BIO214 Lecture 6

Bioinformatics-II

Genomic Data Normalization - 2

Zhen Wei; 2023-Feb-14

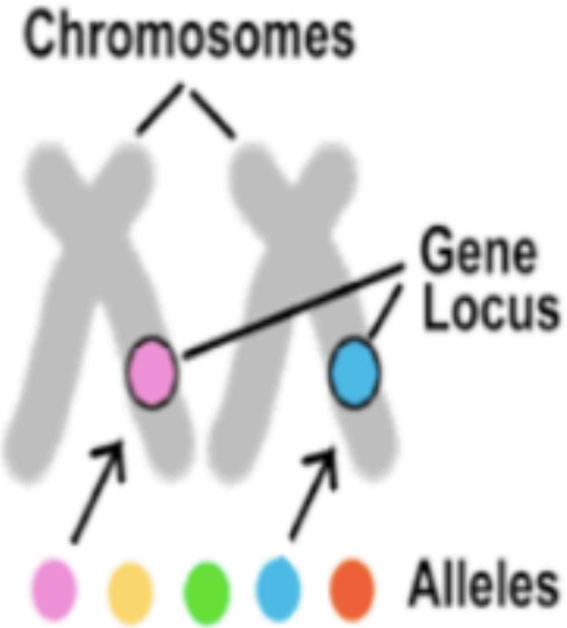
Outline

- Introduction to batch effect
- Batch effect adjustment by feature specific size factors
- Correction for GC content bias
- Combat and SVA
- Control experiment

Introduction to batch effect

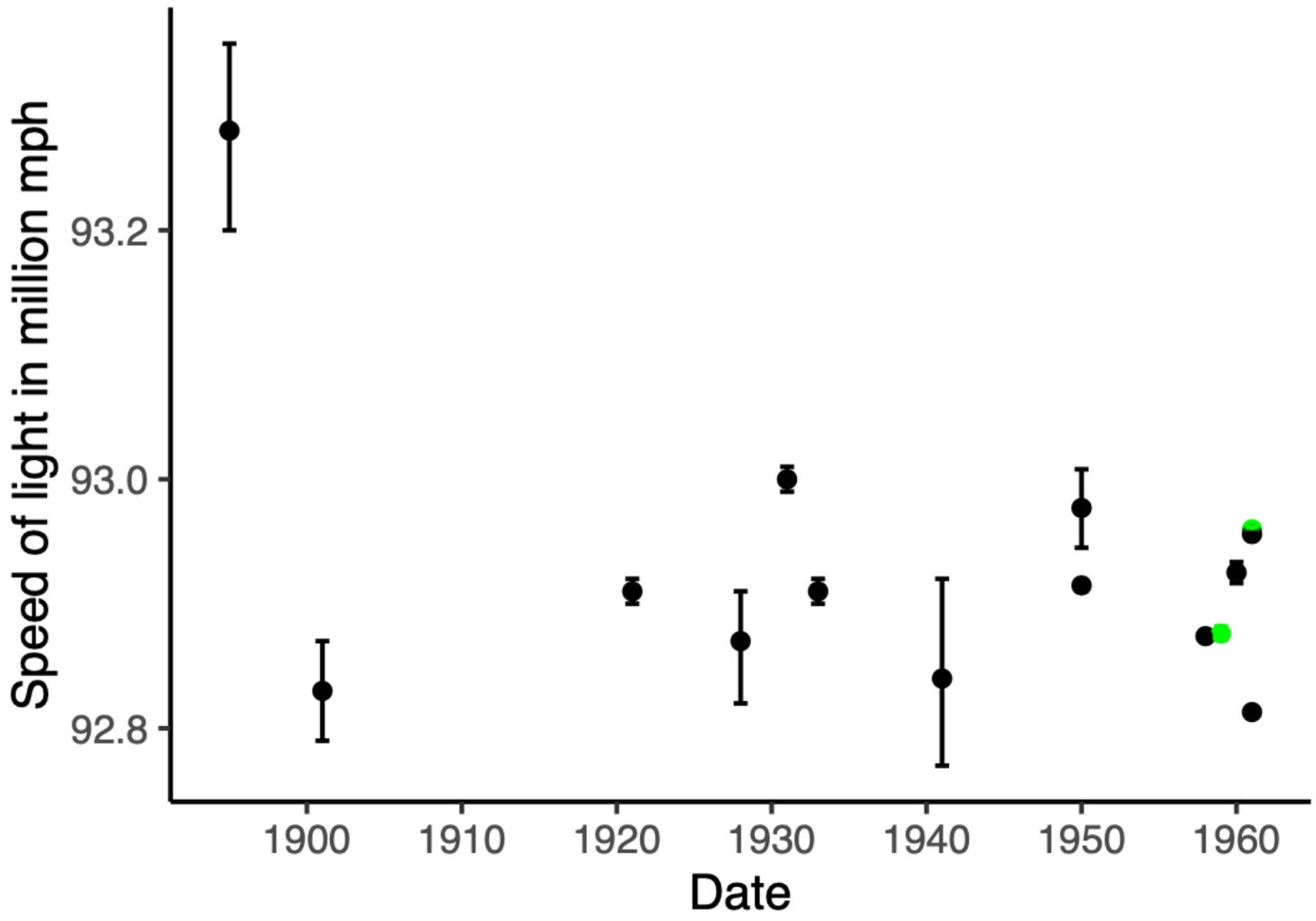
Batch effect?

Unexpected sources of variations between groups of experiments



- **External Factors** (e.x. environment)
- **Genetics / Epigenetics**
- **Technical Factors**

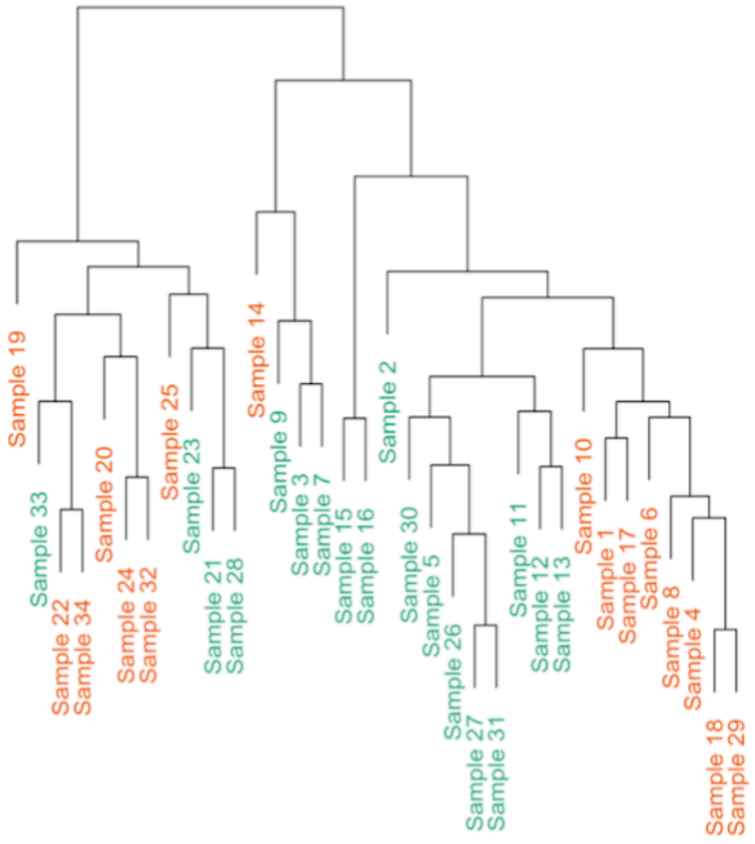
Does batch effect only happen in genomics?



- Speed of light estimates with “Confidence Intervals” (1900-1960)

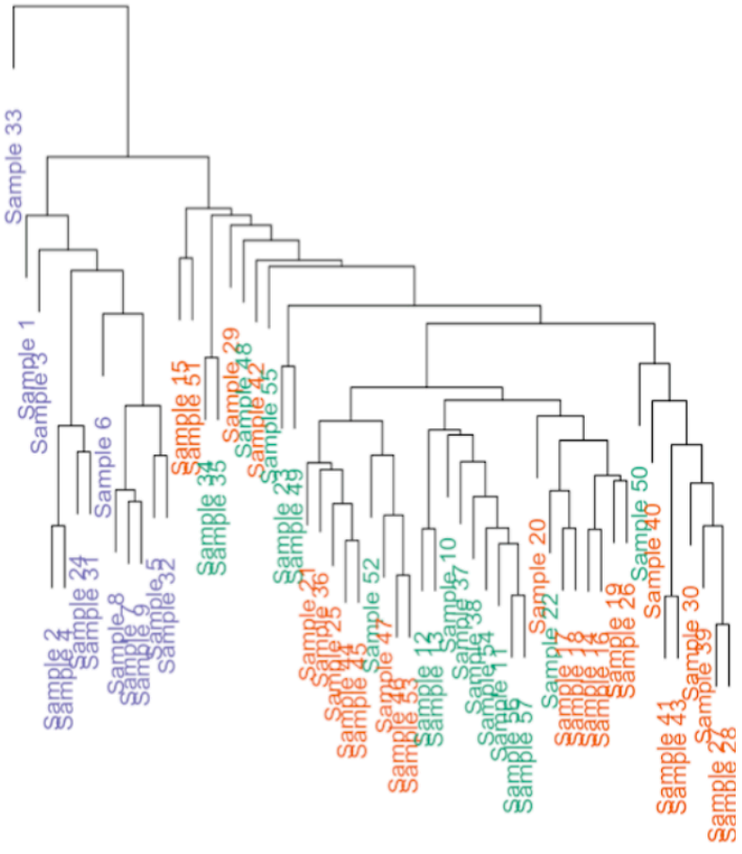
Youden W: Enduring values. Technometrics 1972, 14(1):1-11.

