Analyzing Cross-Plattform Consistency Using Tests Against Ordered Alternatives

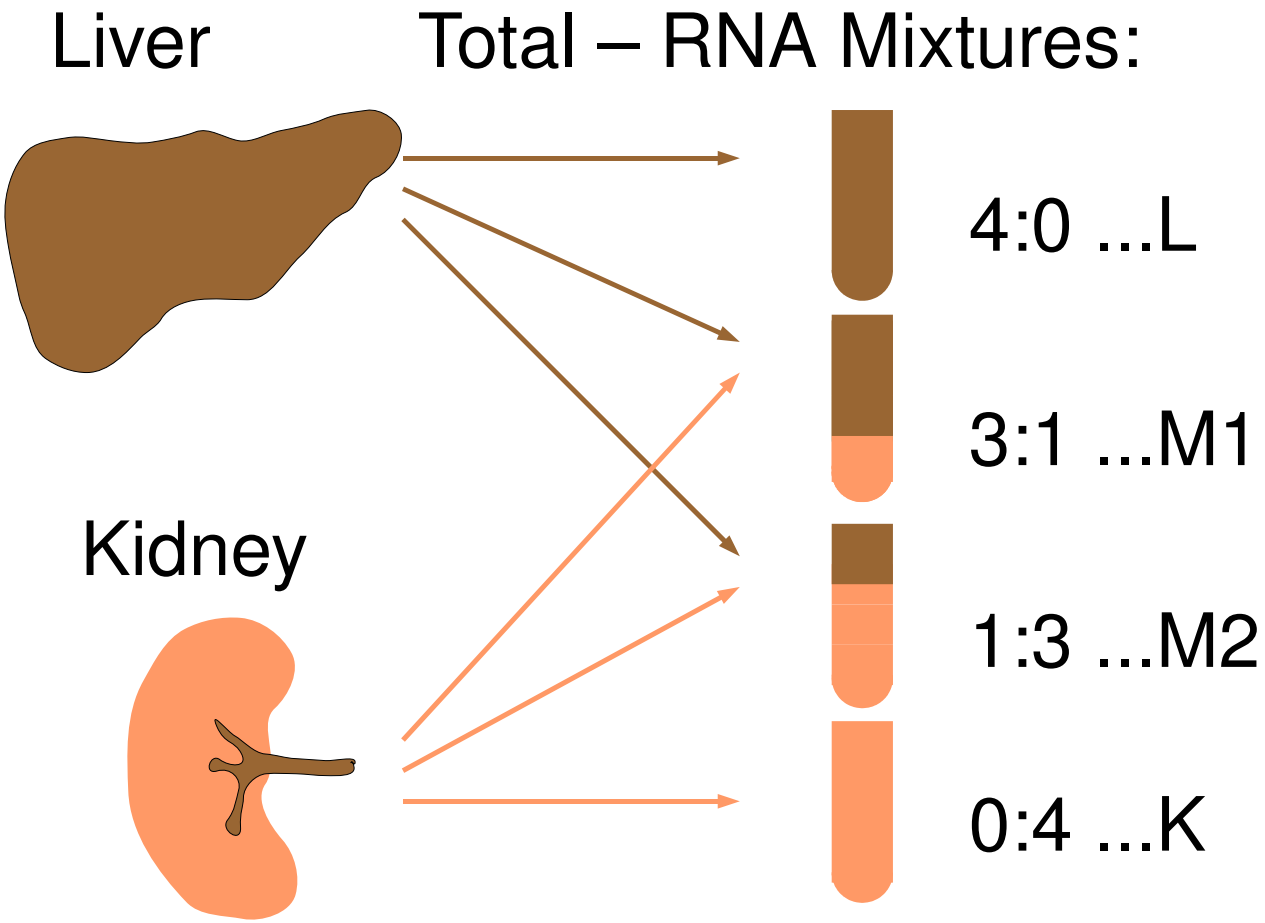
CAMDA Emerald Competition

Florian Klinglmueller[1] Thomas Tuechler[2]

[1]Core Unit for Medical Statistics and Informatics
Medical University of Vienna
florian.klinglmueller@meduniwien.ac.at
[2]WWTF Chair for Bioinformatics
BOKU University
thomas.tuechler@boku.ac.at

