## How to deal and detect batch effects

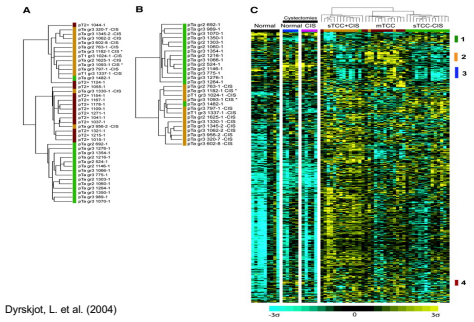
Rafael A Irizarry
@rafalab

## Pro tips

When you find an **unexpected result**, be skeptical, check for systematic errors.

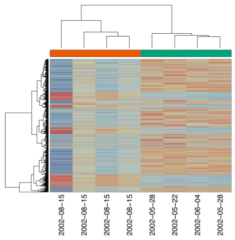
Always, look at the data!

## T

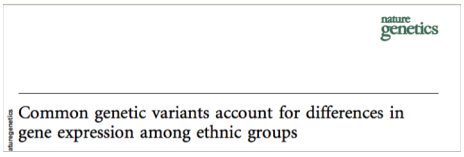



Dyrskjot, L. et al. (2004)

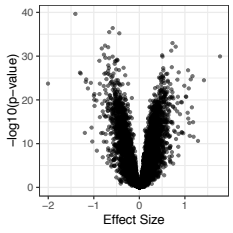
## Clustering of normals




## First example of unexpected result




## After running the standard workflow




**Conclusion:**
"This quantitative phenotype differs significantly between European derived and Asian-derived populations for 1,097 of 4,197 genes tested."