Influence of batch factors in gene expression analysis (I)



Color: Environments

Idaghdour et al. 2008



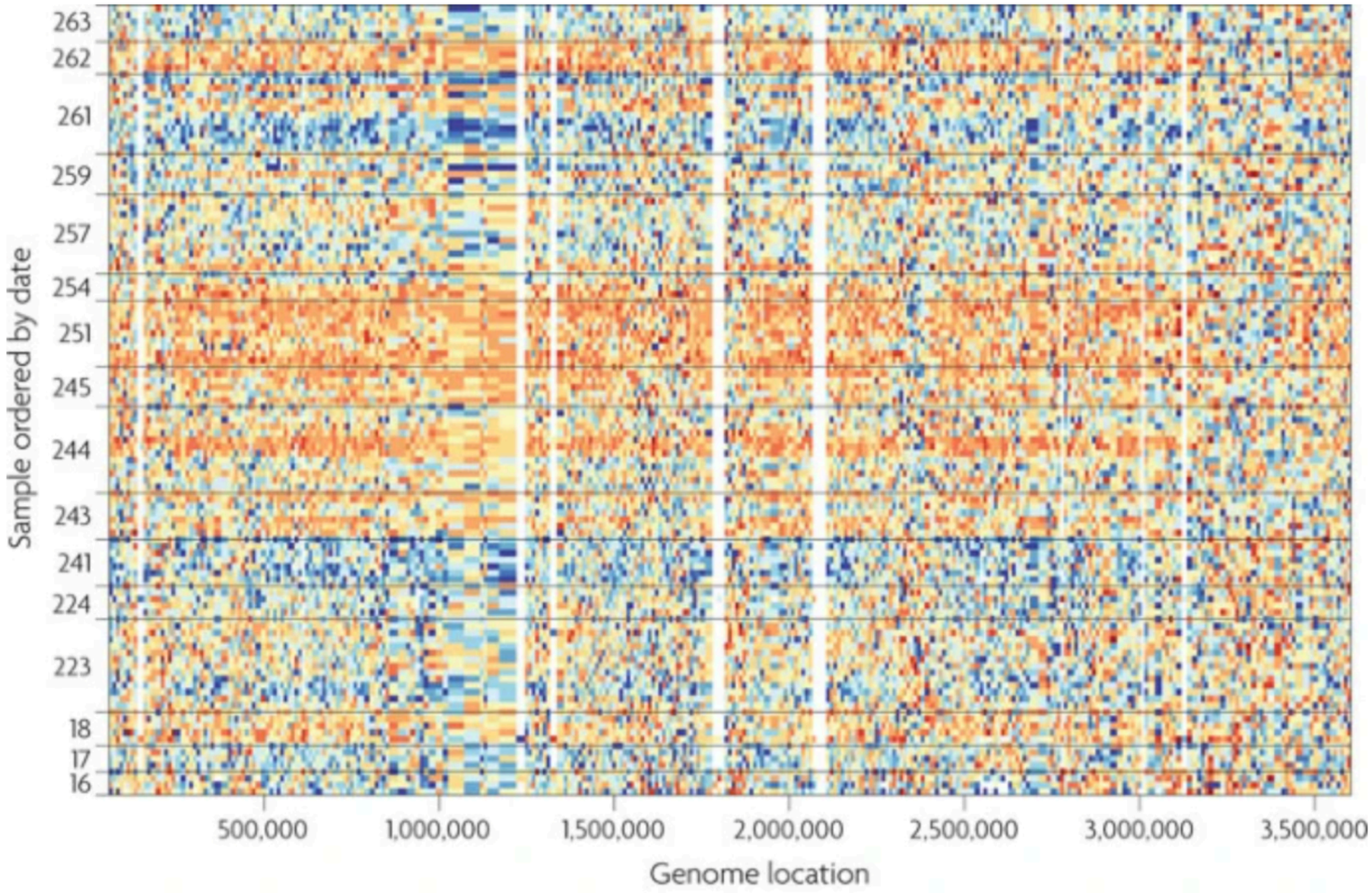
Color: Processing Year

Cheung et al. 2008

Hierarchical clustering dendrograms over gene expression samples.

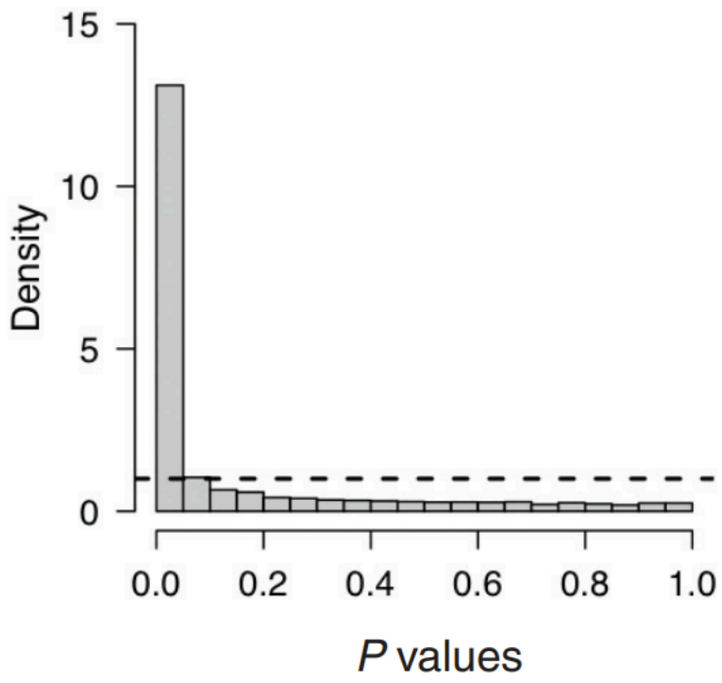
Influence of batch factors in gene expression analysis (II)

sorted by the dates which the samples are generated

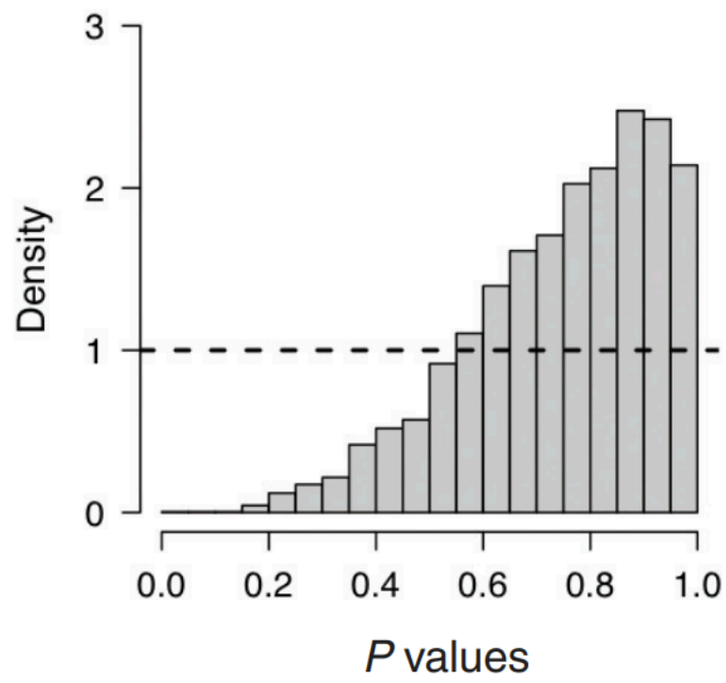


1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), 68.

False discoveries due to confounding



P values distribution for tests of differential expression between CEU and ASN samples



