STAT 540
Class meeting 05
Monday, January 19, 2015

Dr. Gabriela Cohen Freue

Department of Statistics

Data cleaning and Normalization

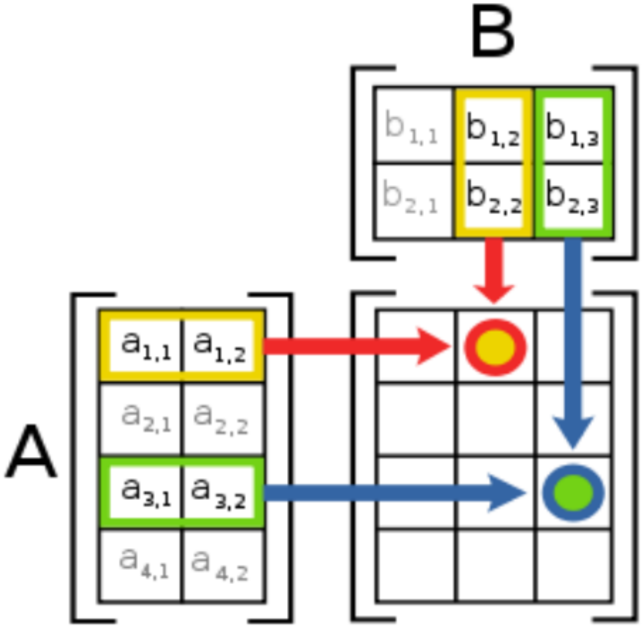
Outline

- Matrix operations
- Data cleaning and QC
- General comments on QC
- Outliers
- Normalization
- Batches
- Filtering
- (Missing values)

Matrix multiplication

- "Dot products" done in a systematic way on two sets of vectors (matrices). In R, use `%*%` (`?matmult`)
- If $\mathbf{A}$ is a $n \times m$ matrix, $\mathbf{AB}$ is defined only if $\mathbf{B}$ has m rows (number of columns in B doesn't matter)

$$\text{Dot product}: a \cdot b = \sum a_i b_i$$



http://en.wikipedia.org/wiki/Matrix_multiplication

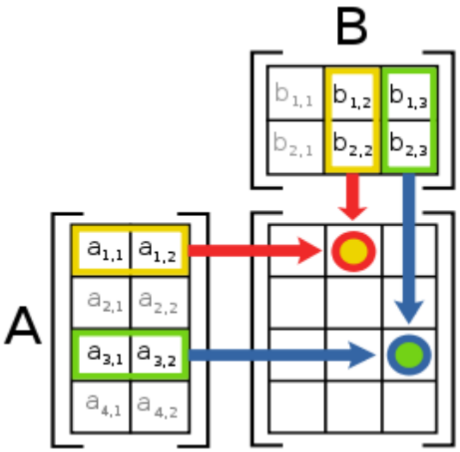
Example

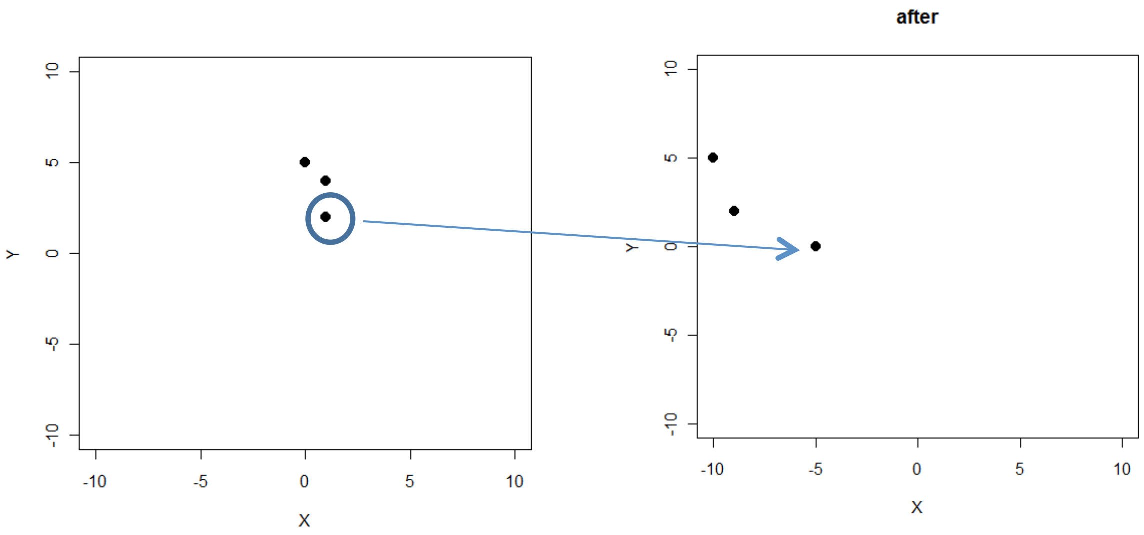
$$\begin{bmatrix}1 & 2\\1 & 4\\0 & 5\end{bmatrix} * \begin{bmatrix}-1 & -2\\-2 & 1\end{bmatrix} = \begin{bmatrix}-5 & 0\\-9 & 2\\-10 & 5\end{bmatrix}$$

(3x2) (2x2) (3x2)

```
m<-rbind(c(1,2),c(1,4),c(0,5))
p<-cbind(c(-1,-2),c(-2,1))
m %*% p

##Note: m %*% p ≠ m * p
```



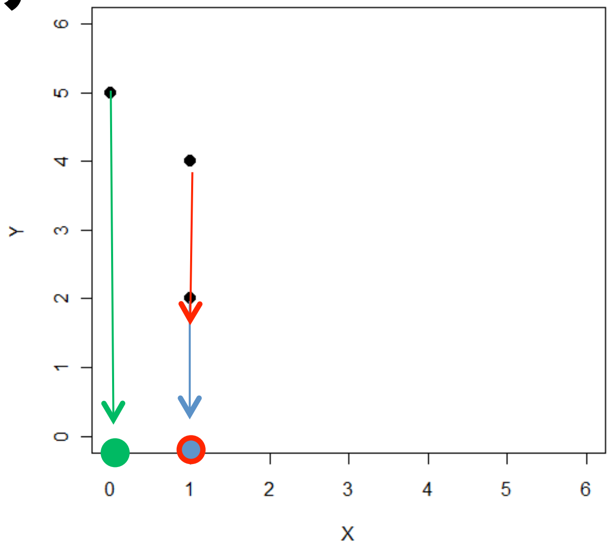


Matrix operations as transformations in space

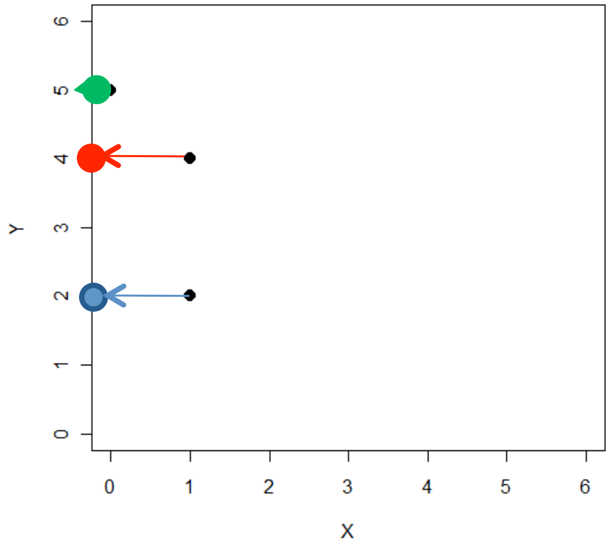
- Multiply by a scalar: moving the points further apart in space (or closer together).
- Multiply by another matrix: e.g., rotation, or projection
- Projection in particular is a fundamental operation: often want to project from original space to a reduced space that is "explanatory".