Summaries for one gene

p-value = $1.146925 \times 10^{-29}$

Fold change = 83%

Summaries for one gene

p-value = $10^{-29}$

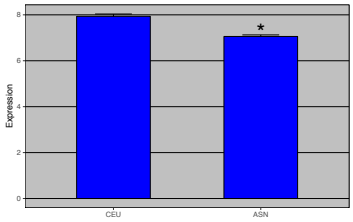
Fold change = 35%

Summaries for one gene

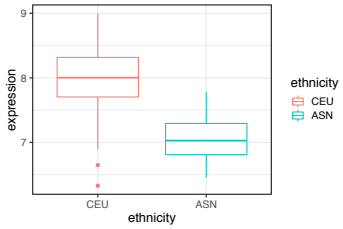
p-value is really really small

Fold change = 35%

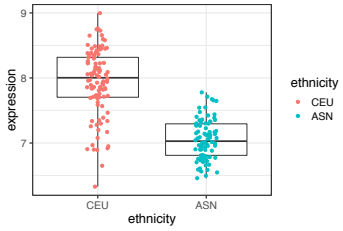
## First, dynamite plots needs to die



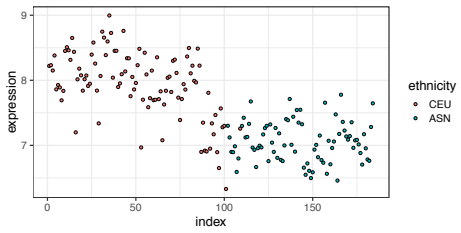

## Boxplots are better



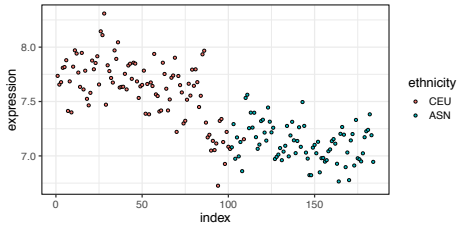

## Showing the data is even better



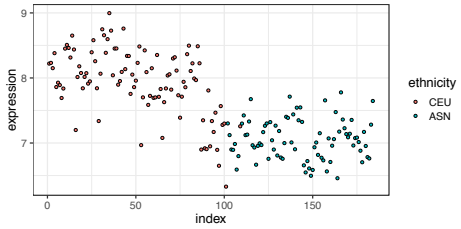
## Show the raw data



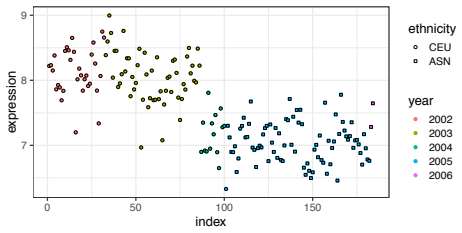

## Here is another of the top genes



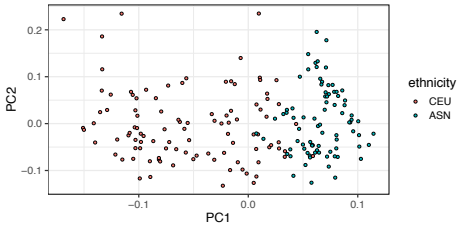

## And another



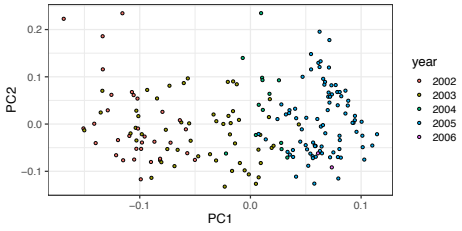
## Check for confounders



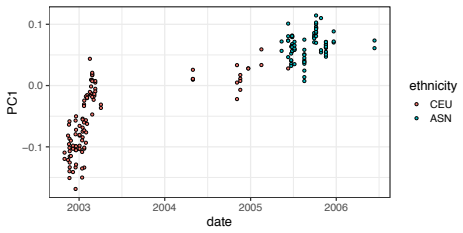

## PCA are a summary not raw data, but still useful



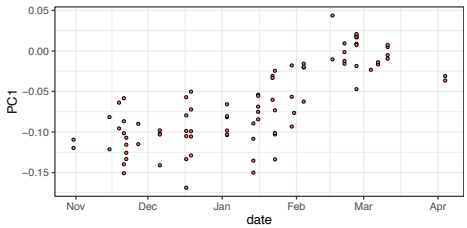

## PCA are a summary not raw data, but still useful



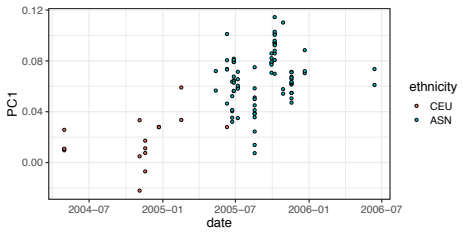
## Some further exploration: PC 1 versus date



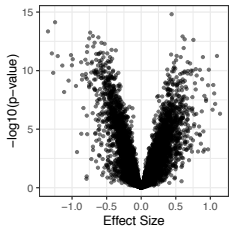

## PC1 versus date in 2003




## PC 1 versus date (after 2004)



## After looking at the data




**Conclusion:**

"A possible explanation for the pervasive signature of differential expression observed… is a systematic bias introduced during sample preparation or microarray expression measurements. ."

Akey, Biswas, Leek & Storey (2007) Nature Genetics

## Experimental design flaw

| Year | ASN | CEU |
|---|---|---|
| 2002 | 0 | 32 |
| 2003 | 0 | 54 |
| 2004 | 0 | 13 |
| 2005 | 80 | 3 |
| 2006 | 3 | 0 |

## Not that hard to look at data

```
library(GSE5859)
data("GSE5859")

ind <- which(e$ethnicity %in% c("ASN", "CEU"))
e <- e[,ind]
e$ethnicity <- droplevels(e$ethnicity)

tt <- rowttests(e, e$ethnicity)
```

## Not that hard to look at data

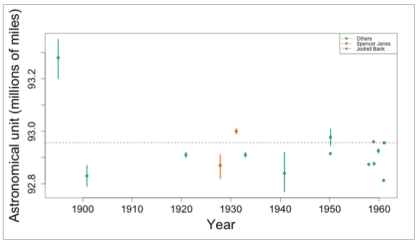
```
library(GSE5859)
data("GSE5859")

ind <- which(e$ethnicity %in% c("ASN", "CEU"))
e <- e[,ind]
e$ethnicity <- droplevels(e$ethnicity)

tt <- rowttests(e, e$ethnicity)

o <- order(tt$p.value)

for(i in o[1:25]){
  plot(exprs(e)[i,])
}
```