05.12.2008 / CAMDA@Boku University

Titration



Design Hierarchy

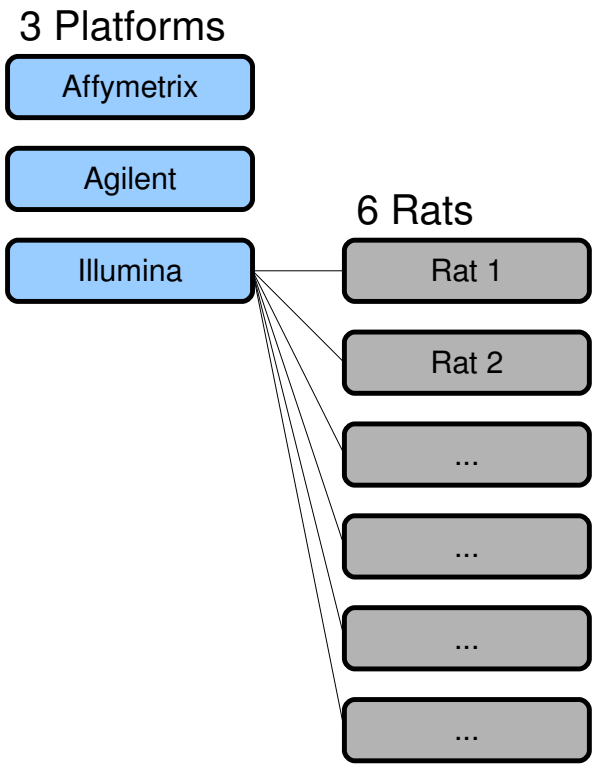
Experimental Design:

3 Platforms

- Affymetrix
- Agilent
- Illumina

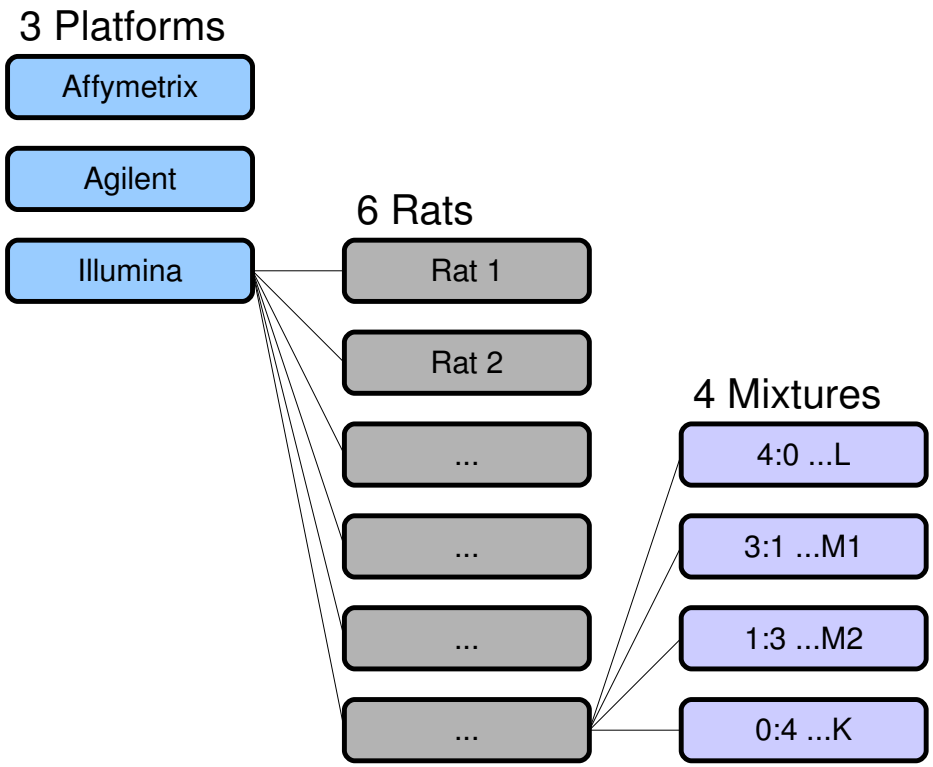
Design Hierarchy

Experimental Design:



