V2 – missing values + batch effect correction

- BEclear / applies latent factor model to predict missing values and to remove batch effects
  - DNA microarray
  - DNA methylation
- Functional Normalization
- gPCA
- Review of Probability Theory Basics

Process *S. aureus* microarray data – part II

StaphyType Test Report

Operator	
Sample ID	2192119
Experiment ID	2192119 - {4083AD2C-7D42-4FB9-82D5-E50CC0FD6206}
Date of Result	Thu Apr 14 10:46:01 2011
Assay Name	StaphyType
Assay ID	10248
Well Position	01 (01-A)
Software Version	2009-07-09
Device	04a0022

Internal Controls

Data Quality	passed

Genetic markers for S. aureus / MRSA / PVL

Taxonomy	Species Marker (*S.aureus*) positive
MRSA (mecA)	positive
PVL	negative

Resistance Genotype

Hybridisation (Gene)	Result	Expected Resistance
mecA	positive	Methicillin, Oxacillin and all Beta-Lactams, defining MRSA
blaZ	negative	Beta-Laktamase
ermA	positive	Macrolide, Lincosamide, Streptogramin
ermB	negative	Macrolide, Lincosamide, Streptogramin
ermC	negative	Macrolide, Lincosamide, Streptogramin
linA	negative	Lincosamides

	11	46	10	33	28
MRSA (mecA)	0	0	0	0	0
PVL	0	0	0	0	0
23S-rRNA	1	1	1	1	1
gapA	1	1	1	1	1
katA	1	1	1	1	1
coA	1	0	1	1	1
Protein A	1	1	1	1	1
sbi	1	1	1	1	1
nuc	1	1	1	1	1
fnbA	1	1	1	1	1
vraS	1	1	1	1	1
sarA	1	1	1	1	1
eno	1	1	1	1	1
saeS	1	1	1	1	1
mecA	0	0	0	0	0
blaZ	0	1	0	0	0
blaI	0	1	0	0	0
blaR	0	1	0	0	0
ermA	0	0	0	0	0
ermB	0	0	0	0	0
ermC	0	0	0	0	0
linA	0	0	0	0	0

Compute Euclidian distance between samples

$$\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

Ambiguous values

In the *S. aureus* genotyping test report, individual markers can be "**positive**" or "**negative**" and also "**ambiguous**".

Such ambiguous classifications can be caused by:

- poor sample quality, or
- poor signal quality, or
- by the presence of plasmids in low copy numbers.

www.alere-technologies.com

Re-Assign ambiguous values in DNA microarray

Task – predict ambiguous values.

Simple idea: **baseline prediction** using average values

total average

$$\mu = \frac{1}{N} \sum_{(i,j)\in\Omega} D_{ij}$$

sample average

$$b_i = \frac{1}{N_i} \sum_{(i,j)\in\Omega} D_{ij} - \mu$$

gene average

$$b_j = \frac{1}{N_j} \sum_{(i,j)\in\Omega} D_{ij} - \mu$$

$$b_{prediction} = \frac{1}{3}(\mu + b_i + b_j)$$

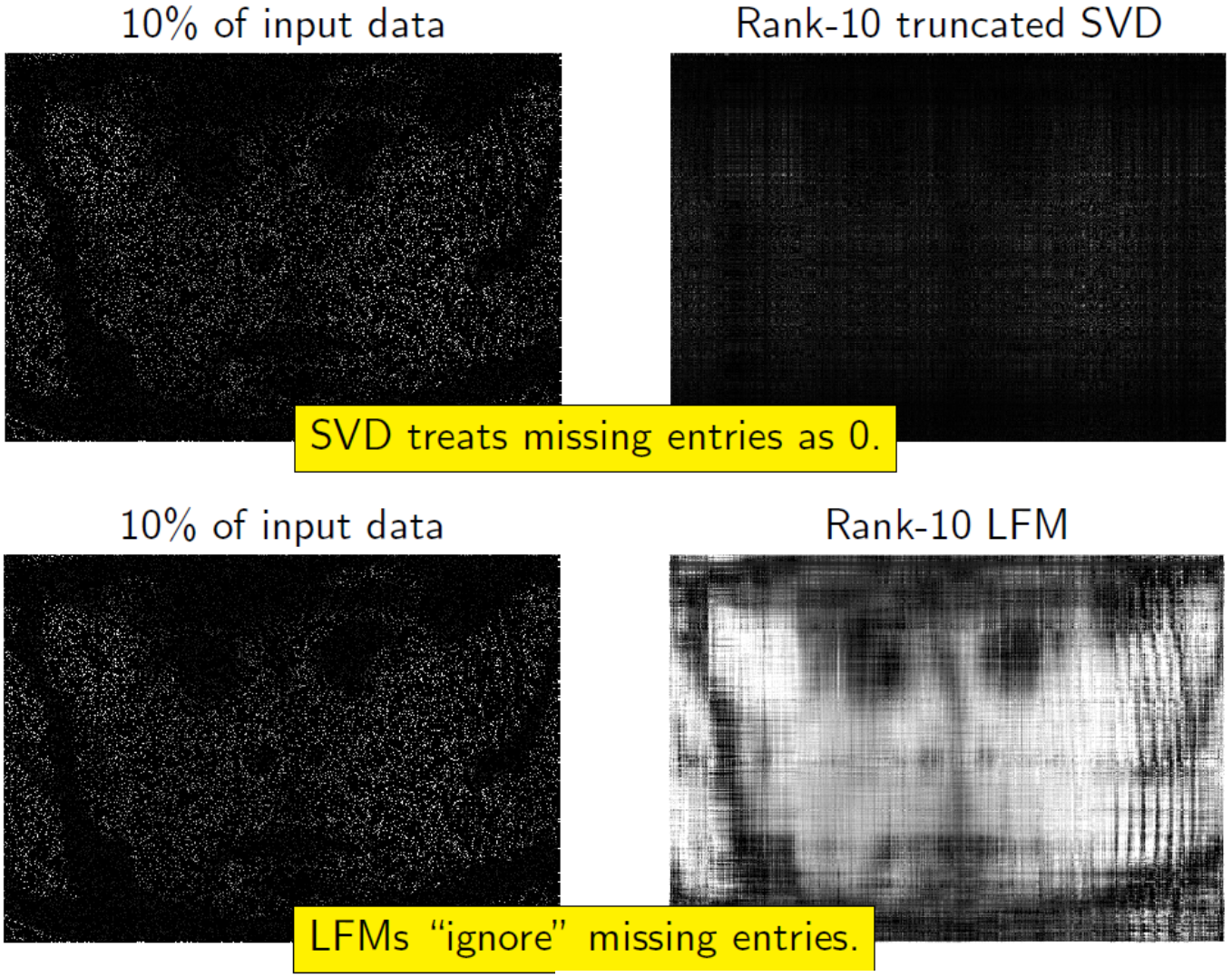
replace small fraction of known values by (thresholded) baseline values -> ~85% correct predictions

But better results are obtained with:

Latent Factor Model (LFM)

~95% correct predictions

Latent Factor Models in image reconstruction



DMM course by R. Gemulla and P. Miettinen

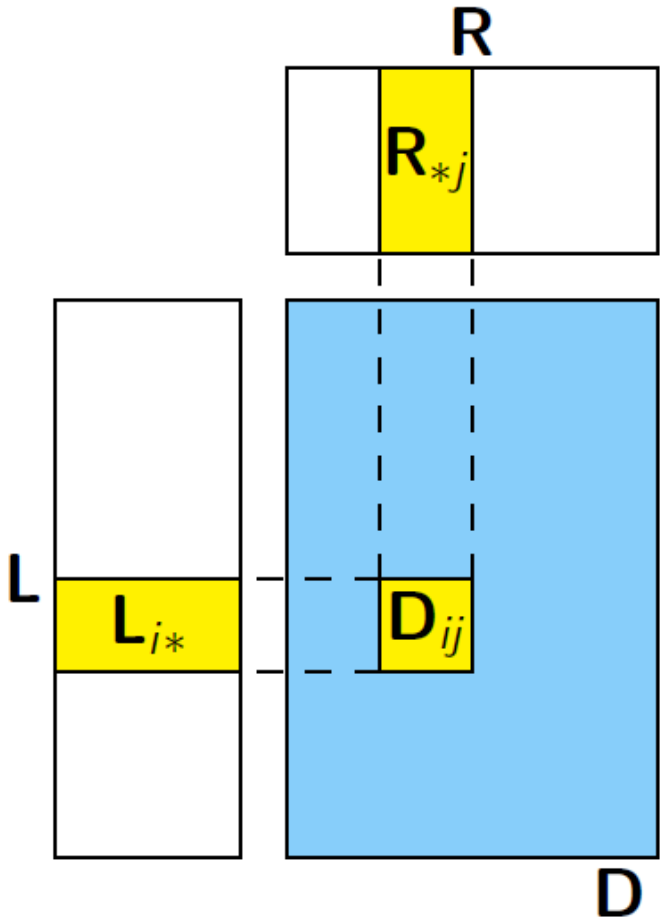
LFM: mathematical background

$$L = \sum_{(i,j)\in\Omega} \left(D_{ij} - [LR]_{ij}\right)^2 + \lambda\left(\|L\|_F^2 + \|R\|_F^2\right)$$

L ($m \times r$) and R ($r \times n$) are sought matrices of rank r

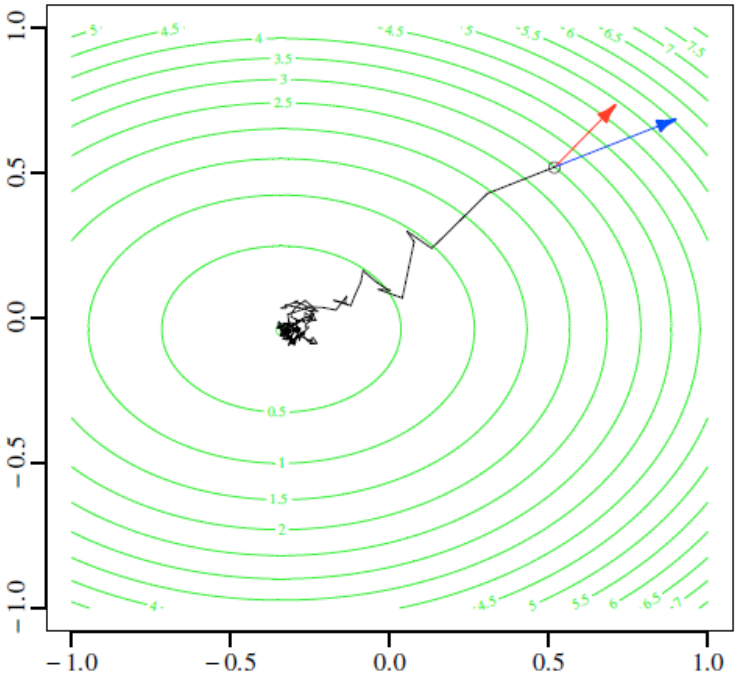
D ($m \times n$) is a given matrix

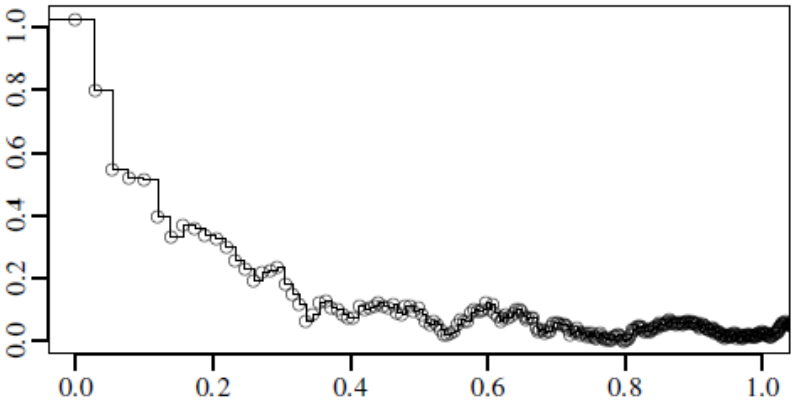
Idea: construct L and R from known data; use them to reconstruct the missing data.



