



AlgoBio WS 16/17

Genexpressionanalyse

Annalisa Marsico

14.12.2016



MILESTONE 21

An array of opportunities

Until the mid-1990s, studies of gene expression were limited to measuring transcription from one or a few genes. But then a tool arrived that changed all this, allowing the study of hundreds or thousands of transcripts at a time. This technology — the expression microarray — has revolutionized many areas of biology, from basic research to the understanding and treatment of human disease.

The study that brought microarrays to the attention of most researchers was published in 1995 by Patrick Brown and colleagues at Stanford University in California. They used an automated method to print 45 *Arabidopsis thaliana* cDNAs onto a glass slide. This 'array' was probed with a mixture of fluorescently labelled cDNAs that were derived from the reverse transcription of mRNAs extracted from a tissue sample. The amount of hybridization

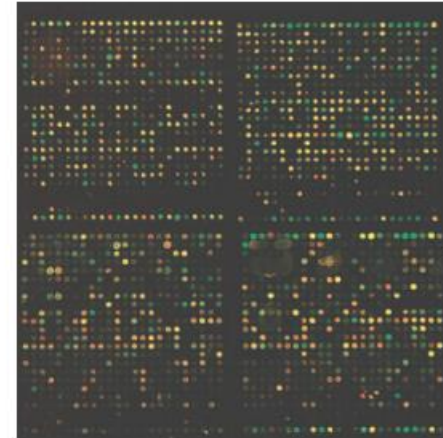
— and, therefore, the level of expression — was determined by measuring the intensity of fluorescence for each gene.

Although this was considered important in terms of the numbers of genes that could be studied, the fundamental implications for interpreting gene-expression patterns were not immediately appreciated. These became clear from later studies, such as a paper published in 1998 that used microarrays to monitor the expression of 800 genes throughout the yeast cell cycle. Subtle changes in overall gene-expression patterns over time were revealed that could not have been detected by other methods, even if dozens of genes had been studied.

By allowing different cell types to be accurately distinguished on the basis of expression patterns, microarrays have also had a

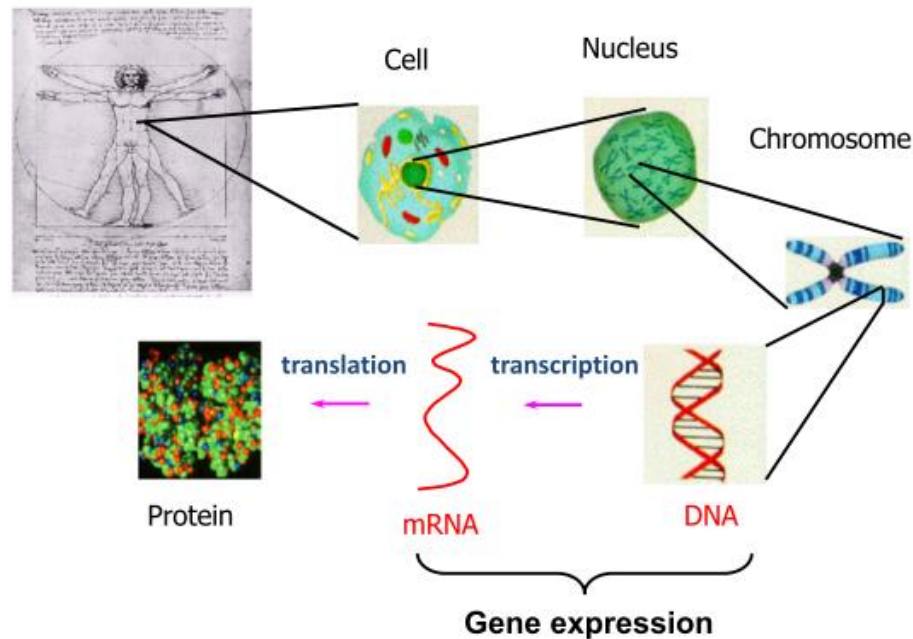
"The development of microarray technology has unequivocally revolutionized the way we examine and interpret gene expression."

James Broach



huge impact on medical research. In 2000, an impressive early example of this was reported, in which microarrays were used to study the expression of more than 8,000 genes in 65 human breast tumours. The authors generated 'molecular portraits' of gene expression that allowed them to distinguish between different classes of breast tumour and to identify two new categories that had been overlooked by traditional classification tools. They were also

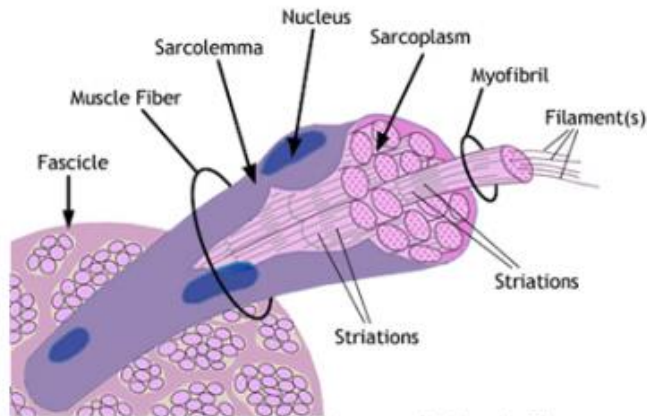
Mikroarrays messen die Genexpression



2

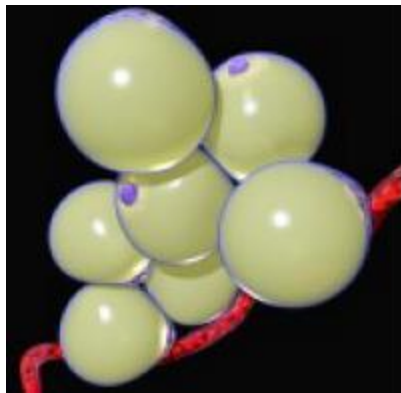
Warum ist es wichtig, die Genexpression zu messen?

Die Vielfalt der Zellen in einem Organismus

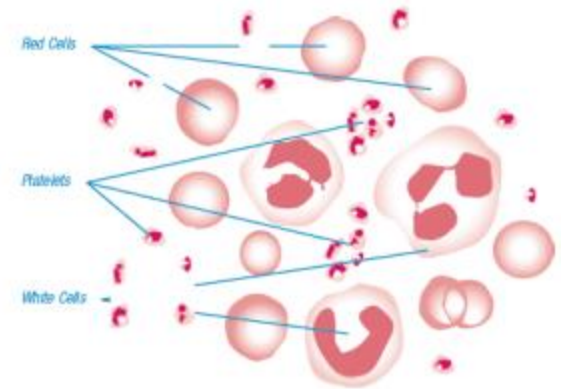


Above: Diagram of the Structure of a Muscle Fibre.

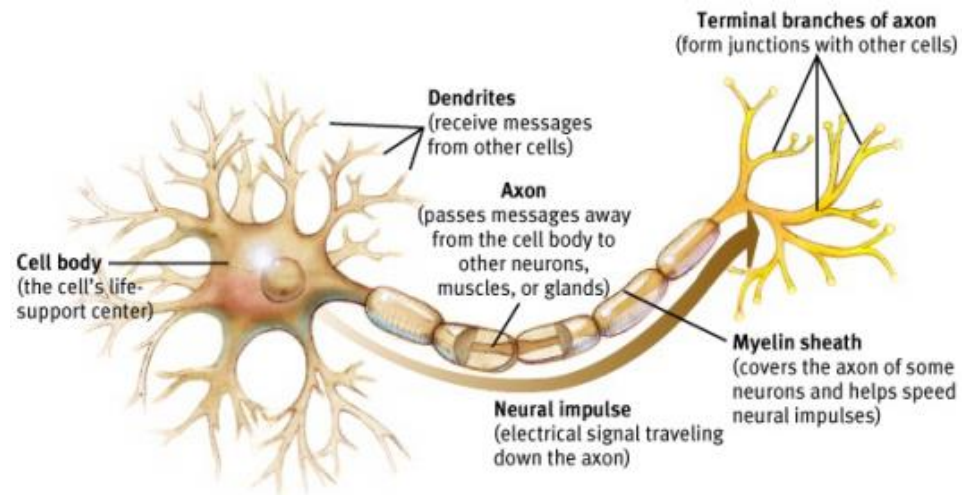
Muscle (fiber) cells



Fat cells



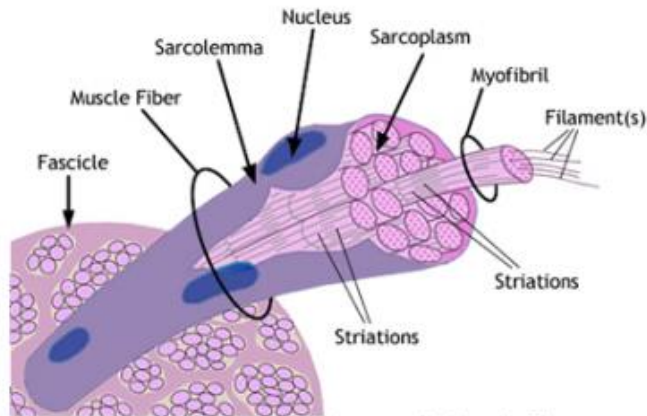
White blood cells



Neuron (nerve) cells

Wie produzieren verschiedene Zellen mit identischer DNA unterschiedlichen Proteinen?

The diversity of cells in an organism

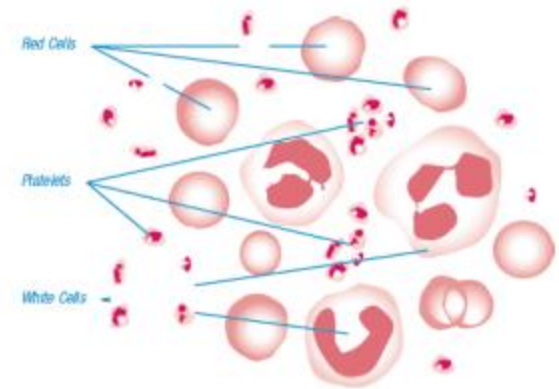


Above: Diagram of the Structure of a Muscle Fibre.