Examples: Projections

$$\begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 0 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$



$$\begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 0 & 5 \end{bmatrix} \times \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 0 & 4 \\ 0 & 5 \end{bmatrix}$$



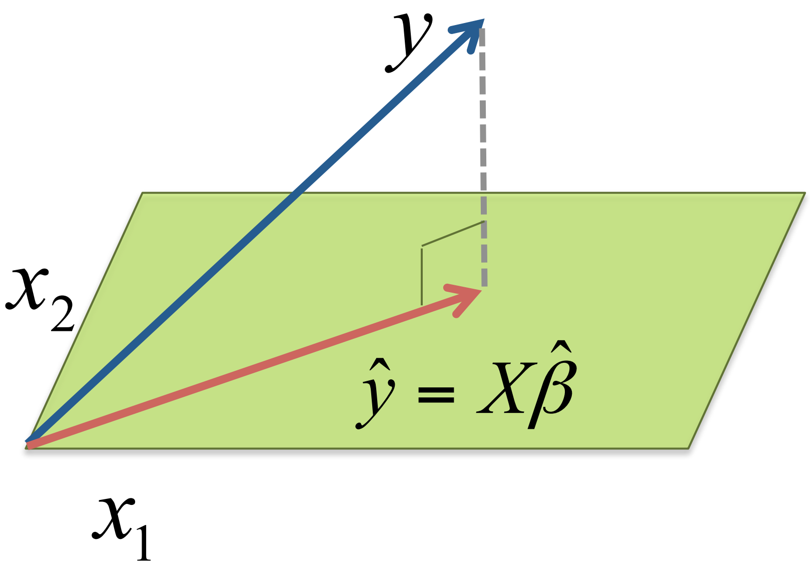
Important: We can project onto any line we like.

```
m<-rbind(c(1,2),c(1,4),c(0,5))
plot(m, xlab="X", ylab="Y", pch=20, ylim=c(0,6), xlim=c(0,6), cex=2)


p=cbind(c(1,0),c(0,0))

points(cm %*% p, pch=19, cex=2)
```

Examples: a wide used projection



$$y = X\beta + \varepsilon$$

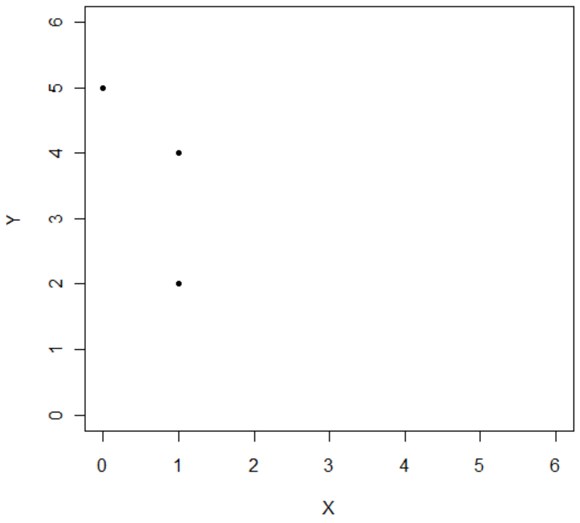
$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \dots & \dots \\ x_{n1} & x_{n2} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

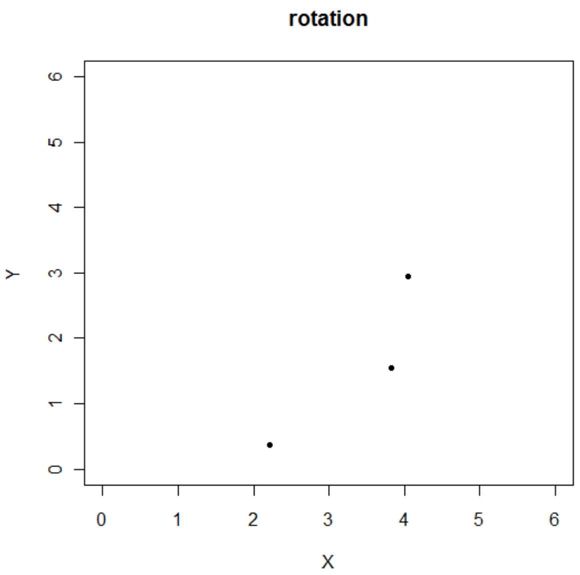
$$\hat{y} = X(X'X)^{-1}X'y$$

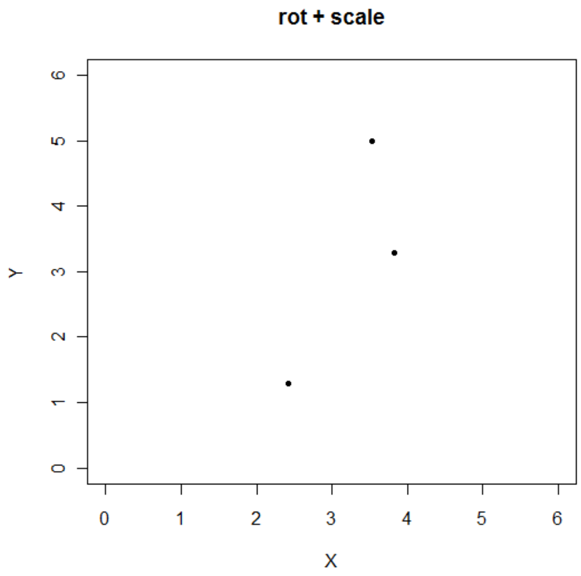
$$P = X(X'X)^{-1}X'$$

The matrix P projects y onto X

Examples: Transformations







$$\begin{bmatrix}1 & 2\\1 & 4\\0 & 5\end{bmatrix} * \begin{bmatrix}0.58 & -0.81\\0.81 & 0.58\end{bmatrix}$$

$$\begin{bmatrix}1 & 2\\1 & 4\\0 & 5\end{bmatrix} * \begin{bmatrix}1 & -0.71\\0.71 & 1\end{bmatrix}$$

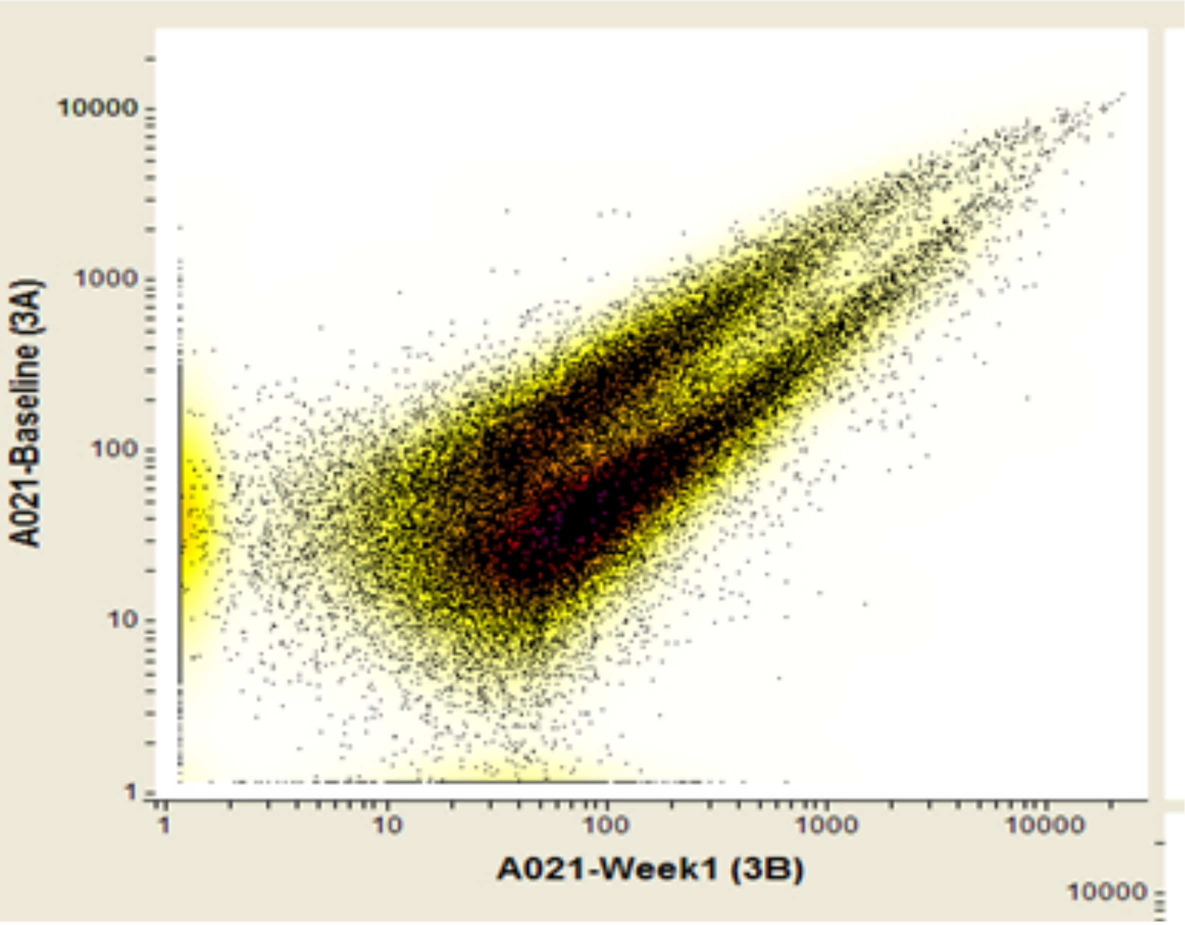
```
a<-pi * 0.3
t0<-cbind(c(cos(a), sin(a) ),c(-sin(a), cos(a)))
p<- m %*% t0
plot(p, xlab="X", ylab="Y", pch=20, ylim=c(0,6), xlim=c(0,6),
main="rotation", cex=2)
```

```
t1<-cbind(c(1,1/sqrt(2)),c(-1/sqrt(2), 1))
t1
n<- m %*% t1
plot(n, xlab="X", ylab="Y", pch=20, ylim=c(0,6), xlim=c(0,6), main="rot
+ scale", cex=2)
```