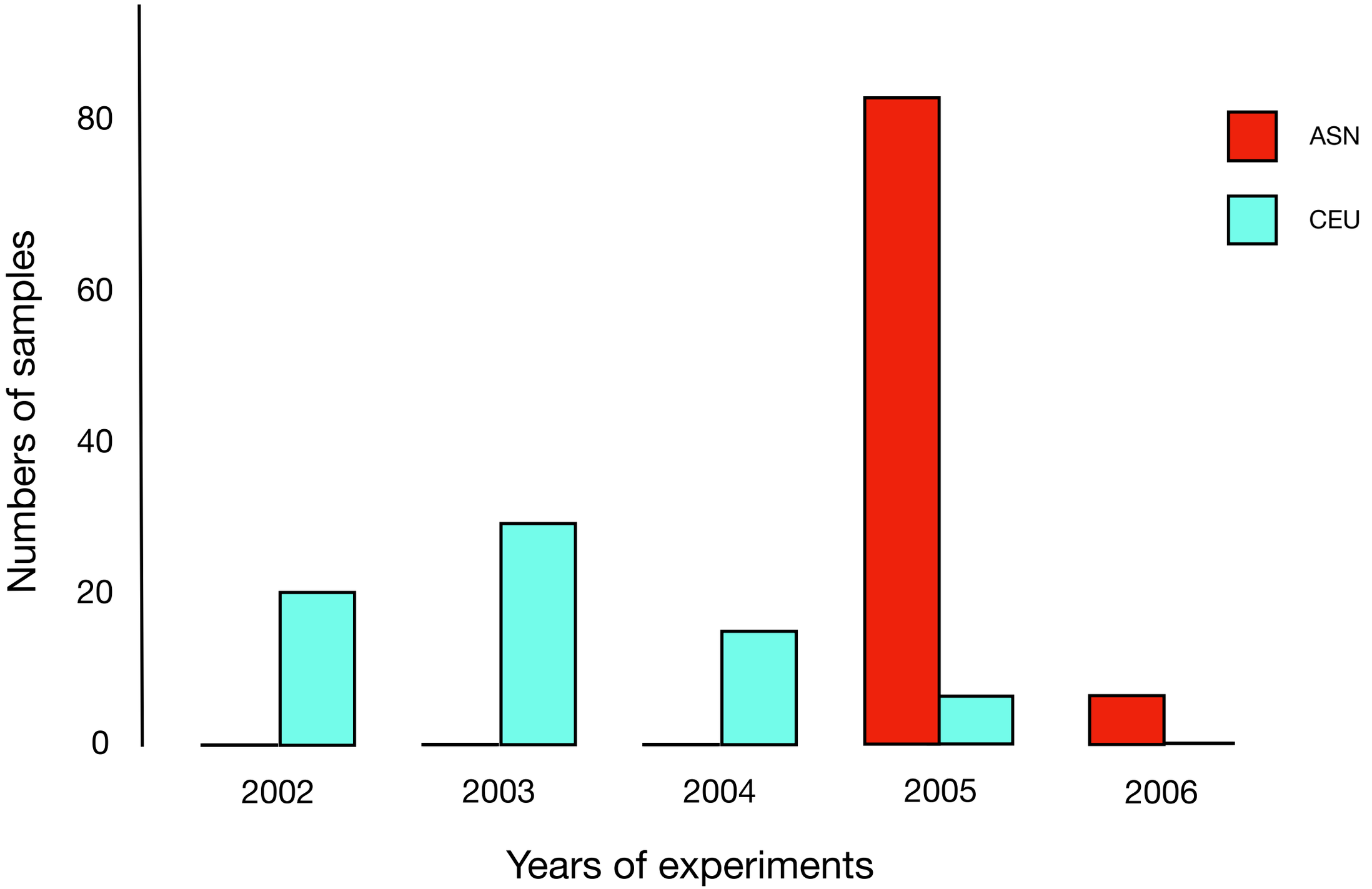
P values distribution after controlling the year in which the microarrays were processed

nature genetics

Published: 07 January 2007

Common genetic variants account for differences in gene expression among ethnic groups

The confounded experimental design



Other problematic conclusions made from big journal articles

- Proteomics test can predict ovarian cancer from serum (Lancet)
- 50% of genes are differentially expressed between ASN and CEU blood (Nature Genetics)
- SNPs associated with longevity (Science)
- Species explain more variability than tissue in gene expression (Genomics, PNAS)
- Age Related Methylation Profiles (Genome Research and others)
- Some proportion of single cell RNA-Seq results (Nature and others)

Batch effect adjustment by feature specific size factors

How to computationally adjust batch effect given the count matrix?

Dividing by more feature specific size factors

	Sample 1	Sample 2
Gene A	16/(s_j*l_i*gc_ij*M_i*…)	5/(s_j*l_i*gc_ij*M_i*…)
Gene B	13/(s_j*l_i*gc_ij*M_i*…)	3/(s_j*l_i*gc_ij*M_i*…)
Gene C	7/(s_j*l_i*gc_ij*M_i*…)	0/(s_j*l_i*gc_ij*M_i*…)
Gene D	28/(s_j*l_i*gc_ij*M_i*…)	12/(s_j*l_i*gc_ij*M_i*…)

Multiplicative model behind, K_{ij} is the read count for the i th gene and j th sample.

$$K_{ij} = \theta_{ij} \times s_j \times l_i \times f_j(\mathbf{gc}_i) \times M_i \times \ldots$$

- θ_{ij} : the true gene expression level (target of estimation).
- s_j : sequencing depth.
- l_i : gene length.
- $f_j(gc_i)$: **GC content bias.**
- M_i : **read mappability.**

Read genome mappability

Fast Computation of Genome Mappability

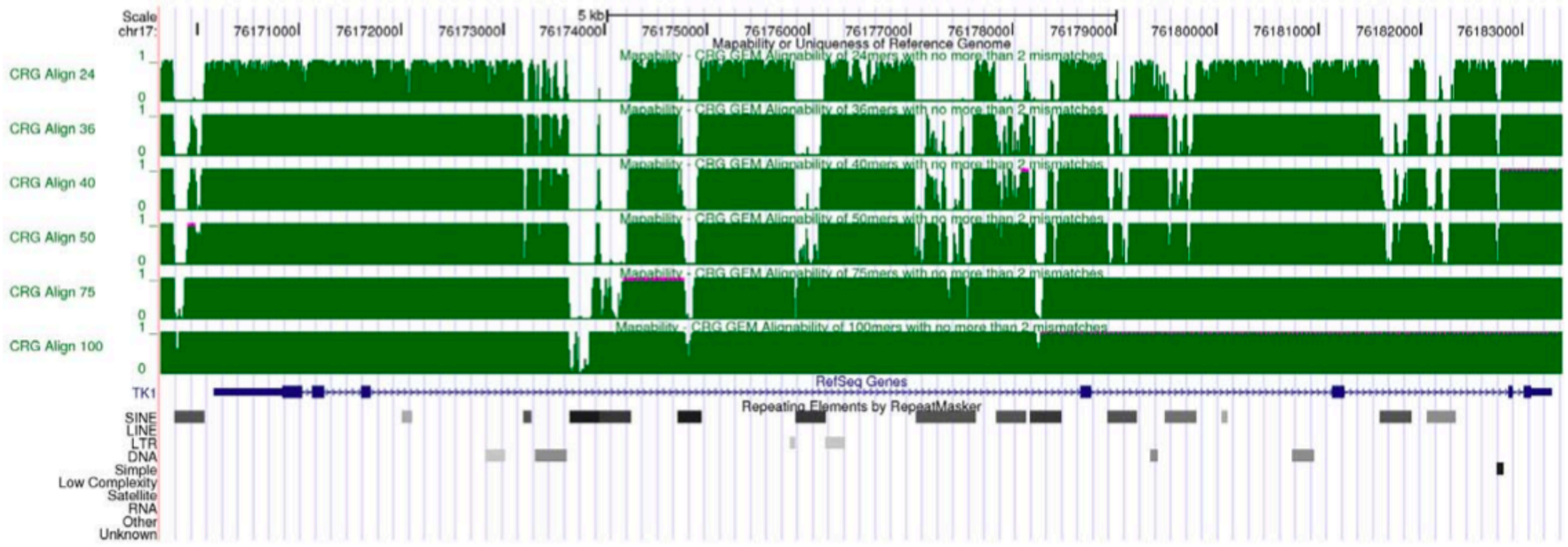
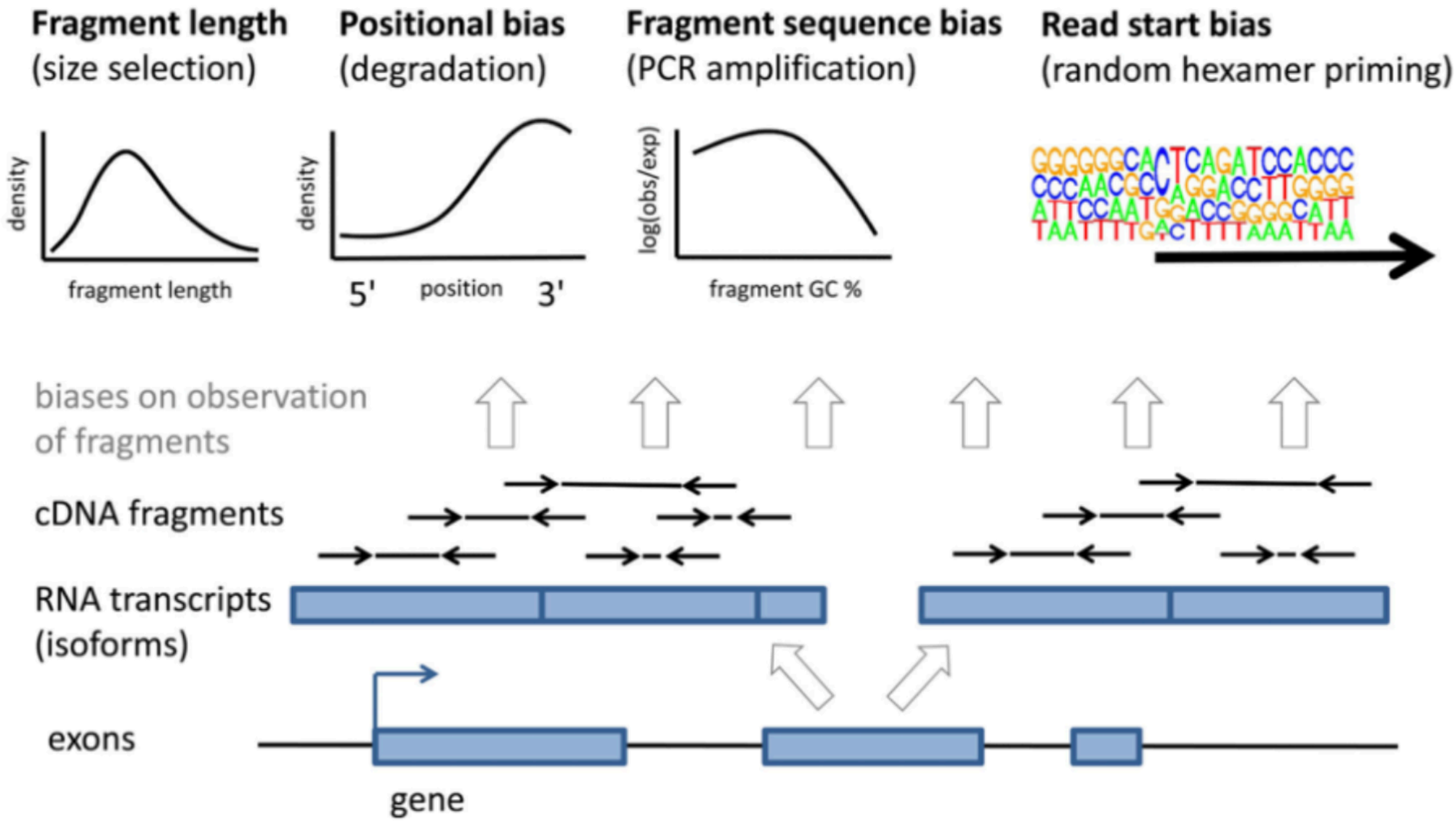


