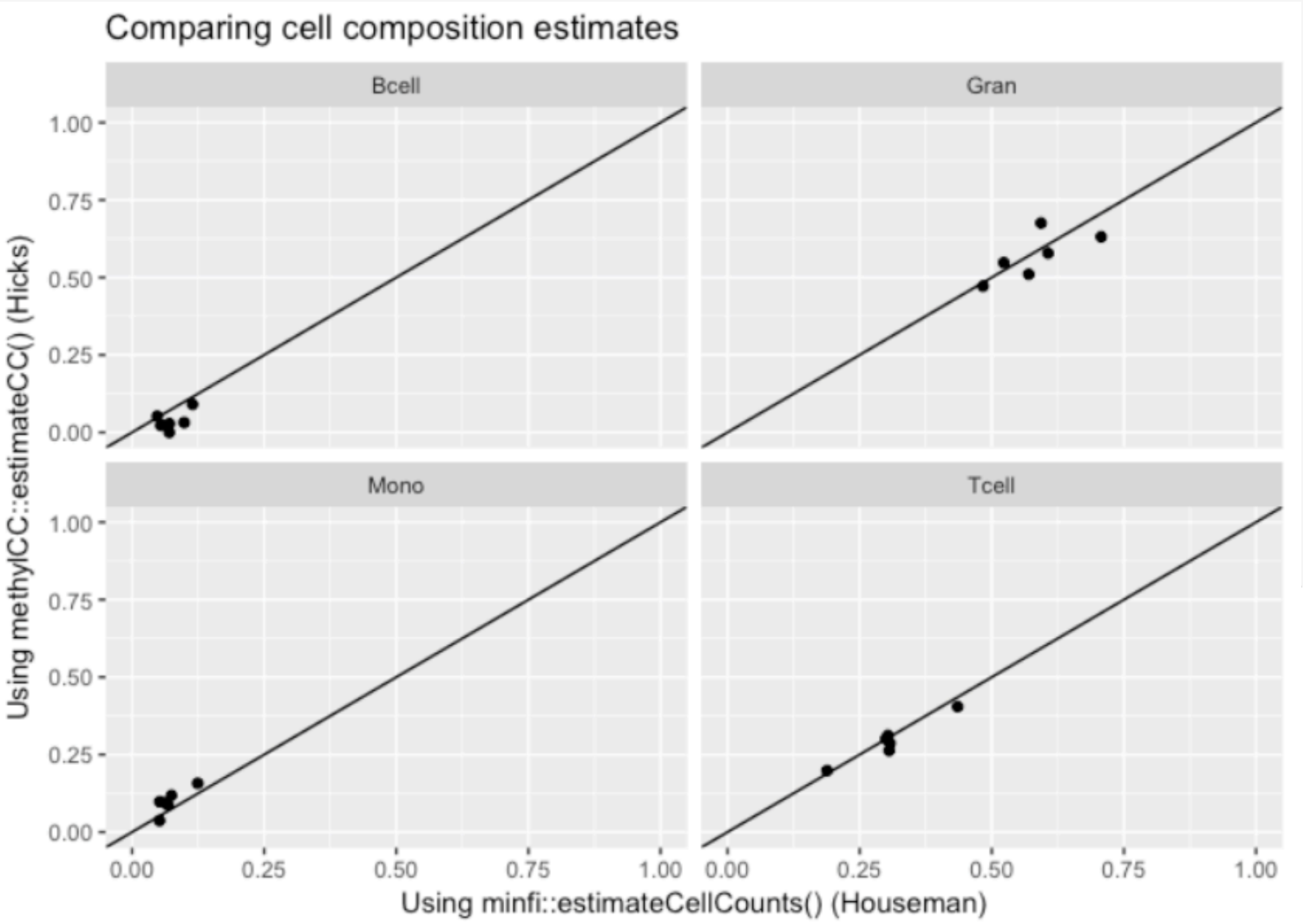
3.1 Demo for BioC 2017

```
library(methylCC)
library(minfi)
library(FlowSorted.Blood.450k)
data(FlowSorted.Blood.450k)

# Subset RGChan
rgset <- FlowSo


# Use methylCC:
est.methylCC <-
counts.methylCC

# Compare to mi
sampleNames(rgs
counts.minfi <-
```



For more information

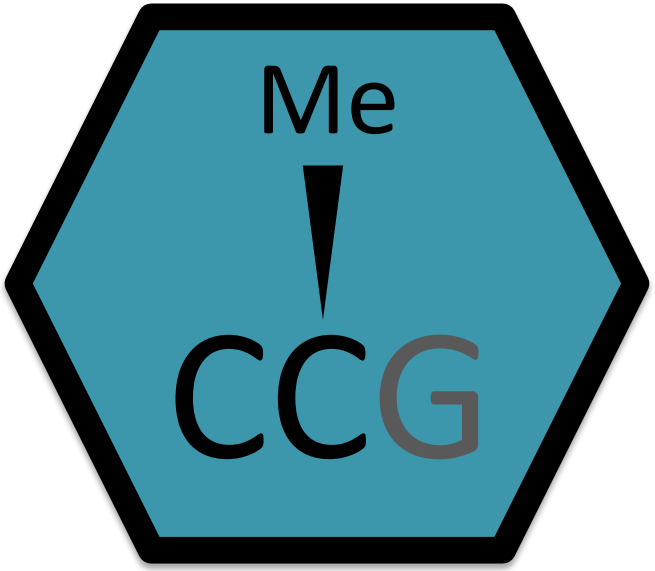
methylCC:

https://github.com/stephaniehicks/methylCC

Comments/Suggestions:

email: shicks@jimmy.harvard.edu

GitHub & Twitter: @stephaniehicks



#BioC2017
#RLadies
#dataparasite