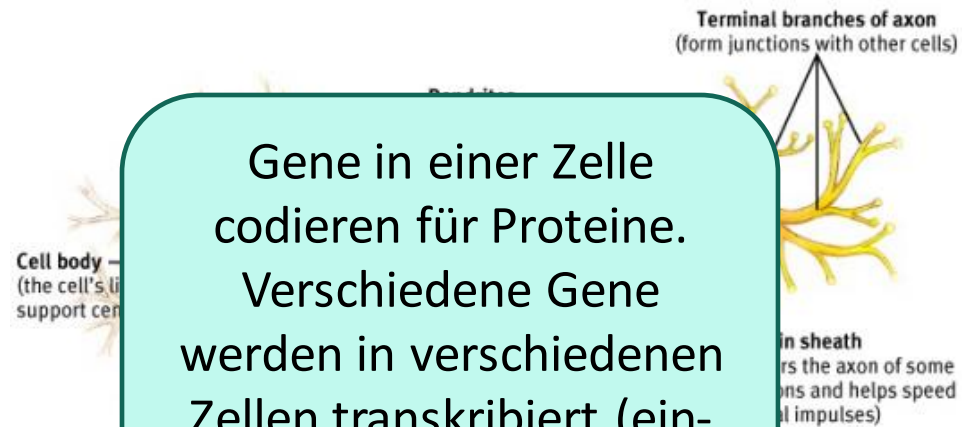
Muscle (fiber) cells



Fat cells



White blood cells



Gene in einer Zelle codieren für Proteine. Verschiedene Gene werden in verschiedenen Zellen transkribiert (ein- und ausgeschaltet)

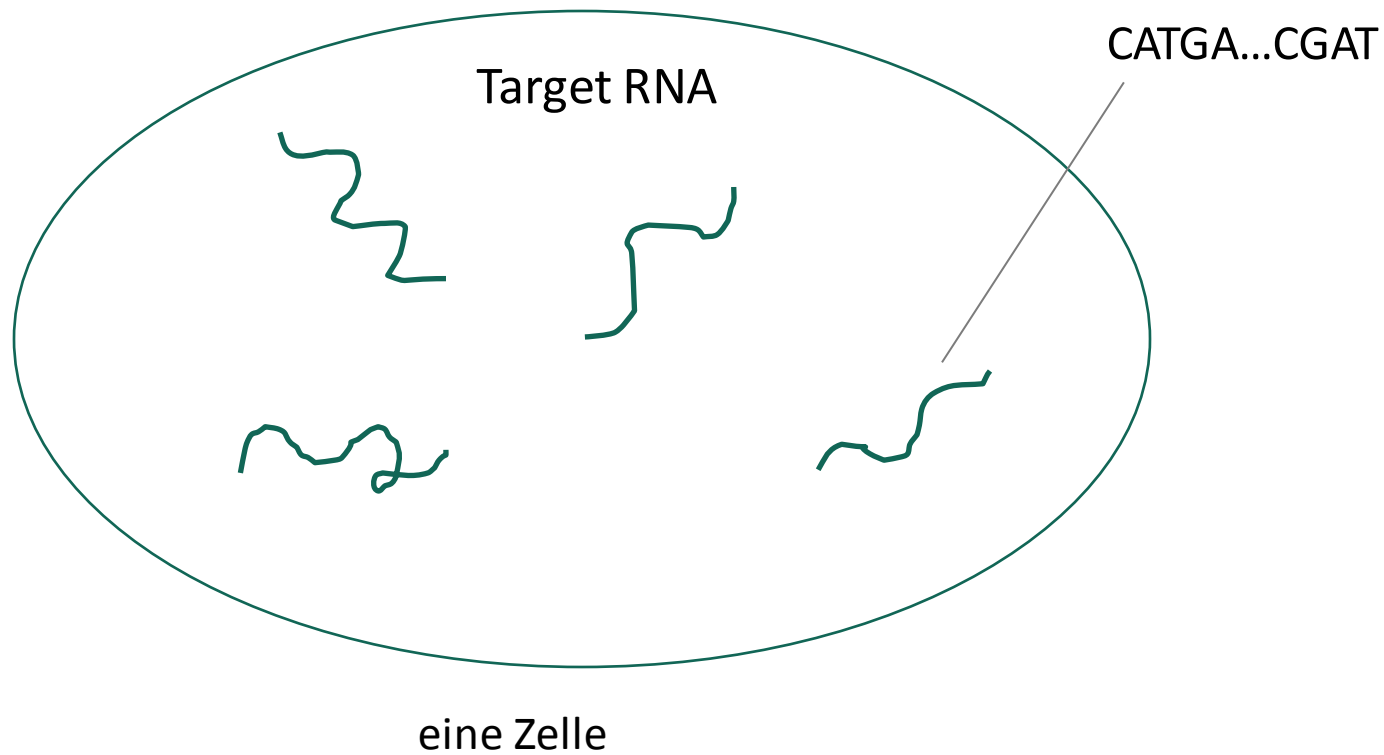
Neuron (nerve) cells

Wie produzieren verschiedene Zellen mit identischer DNA unterschiedlichen Proteinen?

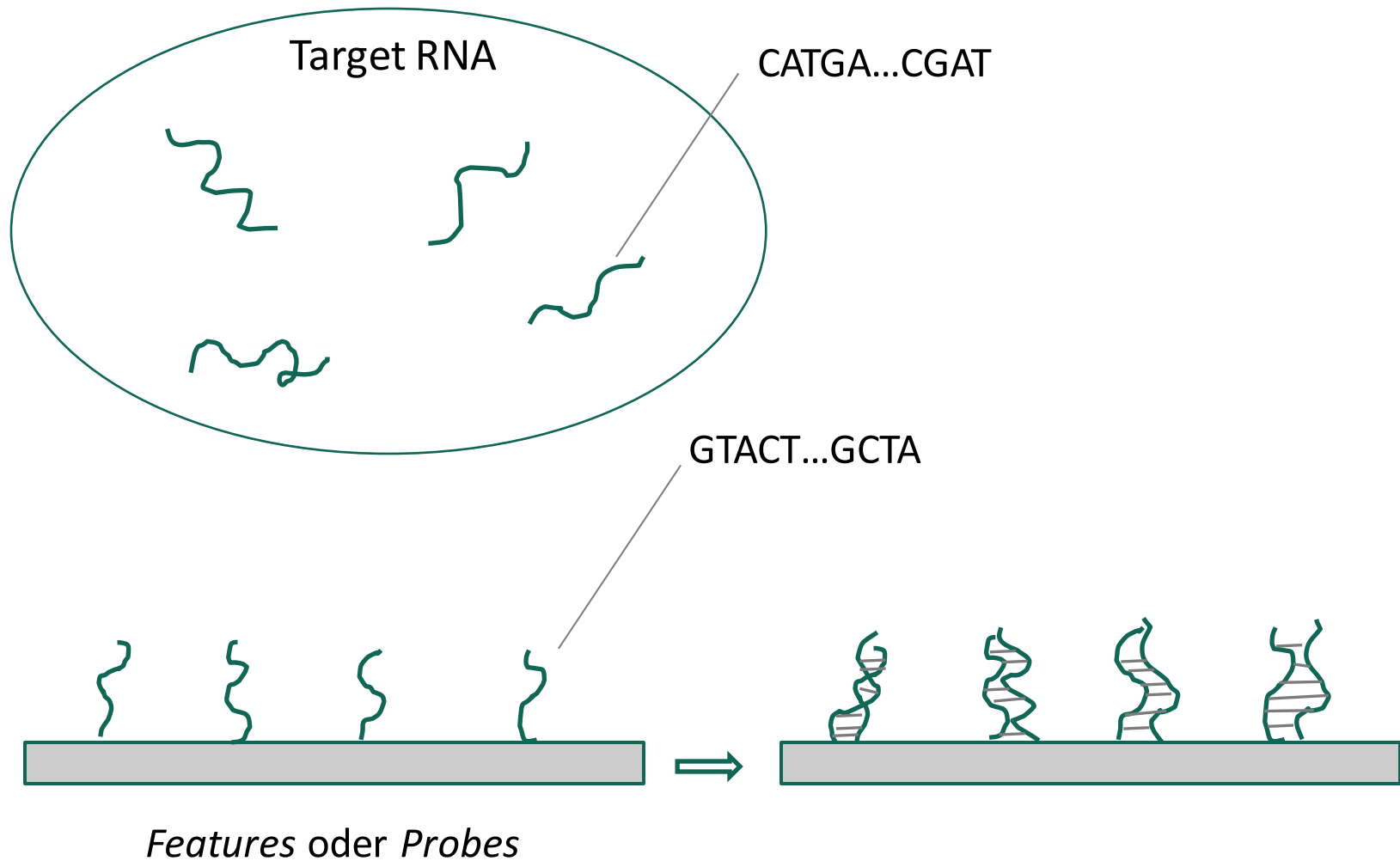


- Zellen unterscheiden sich aufgrund der differentiellen Genexpression (in verschiedenen Organen verschiedene Gene exprimiert werden)
- Ein Gen wird durch Transkribieren von DNA in viele Kopien von mRNA exprimiert
- mRNAs werden dann in Proteinmoleküle translatiert
- **Mikroarrays messen den mRNAgehalt (d.h. Konzentration der mRNA) und somit der Ebene der Genexpression**

Grundidee: Hybridisierung



Hybridisierung





Einige Terminologie:

- **Features or probes** bezieht sich auf die Moleküle an der Oberfläche befestigt
- **Target** bezieht sich auf die Probe, die die Transkripte die wir messen wollen enthält

Verschiedene Arten von Microarray-Technologie. Hauptunterschied:

- **One color microarray** (Ein Sample zu einer Zeit wird hybridisiert)
- **Two-colors microarray**



Single-channel arrays

Affymetrix GeneChip



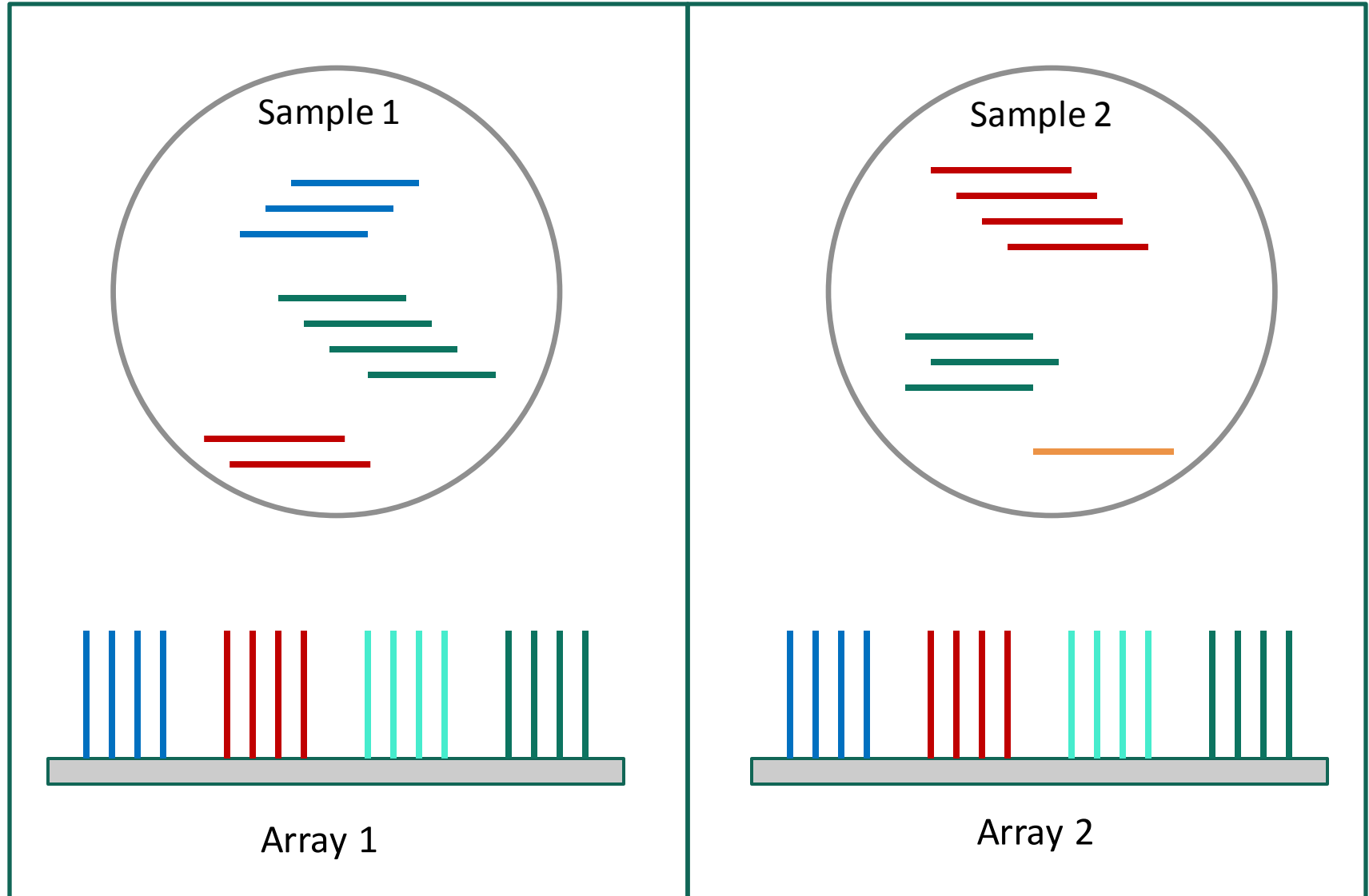
- Gewerbliche Massenproduktion
Oligonucleotid-Array mit hoher Dichte
Technologie von Affymetrix entwickelt