Figure 3. Visualization of mappability on the UCSC browser [7]: the example of the human TK1 gene. Six mappability tracks (green) are shown here corresponding to *k*-mer sizes 24, 36, 50, 75 and 100 bp (from top to bottom of the figure). Regions with low mappability score have high frequencies, and conversely. This example illustrates that the uniqueness of the TK1 locus (especially within the introns) could be inversely correlated with the presence of some repetitive elements as identified by RepeatMasker [37].
doi:10.1371/journal.pone.0030377.g003

- The idea is that some regions along the genome are harder to be (uniquely) mapped due to the presence of repetitive sequences.
- One can use specialized tool to estimate mappability across any genomes: https://evodify.com/gem-mappability/

Sequencing artifacts

The artifact generating mechanism of RNA-Seq

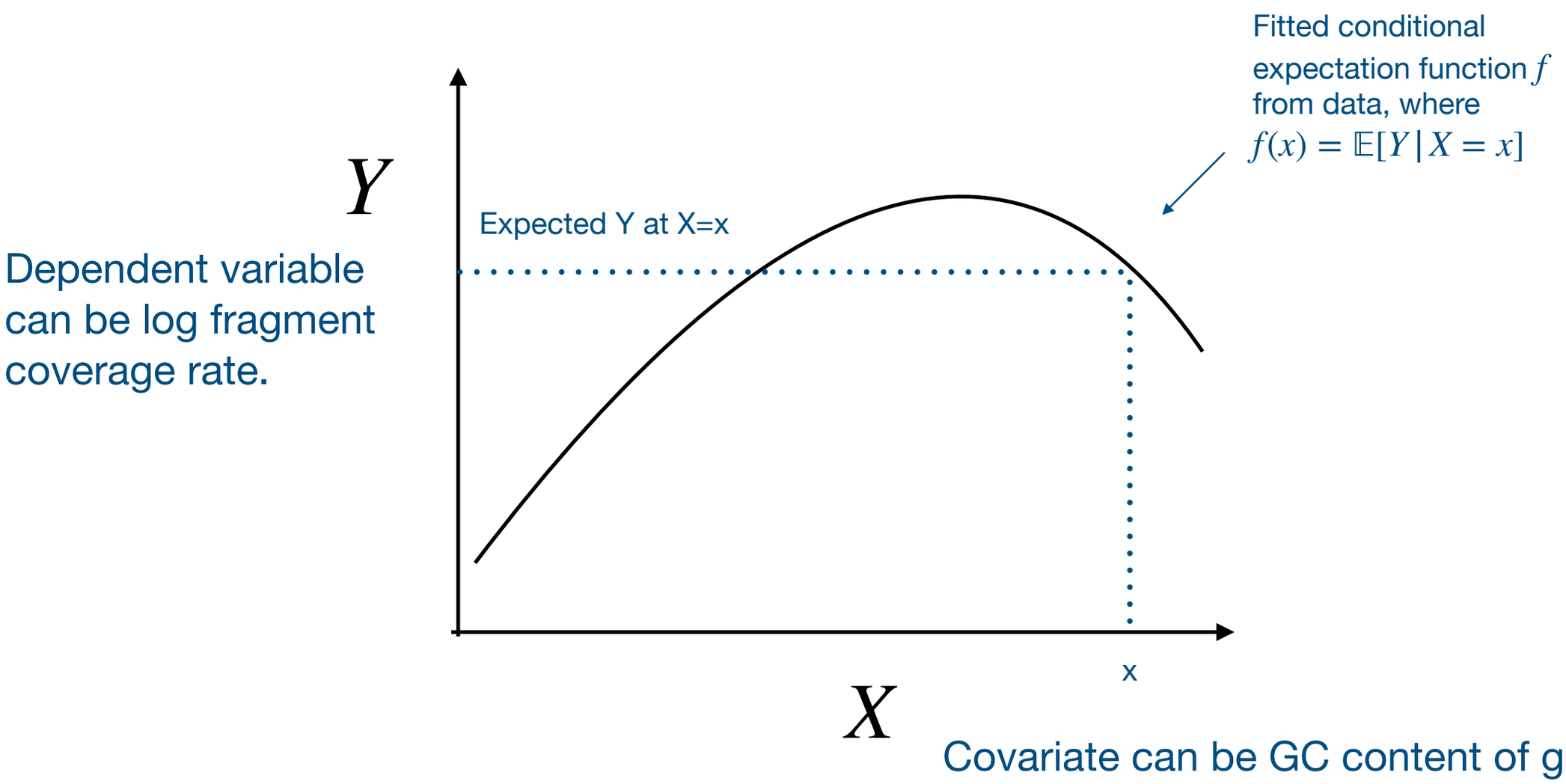


- **Among them, the fragment GC content bias is the leading component of technical variation.**

Love, M. I., Hogenesch, J. B., & Irizarry, R. A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nature biotechnology, 34(12), 1287.

Correction for GC content bias

Review on linear regression



- The purpose of linear regression is to fit the joint relationship between response variable (Y) and covariates (X).
- The output of linear regression fit is a mathematical function of conditional expectation, which can return the expected value of Y given a specific value of X.

Feature expansion: smoothing splines

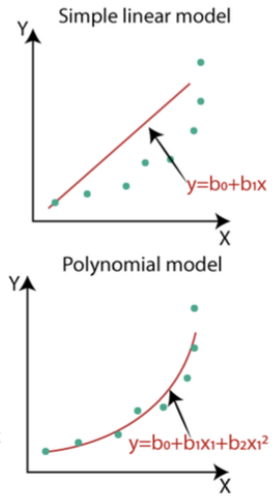
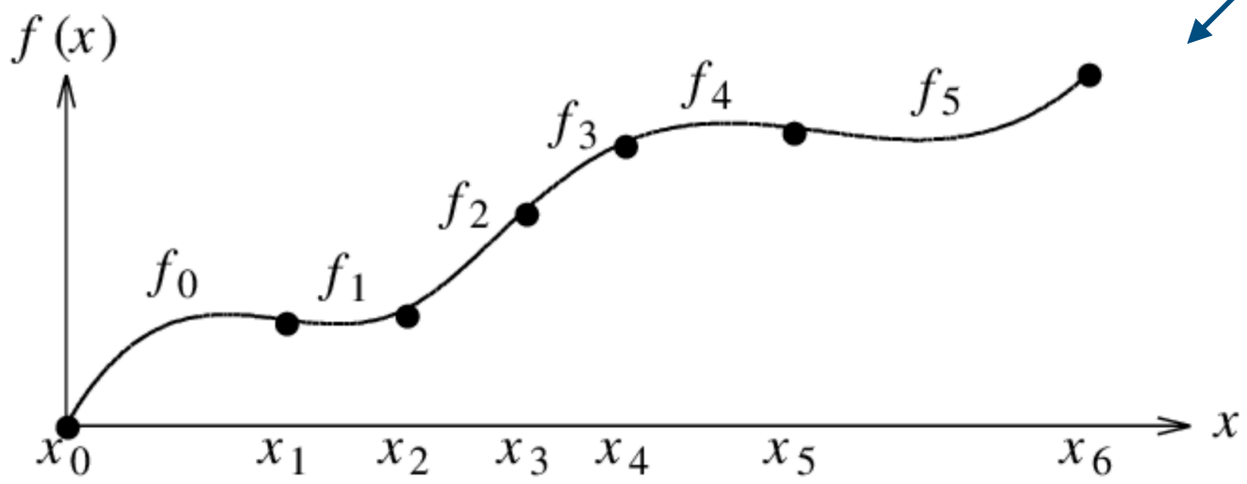