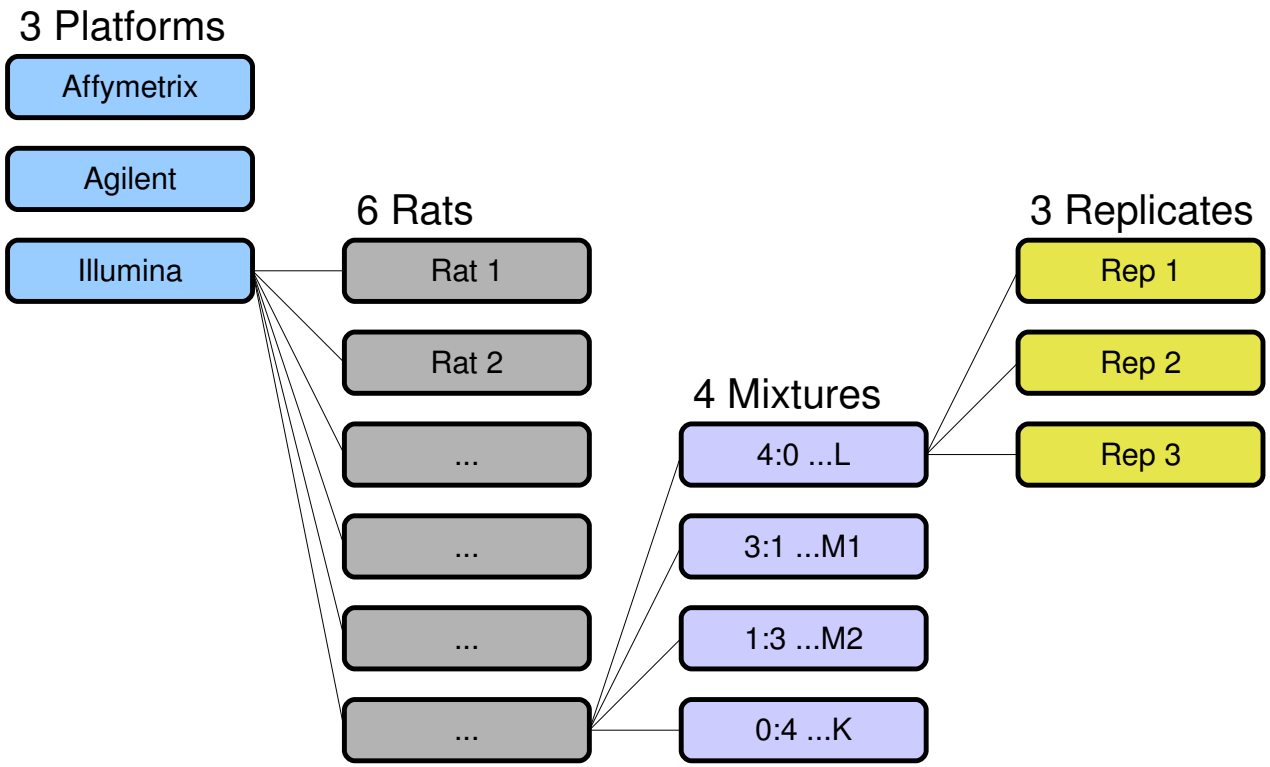
Design Hierarchy

Experimental Design:



Design Hierarchy

Experimental Design:



Main Questions

- Do the measured intensities reflect the titration?
- Agreement across platforms.
- Influence of normalization.

Tests Against Order-Restricted Alternatives

- Dose-response studies
- 70's and 80's literature:
  - Barlow [1]
  - Robertson et al. [3]
- Microarray Application: Lin et al. [2]
- 5 Statistics: Marcus, Wilson, E2, M, ModifiedM
- E2 most powerful $\Rightarrow$ we use E2

Test

Null Hypothesis

We test the null hypothesis of equal means

$$H_{0,g} : \mu_{L,g} = \mu_{M1,g} = \mu_{M2,g} = \mu_{K,g}, \tag{1}$$

against the ordered alternatives

$$H_{1,g}^{up} : \quad \mu_{L,g} \leq \mu_{M1,g} \leq \mu_{M2,g} \leq \mu_{K,g}, \tag{2}$$

$$H_{1,g}^{down} : \quad \mu_{L,g} \geq \mu_{M1,g} \geq \mu_{M2,g} \geq \mu_{K,g}, \tag{3}$$

with at least one strict inequality.