<http://www.affymetrix.com>

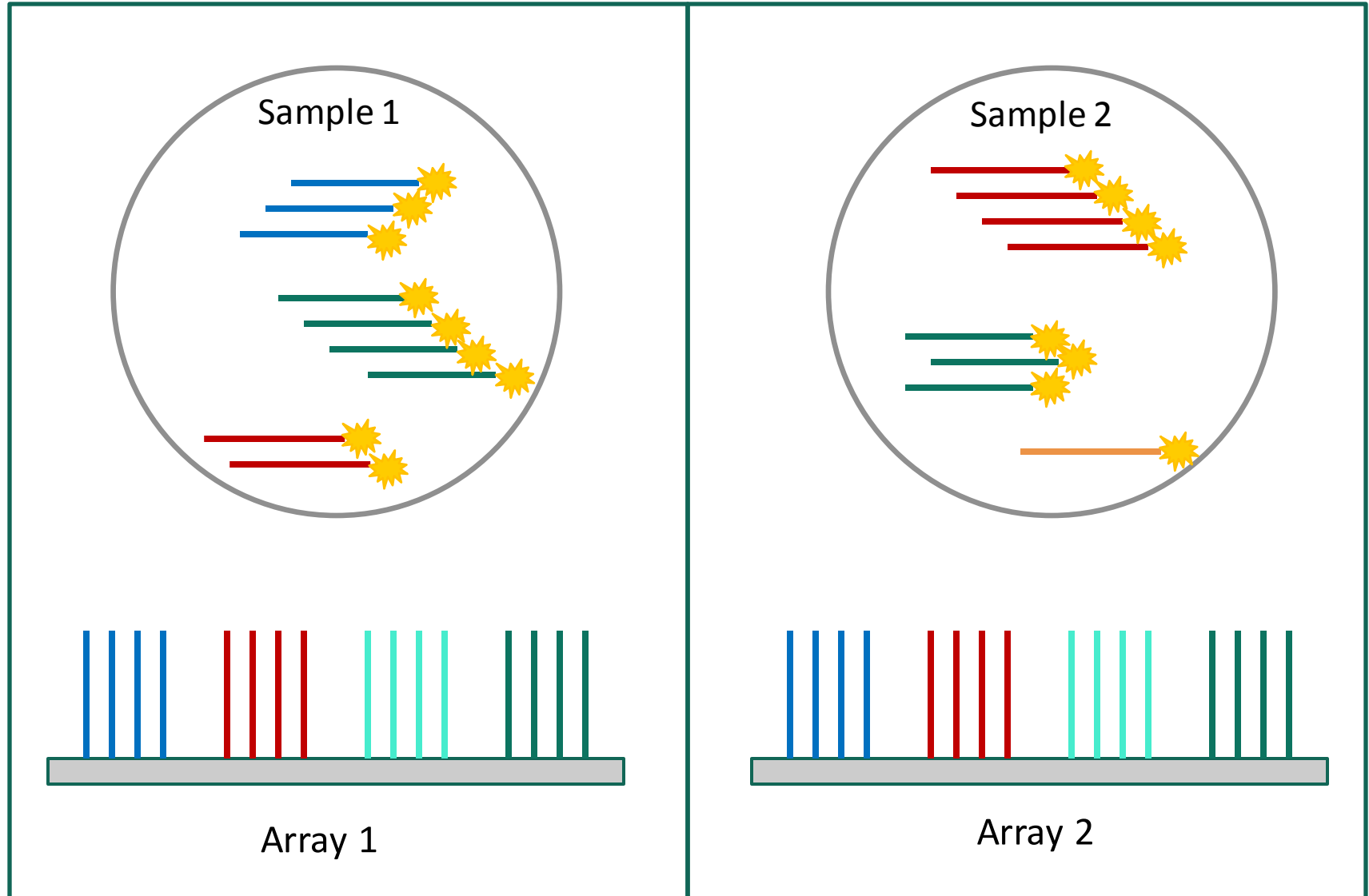
- Single channel microarray



Vor der Kennzeichnung



Kennzeichnung mit fluoreszierendem Farbstoff

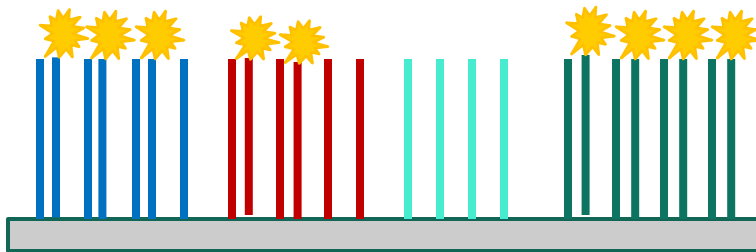


Hybridisierung

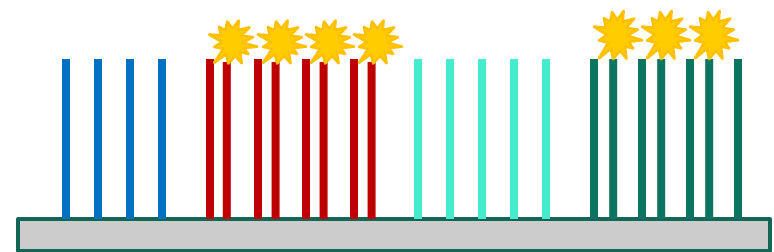


Sample 1

Sample 2



Array 1

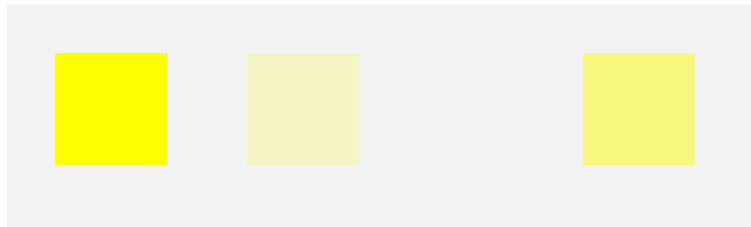


Array 2

Scannerbild

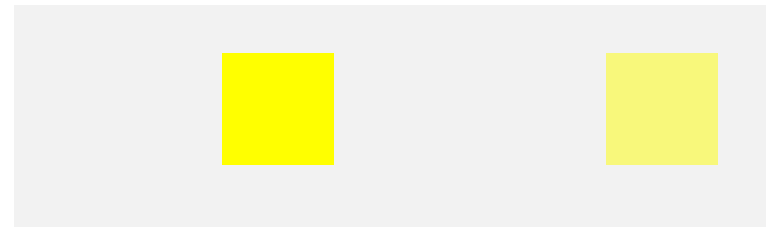


Sample 1



Array 1

Sample 2

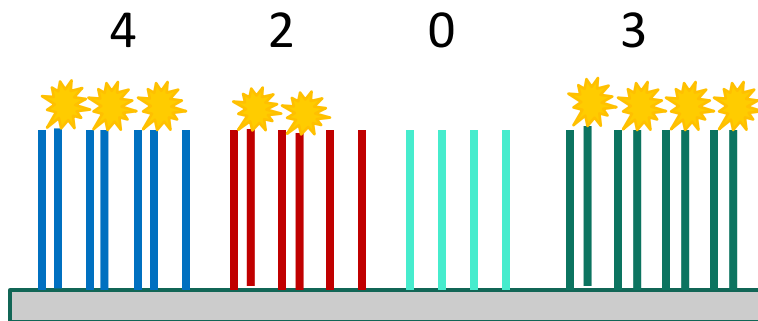


Array 2

Quantifizierung

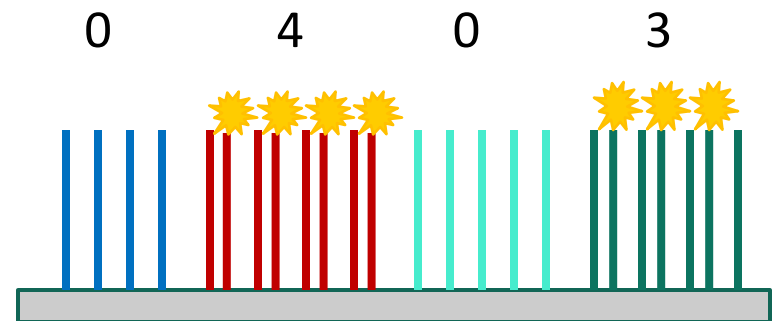


Sample 1

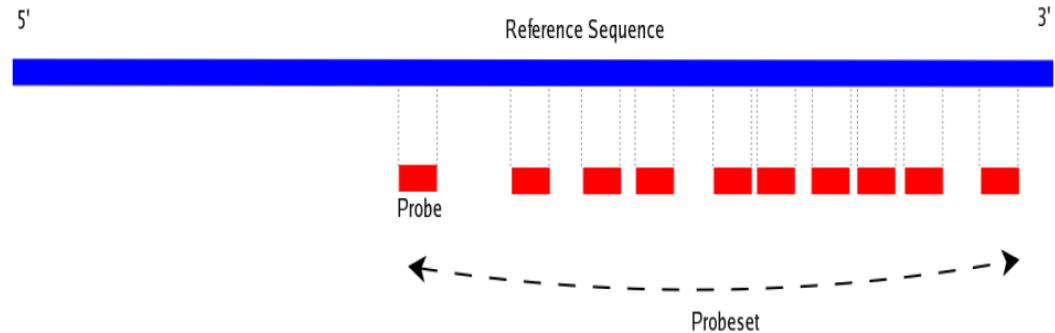
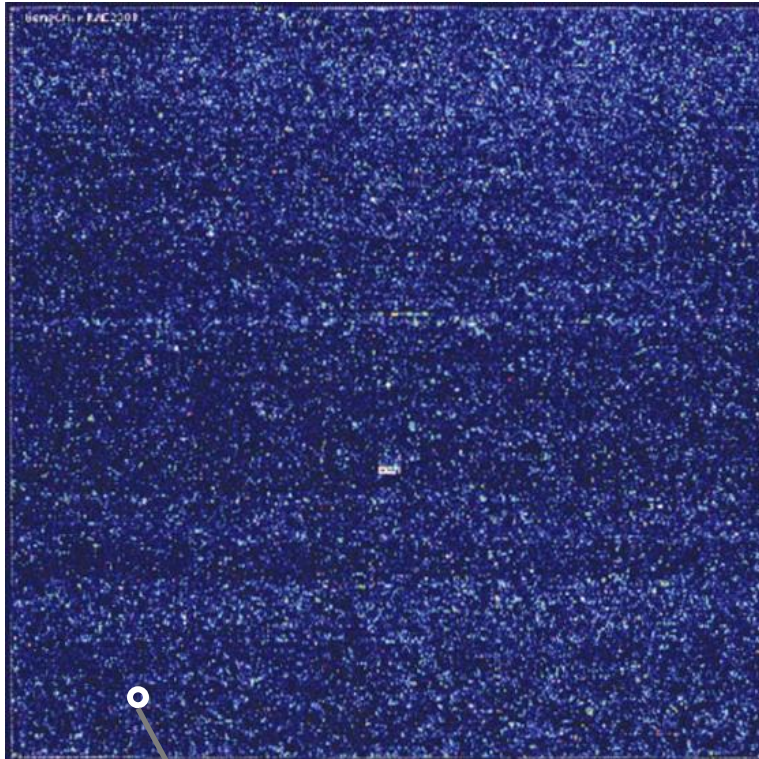