R suggestion

- Reproduce the examples
- Try multiplying the starting matrix by different scalars, vectors and matrices + plotting
  - What works?
  - What doesn’t?
  - What happens?
- Repeat with a non-2D example (>2 columns)

Data Quality and Sanity Checks



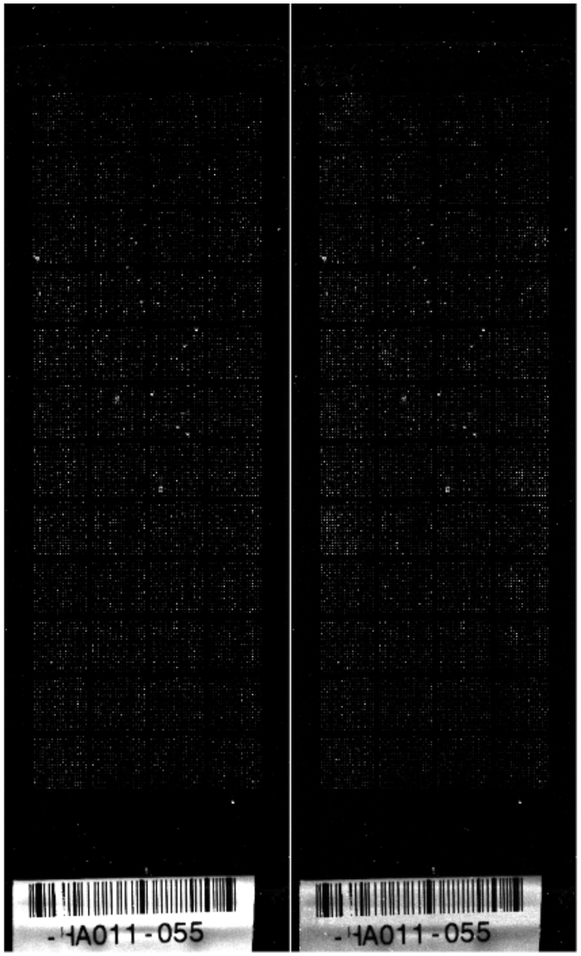
Data Quality:
General types of issues

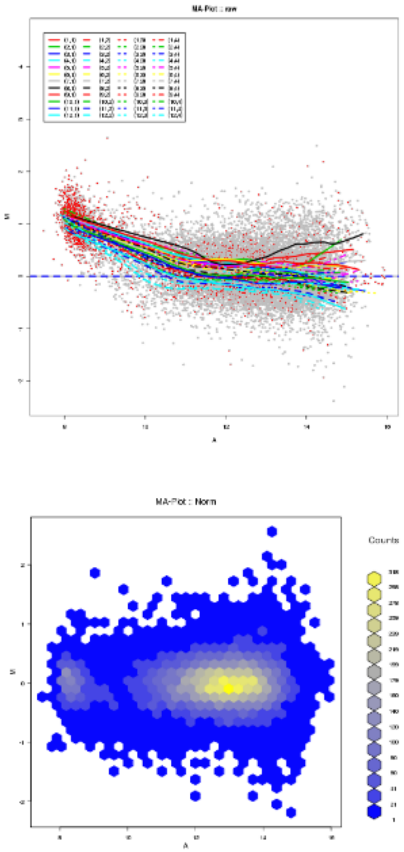
- High technical variability
  - Outliers
- Defects on the assays
- Batch effects (or other systematic trends)

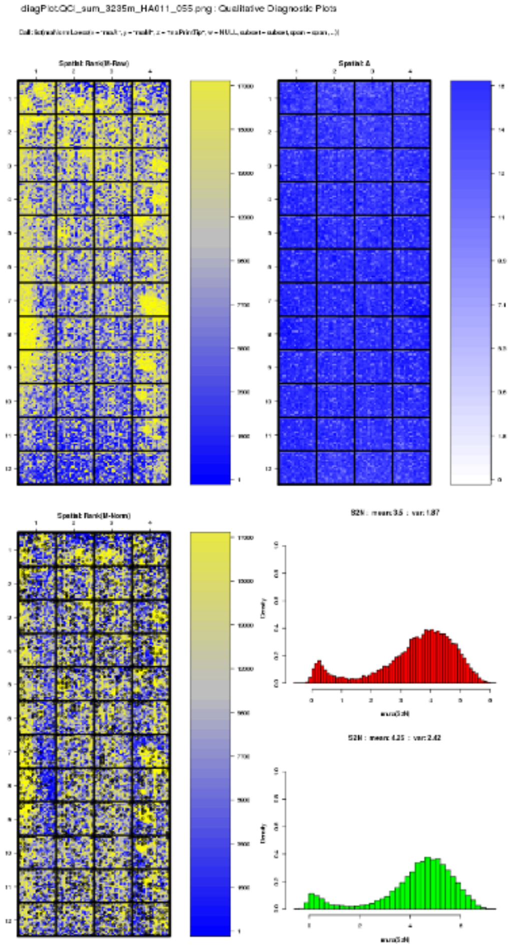
Effects of low quality

- Depends on what you are doing (and the nature of the issue)
  - Some will yield false positives (confounds, "false signals")
  - Others will yield false negatives (by decreasing signal-to-noise)