- Main Principle: Isotonic Regression

Isotonic Regression

Fitting Monotone Functions

Isotonic Regression: Formulation

Isotonic Function

- Set $\mathcal{T} := \{t_1, ..., t_n\}$ with order relation
- $m(t_i)$ is called isotonic if
  $t_i \leq t_j \Rightarrow m(t_i) \leq m(t_j)$
- $\mathcal{F}(\mathcal{T})$: all isotonic functions on $\mathcal{T}$
- Direction has to be specified

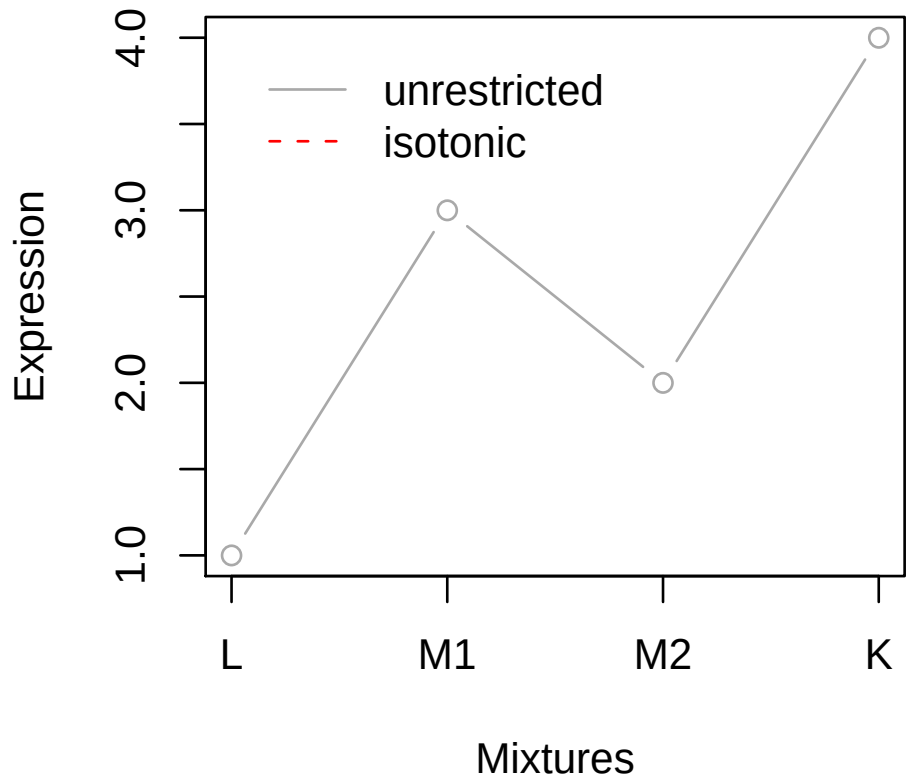
Isotonic Regression

- $y_i = m(t_i) + \epsilon_i$, $m \in \mathcal{F}(\mathcal{T})$
- Least-squares fit:
  $\hat{m} = \text{argmin}_{m \in \mathcal{F}(\mathcal{T})} \sum_{i=1}^{n} (y_i - m(t_i))^2$.