## Not that hard to look at data

```
library(GSE5859)
data("GSE5859")

ind <- which(e$ethnicity %in% c("ASN", "CEU"))
e <- e[,ind]
e$ethnicity <- droplevels(e$ethnicity)

tt <- rowttests(e, e$ethnicity)

o <- order(tt$p.value)

for(i in o[1:25]){
  plot(exprs(e)[i,])
}
```

## Batch effects are not new

## Batch Effects are not new nor specific to genomics




Youden (1972) Enduring Values

## General problem

$$Y_{1,i} = \alpha + \varepsilon_{1,i}$$
$$Y_{2,i} = \alpha + \beta + \varepsilon_{2,i}$$

$$\bar{Y}_2 - \bar{Y}_1 \approx \beta$$
$$\bar{Y}_2 - \bar{Y}_1 \approx \beta \pm \sigma/\sqrt{N}$$

## With “batch effects” Z we may have:

$$Y_{1,i} = \alpha + Z_1 + \varepsilon_{1,i}$$
$$Y_{2,i} = \alpha + \beta + Z_2 + \varepsilon_{2,i}$$

$$\bar{Y}_2 - \bar{Y}_1 \not\approx \beta \pm \sigma/\sqrt{N}$$

With “batch effects” we have (Zs are batch effects):

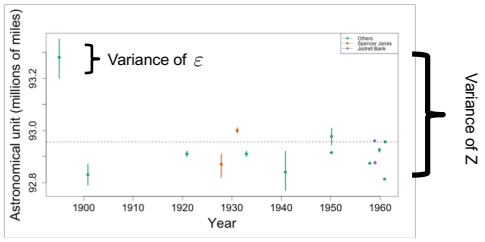
$$Y_{1,i} = \alpha + Z_1 + \varepsilon_{1,i}$$
$$Y_{2,i} = \alpha + \beta + Z_2 + \varepsilon_{2,i}$$

$$\bar{Y}_2 - \bar{Y}_1 \approx \beta + (Z_2 - Z_1) \pm \sigma/\sqrt{N}$$

or

$$\bar{Y}_2 - \bar{Y}_1 \approx \beta \pm (\sigma/\sqrt{N} + \text{variance due to } Z)$$

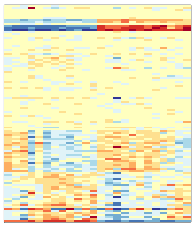
Batch Effects are not new nor specific to genomics