Assessing quality: Each technology is different

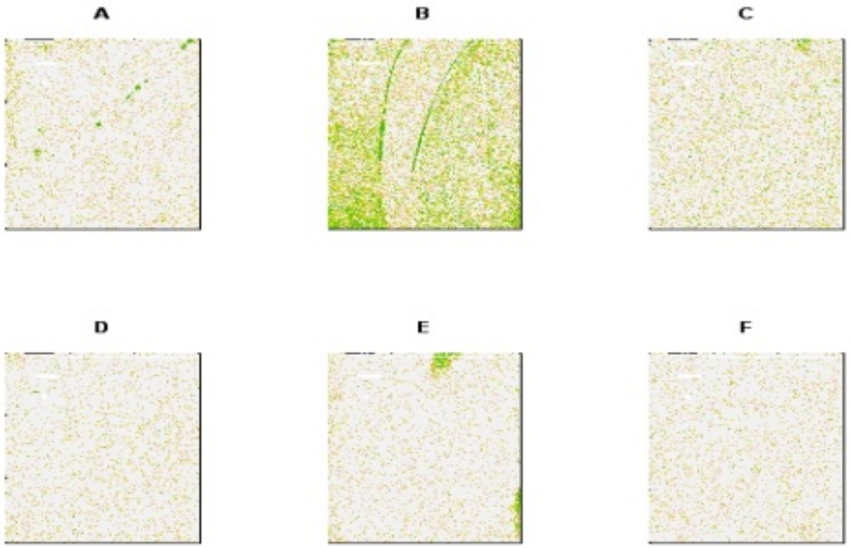


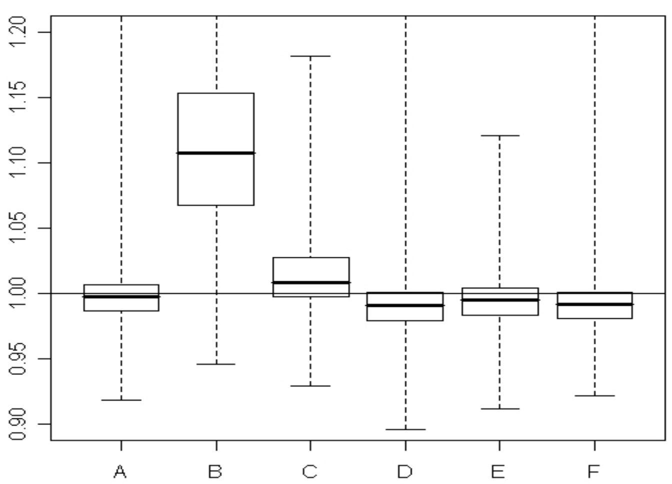


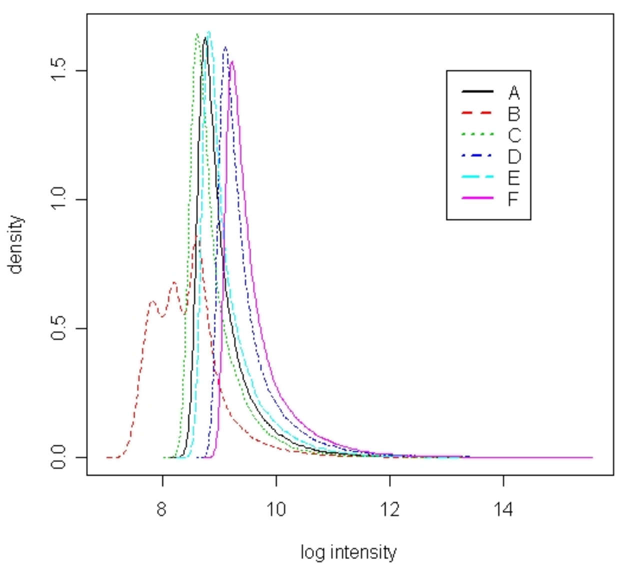


QC package in Bioconductor: arrayQuality

Assessing quality: Each technology is different



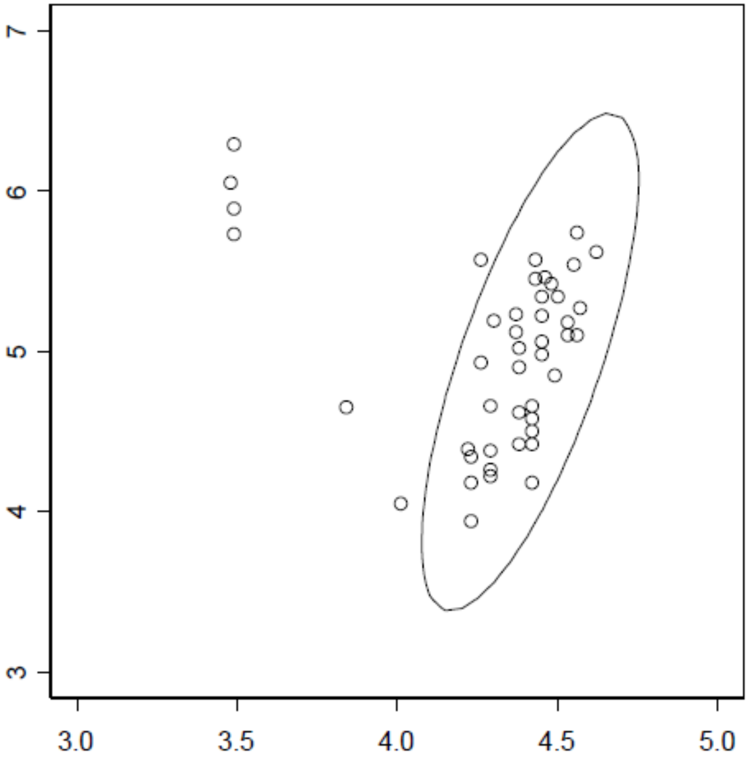

NUSE for jsmHyperdip dataset


Histogram of log2 PM probe intensities

QC packages in Bioconductor: affyPLM
From jsmHyperdim dataset

Consistency and Outliers

- Consistency across QC measure is important
- Outlier: "A sample that deviates significantly from the rest of the samples in its class"



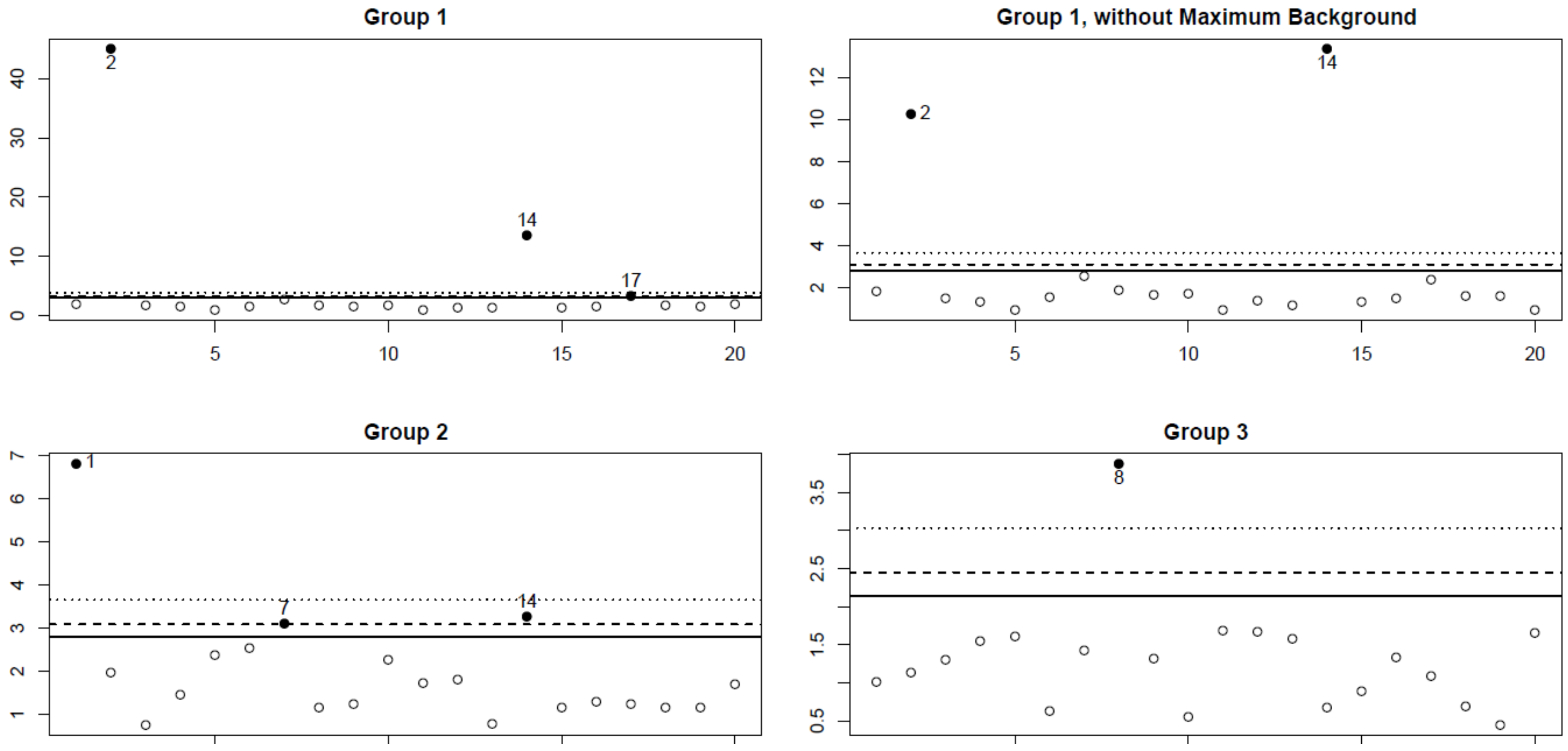
Mahalanobis Distance of each point to a robust location with respect to a robust scatter matrix.

$$d = \sqrt{(x-m)' S^{-1} (x-m)}$$

Shieh and Hung, Statistical Applications in Genetics and Molecular Biology (2009) 8:
en.wiktionary.org/wiki/outlier

Consistency over perfection

- Consistency across QC measure is more important than inspecting each in isolation



From Cohen Freue et al., (2007) *Bioinformatics* **23** 3162–3169

Identifying outlying samples

- Relative vs. absolute quality is important
- Usually, we consider a sample an outlier if:
  - It has "very low" signals
  - "High" background
  - "Low" correlation with most (or all) other samples from the same group
- If a sample is questionable, we might ask: Is there anything in the notebook that would make us suspect it? "*Sample dropped on floor*"
- Outliers can be due to interesting biology too, but you still might not want them in your study.

Entire features (genes) are not *usually* called "outliers"

Exception would be a flawed or "inappropriate" probe design – for example, you would remove the "control spots" from an expression study

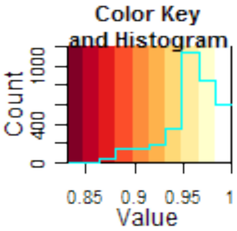
But this is based on prior knowledge, not the data.

Individual points can be “outliers”

- Microarray spot readers will flag “bad spots”
  - Signal saturation
  - Dust, scratches, manufacturing defects
- Other technologies have analogous issues, such as DNA sequences with low quality

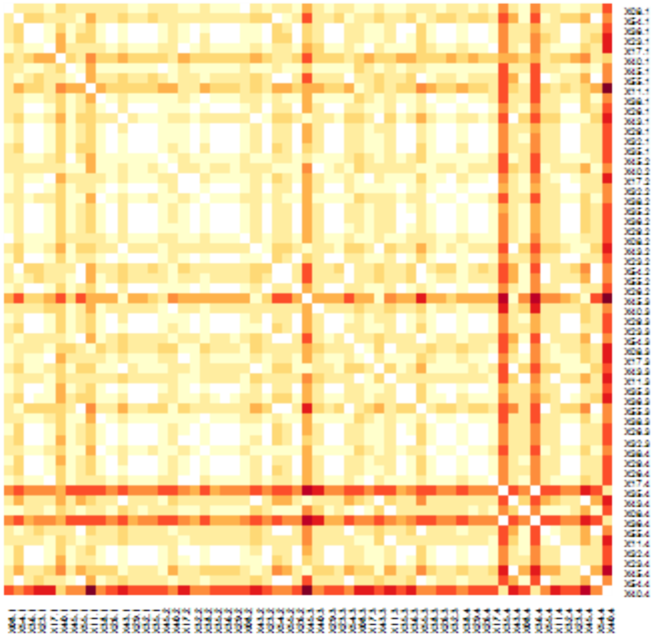
Using heat maps to compare assays

- Pairwise Pearson correlations of entire sample expression profiles
- Expect it to be tighter **within** experimental groups than **across** groups.
- There are no firm guidelines for evaluating this.