LFM: Stochastic Gradient Descent

- Pick a random entry;
- Compute approximate gradient;
- Update parameters *L* and *R*
- Repeat *N* times.





We implemented LFM-completion of missing values in the Bioconductor package **BEclear**.

Akulenko, R., Merl, M., Helms, V. (2016) PLoS ONE, 11:e0159921

MA assignment to clonal complexes + LFM predictions confirmed by WGS

154 *S. aureus* isolates (182 target genes) from Germany-vs-Africa study

Table 1A

Result Category		Result caused by		Functional Category of genes: Identification	Regulation	Resistance	Virulence	Total	% Total
Concordant	**Positive**	Microarray and WGS (*de novo*)		829	990	1,060	8,495	11,374	40.6%
n=27,119	**Negative**	Microarray and WGS (*de novo*)		0	1,159	8,100	6,486	15,745	56.2%
(96.8 %)									
Discrepant	**False Positive**	Microarray	Mishybridizations	0	78	21	103	202	0.7%
n=909 (3.2 %)		LFM	Misprediction	0	17	2	9	28	0.1%
	False Negative	Microarray	Polymorphisms	0	3	14	140	157	0.6%
		LFM	Misprediction	0	0	0	5	5	< 0.1%
		WGS	Assembly error	88	42	16	164	310	1.1%
			Cropped contig	1	12	15	28	56	0.2%
			Not sequenced or aberrant allele	6	9	8	100	123	0.4%
	Unknown			0	0	4	24	28	0.1%
		Total number of typing results		924	2,310	9,235	15,554	28,028	100%

Very few errors due to LFM mis-predictions.

Strauss et al. J Clin Microbiol (2016)

Batch effects

Batch effects are:

Subgroups of measurements that show qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.

For a **microarray experiment**, batch effects may occur due to:

- Chip type/lot/platform
- Different laboratories may have different standard operating procedures
- Sample/preservation protocols (procedures of drawing biological samples may vary from center to center and over time within center, relevant to retrospective studies)
- Storage/shipment conditions
- RNA isolation (different laboratories may use different extraction procedures or kits, and different lots of reagents may perform differently)
- cRNA/cDNA synthesis
- Amplification/labeling/hybridization protocol (different reagents or lots may be used)
- Wash conditions (temperature, ionic strength, fluidics modules/stations; cleaning schedules)
- Ambient conditions during sample preparation/handling, such as room temperature and ozone levels
- Scanner (types, settings, calibration drift over prolonged studies; scheduled maintenance)

Luo et al. Pharmacogenomics J. (2010) 10: 278–291.

Global methods to correct batch effects

Mean-centering : after the transformation, the mean of each feature across all the samples within each batch is set to zero.

Standardization: Beyond mean-centering, this approach normalizes the standard deviation of all features across samples within each batch to unity.

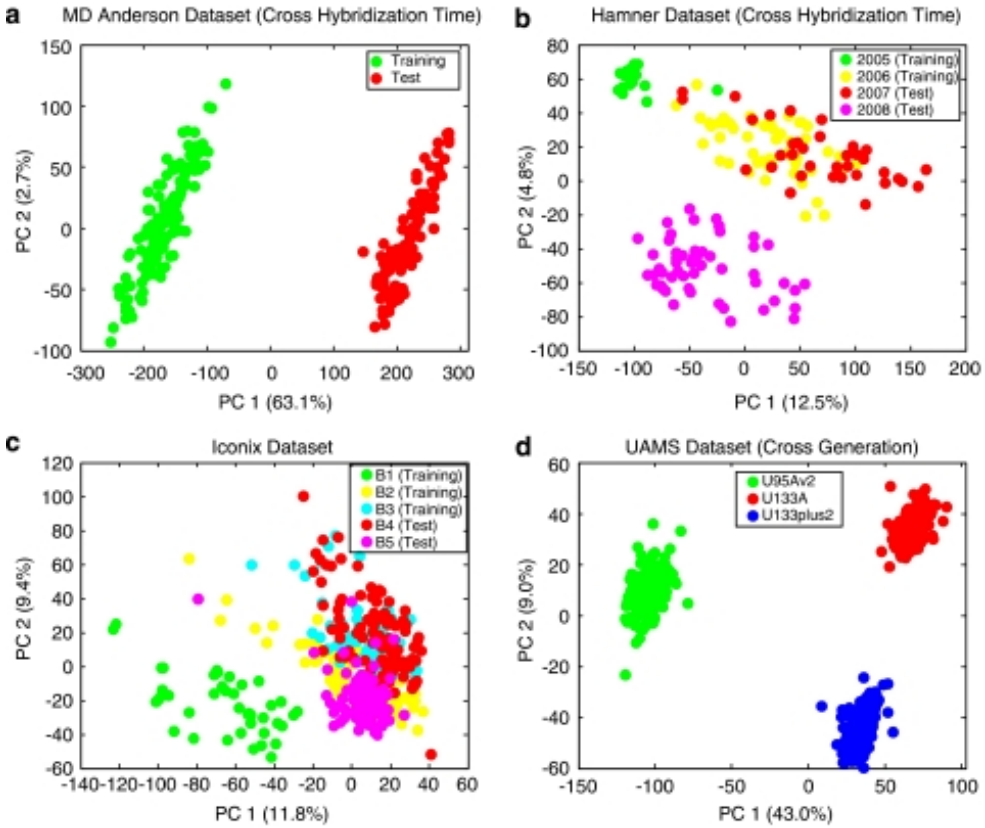
Ratio-based: All samples are scaled by a reference array.

This can be the average of multiple reference arrays, such as the measurement of universal human reference RNA samples for clinical data and vehicle control samples for toxicogenomics data.

Such global **normalization** methods do not remove batch effects if these affect specific subsets of genes so that different genes are affected in different ways.

Luo et al. Pharmacogenomics J. (2010) 10: 278–291.

Batch effects in public MA data sets



Score plot of the first two principal components. Batches (groups) are indicated by colors.

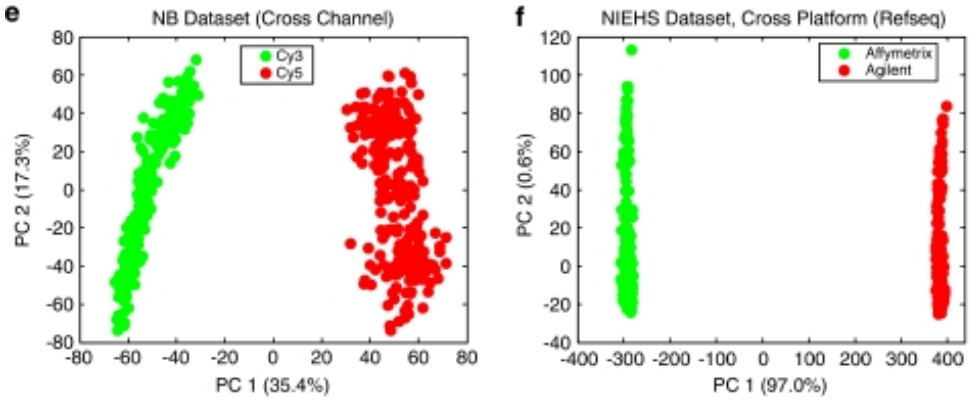
(a) MD Anderson breast cancer data set.

(b) Hamner lung carcinogen data set (two batches in training set hybridized in 2005 and 2006, and two batches in test set hybridized in 2007 and 2008)

(d) UAMS multiple myeloma data set (the three batches represent three generations of Affymetrix chips on *Homo Sapiens*).

Luo et al. Pharmacogenomics J. (2010) 10: 278–291.

Batch effects in public MA data sets

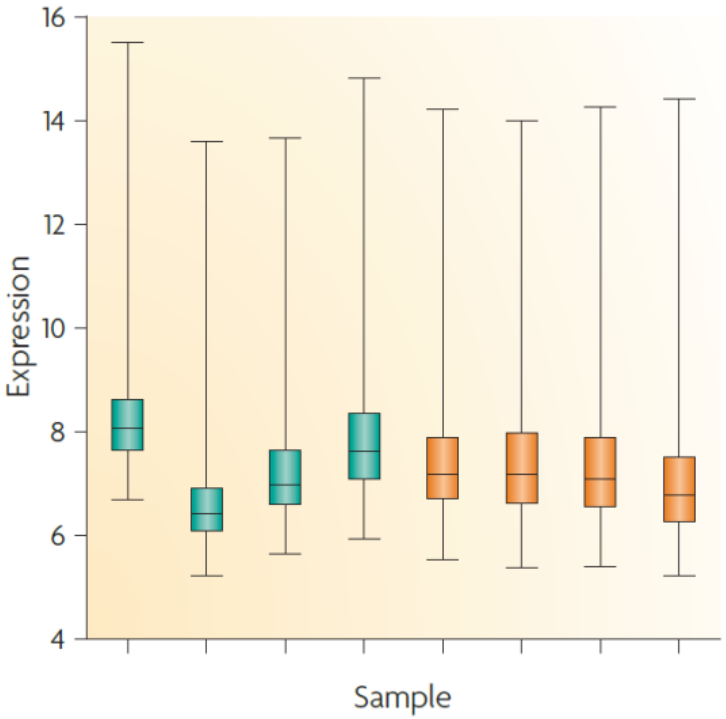


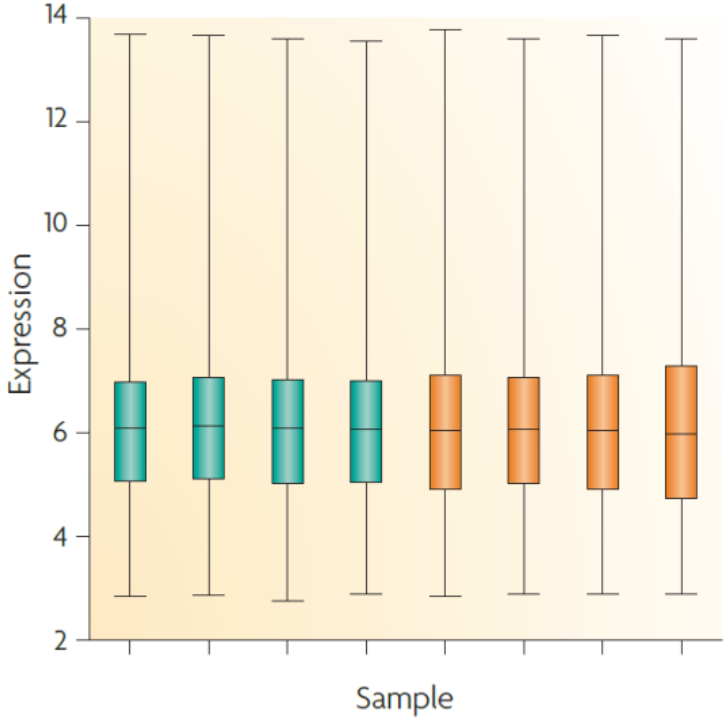
(**e**) Cologne neuroblastoma data set (the two batches represent the two channels of Agilent arrays).