illustration for a cubic spline



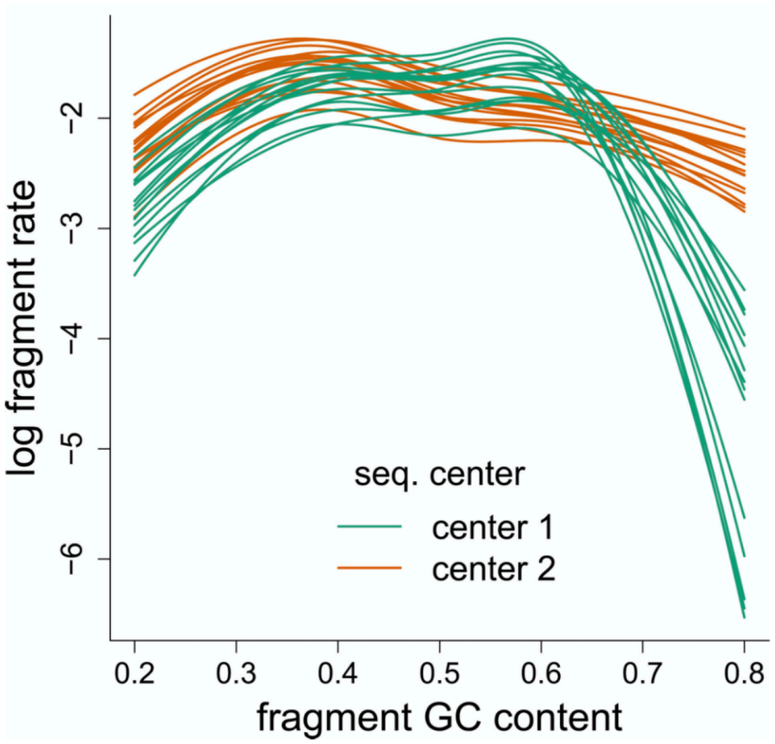
Polynomial regressions $f_0, \ldots, f_5$ are fitted on 6 intervals defined by the knots $x_1, \ldots, x_6$.

- Splines enables fitting most smooth non-linear patter between y and continuous x.
- The **cubic splines** algorithm works as a "piecewise polynomial":
  - Cut the range of x into several intervals, the boundaries of the intervals are called **knots.**
  - For each interval I, fit a polynomial curve with degree=3 for $x \in I$, and fit linear trends for $x \notin I$.
  - The final curve is obtained by the sum of f_I for all intervals together with the basic linear fit of $\beta_0 + \beta_1 x$.

```
#Fitting glm with natural cubic splines of 5 knots
glm(y~splines::ns(x,df=5), data = model_matrix, family = "Poisson")
```

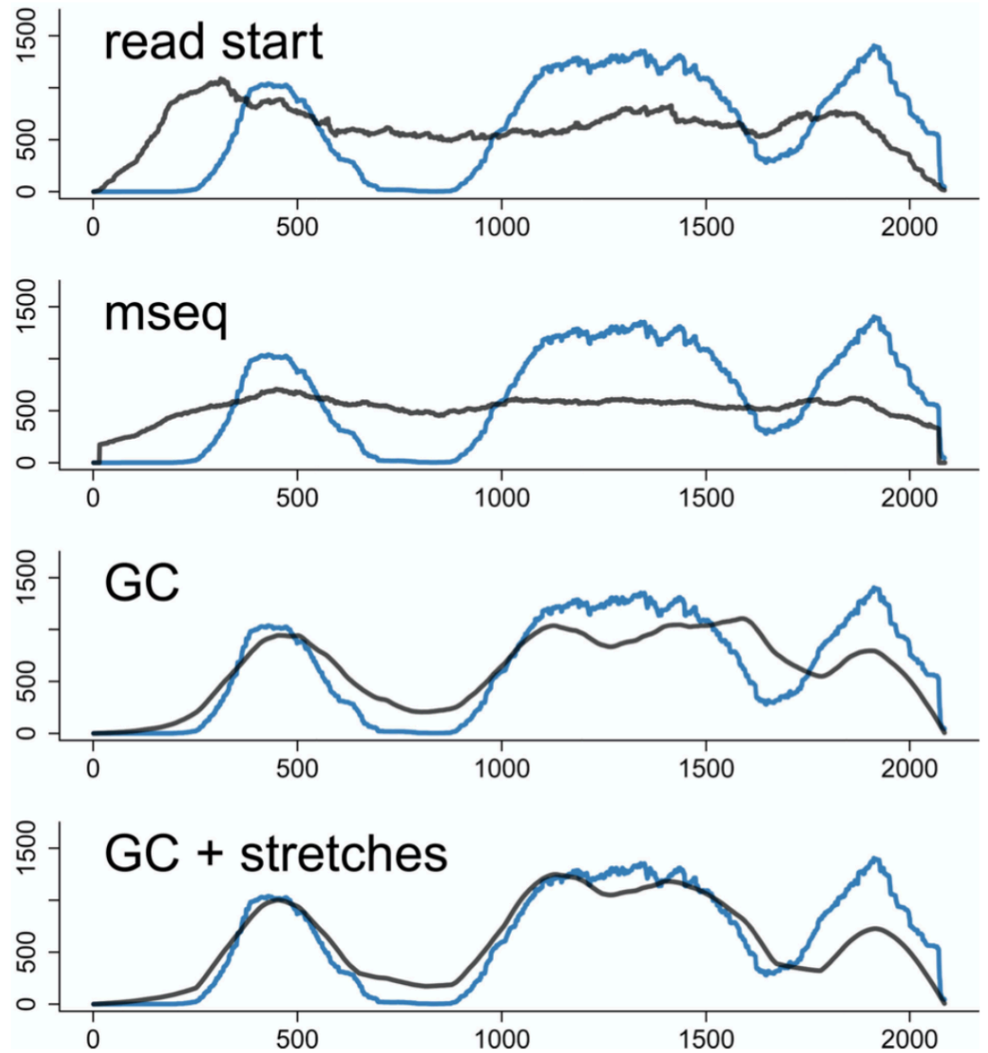
Estimate GC content bias ($f_j(\mathbf{gc}_i)$) with smooth linear regression

Poisson GLM Fits ($\hat{f}_j$) with cubic splines



We often center the correction offsets at 1 by dividing the sample mean $(\hat{f}_j(gc_i)/\text{mean}(\hat{f}_j(gc_i)))$ for identifiability.

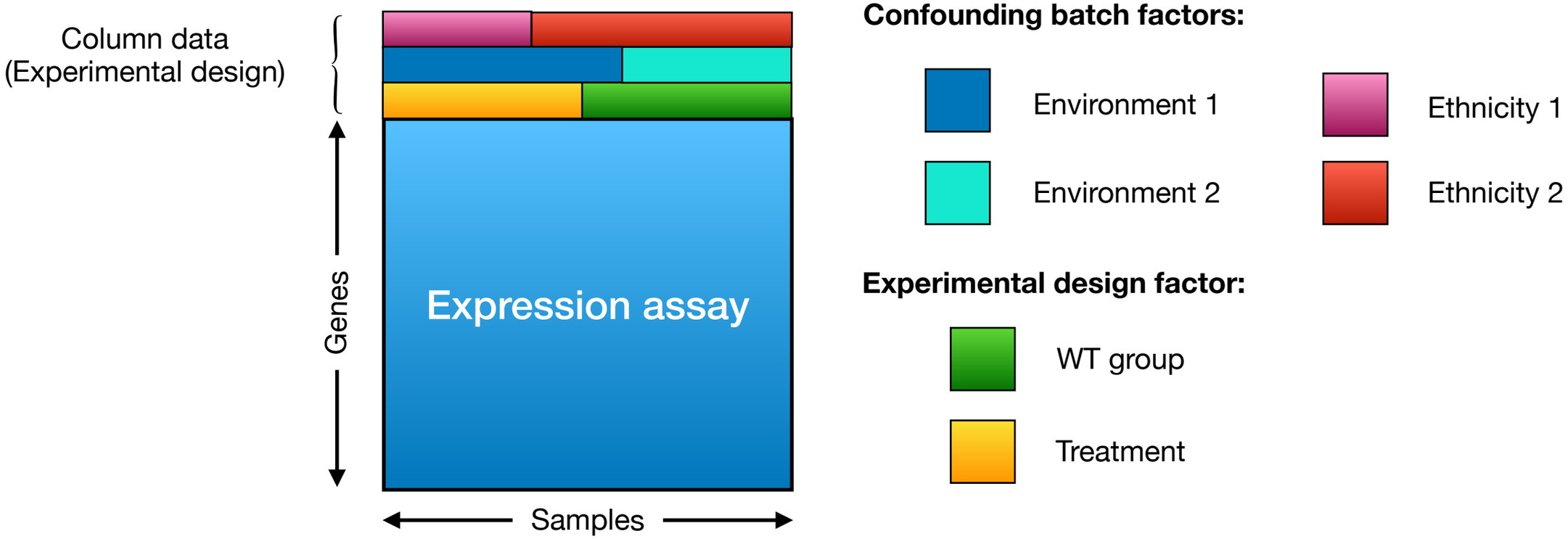
Fragment coverage predicted by different GLM models



Love, M. I., Hogenesch, J. B., & Irizarry, R. A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nature biotechnology, 34(12), 1287.

Combat and SVA

Batch effect factors may beyond accountable technical artifacts



- Batch effects in genomics can be caused by both technical factors and untracked biological factors.
- Technical factors are easier to adjust after understanding the generation mechanism of technical artifacts.
- Untracked biological factors, such as age, ethnicity, environmental factors, and epigenetic differences, can confound with the factor of experimental design.
- Adjusting for bio-based confounding factors is harder since they affect the true biological signals.