- This method loses some information, but is suitable for quite large data sets.
- Also for comparing features (e.g., genes or covariates) if there are not too many.

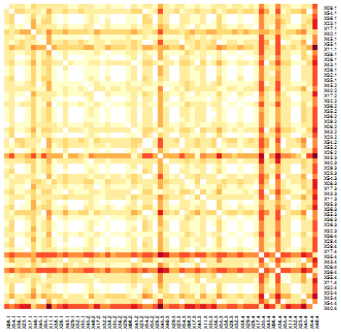
```
library(gplots)
library(RColorBrewer)
cols<-c(rev(brewer.pal(9,"YlOrRd")), "#FFFFFF")
heatmap.2(cor(dat), Rowv=NA, Colv=NA, symm=T,
trace="none", dendrogram="none", col=cols, cexCol=0.5,
cexRow=0.5)
dev.print("heatmap.png", device=png, width=500)
```

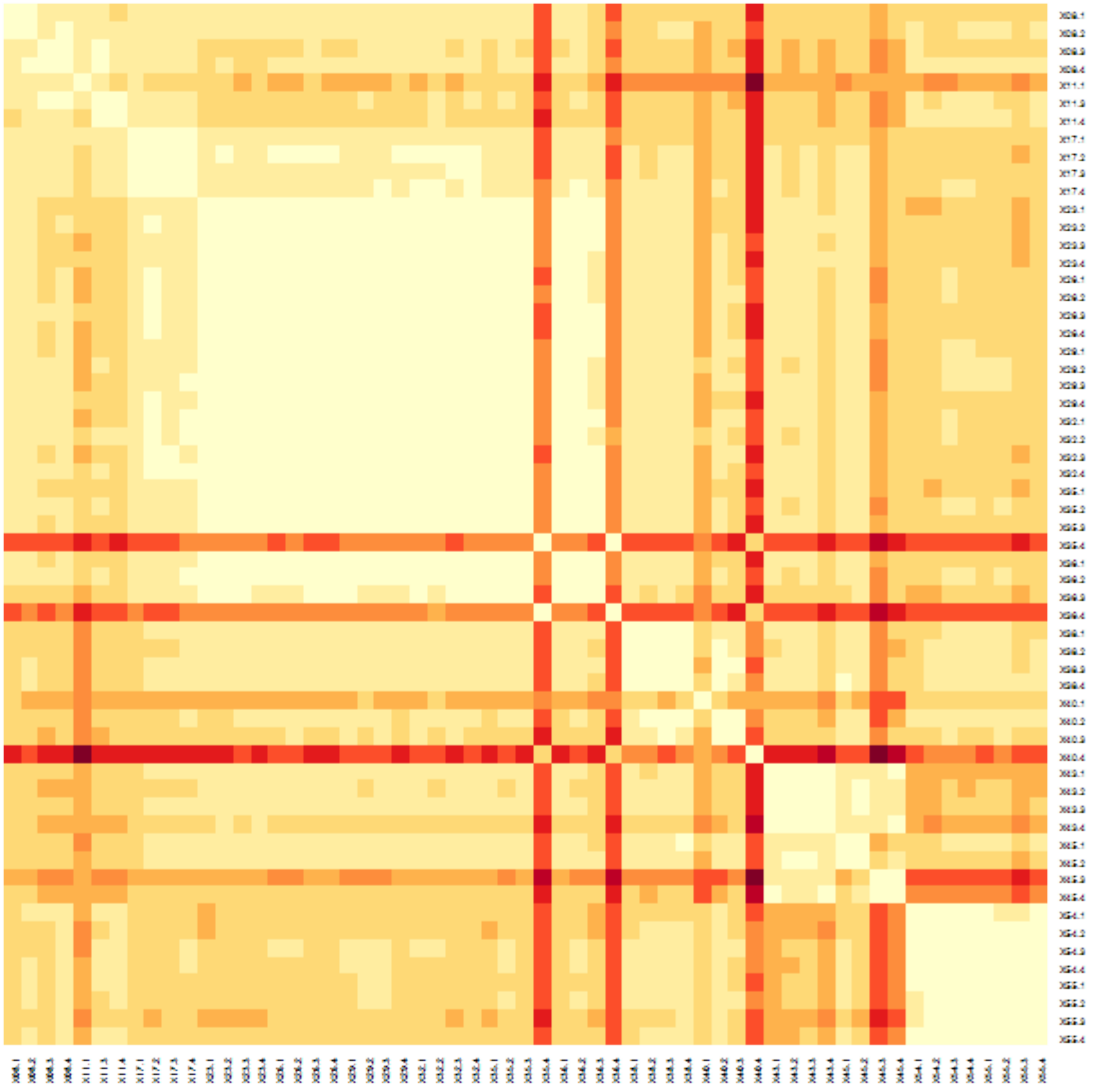


Alternative: matrix2png (http://www.chibi.ubc.ca/matrix2png)

Arrange data by Patient

Original



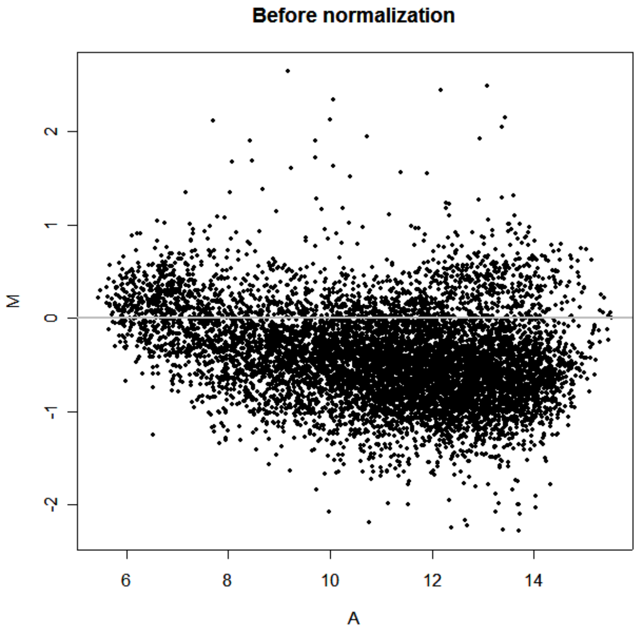


Recommendation

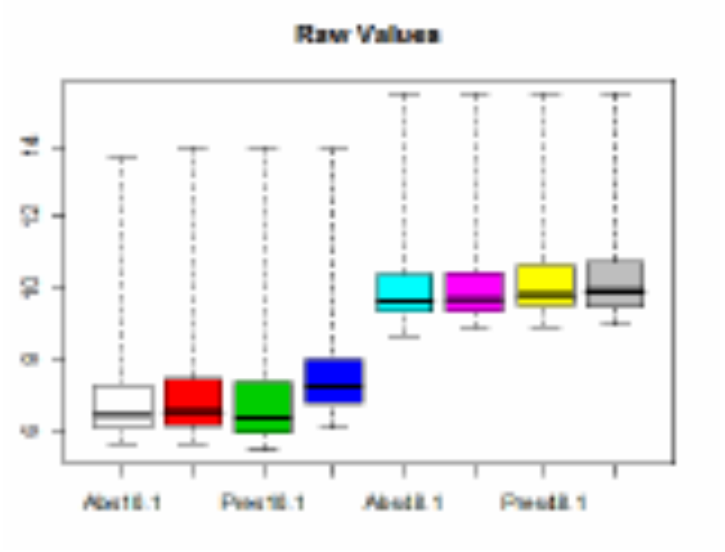
- Make sample sizes large enough that your study won't fail if you have a couple of outliers to remove.
- Determine criteria for identifying outliers and stick with them.
- If outliers are identified, check experimental conditions, physical assays, data pre-processing analysis.
  - In the example of of the "lobster claw", there was no indication of QC problems
  - We later noticed that the "normalization" method used was not appropriate for the Affymetric chips in the study.

Normalization

- Remove as much of the systematic technical variability as we can.
- Example: pipetting errors cause more label to be used in one reaction, causing all signals to be higher.



- Batch effects can still persist.
- When signal-to-noise ratios are wildly different, normalization will not help (enough). Throw out the bad data.



Normalization strategies

- **Invariant set:** identify a set of genes which are known not to change expression levels, and use these as "benchmarks". Difficult in practice because identifying those genes is difficult.
- **Spike-in:** Add a set of "control spots" at a constant concentration. Spike-ins themselves can be subject to systematic errors, and may not reflect biases from all technical stages.
- **Global intensity:** Make an assumption about what the intensity distribution should look like and correct it.