(**f**) NIEHS data set (cross-platform: the two groups represent Affymetrix and Agilent microarray platforms.

Luo et al. Pharmacogenomics J. (2010) 10: 278–291.

Example: bladder cancer microarray data





Raw data for normal samples taken from a bladder cancer microarray data set (Affymetrix chip).

Green and orange represent two different processing dates. Box plot of raw gene expression data ($\log_2$ values)

Same data after processing with RMA, a widely used preprocessing algorithm for Affymetrix data.

RMA applies quantile normalization — a technique that forces the distribution of the raw signal intensities from the microarray data to be the same in all samples.

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

Quantile normalisation: adjusts multiple distributions

Given: 3 measurements of 4 variables A – D.

Aim: all measurements should get identical distributions of values

Original data

A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

→

Determine in each column the rank of each value

A	iv	iii	i
B	i	i	ii
C	ii	iii	iii
D	iii	ii	iv

Sort columns by magnitude

A	2	1	3
B	3	2	4
C	4	4	6
D	5	4	8

→

Compute mean of each row

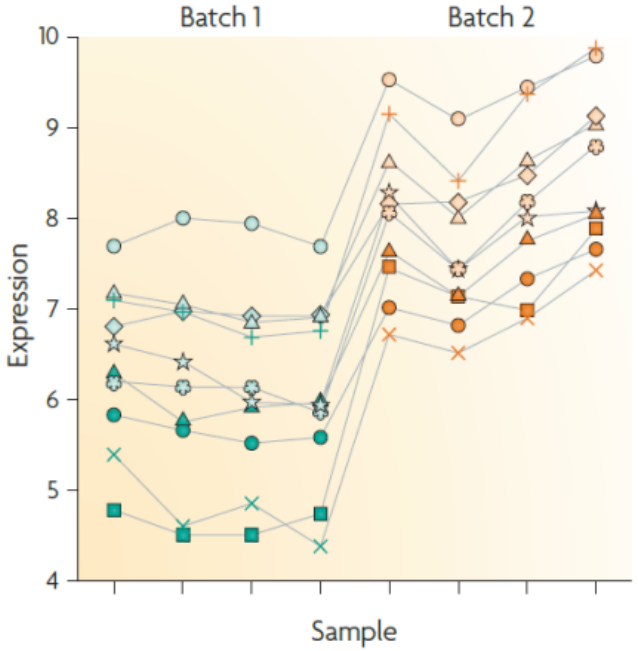
A	2	Rank i
B	3	Rank ii
C	4.67	Rank iii
D	5.67	Rank iv

A	5.67	4.67	2
B	2	2	3
C	3	4.67	4.67
D	4.67	3	5.67

Replace original values by mean values according to the rank of the data field.

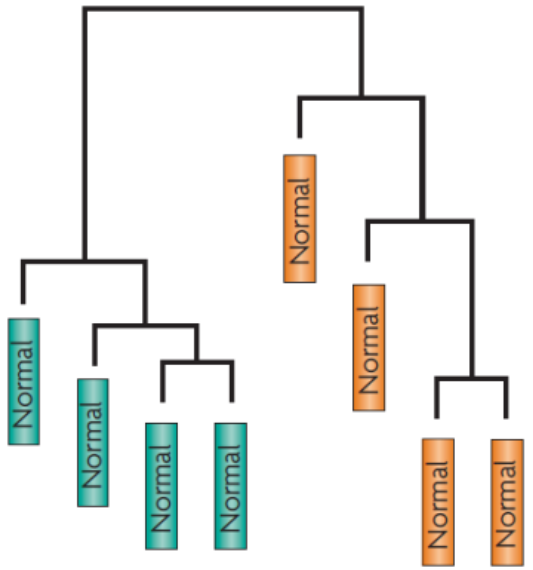
Now all columns contain the same values (except of duplicates) so that they can be easily compared.

Example: same bladder cancer microarray data



Ten particular genes that are susceptible to batch effects even after RMA normalization.

Hundreds of other genes show similar behavior but, for clarity, are not shown.



Clustering of samples after normalization.

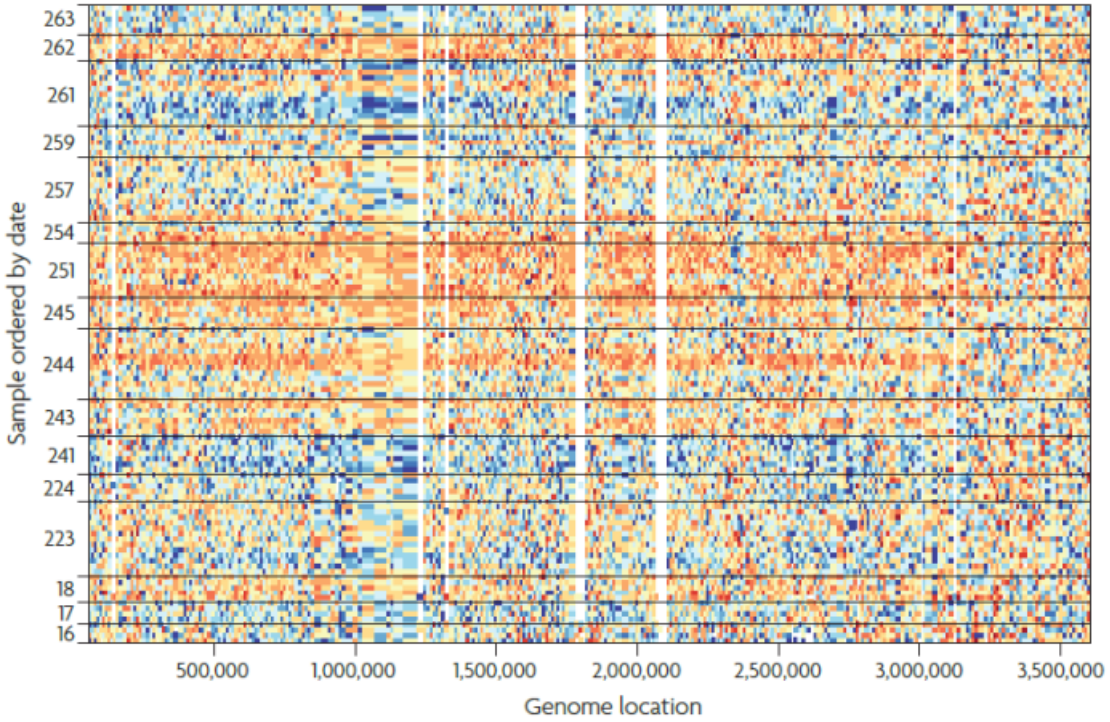
The samples perfectly cluster by processing date.

→ clear evidence of **batch effect**

Processing date is likely a "**surrogate**" for other variations (laboratory temperature, quality of reagents etc.).

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

Example: sequencing data from 1000 Genomes project



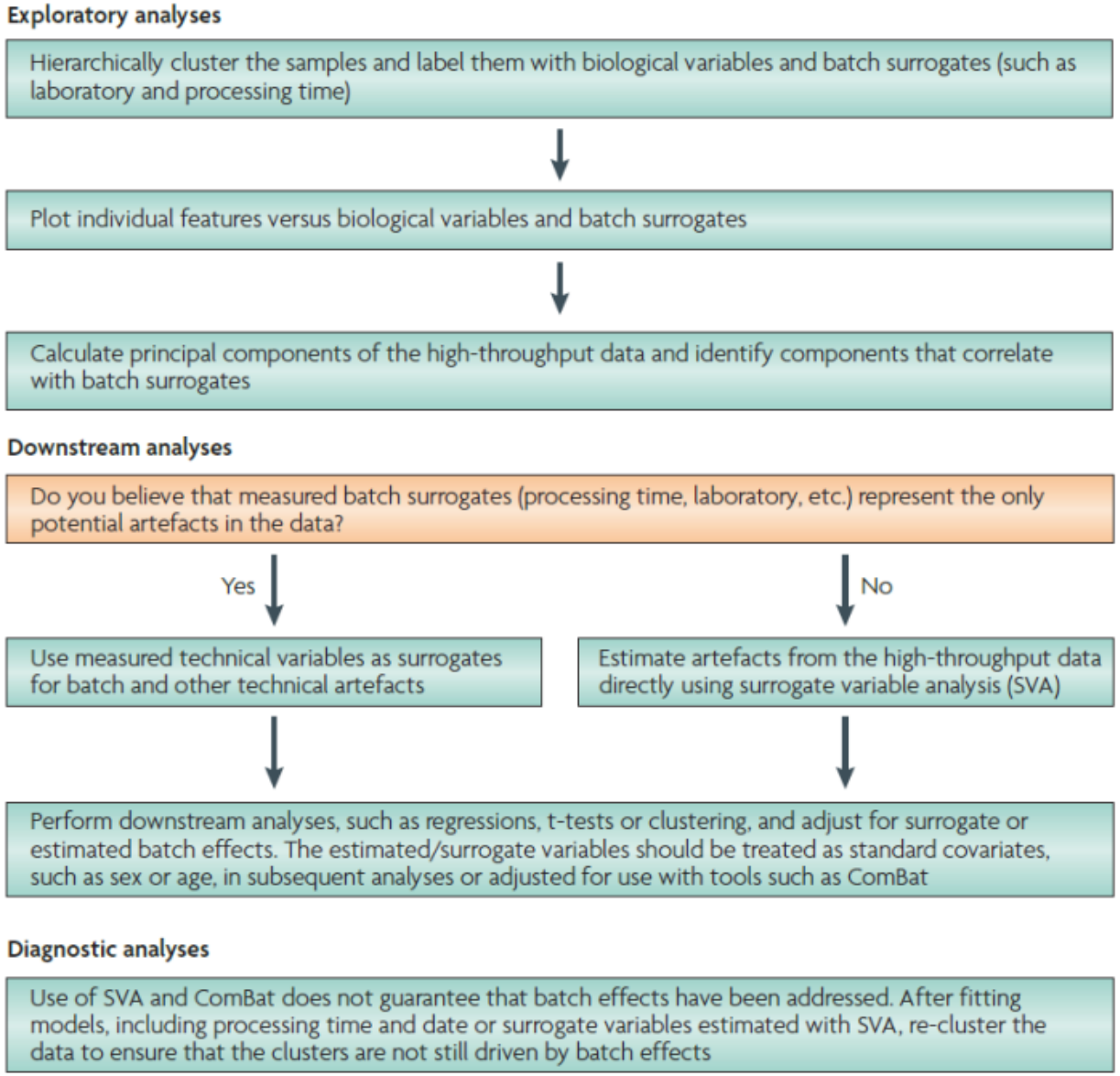
Coverage data were standardized across samples:
blue represents three standard deviations below average and
orange represents three standard deviations above average.