Array 1

Sample 2



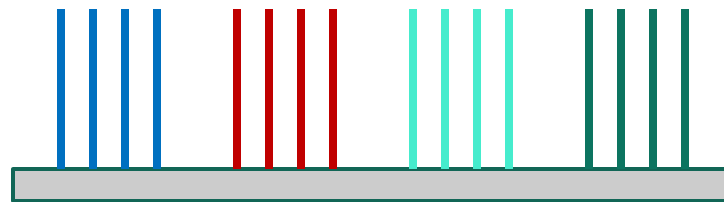
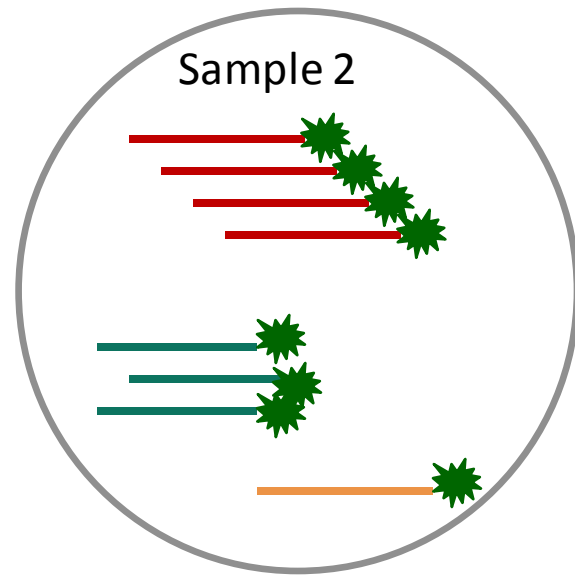
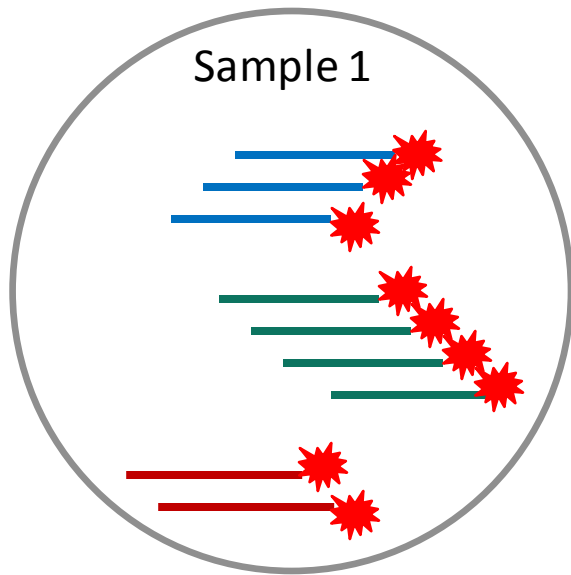
Array 2



Typischerweise 11 Probes in einem Probeset
Neueste GeneChip haben so viele wie:
54,000 probesets
1.3 Million probes

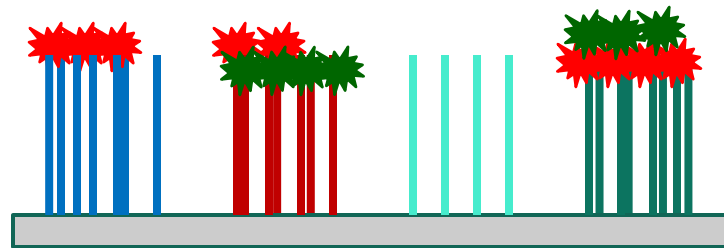
Eine Zelle auf dem Chip ist eine Probe oder ein Probeset.

Vor der Hybridisierung: Labeling



Array 1

After Hybridization

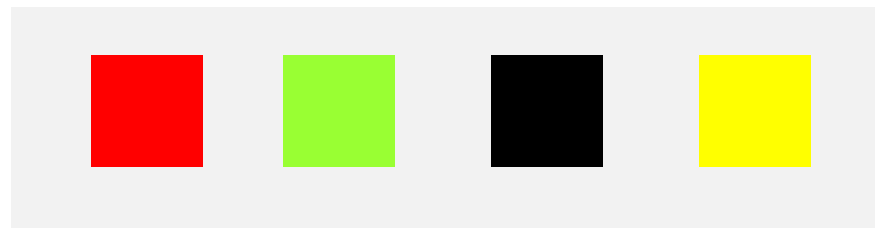


Array 1

Scannerbild



Roten Kanal	3,0	2,0	0,0	4,0
Grünen Kanal	0,0	4,0	0,0	3,0

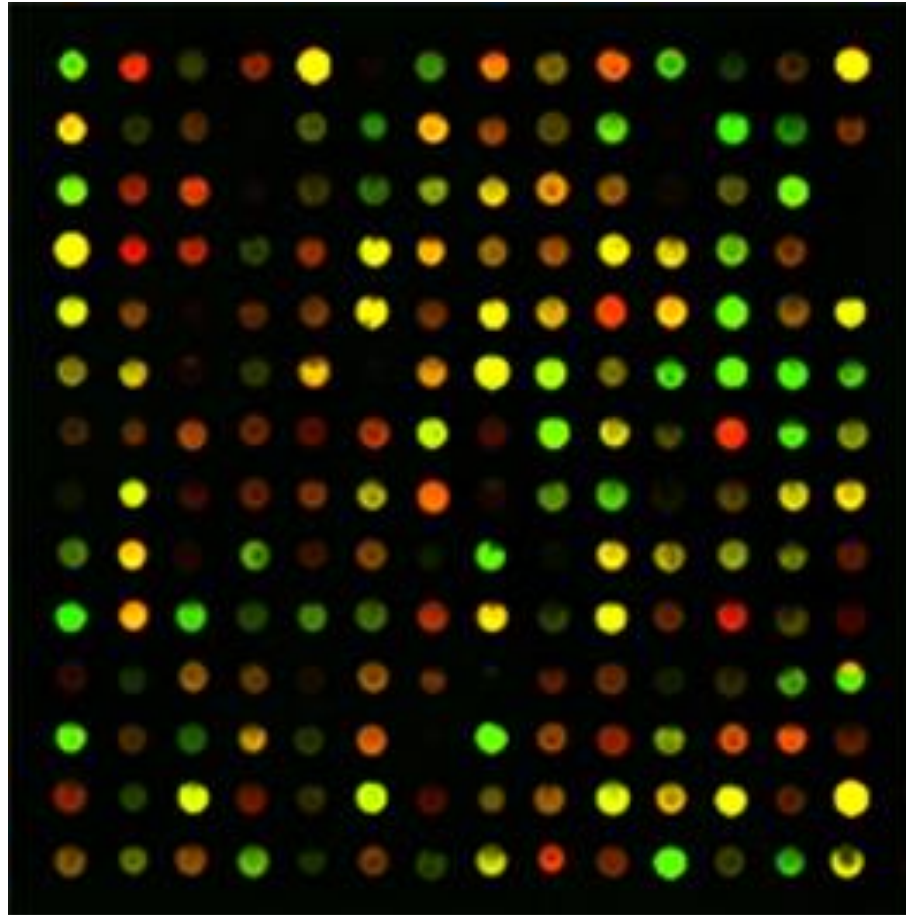


Array 1

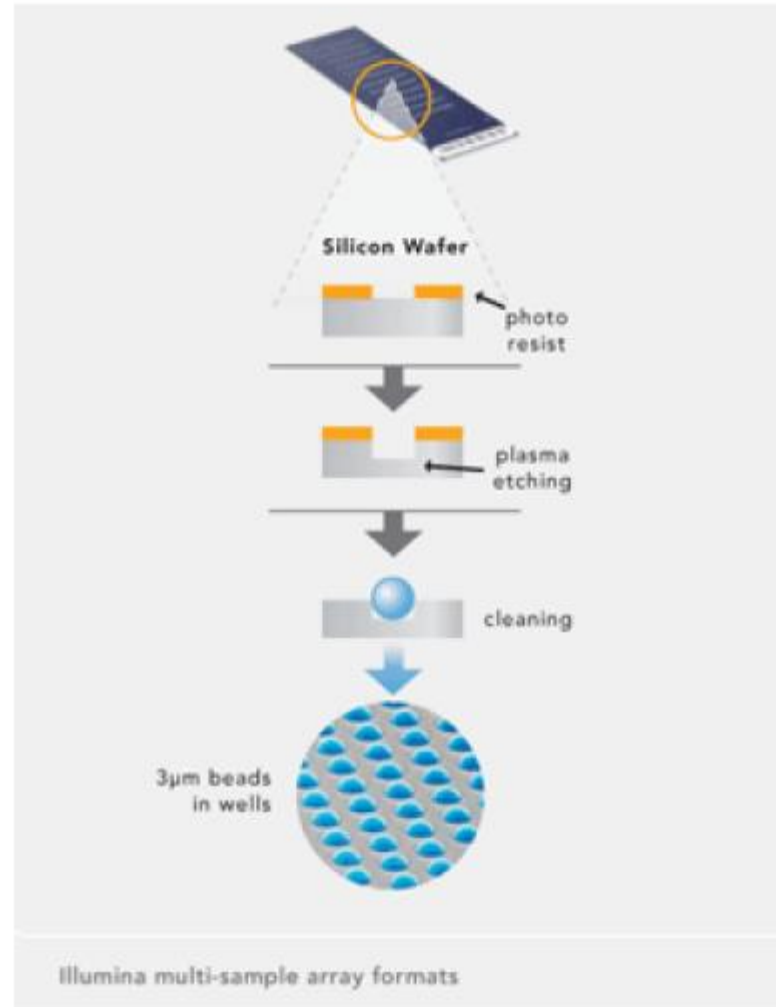
Microarray Image



Typisches Bild eines *Two-channels Microarray*



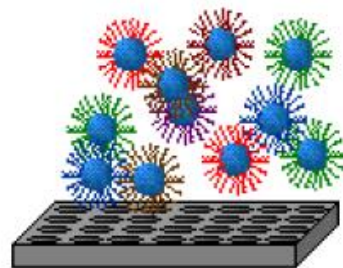
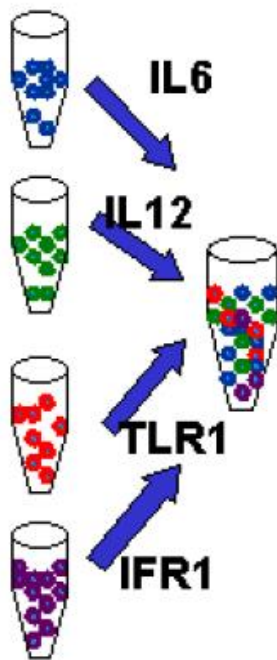
Illumina Bead Arrays



Illumina Bead Arrays



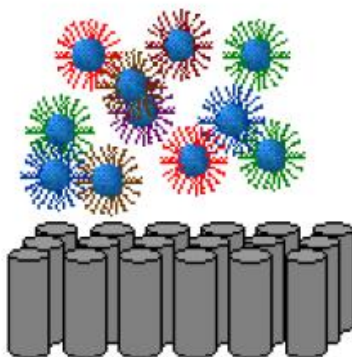
3 micron diameter beads are coated with Oligonucleotides