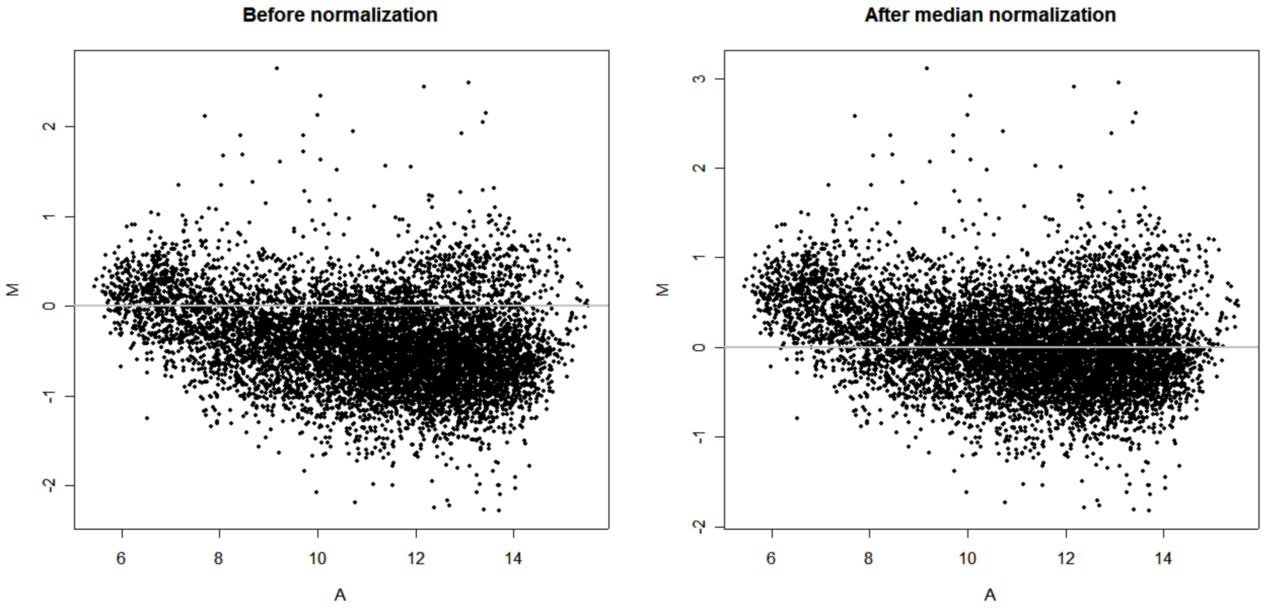
Usual assumptions for “global” normalization schemes

- **Assumption 1:** Most genes don’t show differential expression, so overall shifts in intensity are assumed to be technical in source.
- **Assumption 2:** Similar numbers of genes go “up” or “down”, so generating distribution has the same mean.

(example on next slide)

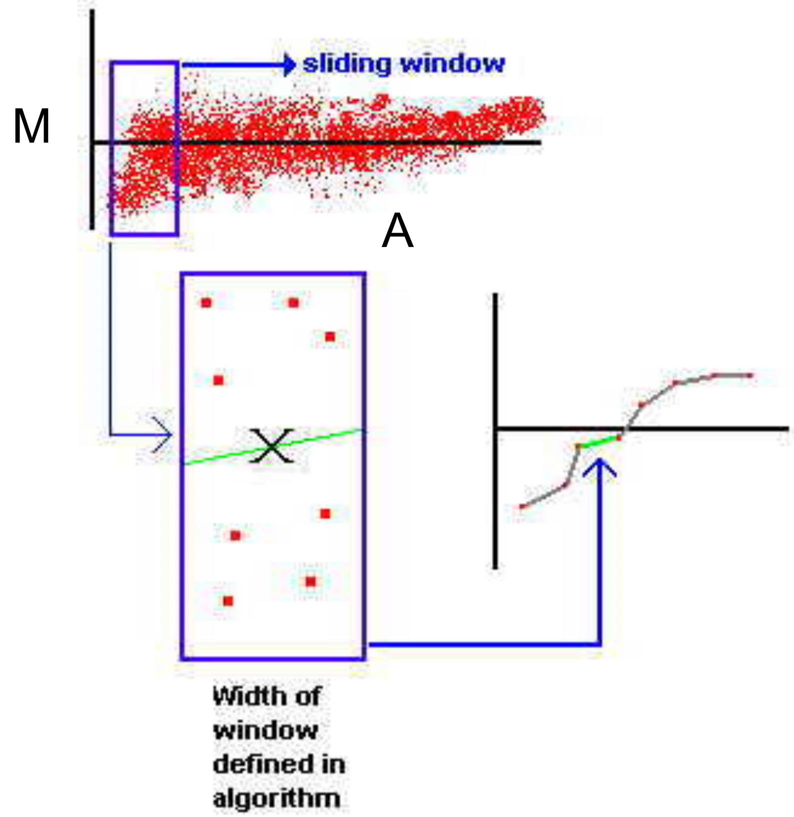
Global normalization

- Median (or mean) set to zero – define a single normalization factor for the array
- Just shifts the distribution so it is centred.
- Problems: nonlinear shapes



LOWESS

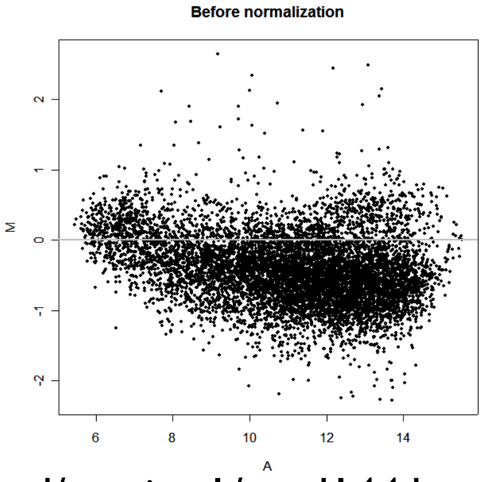
Locally weighted polynomial regression

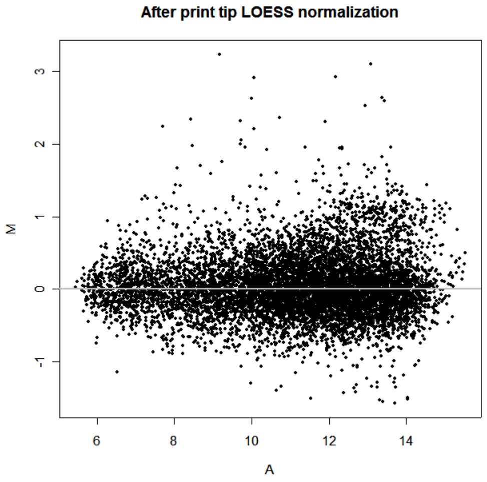


Fit is obtained at each point, considering a fraction of the available data (often 0.3-0.4).

Weighted least square: each data point is assigned a weight proportional to its distance from the current data point.

Finally, $M_i' = M_i - M_{fit}$





http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm

Across-array normalization

- This is a primary issue for one-color arrays, but can be applied to two-color arrays as well (after within-array normalization of the log ratios)
  - Arrays have varying "overall brightness", but no internal normalization is possible

- Methods work either by adjusting each array without reference to other arrays (basically setting the mean)