Each row is a different HapMap sample processed in the same facility with the same platform. The samples are ordered by processing date with horizontal lines dividing the different dates. Shown is a 3.5 Mb region from chromosome 16.

Various batch effects can be observed. The largest one occurs between days 243 and 251 (the large orange horizontal streak).

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

Workflow to identify batch effects



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

Detect batch effects

PCA is commonly used as a visual tool to determine whether batch effects exist after applying a global normalization method.

However, PCA yields linear combinations of the variables that contribute maximum variance and thus will not necessarily detect batch effects if they are not the largest source of variability in the data.

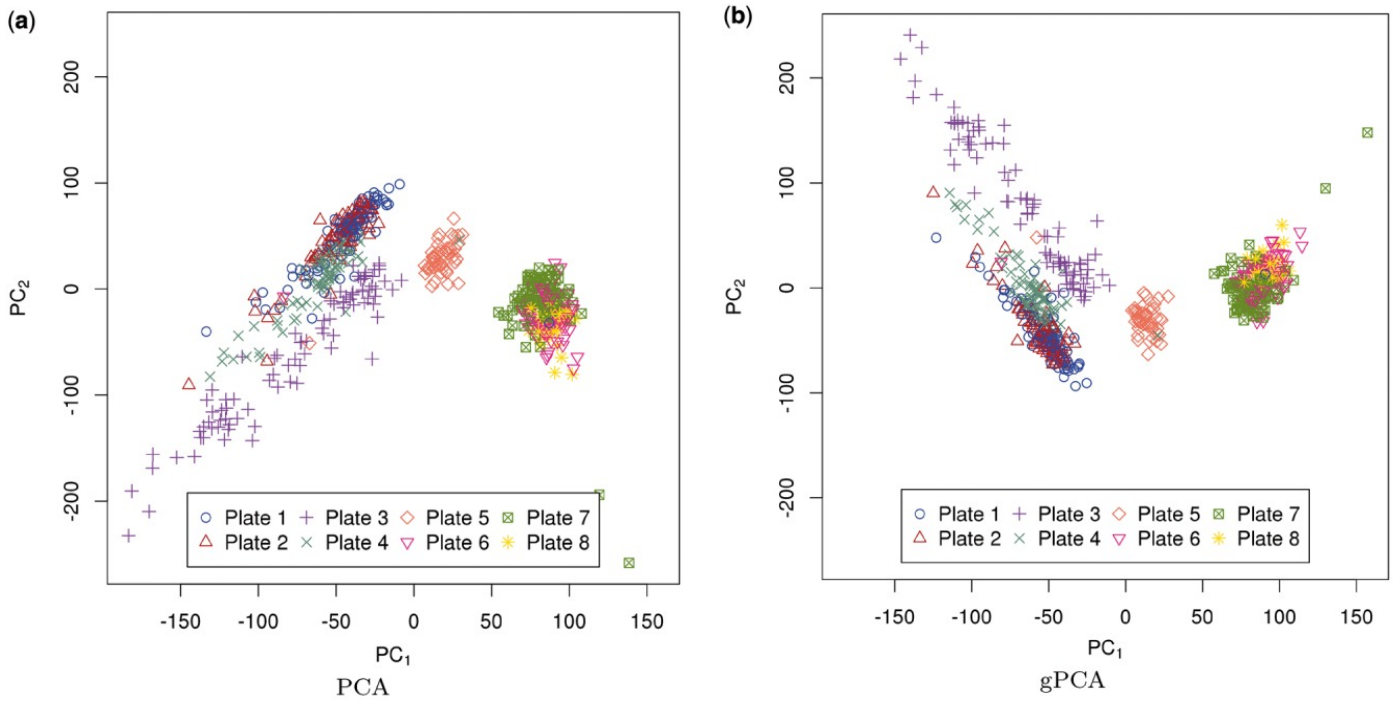
$\rightarrow$ Guided PCA: For detecting batch effects, a more informative version of PCA is to perform SVD on **Y'X**, where **Y** is an $n \times b$ indicator indicator matrix for b batches and n samples.

y_{ik} = 1 if sample is in batch k, y_{ik} = 0 otherwise

Large singular values imply that the batch is important for the corresponding principal component.

Reese et al. Bioinformatics (2013) 29: 2877–2883.

Detect batch effects



GENEMAM data

(a) Unguided PCA of **X** and

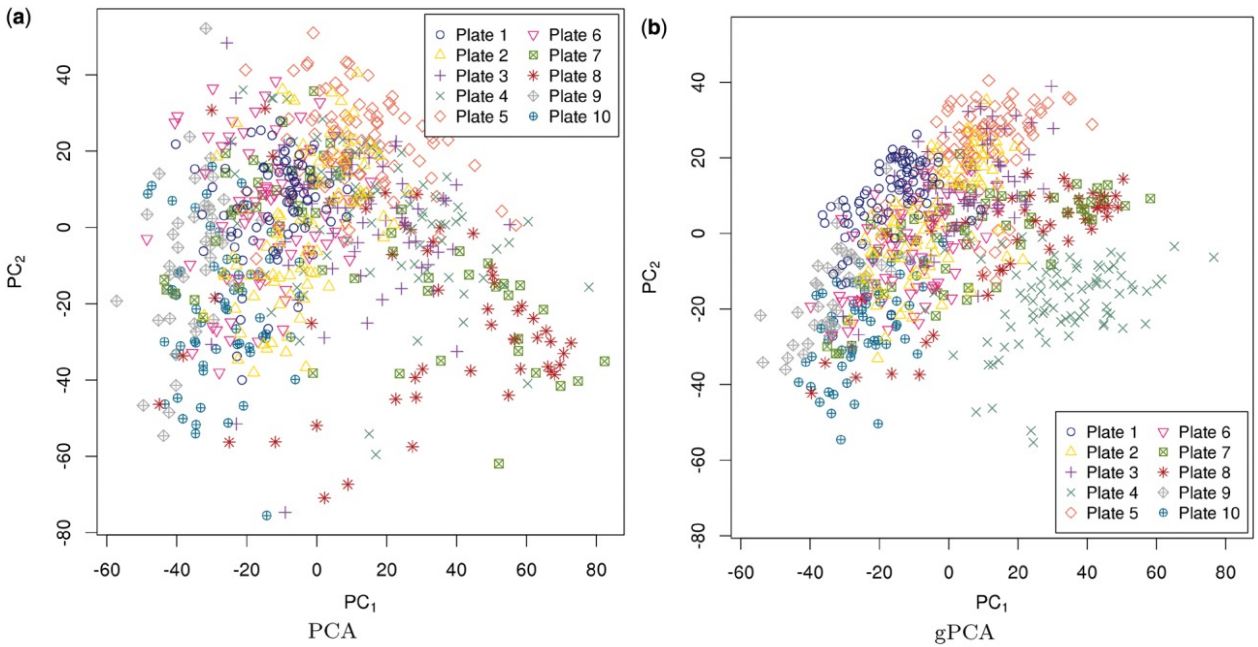
(**b**) gPCA of **Y'X**.

The standard use of PCA is to plot PC1 of the data versus PC2.

The GENEMAM data have an obvious batch effect. The PCA plot shows that this batch effect is due to the plate when colored by plate with three batches consisting of plates 1–4, **5** and 6–8 that were run at different times. The gPCA plot of the first two principal components shows greater separation in the batches, especially of **plate 3 (+)** from plates 1, 2 and 4, than the unguided principal component plot.

Reese et al. Bioinformatics (2013) 29: 2877–2883.

Detect batch effects



GENOA data

(a) Unguided PCA of **X** and

(**b**) gPCA of **Y'X**.

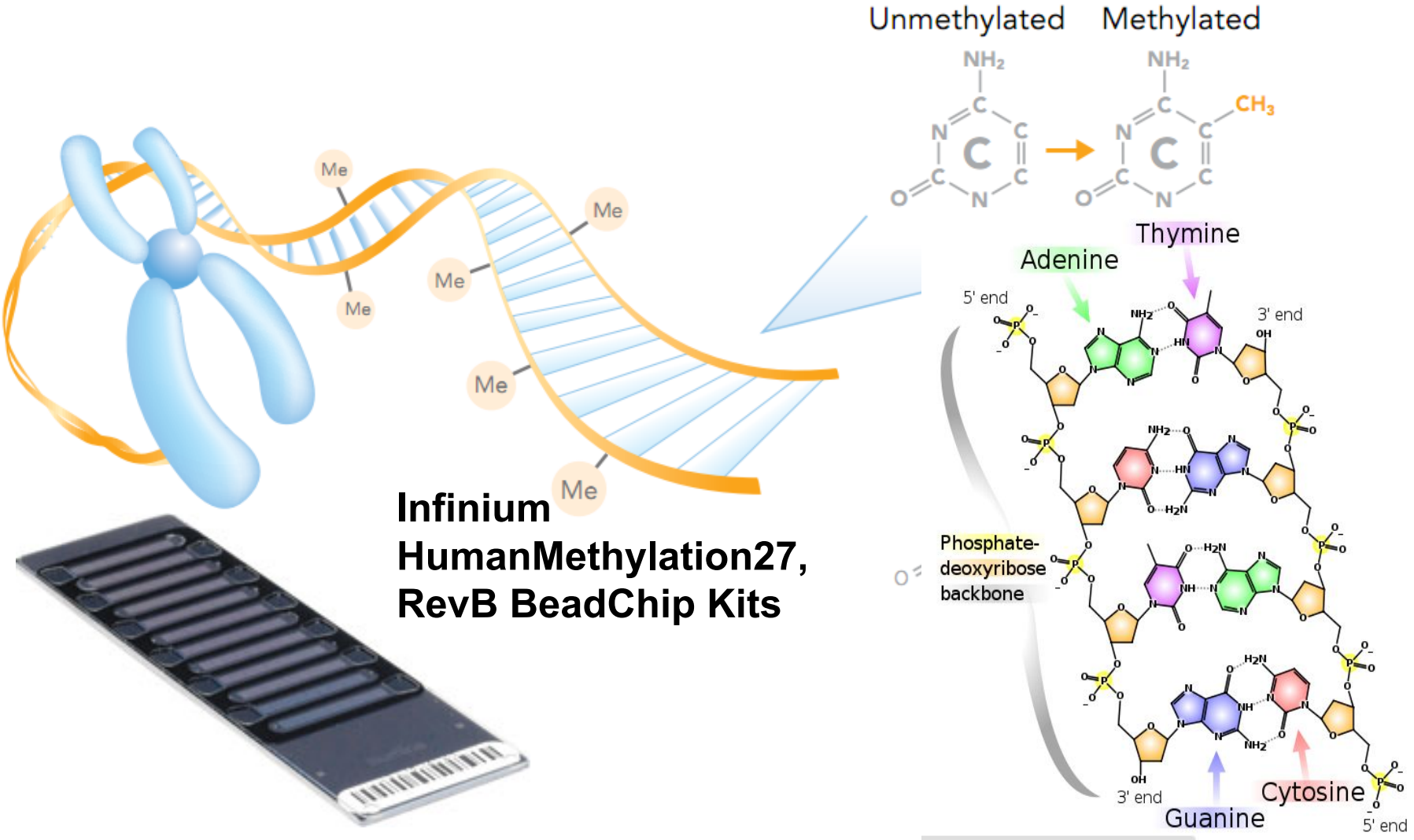
For the GENOA data set, batch is not so easily detected using unguided PCA.