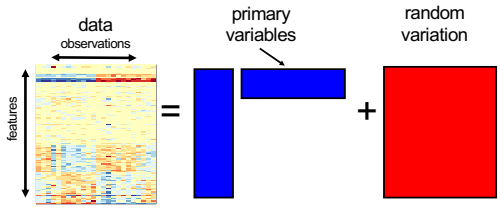

Youden (1972) Enduring Values

12 males, 12 females, two months, 109 genes

| | Female | Male |
|---|---|---|
| June 2005 | 3 | 9 |
| October 2005 | 9 | 3 |

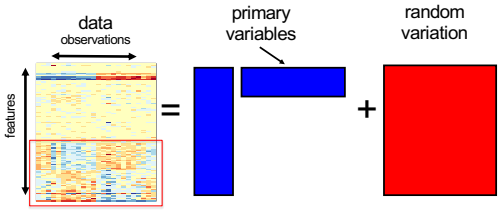


## Decomposing variability



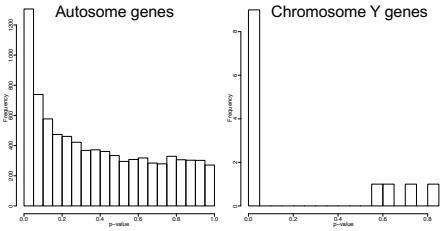

$$Y_{m\times n} = \beta_{m\times p} X_{p\times n} + \varepsilon_{m\times n}$$

## This model does not account for batch

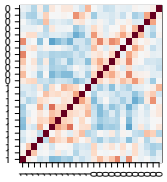



$$Y_{m,n} = \beta_{m,p} X_{p,n} + \varepsilon_{m,n}$$

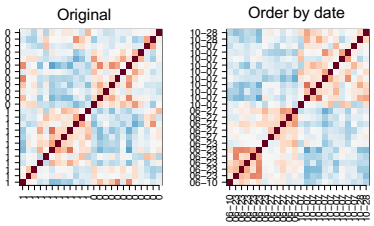
## p-value histograms



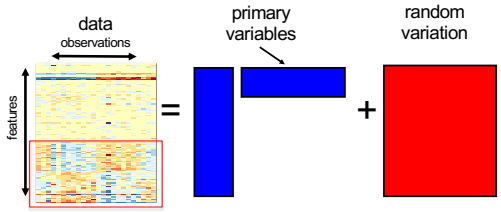
## Sample correlations



## Sample correlations

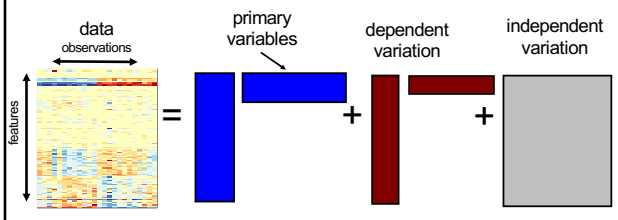



## Note structure




$$Y_{m\times n} = \beta_{m\times p} X_{p\times n} + \varepsilon_{m\times n}$$

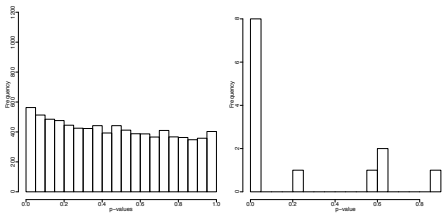
## Factor analysis



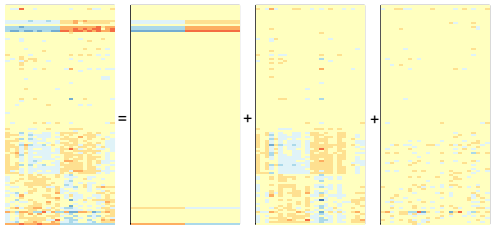

$$Y_{m\times n} = \beta_{m\times p} X_{p\times n} + \alpha_{m\times k} W_{k\times n} + \varepsilon_{m\times n}$$

Leek and Storey 2007,2008

## After Factor Analysis




## Decomposed data



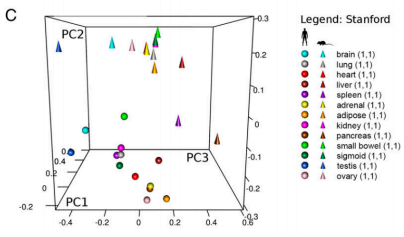
A more subtle example

## Surprising conclusions

Lin et al. (2014) Comparison of the transcriptional landscapes between human and mouse tissues. *PNAS*

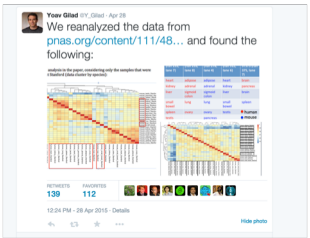
**``tissues appear more similar to one another within the same species than to the comparable organs of other species".**