Beads embedded in a slide with multiple arrays located on each slide



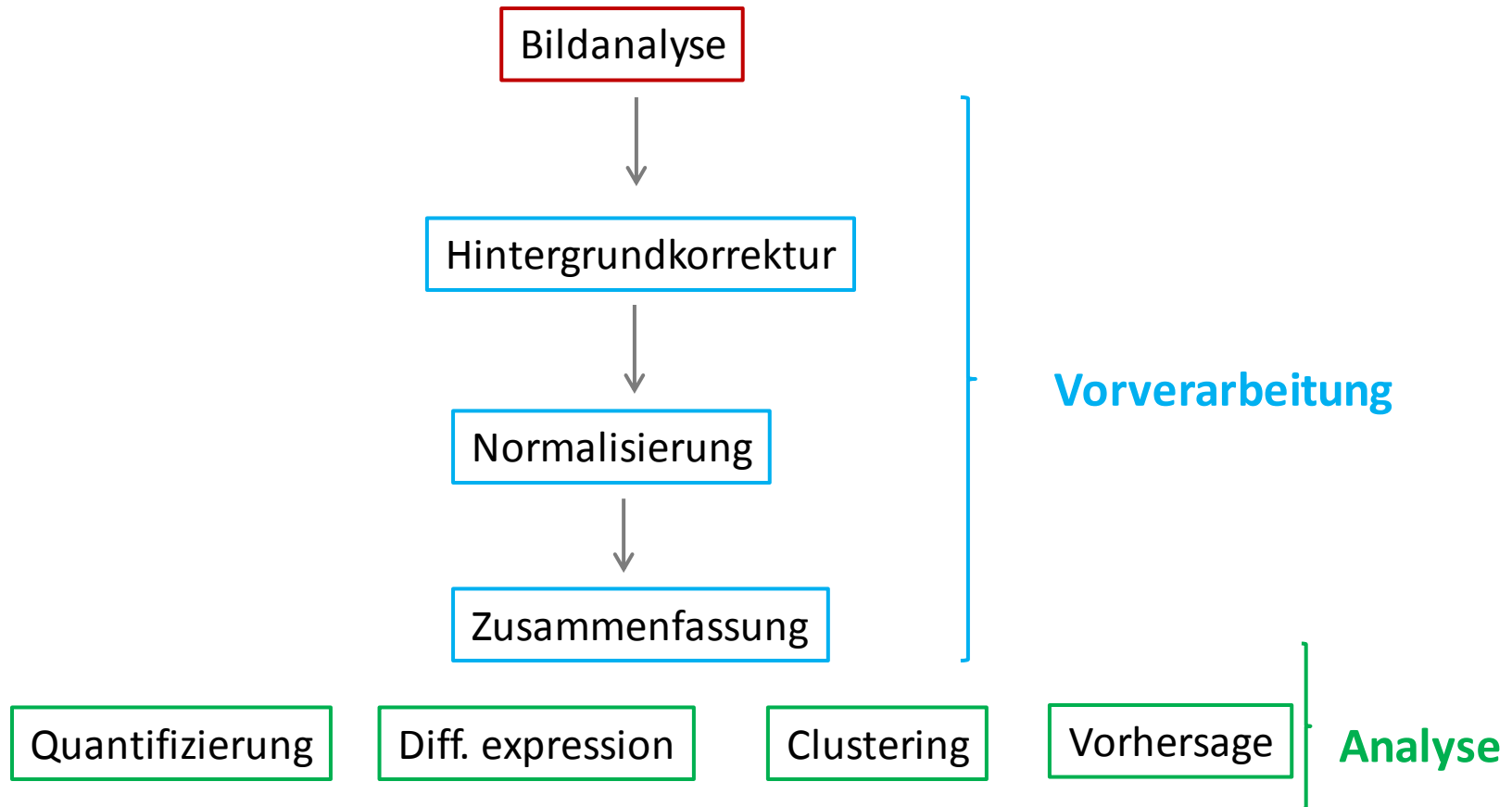
Pooled Beads are randomly located and assayed to identify the location of each bead within the array using a 29 base tag sequence



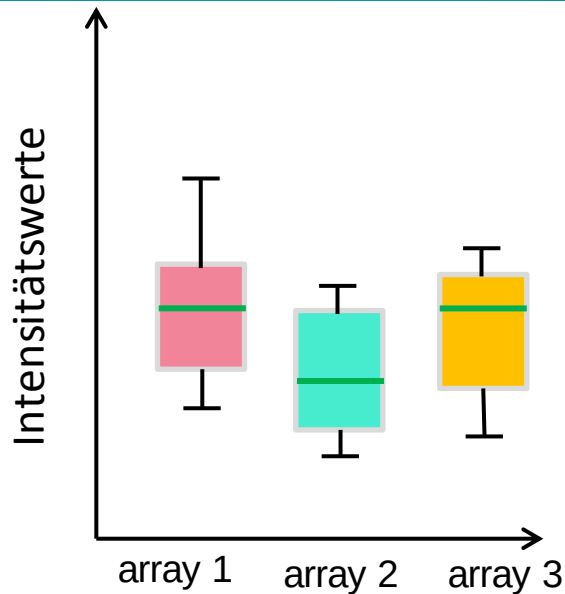
Beads located in the end of a microfibre and collected into a bundle each bundle in a 96 sample/array matrix format



Pipeline für die Mikroarray-Analyse



Systematische Unterschiede zwischen Arrays



Die Boxplots zeigen Verteilungen von Log-Intensitätswerten aus drei Arrays Hybridisierten mit der gleichen RNA.

Und sie sind global verschieden von einander

Experimentelle Variation kann aus vielen Quellen kommen:
verschiedene Scanner, Labeling, Verschiedene RNA-Mengen in Samples

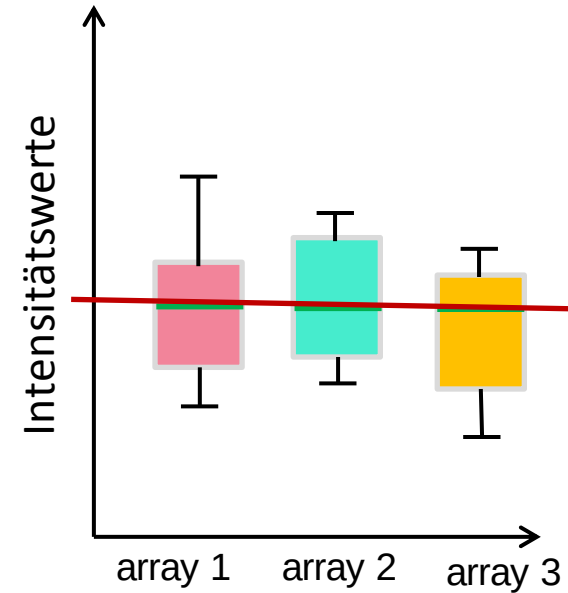
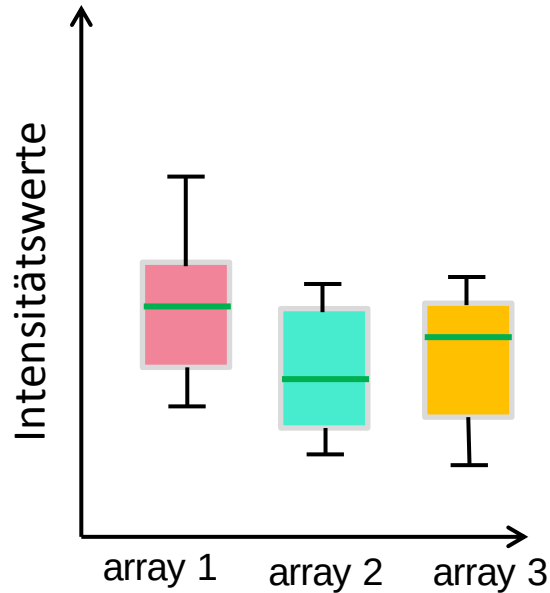
Systematische Variation:

- Ähnliche Effekte auf viele Messungen (viele Gene)
- Korrekturen können auf den Daten abgeschätzt werden



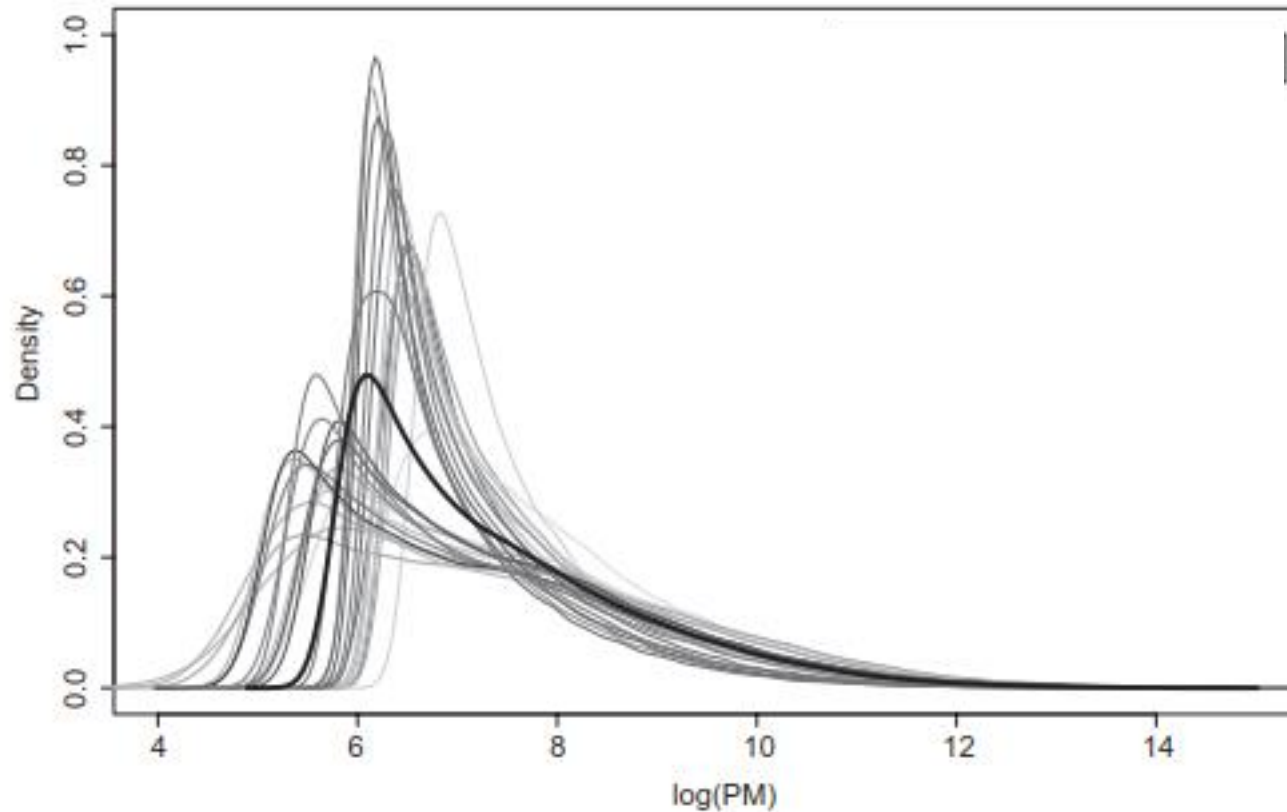
Normalisierung: Korrektur systematischer Effekte aufgrund von Variationen im experimentellen Bereich

Systematische Unterschiede zwischen Arrays



Median normalisierung: Macht die Mediane der Verteilungen von jedem Array gleich. Einfach dividieren jeden Wert im Array durch den Array-Median.

Das Problem bleibt...