Linear regression representation

Samples	Gene i's expression		Intercept		Experimental Treatments		Batch factors (e.g. date, age)		i.i.d Gaussian noise
Treated rep 1	$\begin{bmatrix} 28.23 \\ 15.36 \\ 22.53 \\ 10.11 \\ 8.73 \\ 3.49 \end{bmatrix}$	$=$	$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$	$\mu +$	$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} +$	$\begin{bmatrix} 0 & 18 \\ 1 & 20 \\ 1 & 18 \\ 0 & 30 \\ 0 & 10 \\ 1 & 60 \end{bmatrix}$	$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} +$	$\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$
Treated rep 2									
Treated rep 3									
Control rep 1									
Control rep 2									
Control rep 3									

Matrix format: $y = \mu + \text{Treatment}\ \alpha + \text{Batch}\ \beta + e$

y: a vector of normalized* gene expression levels for gene i.

Treatment : design matrix for experimental treatments (e.g. treatment v.s. control).

α: an unknown vector containing effects of experimental treatments.

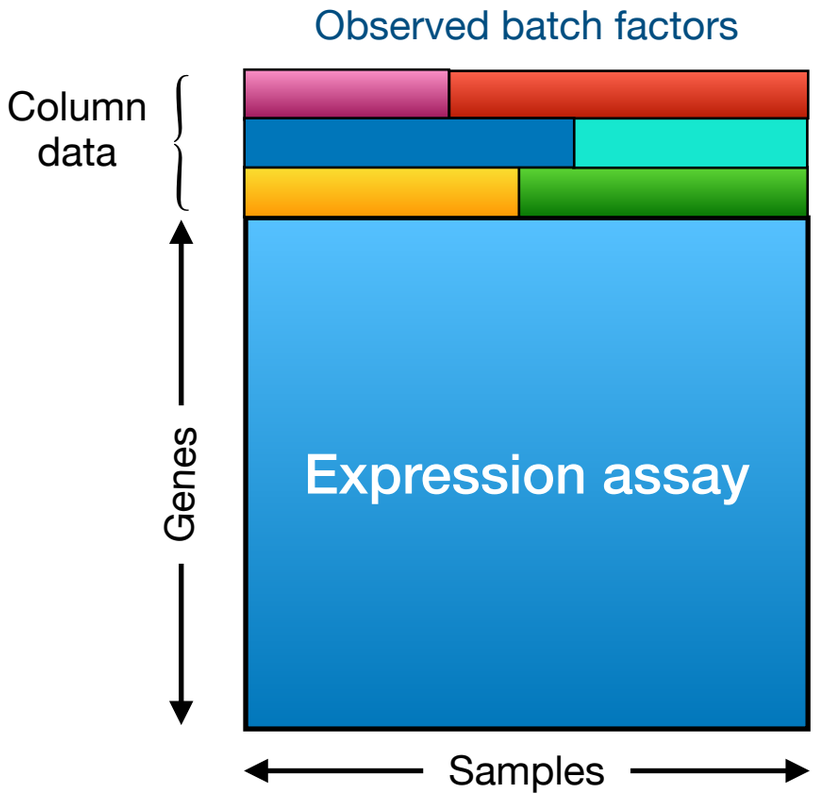
Batch : Design matrix for confounding batch factors.

β: effects of batches.

e: a vector for random Gaussian error.

*For NGS data, minimal normalizations should be: sequencing depth > log > row z-score.

Supervised batch effect modeling: combat



Combat like approach (Regression only):

$$y = \mu + \text{Treatment } \alpha + \text{Batch } \beta + e$$

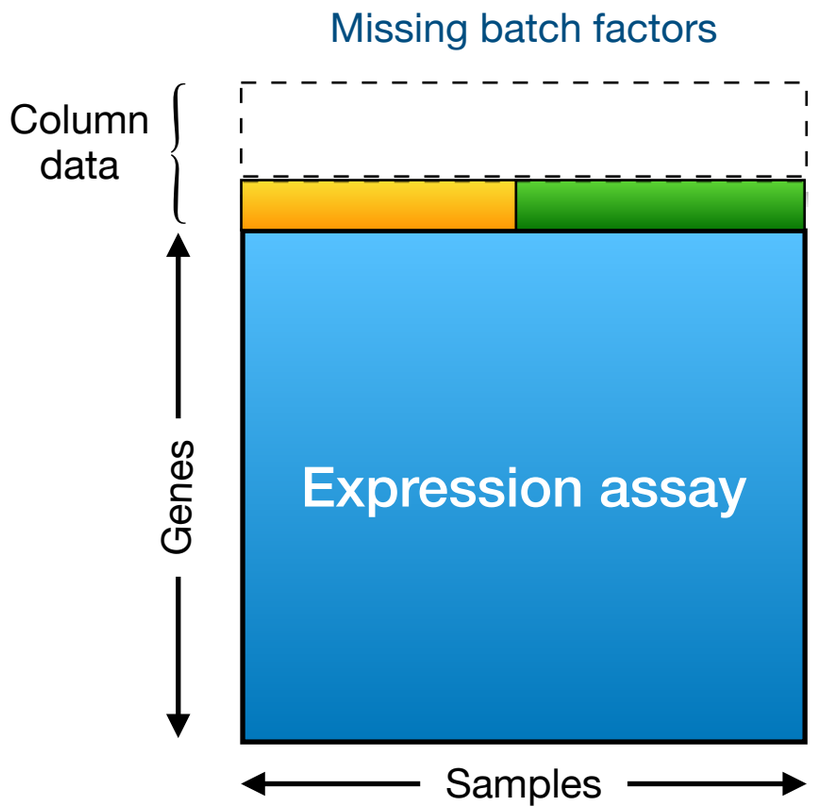
$$\hat{y}^{\text{correct}} = y - \text{Batch } \hat{\beta}$$

Fit linear regression first to estimate $\hat{\beta}$

Subtract the batch effect term from the gene expression vector to get the corrected gene expression

- **Combat** is a method used to correct for batch effects when we know the key confounding factors that are causing the batch effects.
- It works by fitting a multiple linear regression model to the gene expression data, where both the known confounding factors and the experimental design factors are used as covariates in the model.
- The model then estimates the effect of each covariate on the gene expression data and removes the unwanted variation due to the known confounding factors.

Unsupervised batch effect modeling: SVA



SVA like approach (1. PCA, 2. Regression):

Step1:

Use PCA / SVA to estimate Batch from the entire gene expression matrix

Step2:

$$y = \mu + \mathrm{Treatment}\ \alpha + \hat{\mathrm{Batch}}\ \beta + e$$
$$\hat{y}^{\mathrm{correct}} = y - \hat{\mathrm{Batch}}\ \hat{\beta}$$

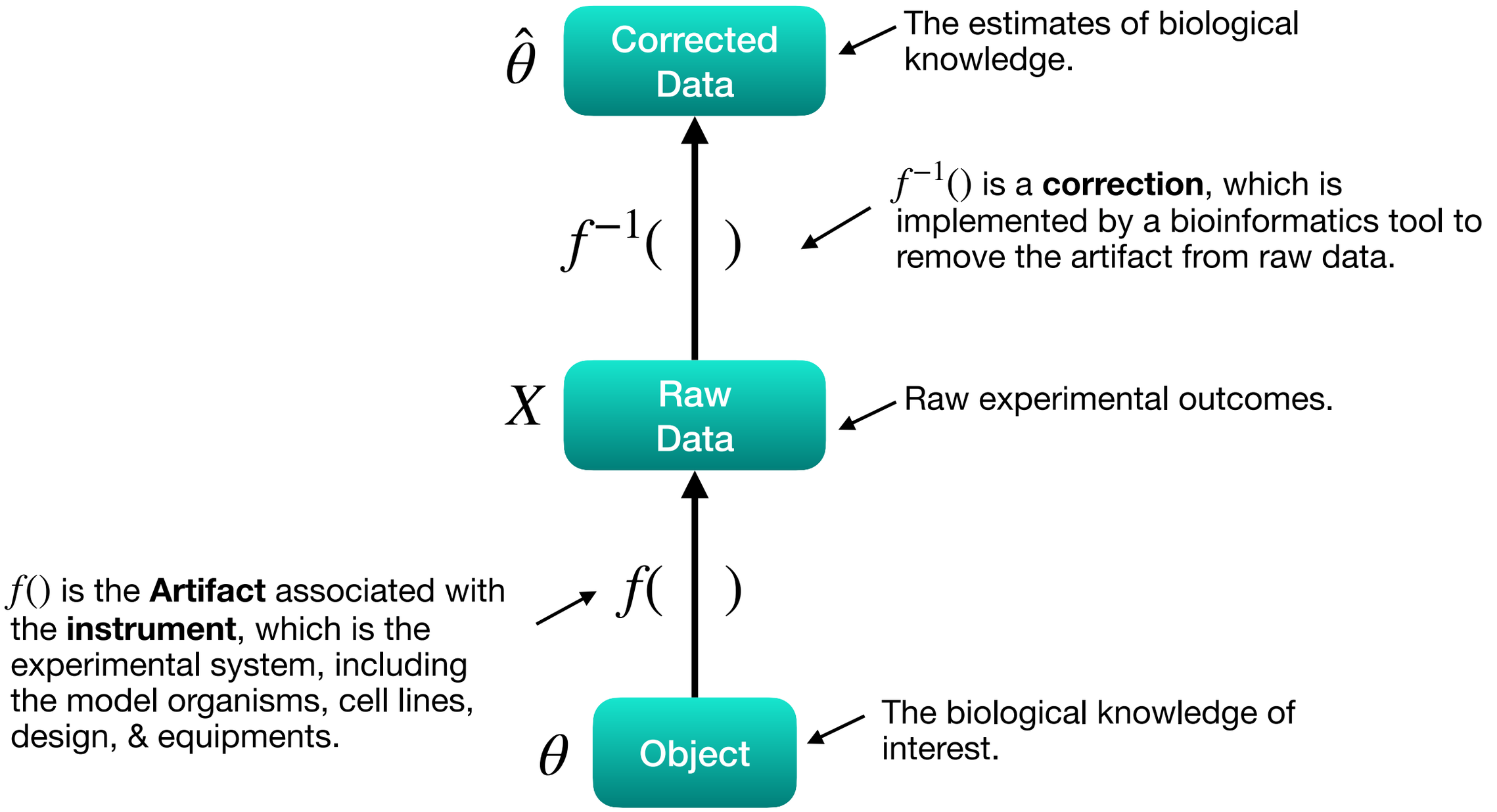
• Unsupervised methods are used when batch factors are unknown and cannot be directly accounted for.