OR

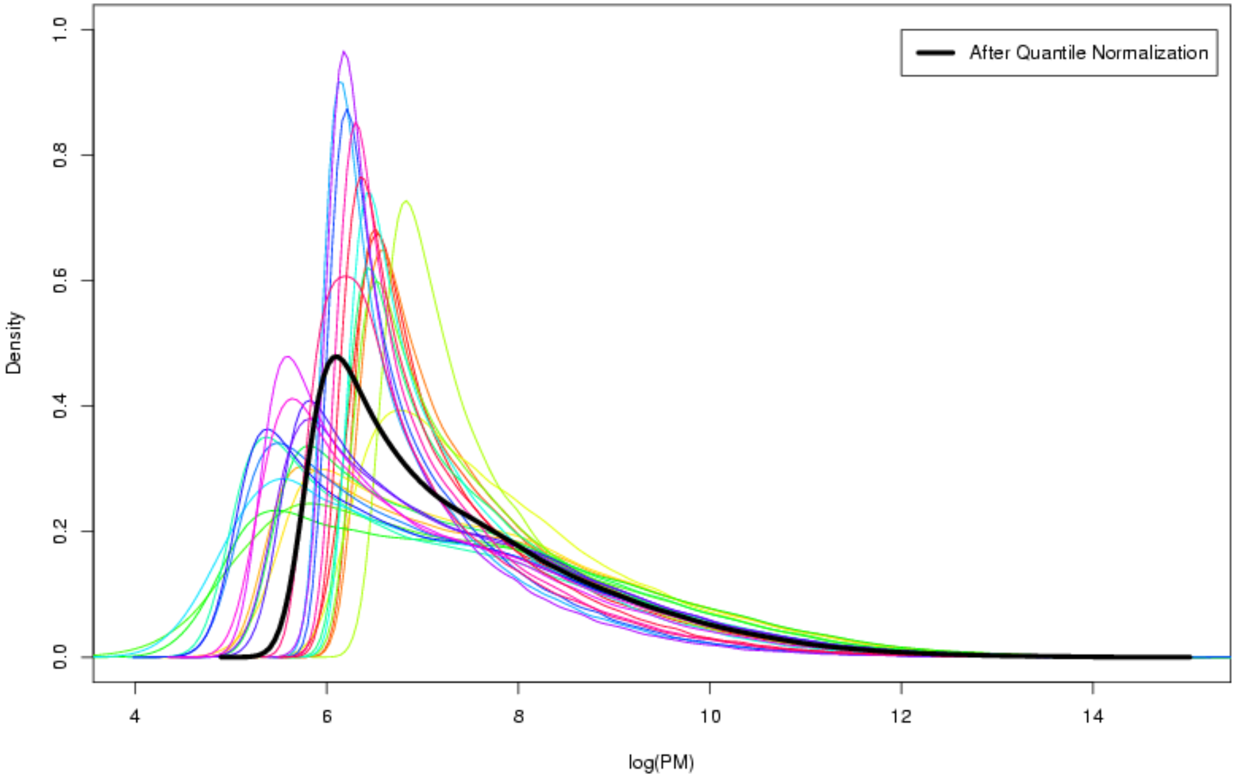
- Using all arrays simultaneously (quantile normalization)

Quantile normalization

- Seems to be accepted as a good, general purpose normalization method.
- It is definitely not limited to microarray data.

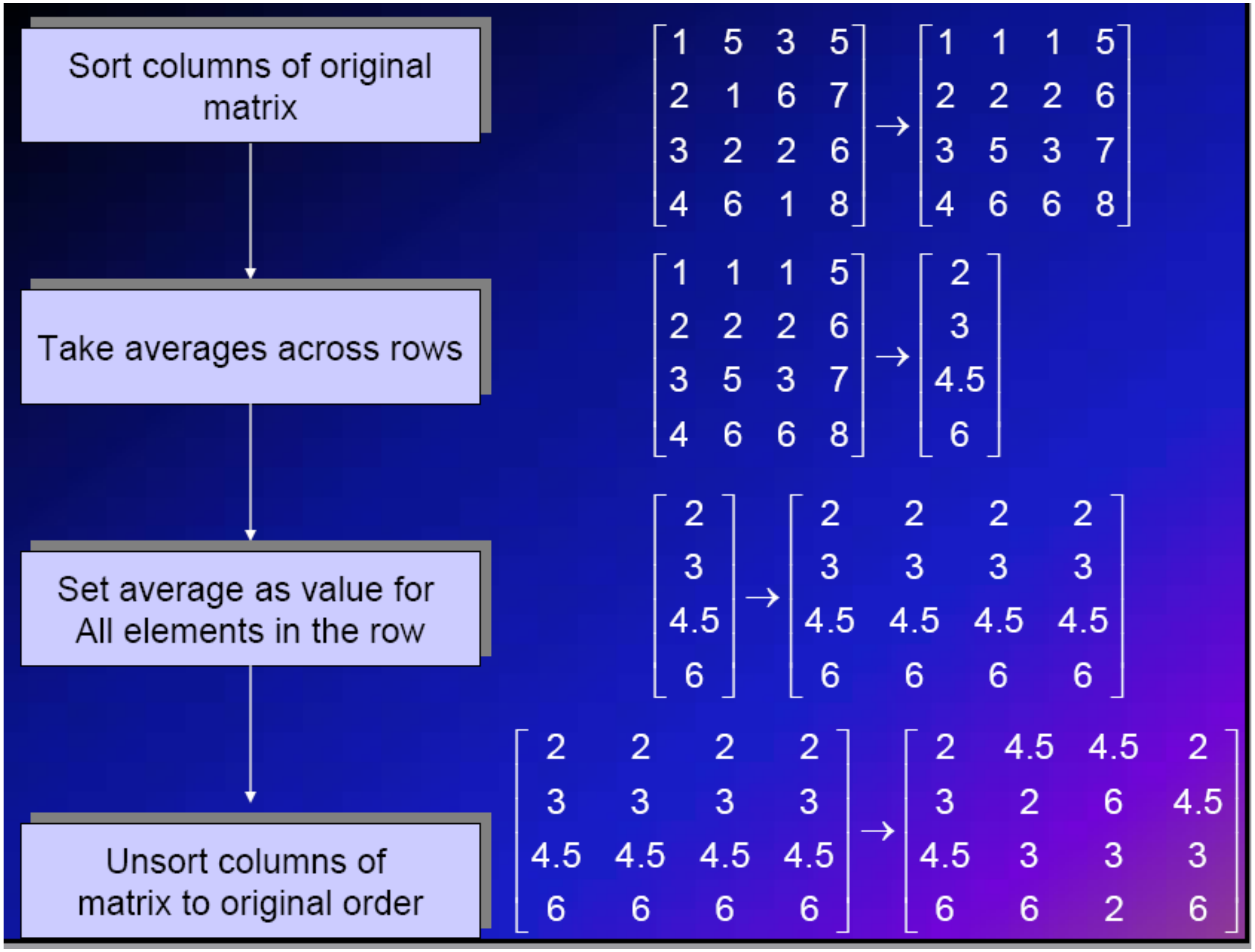
Quantile normalization II

- Force all samples to have the same distribution of values
- Assumption is that they all come from the same distribution
- Adjust both scale AND shape of the intensity distribution.



Bolstad et al. Bioinformatics. 2003 Jan 22;19(2):185-93

Quantile normalization example



Sort columns of original matrix

$$\begin{bmatrix} 1 & 5 & 3 & 5 \\ 2 & 1 & 6 & 7 \\ 3 & 2 & 2 & 6 \\ 4 & 6 & 1 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{bmatrix}$$

Take averages across rows

$$\begin{bmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{bmatrix}$$

Set average as value for All elements in the row

$$\begin{bmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{bmatrix}$$

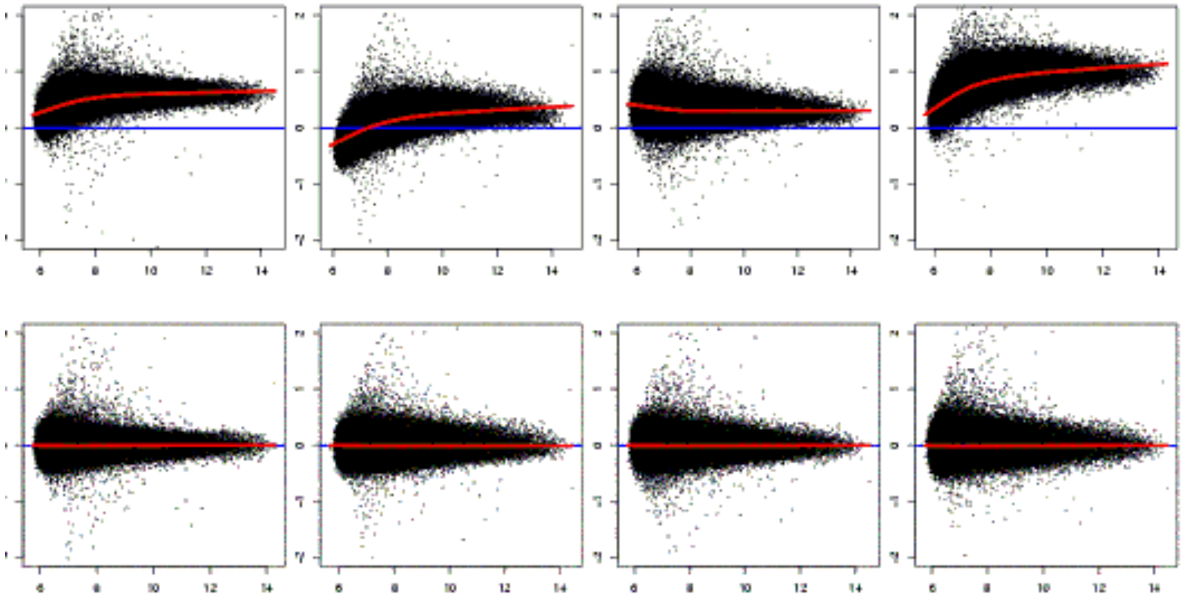
Unsort columns of matrix to original order

$$\begin{bmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 4.5 & 4.5 & 2 \\ 3 & 2 & 6 & 4.5 \\ 4.5 & 3 & 3 & 3 \\ 6 & 6 & 2 & 6 \end{bmatrix}$$

Bolstad

Results

- M-A plots for four *pairs* of Affymetrix arrays.