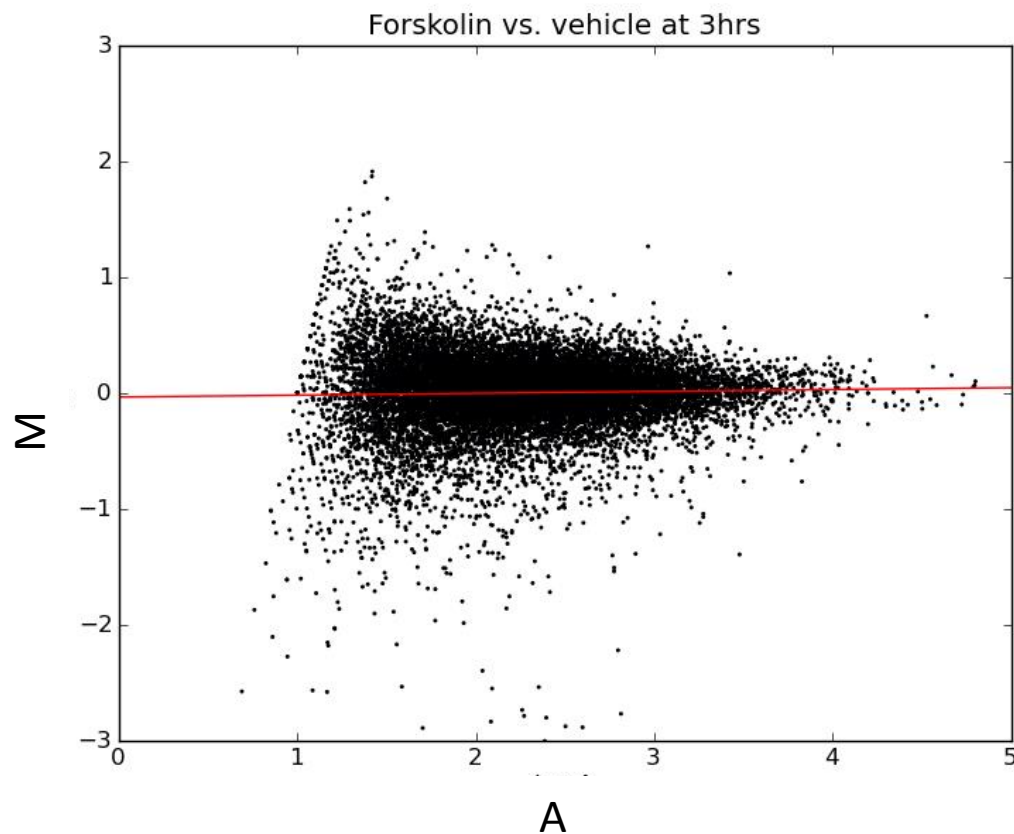


Die mediane Normalisierung ist oft nicht die optimale Lösung

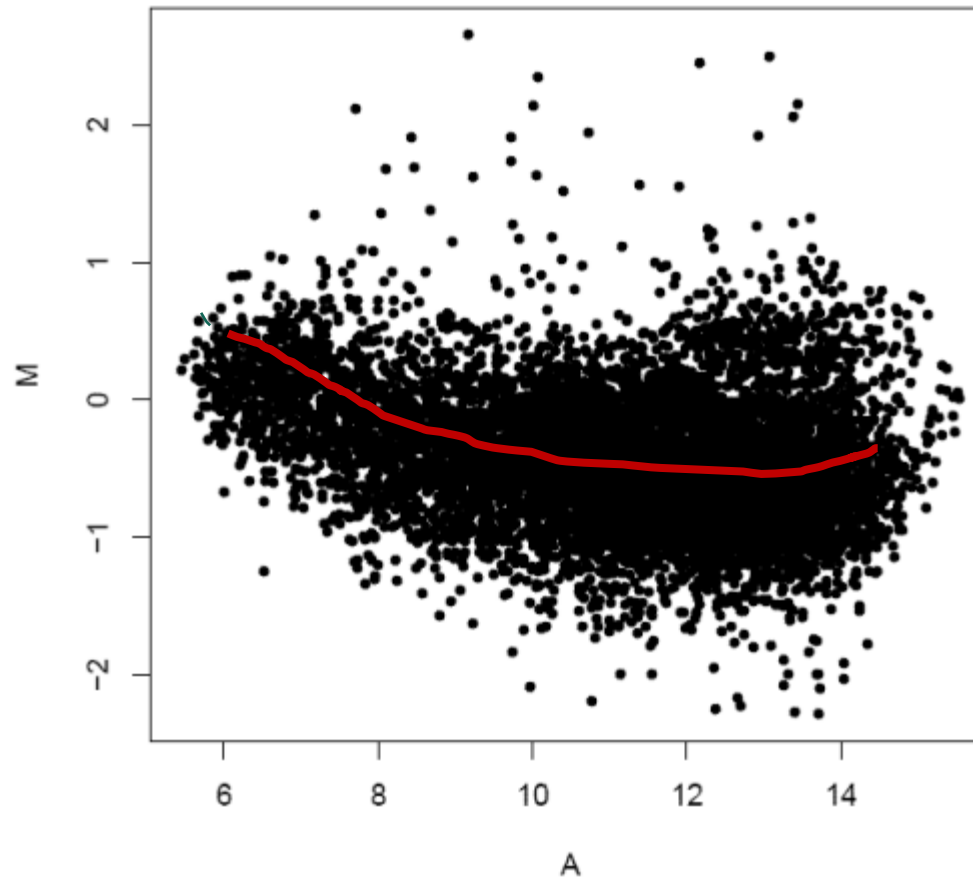
Diagnostisches plot: MA- plot -Minus vs Average



$$M = \log_2(R/G) = \log_2 R - \log_2 G$$
$$A = \frac{1}{2}(\log_2 R + \log_2 G)$$



Häufigstes Problem



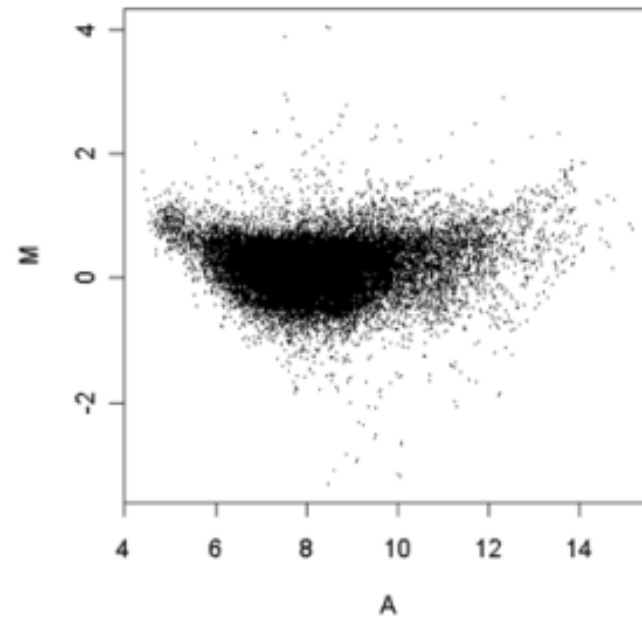
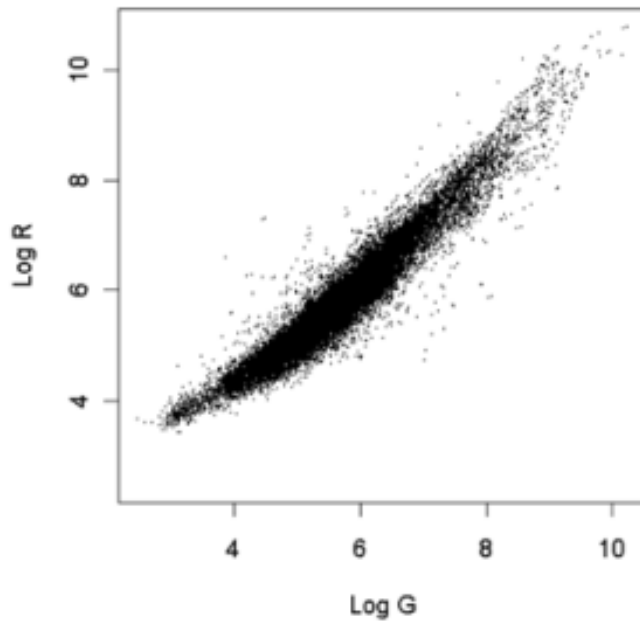
Intensitätsabhängiger Effekt.

Hilft die Median Normalisierung, diese Form zu fixieren?

Systematische bias



$$M = \log_2(R/G) = \log_2 R - \log_2 G$$
$$A = \frac{1}{2}(\log_2 R + \log_2 G)$$