The PCA plot of PC1 and PC2 shows that plates 7 and 8 might be slightly separated from the rest of the plates. A gPCA with batch defined by plate shows that plates 7 and 8, along with **plate 4 (X)**, separate slightly from the other plates.

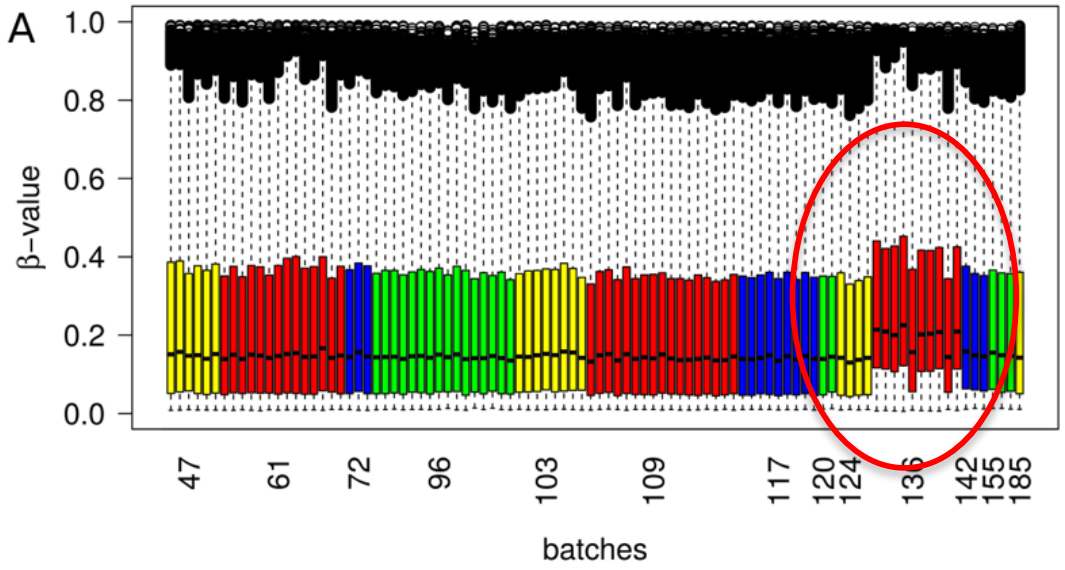
It is not obvious from the unguided PCA that plate 4 is separate from the rest of the plates. However, gPCA shows a separation between plate 4 and the rest of the plates.

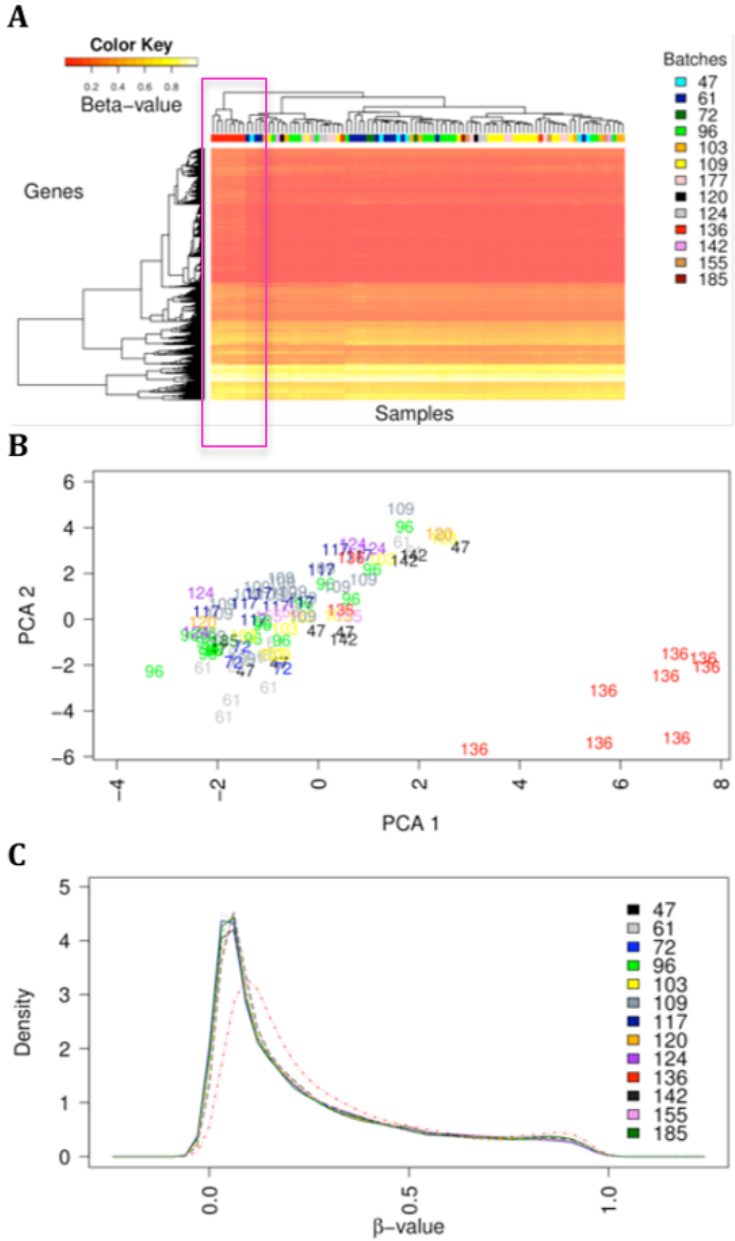
Reese et al. Bioinformatics (2013) 29: 2877–2883.

Correcting batch effects in DNA methylation data



Original DNA methylation data for breast cancer (TCGA)





β: fraction of methylated cytosines in CpG

Clear batch effect in batch 136

Left: box-plot

Right/top: hierarchical clustering

Right/middle: PCA

Right/bottom: density distribution

Batch effect correction with BEclear

(1) Compare the distribution of every gene in one batch to its distribution in all other batches using the nonparametric Kolmogorov-Smirnov (KS) test. P-values are corrected by False Discovery Rate.

(2) To consider only biologically relevant differences in methylation levels, identify the absolute difference between the median of all β-values within a batch for a specific gene and the respective median of the same gene in all other batches.

Those genes that have a FDR-corrected significance p-value below 0.01 (KS-test) AND a median difference larger than 0.05 are considered as batch effected (BE) genes in a specific batch.

Batch effect correction with BEclear

(3) Score severeness of batch effect in single batches by a weighting-scheme :

$$BEscore = \frac{\sum_{i \in mdif_{cat}} (N_{BEgenes_i} \cdot w_i)}{N}$$

N : total number of genes in a current batch,
$mdif_{cat}$: category of median differences,
$N_{BEgenes\_i}$: # BE-genes in *mdif* category *i*
w_i : weight of *mdif* category *i*
Weight categories:

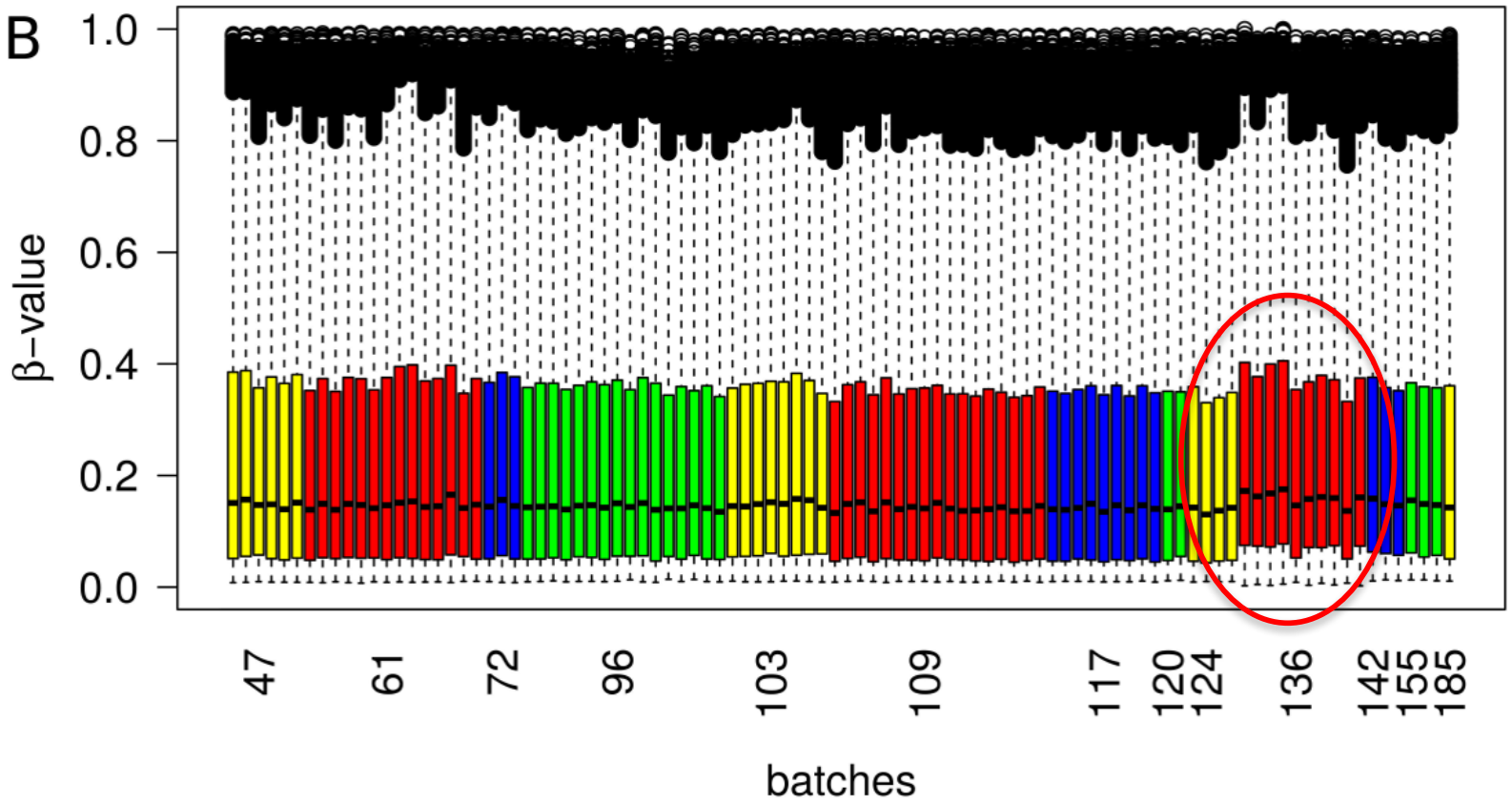
if *mdif* < 0.05, then weight = 0;
if 0.05 ≤ *mdif* < 0.1 weight = 1;
if $m \times 0.1 \leq mdif < (m + 1) \times 0.1$, $m \in N^+$

Scoring scheme considers number of BE-genes in the batch + magnitude of deviation of the medians of BE-genes in one batch compared to all other batches.

Based on the BE-scores of all batches, identify using the Dixon test which batches have BE-scores that deviate significantly from the BE-scores of the other batches.

All BE-gene entries in these affected batches are **replaced** by **LFM predictions**.