## Here is the plot in which this is based on

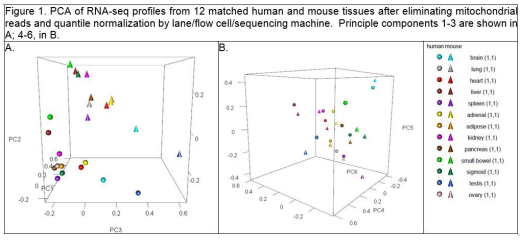



NB: The use of pseudo-3D is discouraged by data visualization experts

## The internet reacted

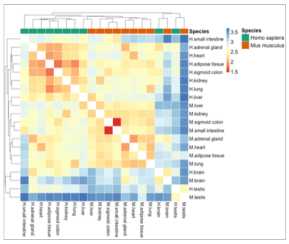



## Lin et al re-ran the experiment




**It's not the sequencer!**

## A better version of that plot confirms the result (sort of)

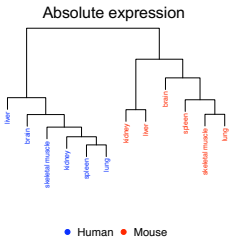


## Over Ten Years Ago

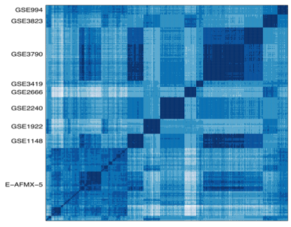
Yanai at al 2004 came to a very similar conclusion using microarrays:

**``any tissue is more similar to any other human tissue examined than to its corresponding mouse tissue".**

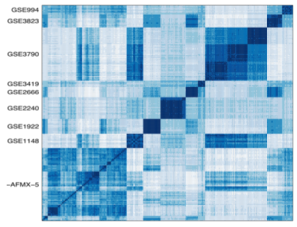
## Consequences



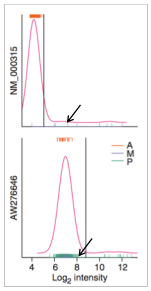

## Correlation between samples before centering



## After centering



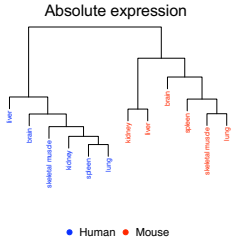

## Probe effect is strong



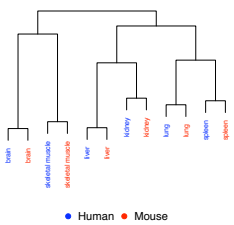

Gene Expression Bar Code work:
Zilliox and Irizarry Nature Methods 2007
McCall et al. NAR 2011

## Consequences

Absolute expression

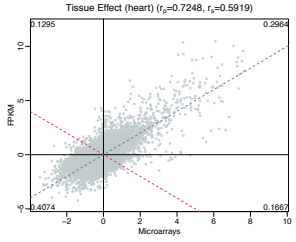



Gene Expression barcode



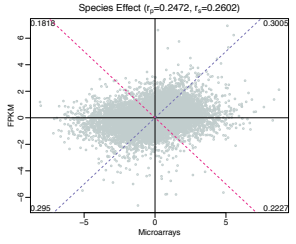
Is there a “probe effect” in RNAseq?

## Arrays and RNA-Seq agree on tissue specific genes



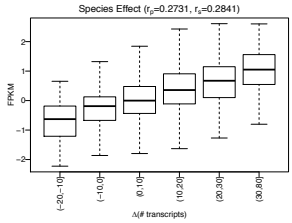

Comparing log-fold change for heart versus all other tissues

## But for the species effect they **do not** agree

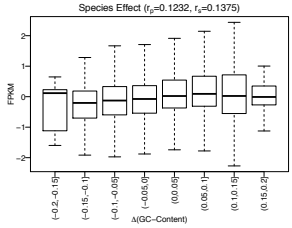


Are there “probe effects” in RNA-Seq?