• These methods estimate the "latent" batch factors using techniques like PCA or other factor analysis algorithms.

• **Surrogate Variable Analysis** (SVA) is a sophisticated form of PCA that can estimate batch effect factors while also isolating the influence of experimental design factors.

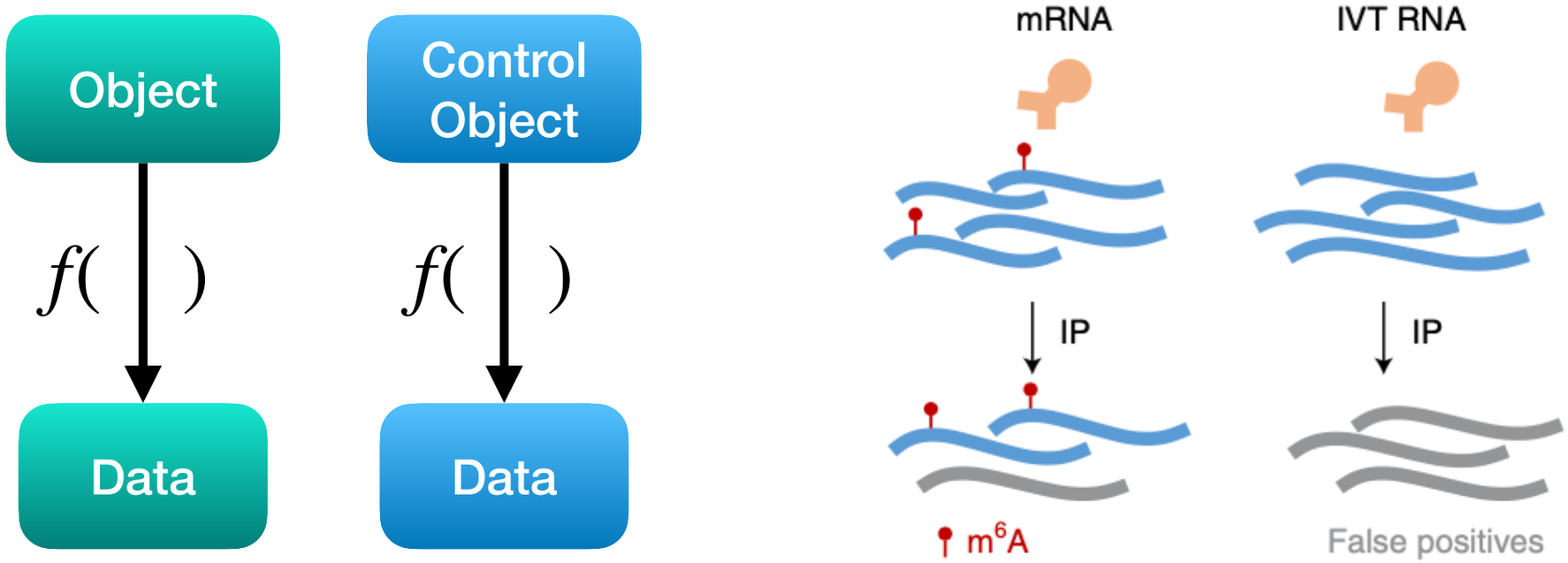
Control experiment

How to understand artifacts under a big picture?



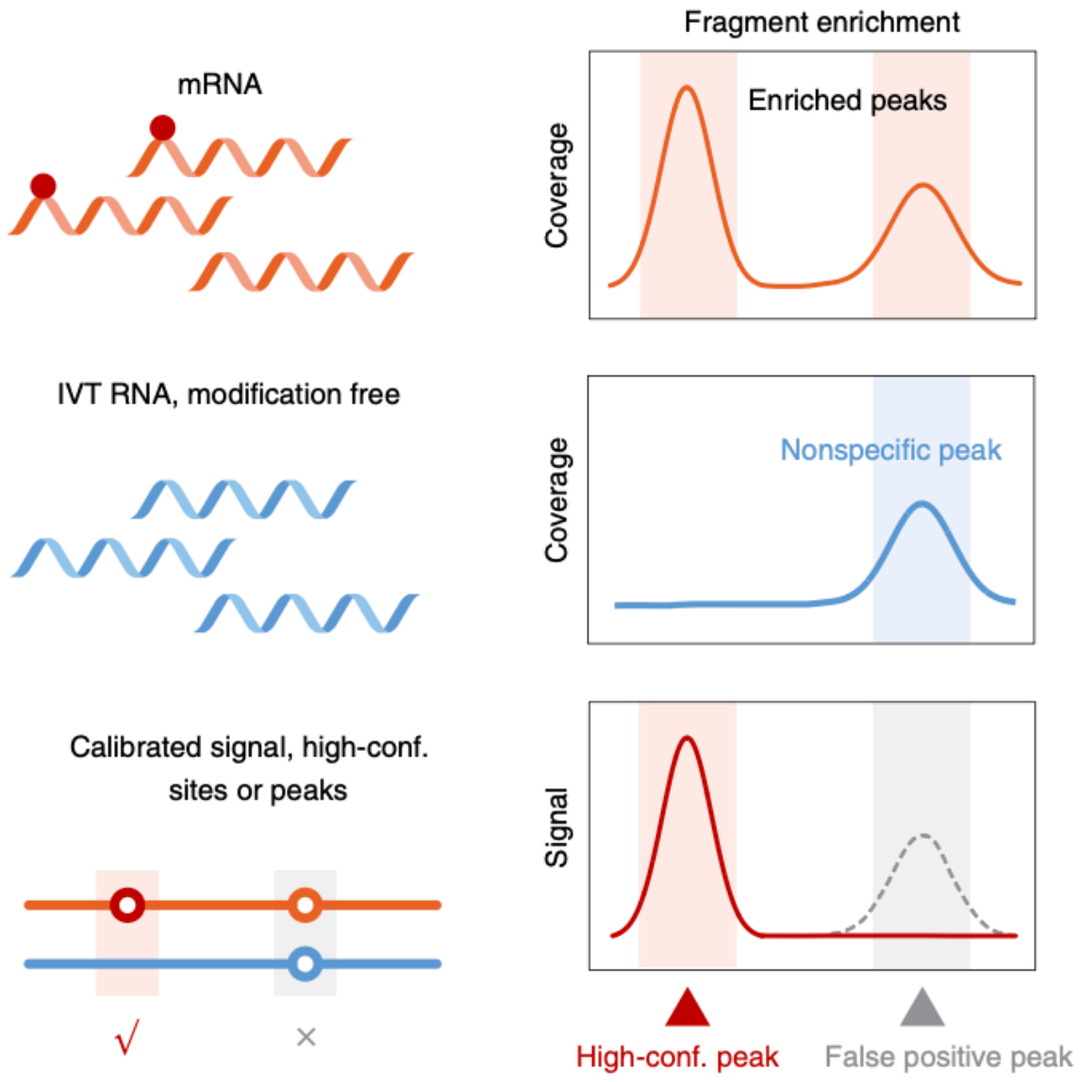
How can we know $f()$ exactly?

Control experiments



- A strong approach to identify artifact is to run a control experiment.
- When object is known, we can learn artifact $f()$ by observing the deviations in data.

Calibration of antibody unspecific binding by control experiment



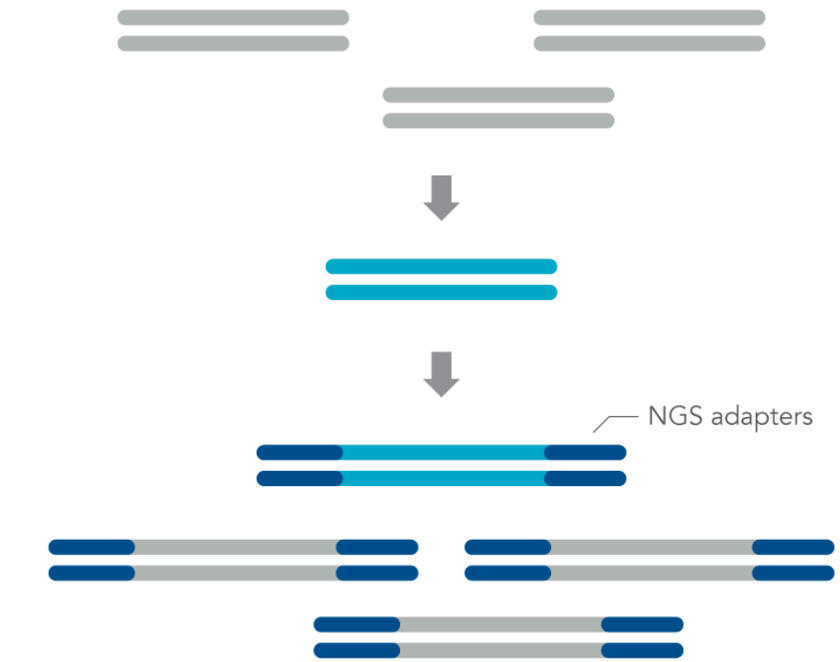
Zhang Z, Chen T, Chen H X, et al. Systematic calibration of epitranscriptomic maps using a synthetic modification-free RNA library[J]. Nature Methods, 2021, 18(10): 1213-1222.