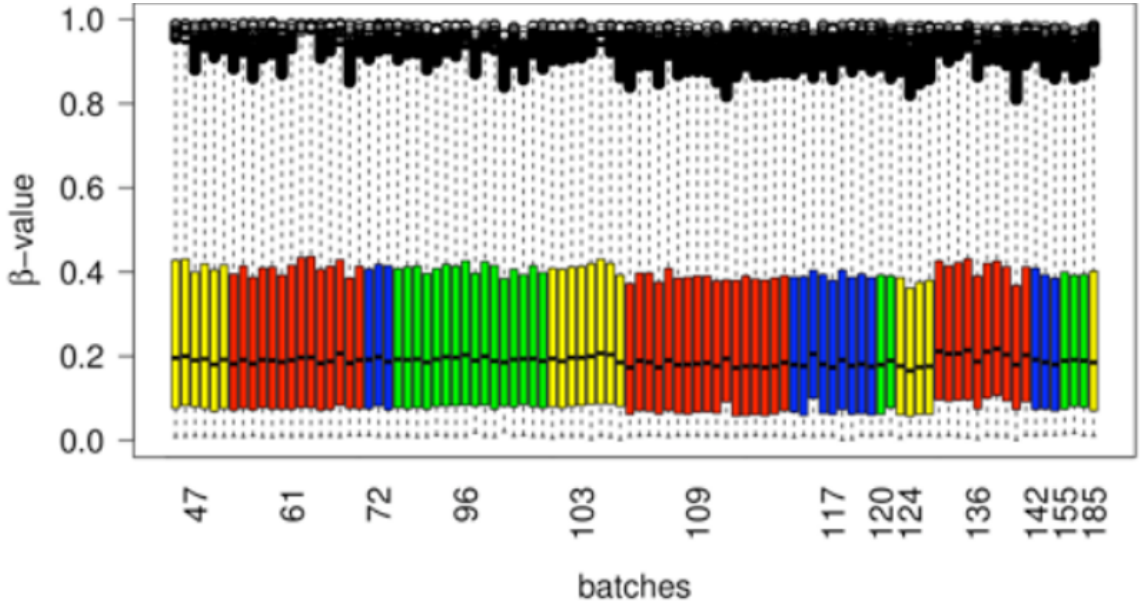
TCGA data for breast cancer – batch affected entries predicted by LFM/BEclear



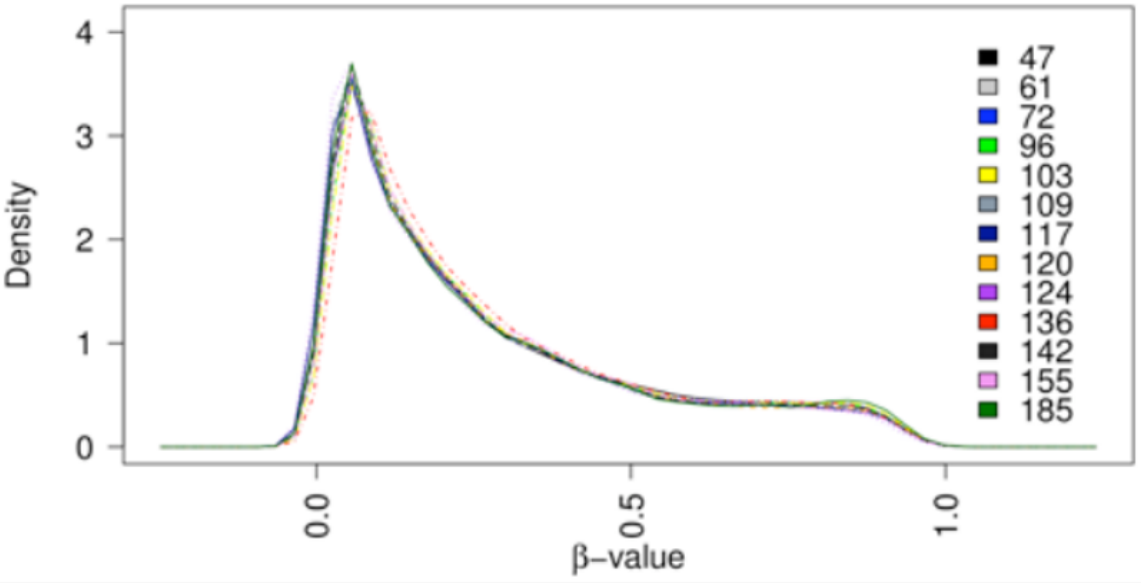
Batch 136 has still slightly larger values than other batches,

but the deviation is no longer statistically significant.

TCGA data for breast cancer – data corrected by FunNorm



B



A. Per sample boxplot

B. Density plot.

Functional normalization was able to adjust the batch effect equally well as BEclear

Functional Normalization

Functional normalization uses information from 848 control probes on 450k array.

The method extends the idea of quantile normalization by adjusting for known covariates measuring unwanted variation.

Consider $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ high-dimensional vectors each associated with a set of scalar **covariates** $Z_{i,j}$ with $i = 1,\ldots,n$ indexing samples and $j = 1,\ldots,m$ indexing covariates.

Ideally these known covariates are associated with unwanted variation and unassociated with biological variation.

Functional normalization attempts to remove their influence.

Functional Normalization

For each high-dimensional observation $\mathbf{Y}_i$, we form the empirical quantile function $r \in [0,1]$ for its marginal distribution, and denote it by q_i^{emp} .

We assume the following model $$q_i^{\text{emp}}(r) = \alpha(r) + \sum_{j=1}^{m} Z_{i,j}\beta_j(r) + \epsilon_i(r)$$

α : mean of the quantile functions across all samples,

β_j : coefficient functions associated with the covariates and

ε_i : error functions, which are assumed to be independent and centered around 0.

In this model, the term $$\sum_{j=1}^{m} Z_{i,j}\beta_j$$

represents variation in the quantile functions explained by the covariates.

Functional normalization removes unwanted variation by regressing out this term.

Functional Normalization

$\hat{\beta}_j$ for j = 1, …., m

are estimated using regression from the values observed for the control probes.

Assuming we have obtained estimates $\hat{\beta}_j$ for $j = 1, \ldots, m$, we form the functional normalized quantiles by

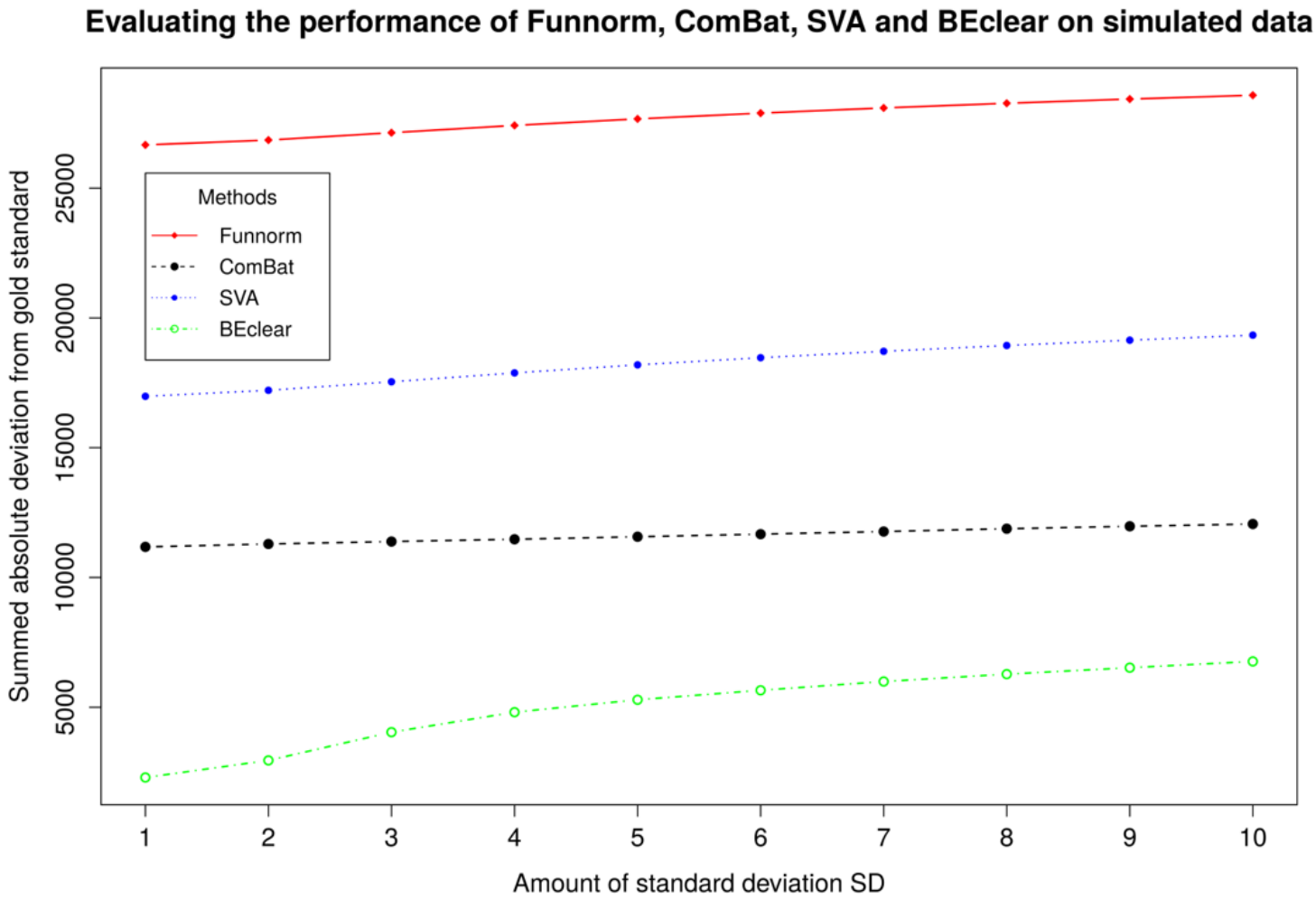
$$q_i^{\text{Funnorm}}(r) = q_i^{\text{emp}}(r) - \sum_{j=1}^{m} Z_{i,j}\hat{\beta}_j(r)$$

We then transform $\mathbf{Y}_i$ into the functional normalized quantity $\tilde{\mathbf{Y}}_i$ using the formula

$$\tilde{\mathbf{Y}}_i = q_i^{\text{Funnorm}}\left(\left(q_i^{\text{emp}}\right)^{-1}(\mathbf{Y}_i)\right)$$

This ensures that the marginal distribution of $\tilde{\mathbf{Y}}_i$ has q_i^{Funnorm} as its quantile function.