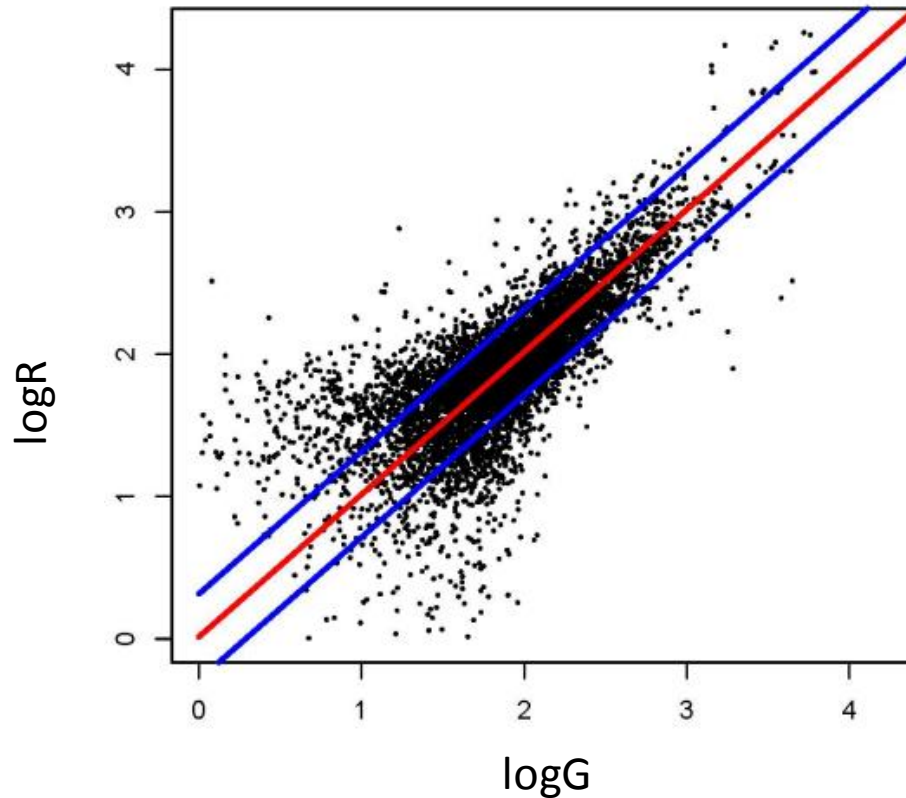


Vergleich zwischen Samples



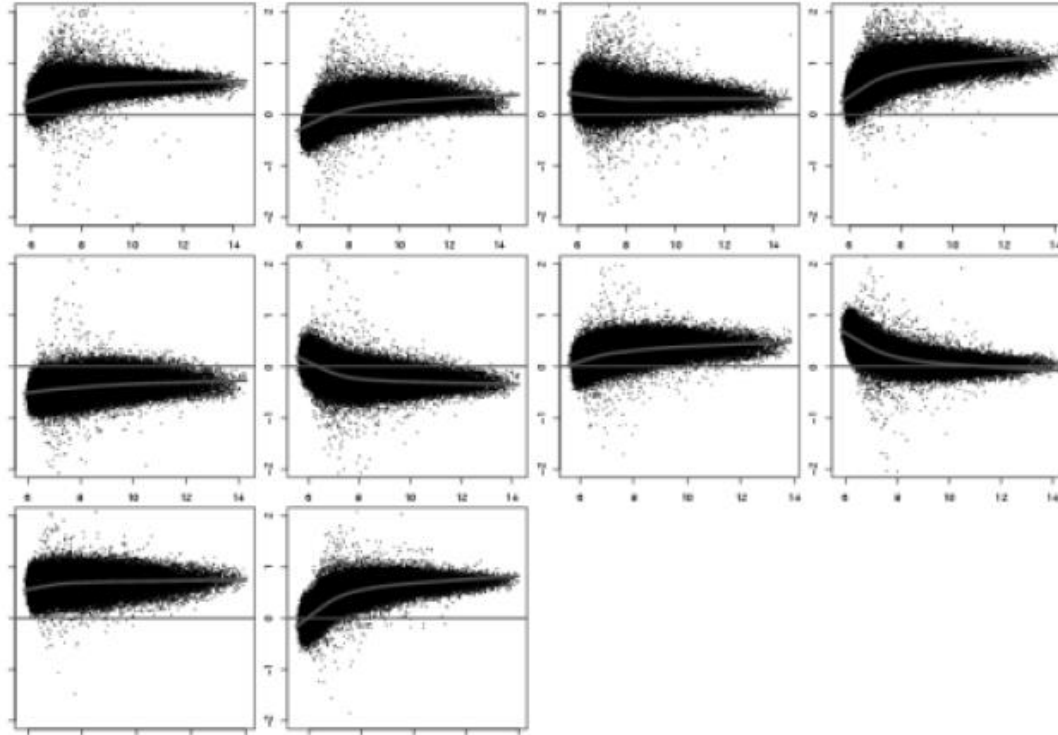
e.g. Krebs vs Normal für a *two channel microarray*

Auf der Achse kann ich auch zwei technische Replikate aus einer Single-channel Mikroarray haben



Blaue Linien zeigen zwei Fold change.
Die rote Linie zeigt kein fold change

Systematische Unterschiede zwischen Arrays

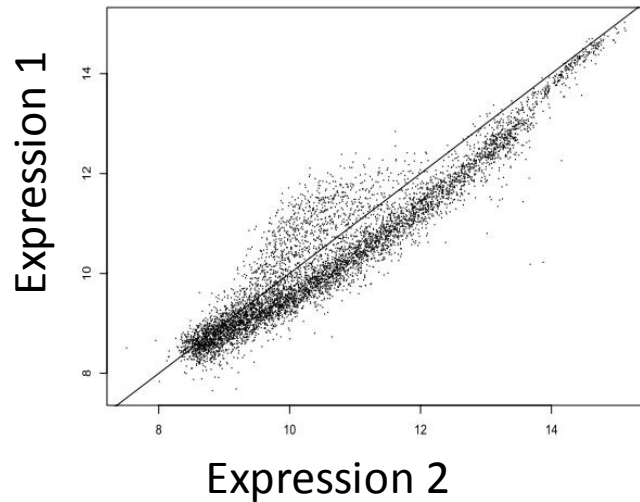


Die meisten der MA-Plots sehen so aus, bevor Normalisierung!

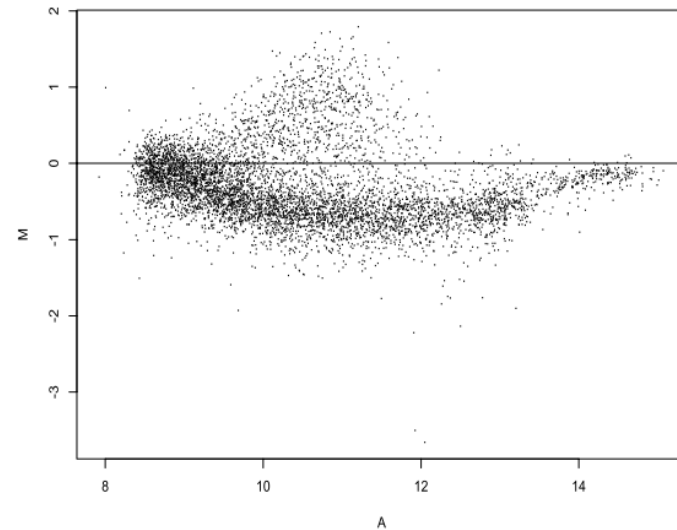
Häufigstes Problem



Scatter-plot vs MA plot: die Macht von MA plots



Hier ist das Bias kaum sichtbar



Hier ist das Bias viel sichtbarer



Vorgeschlagene Lösungen

- Machen die Verteilungen (nicht nur Median) gleich
 - **Quantile Normalisierung**
- Verwenden Sie Kurvenschätzer, z.B. loess auf Bias-Effekte zu korregieren
 - **Loess Normalisierung**



Annahme zur Normalisierung:

- Wenn wir auf der Grundlage von der beobachteten Daten normalisieren, gehen wir davon aus, dass die meisten Gene unverändert zwischen zwei Bedingungen sind oder dass es gibt eine Symmetrie zwischen up- und down-Regulation
- In einigen Fällen kann dies nicht wahr sein. Alternative: Haushalt Gene benutzen, die angeblich nicht ändern oder (*spiked*) Kontrollen und Normalisierung auf sie basieren.

Quantile Normalisierung: Example



Originaldaten

2	4	4
5	4	14
4	6	8
3	5	8
3	3	9

Zeilen sind Gene, Spalten sind Proben (oder Bedingungen)

Quantile Normalisierung: Example



Originaldaten

2	4	4
5	4	14
4	6	8
3	5	8
3	3	9



Anordnung

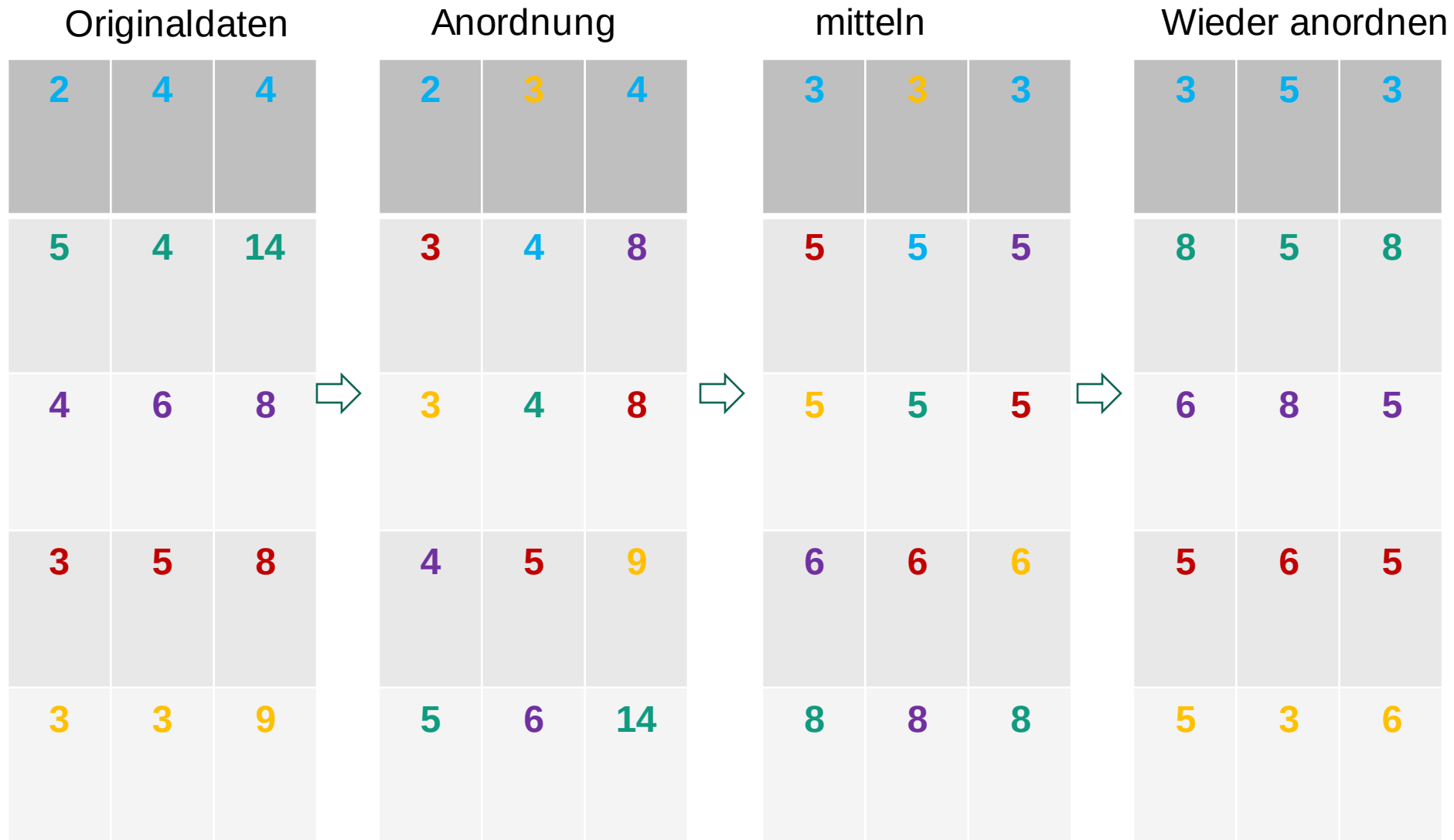
2	3	4
3	4	8
3	4	8
4	5	9
5	6	14

Quantile Normalisierung: Example



Originaldaten			Anordnung			mitteln		
2	4	4	2	3	4	3	3	3
5	4	14	3	4	8	5	5	5
4	6	8	3	4	8	5	5	5
3	5	8	4	5	9	6	6	6
3	3	9	5	6	14	8	8	8

Quantile Normalisierung: Example

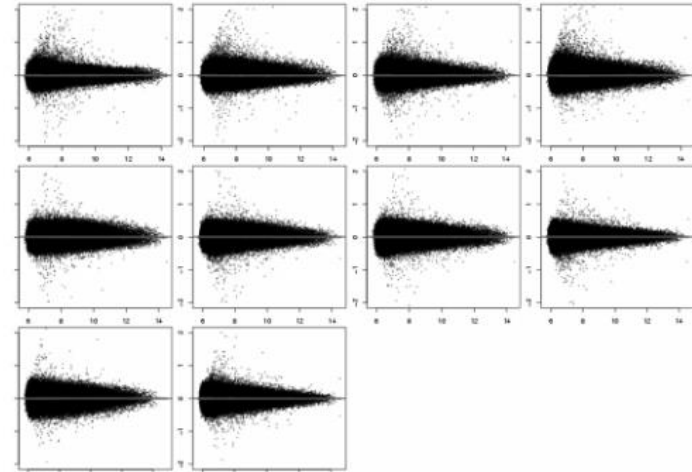
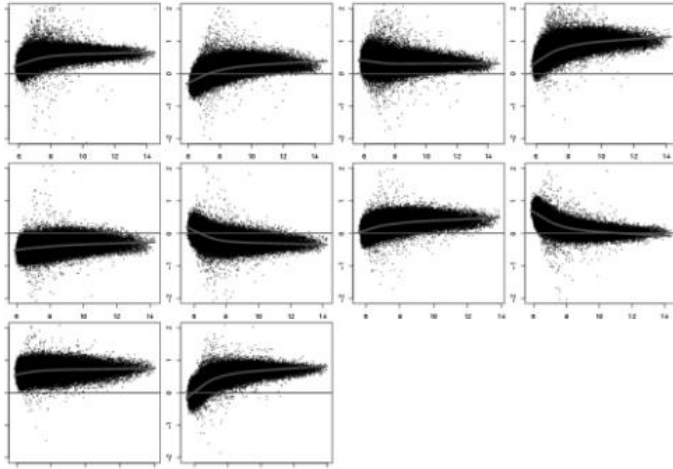