Before

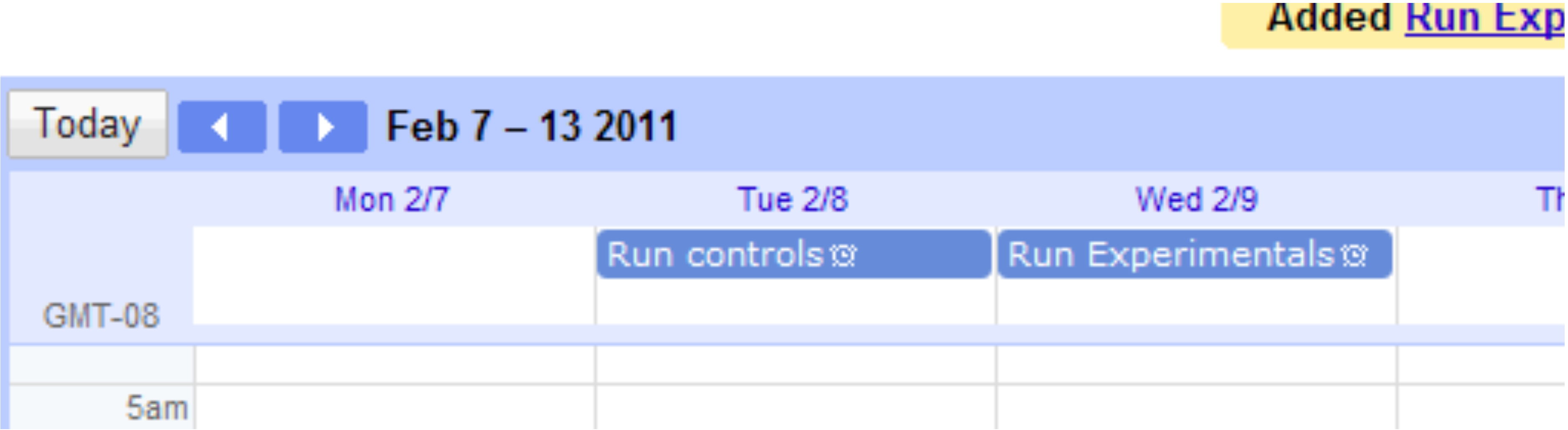
After quantile normalization

Bolstad, Speed

Issues for quantile normalization

- Is assumption that distribution of values should be the same for all samples valid?
  - Most likely violated in the tails, where sparsity of data means it is possible for all samples to be pushed to the same value.
- Missing values are a problem
- If you remove outliers, do it before your final quantile normalization.

Batches



Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study

Leek et al., (2010) Nature Rev. Genetics 11: 733

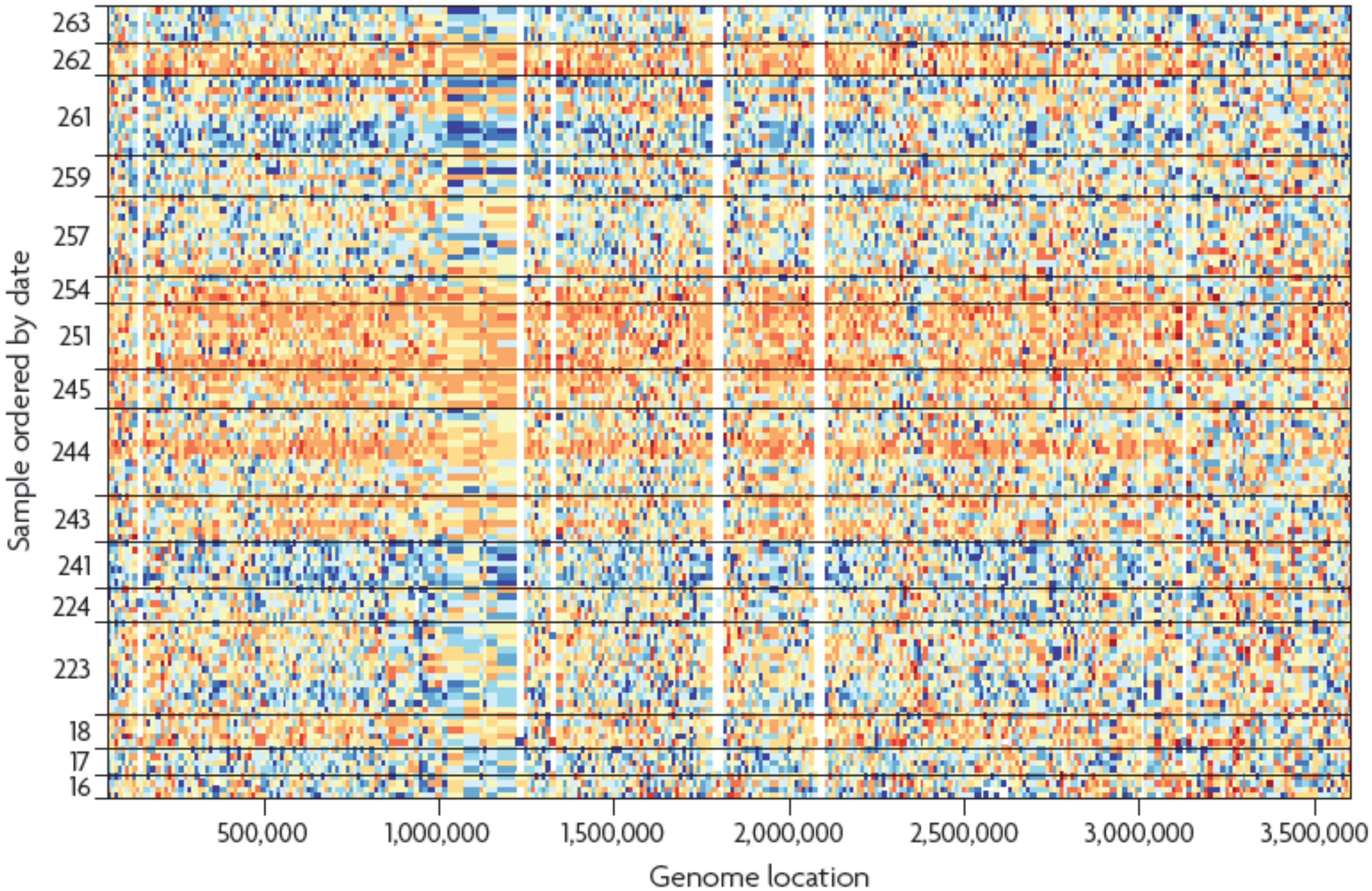


Figure 2 | **Batch effects for second-generation sequencing data from the 1000 Genomes Project.** Each row is a different HapMap sample processed in the same facility with the same platform. See Supplementary information S1 (box) for a description of the data represented here. The samples are ordered by processing date with horizontal lines dividing the different dates. We show a 3.5 Mb region from chromosome 16. Coverage data from each feature were standardized across samples: blue represents three standard deviations below average and orange represents three standard deviations above average. Various batch effects can be observed, and the largest one occurs between days 243 and 251 (the large orange horizontal streak).

Leek et al., (2010) Nature Rev. Genetics 11: 733

Avoiding batch artifacts

- Don’t run in batches (can be hard to avoid)
- Balance or randomize design with respect to batches.
- Avoid (or at least record) obvious potential sources of batch variability, such as a new batch of reagent.
- Run some technical replicates across your batches.

- Batch effects are not always large.
- Consider correcting for batch effects.

Removing batch effects

- Include batch as a covariate in per-gene regression
- SVD – Not an explicit batch correction. Danger of eliminating biological signals
- DWD – Distance weighted discrimination, needs large batches (>20 samples; MatLab or Java impl.) (in Bioinformatics (2004) 20 (1): 105-114).
  - https://genome.unc.edu/pubsup/dwd/
- ComBat – treat batch as a covariate in a linear model. Uses empirical Bayes estimation to allow batch correction in small data sets (R impl.)
  - http://jlab.byu.edu/ComBat

Filtering

- In gene expression studies, it is common to remove genes that are “not expressed” at an early stage of the analysis.
  - Similarly: remove genes that have too many missing values, e.g. >25%
- This makes sense. If you are sure the data are garbage, why analyze it? Hope is to increase power.
- Deciding what is useless is not so easy.
- Bourgon et al. PNAS, 107(21):9546-9551, 2010.

Filters must be unsupervised

- Common supervised method: "At least N 'positive detection calls' in one condition"
- This could bias towards retention of genes that have differences between conditions.
- More generally, filtering method and statistical analysis must be independent.
  - Bourgon et al. show that even unsupervised filters can still cause distortion of Type I error rates

Coming up

- We have a cleaned, normalized data matrix.

- How do we find genes with interesting patterns with respect to our experimental design?