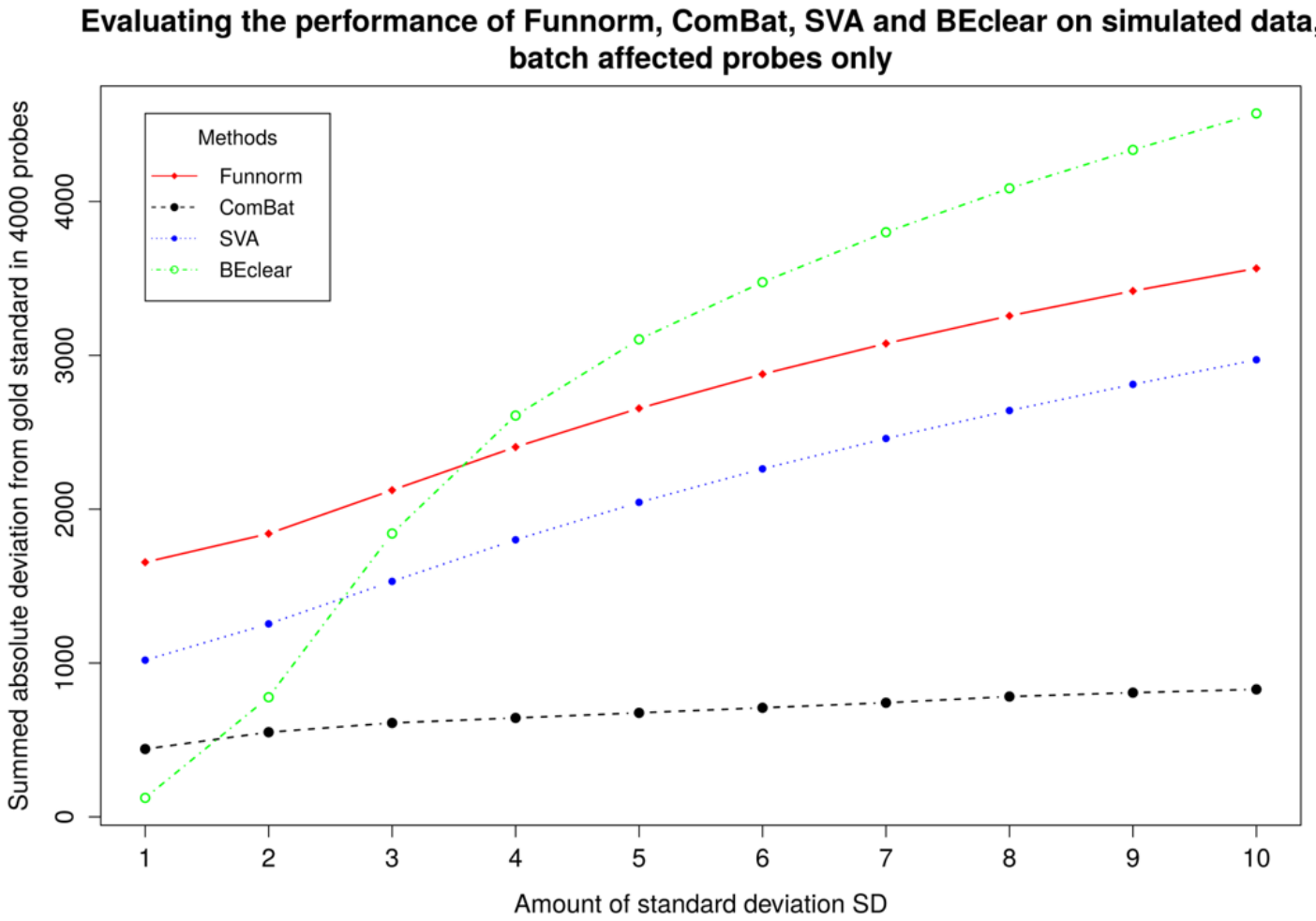
Benchmarking BEclear



Funnorm, ComBat and SVA scale all values
-> large total deviation

BEclear corrects only affected entries

Effect on corrected entries only



Even for affected entries,

BEclear predicts smallest changes for batch effects

up to 2 s.dev.

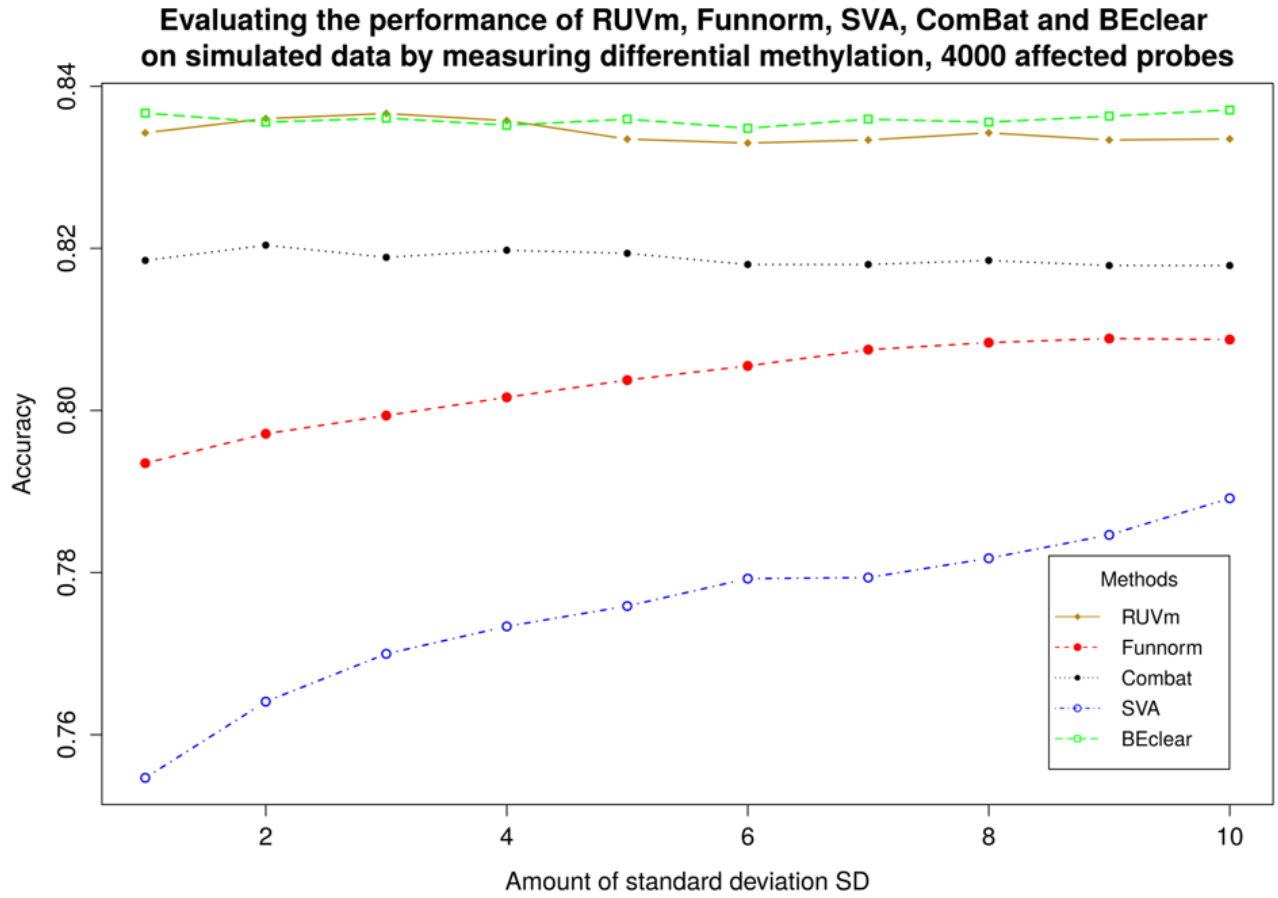
which is a typical magnitude of batch effects.

Accuracy of differential methylation analysis

Identify differentially methylated CpG probes (tumor vs. normal) in original data

Then introduce synthetic batch effect (*n* x st.dev.) + noise term

Identify differentially methylated CpG probes again + compare to reference



Conclusions

Predicting **missing values or batch-effected values** by **Latent Factor Model** (**BEClear software**):

- Accuracy of MA hybridization prediction confirmed by WGS (97%),

  low LFM error
- Superior accuracy of predicting DNA methylation levels by LFM confirmed in benchmark against SVA, Combat, FunNorm softwares

Review: Foundations of Probability Theory

„**Probability**“ : degree of confidence that an event of an uncertain nature will occur.

„Events“ : we will assume that there is
an agreed upon **space Ω** of possible outcomes („events“).

E.g. a normal die (*dt. Würfel*) has a space $\Omega = \{1,2,3,4,5,6\}$

Also we assume that there is a set of **measurable events S**
to which we are willing to assign probabilities.

In the die example, the event $\{6\}$ is the case where the die shows 6.

The event $\{1,3,5\}$ represents the case of an odd outcome.

Foundations of Probability Theory

Probability theory requires that the **event space** satisfies 3 basic properties:

- It contains the **empty event** $\varnothing$ and the **trivial event** Ω.

- It is **closed under union** $\rightarrow$ if $\alpha, \beta \in S$, then so is $\alpha \cup \beta \in S$,

- It is **closed under complementation** $\rightarrow$ if $\alpha \in S$, then so is $\Omega - \alpha \in S$

The requirement that the event space is closed under union and complementation implies that it is also closed under other Boolean operations, such as intersection and set difference.

Probability distributions

A **probability distribution** P *over* (Ω, S) is a mapping from events in S to real values. The mapping must satisfy the following conditions:

(1) $P(\alpha) \geq 0$ for all $\alpha \in$ S $\rightarrow$ *Probabilities are not negative*

(2) $P(\Omega) = 1$ $\rightarrow$ *The probability of the trivial event which allows all possible outcomes has the maximal possible probability of 1.*

(3) If $\alpha, \beta \in$ S and $\alpha \cap \beta = 0$ then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$

Interpretation of probabilities

The **frequentist's** interpretation:

The **probability** of an event is the **fraction of times**

the event occurs if we repeat the experiment indefinitely.

E.g. throwing of dice, coin flips, card games, …

where frequencies will satisfy the requirements of proper distributions.

For an event such as „It will rain tomorrow afternoon“,

the frequentist approach does not provide a satisfactory interpretation.

Interpretation of probabilities

An alternative interpretation views probabilities as **subjective degrees of belief**.

E.g. the statement „the probability of rain tomorrow afternoon is 50 percent“ tells us that - in the opinion of the speaker - the chances of rain and no rain tomorrow afternoon are the same.

When we discuss probabilities in the following we usually do not explicitly state their interpretation since both interpretations lead to the same mathematical rules.

Conditional probability

The **conditional probability** of β given α is defined as

$$P(\beta|\alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}$$

The probability that β is true given that we know α is the relative proportion of outcomes satisfying β among these that satisfy α.

From this we immediately see that

$$P(\alpha \cap \beta) = P(\alpha)P(\beta|\alpha)$$

This equality is know as the **chain rule** of conditional probabilities.

More generally, if α_1, α_2, … α_k are events, we can write

$$P(\alpha_1 \cap \alpha_2 \cap \cdots \cap \alpha_k) = P(\alpha_1)P(\alpha_2|\alpha_1) \ldots P(\alpha_k|\alpha_1 \cap \cdots \cap \alpha_{k-1})$$

Bayes rule

Another immediate consequence of the definition of conditional probability is **Bayes‘ rule**.

Due to symmetry, we can swap the 2 variables α and β in the definition $P(\beta|\alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}$ and get the equivalent expression $P(\alpha|\beta) = \frac{P(\beta \cap \alpha)}{P(\beta)}$

If we rearrange, we get Bayes‘ rule $P(\beta|\alpha)P(\alpha) = P(\alpha|\beta)P(\beta)$ or

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)}$$

A more general conditional version of Bayes‘ rule where all probabilities are conditioned on some background event γ also holds:

$$P(\alpha|\beta \cap \gamma) = \frac{P(\beta|\alpha \cap \gamma)P(\alpha|\gamma)}{P(\beta|\gamma)}$$

Example 1 for Bayes rule

Consider a student population.
Let `Smart` denote smart students and `GradeA` denote students who got grade A.