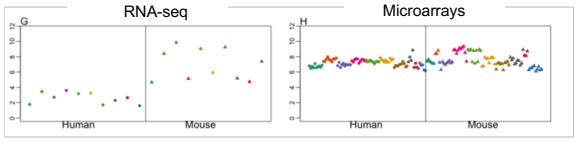
Number of transcripts explains much of the variability



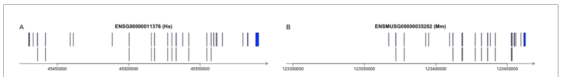

GC content explains some of the variability

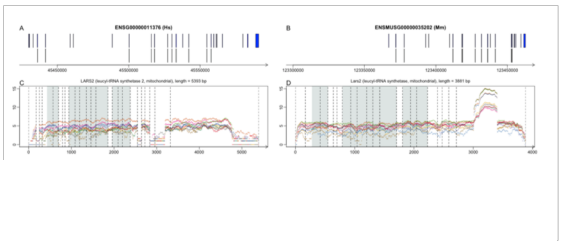


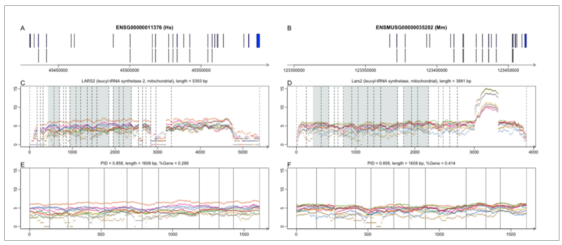
## An example



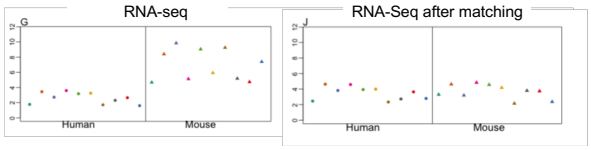

## Look at the data

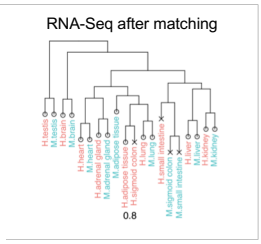






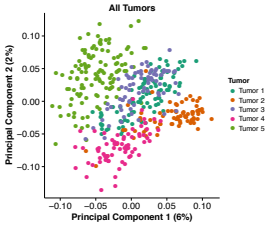
## An example


RNA-seq
RNA-Seq after matching
Human
Mouse
Human
Mouse

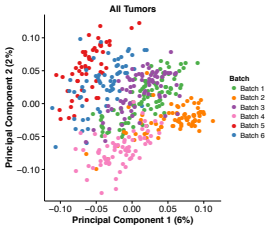

RNA-Seq after matching
0.8

Single Cell

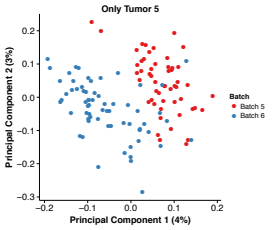
## Discovering new cell types



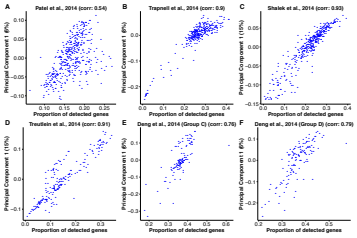

## Or is it a batch effect?



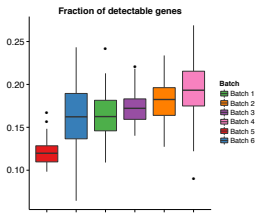
## Same group measured on two sequencers




## There are many zeros and they varies across sample




## The proportion of zeros changes



## Sites that change with age ?


Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, Whittaker P, McCann OT, Finer S, Valdes AM, et al: **Human aging-associated DNA hypermethylatio occurs preferentially at bivalent chromatin domains.** *Genome Res* 2010, **20:**434-439.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, et al: **Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer.** *Genome Res* 2010, **20:**440-446.

Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, Warren ST: **Age-associated DNA methylation in pediatric populations.** *Genome Res* 2012, **22:**623-632.

Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D, Mangino M, Zhai G, Zhang F, Valdes A, et al: **Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population.** *PLoS Genet* 2012, **8:**e1002629.

Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, et al: **Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates.** *Mol Cell* 2012.

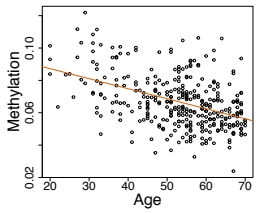
Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Carmona FJ, et al: **Distinct DNA methylomes of newborns and centenarians.** *Proc Natl Acad Sci U S A* 2012, **109:**10522-10527.

Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, van Eijk K, van den Berg LH, Ophoff RA: **Aging effects on DNA methylation modules in human brain and blood tissue.** *Genome Biol* 2012, **13:**R97.

Lee H, Jaffe AE, Feinberg JI, Tryggvadottir R, Brown S, Montano C, Aryee MJ, Irizarry RA, Herbstman J, Witter FR, et al: **DNA methylation shows genome-wide association of NFIX, RAPGEF2 and MSRB3 with gestational age at birth.** *Int J Epidemiol* 2012, **41:**188-199.

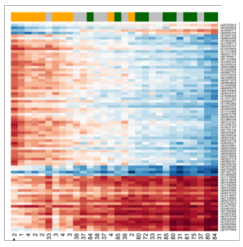
Johansson A, Enroth S, Gyllensten U: **Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan.** *PLoS One* 2013, **8:**e67378.


## DNA methylation correlates with Age for this CpG

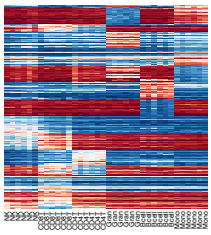



Data from GSE32148

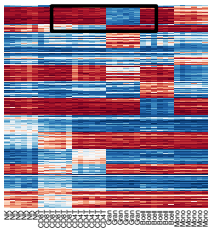
## Whole Blood



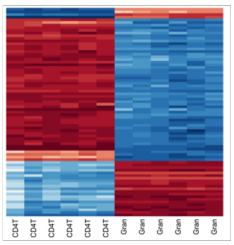
## Blood is a mixture of many cell types



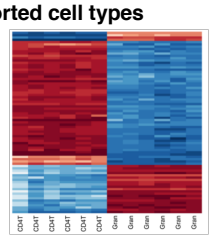
## Blood is a mixture of many cell types



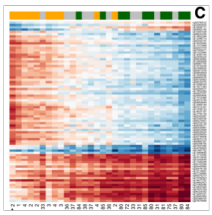
## Close up



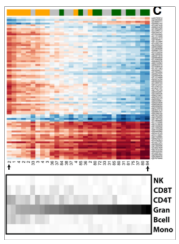
**Sorted cell types**



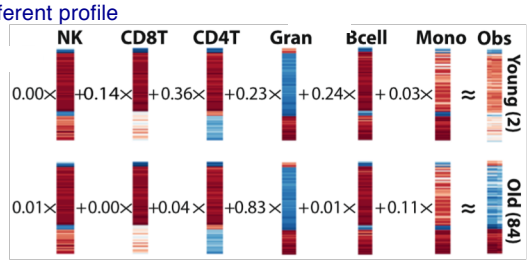
**Whole blood**




## Cell composition changes with age

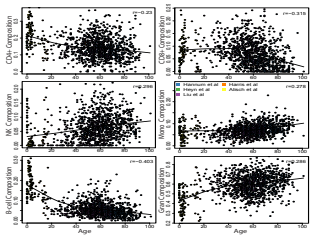



## Different composition implies different profile




Jaffe and Irizarry (2014) Genome Biology

## Cell composition versus age



## Confounding differences in brain methylation

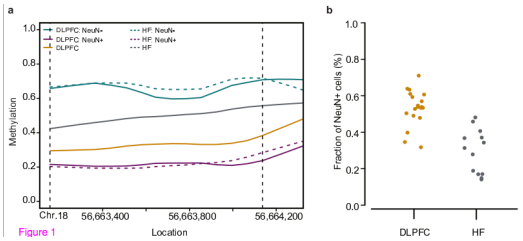
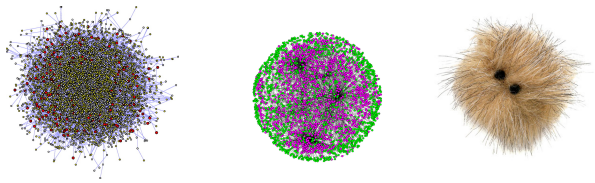



Figure 1

Montaño et al (2013) *Genome Biology*

## Systems Biology



## How do we interpret correlation?

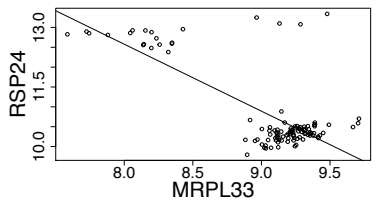
Two random variables are correlated when they vary together

But vary across what?