Spike-in control in NGS

Sample DNA/RNA:

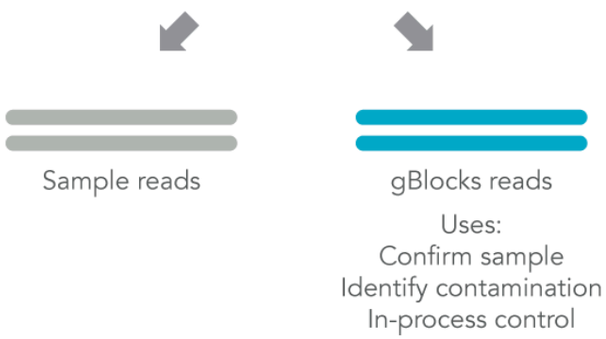
Spike-in control
(known abundance/# of copies):

Library preparation:



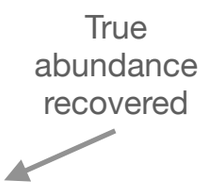
Sequencing:

Bioinformatics analysis
(mapping & normalization):



Using spike-in control to estimate exact sequencing depth

		Region A	RegionB
True relative abundance of DNA regions	Experiment sample:	3	3
	Control sample:	2	2
Raw data before normalization	Experiment sample:		
	Control sample:		
Normalization to total reads being the same for each sample (like RPKM)	Experiment sample:		
	Control sample:		
Normalization to spike-in being the same for each sample	Experiment sample:	$\times \frac{3}{2}$	
	Control sample:	$\times \frac{3}{2}$	



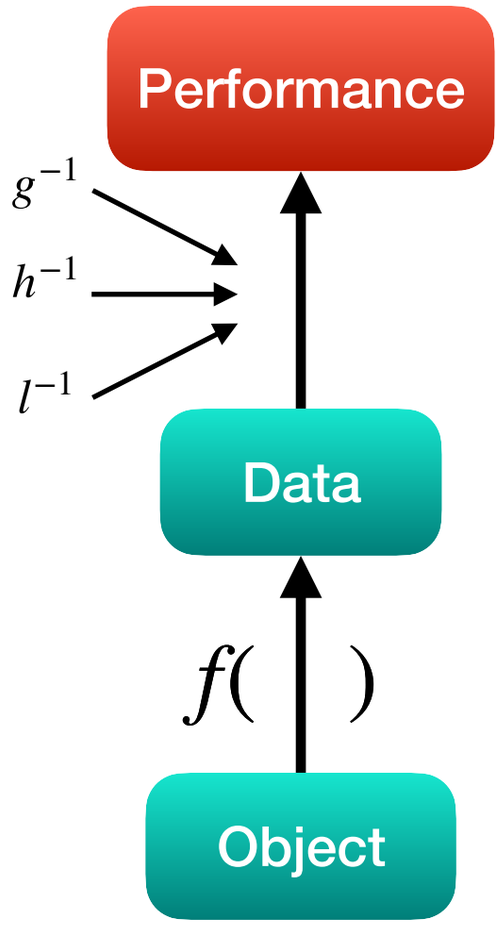
- When the same degree of change happens everywhere on the genome, normalizing total sequencing reads to the same number hides the change, whereas normalizing spike-in reads to the same number reveals the global change of read density.

Chen, Kaifu, et al. "The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses." *Molecular and cellular biology* 36.5 (2016): 662-667.

Most NGS experiments don't have control

Correction by trial and error

When control experiments are not available, we can propose different hypotheses ($g, h, l \ldots$) on the artifact f, and test which one perform best in a downstream analysis.



- When lacking control, the optimal correction pipeline is often discovered by trial and error.
- True theta and f are often not identifiable in such cases, as different combinations of theta and f can generate the same data X.
- As a result, the optimal correction methods are often different by different downstream applications, since they have different tolerances to different types of errors.

Take home messages

GC content bias correction is effective in copy number variation detection (a DNA-Seq application)

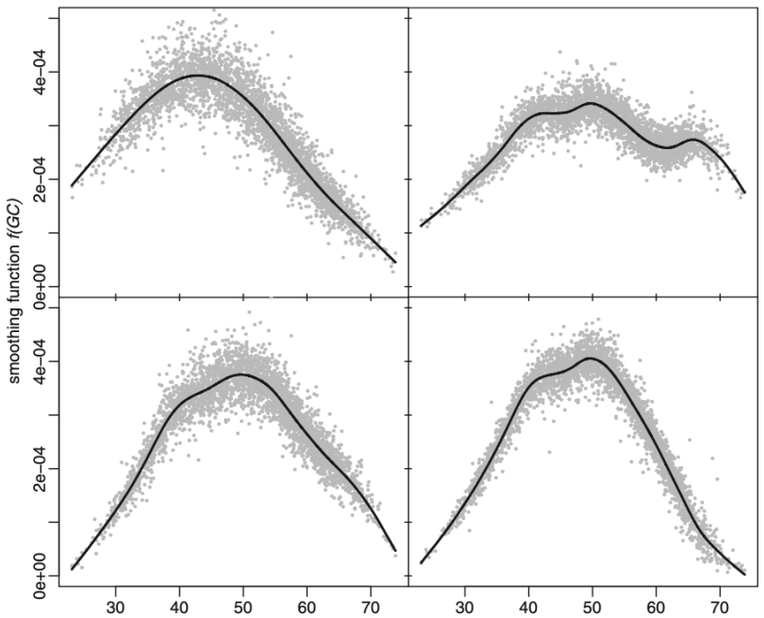
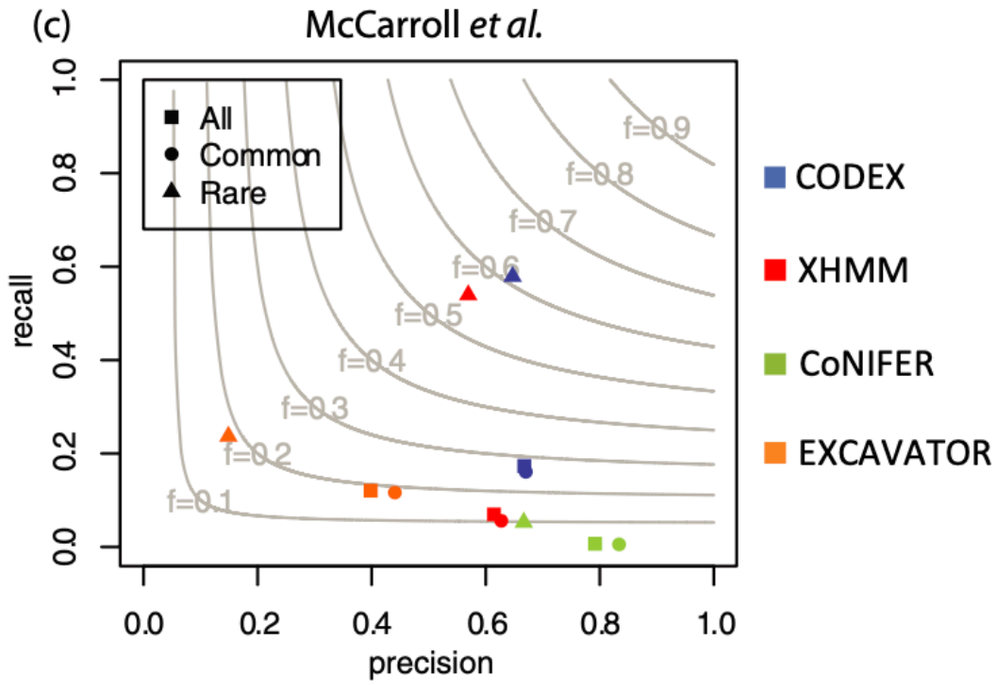


Figure 3. Predicted values of $f(GC)$ for four samples from the 1000 Genomes Project data set. Most patterns agree with previous observations that read depth has a unimodal relationship with GC content. However, dual modality is also observed. Furthermore, the function changes in shape and not just by a scaling factor.



Nucleic Acids Research, 2015, Vol. 43, No. 6 ***e39***
doi: 10.1093/nar/gku1363

1. In practice, choosing the right normalization and batch effect removal methods often lead to the most significant performance boost among all steps.

2. The normalization procedures introduced in Lec 5 and Lec 6 are generally useful for most types of genomic assays.

   E.g. DNA-Seq, RNA-Seq, scRNA-Seq, metagenomic sequencing, and CHIP-Seq can all benefit from GC bias correction and quantile normalization.