Assume we believe that P(`GradeA|Smart`) = 0.6, and that we get to know that a particular student received grade A.

Suppose that P (`Smart`) = 0.3 and P (`GradeA`) = 0.2

Then we have P(`Smart|GradeA`) = 0.6 × 0.3 / 0.2 = 0.9

In this model, an A grade strongly suggests that the student is smart.

On the other hand, if the test was easier and high grades were more common, e.g. P (`GradeA`) = 0.4, then we would get

P(`Smart|GradeA`) = 0.6 × 0.3 / 0.4 = 0.45 which is much less conclusive.

Example 2 for Bayes rule

Suppose that a tuberculosis skin test is 95% percent accurate.

That is, if the patient is TB-infected, then the test will be positive with probability 0.95 and if the patient is not infected, the test will be negative with probability 0.95.

Now suppose that a person gets a positive test result.

What is the probability that the person is infected?

Naive reasoning suggests that if the test result is wrong 5% of the time, then the probability that the subject is infected is 0.95.

That would mean that 95% of subjects with positive results have TB.

Example 2 for Bayes rule

If we consider the problem by applying Bayes‘ rule, we need to consider the prior probability of TB infection, and the probability of getting a positive test result.

Suppose that 1 in 1000 of the subjects who get tested is infected $\rightarrow$ P(TB) = 0.001

We see that 0.001×0.95 infected subjects get a positive result
and 0.999×0.05 uninfected subjects get a positive result.

Thus P(Positive) = $0.001 \times 0.95 + 0.999 \times 0.05 = 0.0509$

Applying Bayes‘ rule, we get P(TB|Positive) = P(TB) $\times$ P(Positive|TB) / P(Positive)
= $0.001 \times 0.95 / 0.0509 \cong 0.0187$

Thus, although a subject with a positive test is much more probable to be TB-infected than is a random subject, fewer than 2% of these subjects are TB-infected.

Random Variables

A **random variable** is defined by a function
that associates with each outcome in Ω a value.

For students in a class, this could be a function f_{grade} that maps
each student in the class (in Ω) to his or her grade (1, …, 5).

The event `grade = A` is a shorthand for the event $\{\omega \in \Omega: f_{grade}(\omega) = A\}$.

There exist **categorical (or discrete) random values** that take on
one of a few values, e.g. intelligence could be „high“ or „low“.

There also exist **integer or real random variable** that can take on
an infinite number of continuous values, e.g. the height of students.

By Val(X) we denote the set of values that a random variable X can take.

Random Variables

In the following, we will either consider categorical random variables or random variables that take real values.

We will use capital letters X, Y, Z to denote random variables.

Lowercase values will refer to the values of random variables.

E.g. $P(X = x) \geq 0 \text{ for all } x \in Val(X)$

When we discuss categorical random numbers, we will denote the *i-th* value as x^i.

Bold capital letters are used for sets of random variables: **X, Y, Z**.

Marginal Distributions

Once we define a random variable X, we can consider the **marginal distribution** *P(X)* over events that can be described using X.

E.g. let us take the two random variables `Intelligence` and `Grade` and their marginal distributions `P(Intelligence)` and `P(Grade)`

Let us suppose that

$$P(\text{Intelligence=high}) = 0.3$$
$$P(\text{Intelligence=low}) = 0.7$$

$$P(\text{Grade=A}) = 0.25$$
$$P(\text{Grade=B}) = 0.37$$
$$P(\text{Grade=C}) = 0.38$$

These marginal distributions are probability distributions satisfying the 3 properties.

Joint Distributions

Often we are interested in questions that
involve the values of several random variables.

E.g. we might be interested in the event „`Intelligence = high` and `Grade = A`“.

In that case we need to consider the **joint distribution** $P(X_1, \dots, X_n)$
over these two random variables.

The joint distribution of 2 random variables has to be consistent
with the marginal distribution in that $P(x) = \sum_y P(x, y)$.

		Intelligence		
		low	high	
	A	0.07	0.18	0.25
Grade	B	0.28	0.09	0.37
	C	0.35	0.03	0.38
		0.7	0.3	1

Conditional Probability

The notion of conditional probability extends to induced distributions over random variables.

$P(\text{Intelligence}|\text{Grade}=\text{A})$ denotes the conditional distribution over the events describable by `Intelligence` given the knowledge that the student‘s grade is A.

Note that the conditional probability $P(\text{Intelligence}=\text{high}|\text{Grade}=\text{A}) = \frac{0.18}{0.25} = 0.72$ is quite different from the marginal distribution $P(\text{Intelligence}=\text{high}) = 0.3$.

We will use the notation $P(X|Y)$ to present a set of conditional probability distributions.

Bayes‘ rule in terms of conditional probability distributions reads

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}$$

Probability Density Functions

A function $p: \mathbb{R} \to \mathbb{R}$

is a **probability density function** (PDF) for X

if it is a nonnegative integrable function so that $\int_{Val(X)} p(x)dx = 1$

The function $\mathrm{P}(X \leq a) = \int_{-\infty}^{a} p(x)dx$ is the **cumulative distribution** for X.

By using the density function we can evaluate the probability of other events. E.g.

$$\mathrm{P}(a \leq X \leq b) = \int_{a}^{b} p(x)dx$$

Uniform distribution

The simplest PDF is the **uniform distribution**

Definition: A variable X has a uniform distribution over [a,b] denoted X ~ Unif[a,b] if it has the PDF

$$p(x) = \begin{cases} \frac{1}{b-a} & b \geq x \geq a \\ 0 & \text{otherwise} \end{cases}$$

Thus the probability of any subinterval of [*a,b*] is proportional to its size relative to the size of [*a,b*].

If $b - a < 1$, the density can be greater than 1.

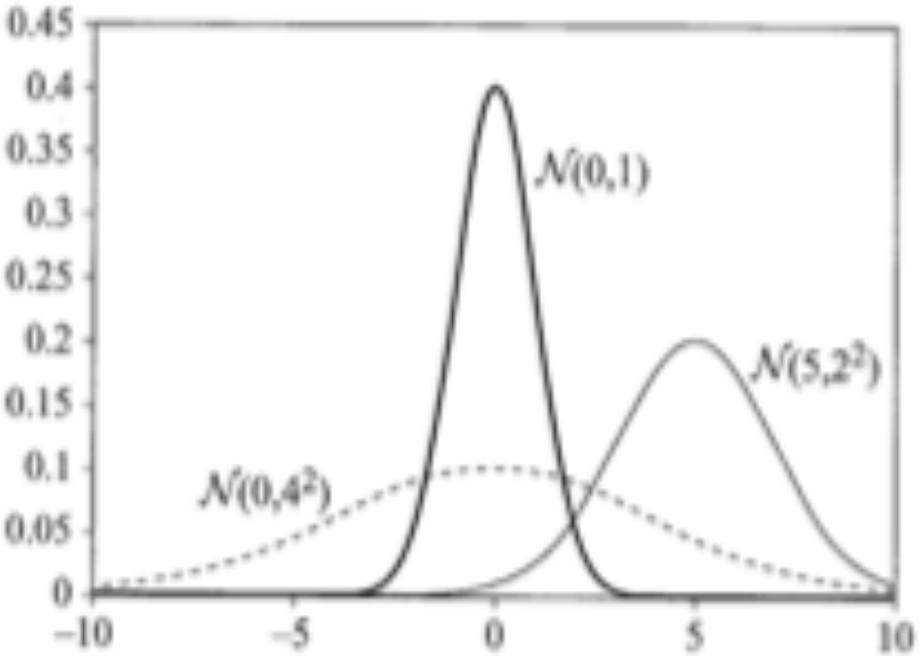
We only have to satisfy the constraint that the total area under the PDF is 1.

Gaussian distribution

A random variable X has a Gaussian distribution with mean μ and variance σ^2 , denoted X ~ N(μ;σ^2) if it has the PDF

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A standard Gaussian has mean 0 and variance 1.



Expectation

Let X be a discrete random variable that takes numerical values.

Then, the **expectation** of X under the distribution P is

$$\mathbf{E}_P[X] = \sum_x x \cdot P(x)$$

If X is a continuous variable,
then we use the density function

$$\mathbf{E}_P[X] = \int x \cdot p(x) dx$$

E.g. if we consider X to be the outcome of rolling a good die with probability 1/6 for each outcome, then **E**[X] = 1 · 1/6 + 2 · 1/6 + … + 6 · 1/6 = 3.5

Properties of the expectation of a random variable

$\mathbf{E}[a \cdot X + b] = a\,\mathbf{E}[X] + b$

Let X and Y be two random variables

$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$

Here, it does not matter whether X and Y are independent or not.

What can be say about the expectation value of a product of two random variables?

In the general case, we can say very little.

Consider 2 variables X and Y that each take on the values +1 and -1 with probabilities 0.5.

If X and Y are independent, then $\mathbf{E}[X \cdot Y] = 0$.

If they always take on the same value (they are correlated), then $\mathbf{E}[X \cdot Y] = 1$.

Properties of the expectation of a random variable

If X and Y are independent then

E[X · Y]= **E**[X] · **E**[Y]

The **conditional expectation** of X given y is

$$E_P[X|y] = \sum_x x \cdot P(x|y)$$

Variance

The expectation of X tells us the mean value of X. However, it does not indicate how far X deviates from this value. A measure of this deviation is the **variance** of X:

$$Var_P[X] = \mathbf{E}_P[(X - \mathbf{E}_P[X])^2]$$

The variance is the **expectation** of the **squared difference** between X and its expected value. An alternative formulation of the variance is

$$Var[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

If X and Y are independent, then $Var[X + Y] = Var[X] + Var[Y]$

$$Var[a \cdot X + b] = a^2 Var[X]$$

For this reason, we are often interested in the square root of the variance, which is called the **standard deviation** of the random variable. We define

$$\sigma_X = \sqrt{Var[X]}$$

Variance

Let X be a random variable with Gaussian distribution $N(\mu;\sigma^2)$.
Then **E**[X] = μ and Var[X] = σ^2.

Thus, the parameters of the Gaussian distribution specify the expectation and the variance of the distribution.
The form of the Gaussian distribution implies that the density of values of X drops exponentially fast in the distance (x - μ) / σ.

Not all distributions show such a rapid decline in the probability of outcomes that are distant from the expectation.
However, even for arbitrary distributions, one can show that there is a decline.

The **Chebyshev inequality** states $P(|X - \mathbf{E}_P[X]| \geq t) \leq \frac{Var_P[X]}{t^2}$

or in terms of σ $P(|X - \mathbf{E}_P[X]| \geq k\sigma_X) \leq \frac{1}{k^2}$

Variance

Let X be a random variable with Gaussian distribution $N(\mu;\sigma^2)$.
Then **E**[X] = μ and Var[X] = σ^2.

Thus, the parameters of the Gaussian distribution specify the expectation and the variance of the distribution.
The form of the Gaussian distribution implies that the density of values of X drops exponentially fast in the distance $(x - \mu) / \sigma$.

Nice **online resources** on statistics:

https://www.khanacademy.org/math/statistics-probability

http://tutorials.istudy.psu.edu/basicstatistics/

https://stattrek.com/statistics/problems.aspx