Quantile Normalisierung: Algorithmus



1. Angenommen n Arrays der Länge p (p Gene) bilden eine Matrix X der Dimension $p \times n$, wobei jedes Array eine Spalte ist
2. Sortieren jede Spalte von X um X_{sort} zu erhalten
3. Nehmen den Mittelwert über Zeilen von X_{sort} und weisen diesen Mittelwert jedem Element in der Zeile zu, um $\bar{X}_{sort}$ zu erhalten
4. Erhalten $X_{normalized}$, und jede Spalte von $\bar{X}_{sort}$ neu anordnen, um die gleiche Reihenfolge wie original X zu haben

Quantile Normalisierung: Algorithmus



Vor der Quantil-Normalisierung

Nach der Quantil-Normalisierung

M+bias or Loess Ansatz

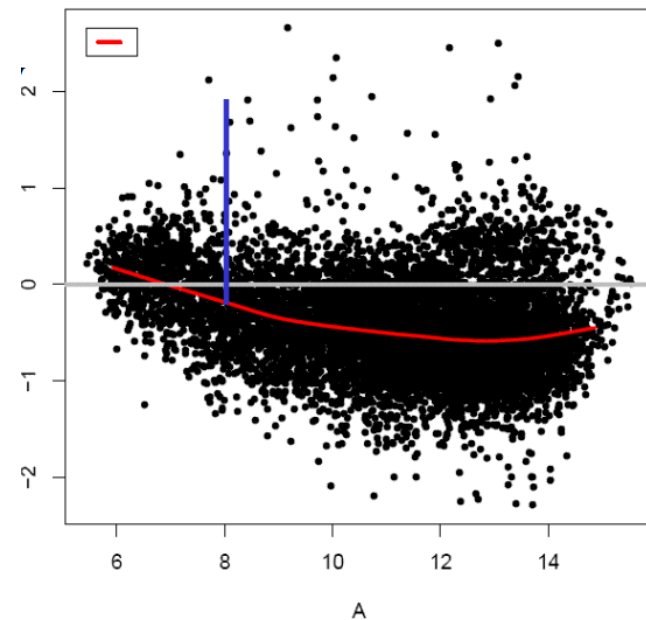


- Ein populärer Ansatz ist, M direkt zu korrigieren
- Wir nehmen an, dass wir **$M + \text{Bias}$** beobachten und dass diese Bias von der Intensität abhängig ist
- Idee: das Bias schätzen und es entfernen
- Für kontinuierliche Variablen gehen wir davon aus, dass die Abhängigkeit ‚smooth‘ ist, und wir verwenden **loess**, um es zu schätzen
- Das normalisierte M ist M - geschätzte Bias
- Vielseitigste Methode

Loess Normalisierung



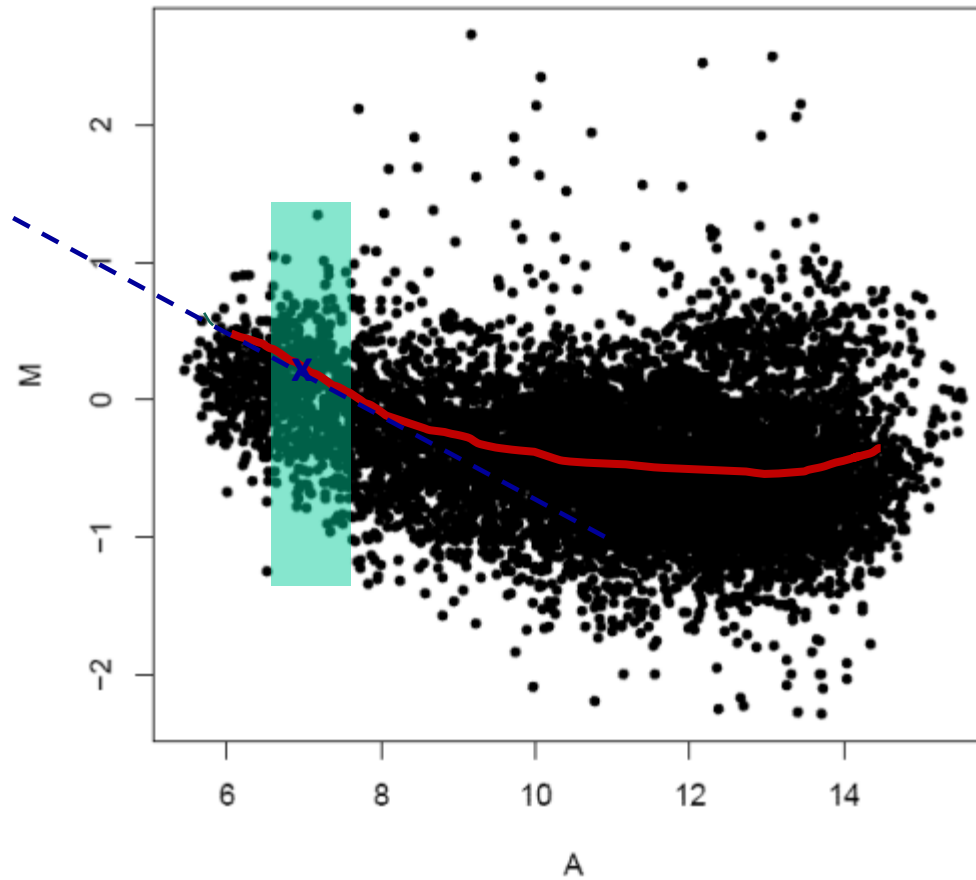
- Intensitätsabhängige Trends werden durch eine Regressionskurve modelliert
 $M = f(A) + \varepsilon$
- Die normalisierten Log-Ratios (M) werden als Residuen der Loess-Regression berechnet



Loess Normalisierung



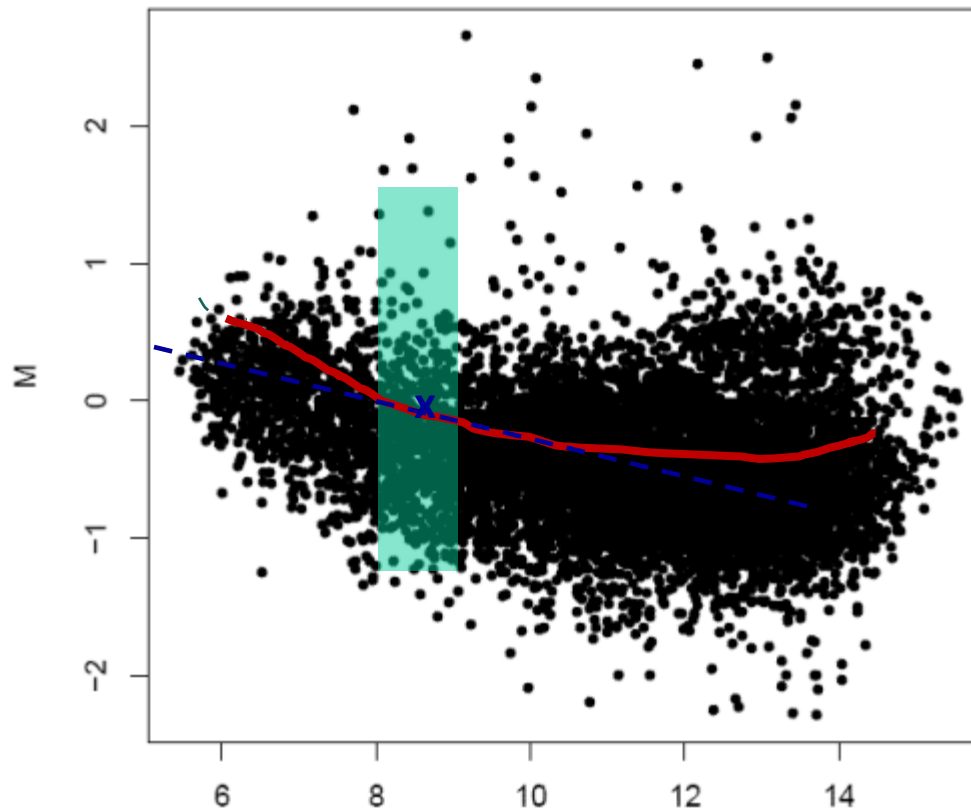
Sliding window Ansatz



Loess Normalisierung



Sliding window Ansatz



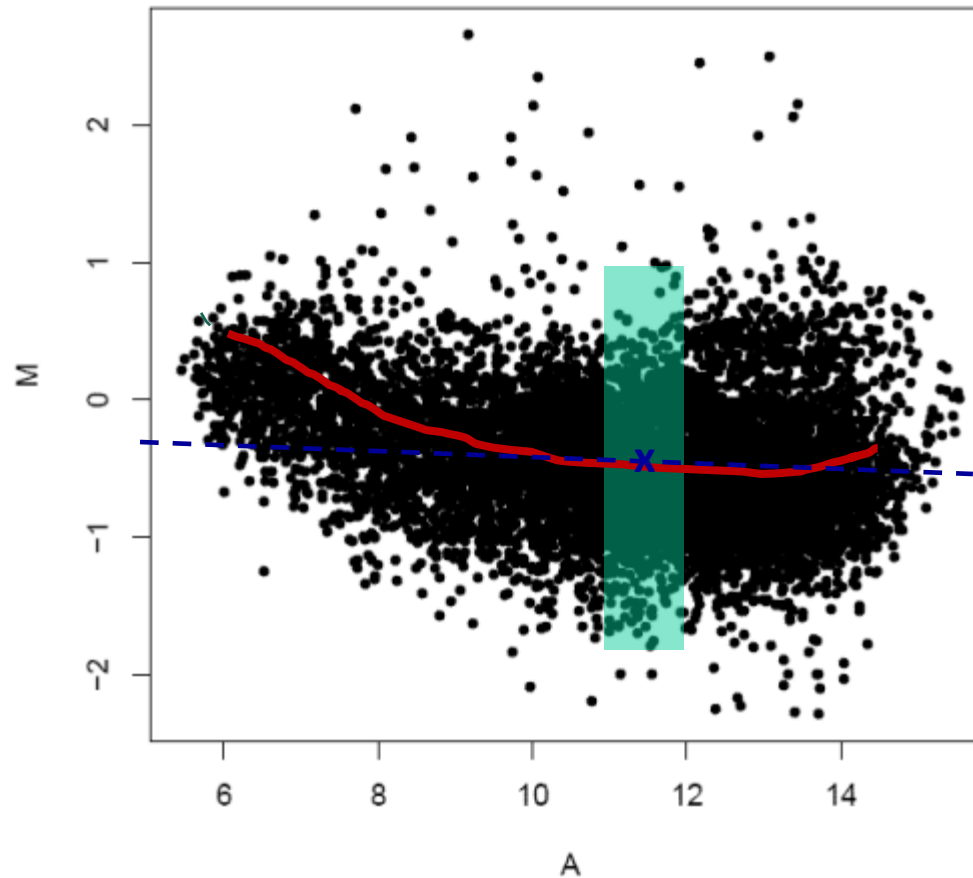
Für jeden Punkt betrachte ich ein Fenster um den Punkt und passe eine Linie innerhalb des Fensters (nur die Punkte innerhalb des Fensters verwenden).

Danach kann ich die kleinen Stücke verbinden, um die finale Loesskurve zu bilden

Loess Normalisierung



Sliding window Ansatz



Innerhalb jedes Fensters f ist eine Linie mit Formula $f_i(A) = \beta_0 + \beta_1 A$