## You can also have batch effects

RSP24 and MRPL33 are negatively correlated

## Raw data confirms negative correlation

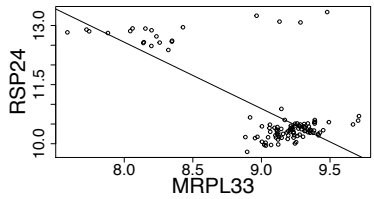


## Batch Effects

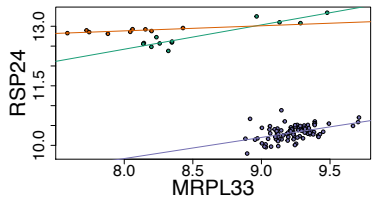
| Experiment | Correlation | p-value |
|---|---|---|
| GSE25076 | 0.93 | <0.01 |
| GSE33372 | 0.71 | 0.03 |
| GSE36674 | 0.46 | <0.01 |
| Overall | -0.74 | <0.01 |

Simpons's Paradox!

## Raw data confirms negative correlation



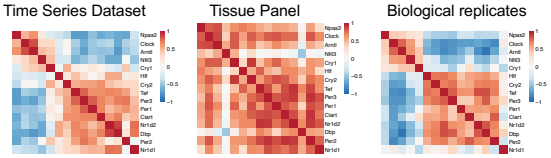

## Correlation positive once we account for batch effects



Thanks!

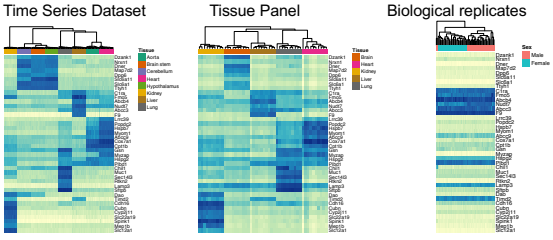
## Circadian genes

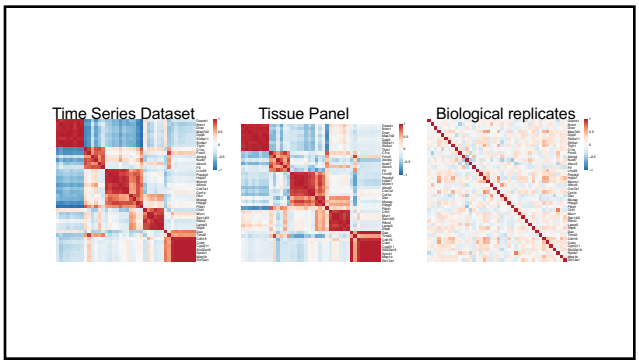
Time Series Dataset | Tissue Panel | Biological replicates

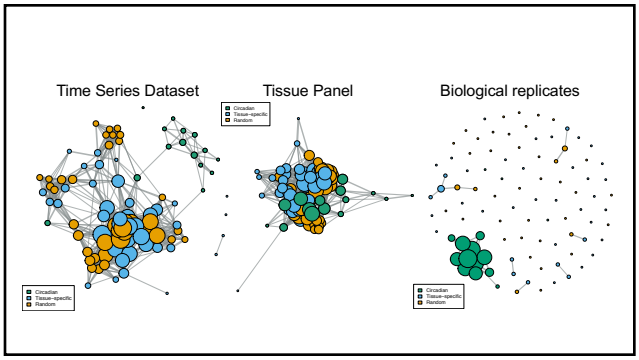


## Tissue specific genes

Time Series Dataset | Tissue Panel | Biological replicates




Time Series Dataset
Tissue Panel
Biological replicates


Time Series Dataset
Tissue Panel
Biological replicates
Circadian
Tissue−specific
Random