Isotonic Regression

Example

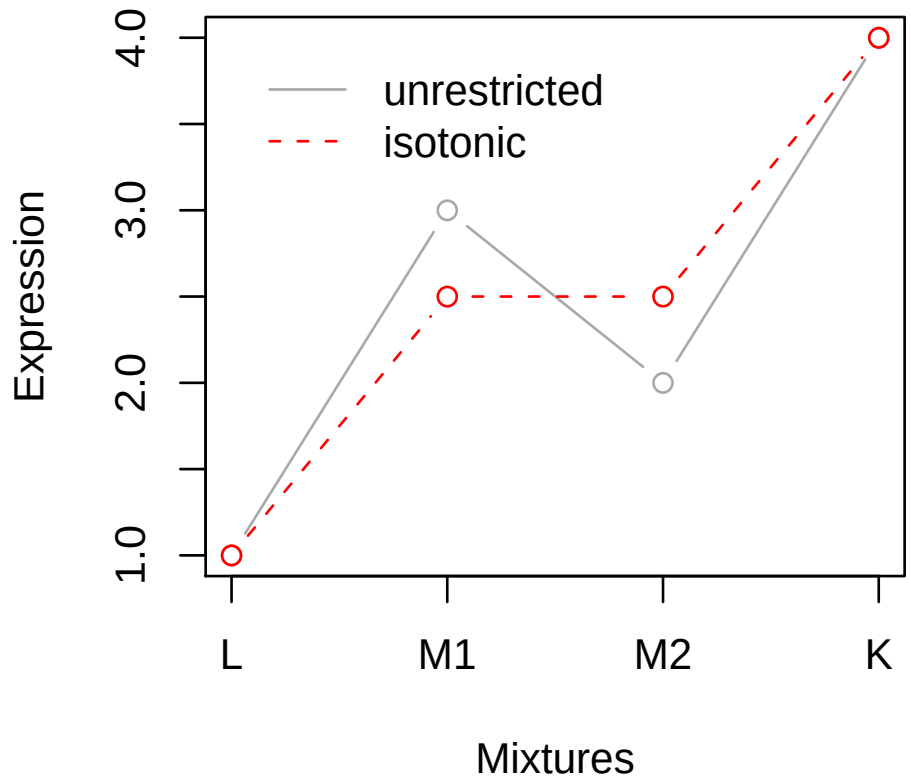
- $\mathcal{T} = \{L \leq M1 \leq M2 \leq K\}$
- $\overline{y}_g(t_i) = m^{up}(t_i) + \epsilon_i$
- Some gene expressions:



Isotonic Regression

Upwards Trend

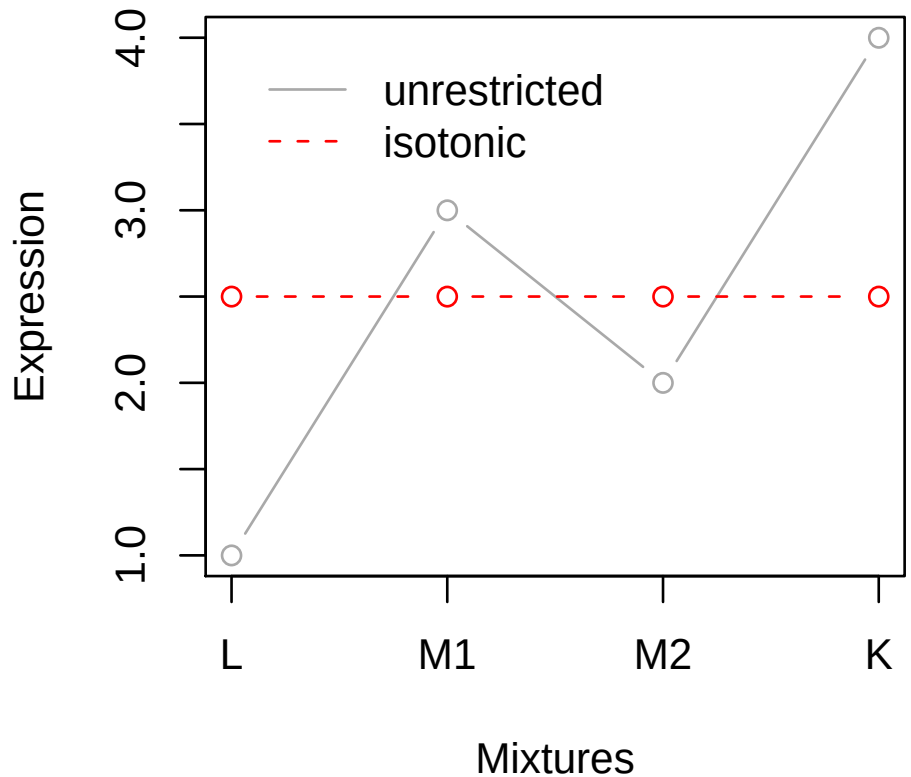
- $\mathcal{T} = \{L \leq M1 \leq M2 \leq K\}$
- $\overline{y}_g(t_i) = m^{up}(t_i) + \epsilon_i$
- Isotonic Regression for upwards trend:



Isotonic Regression

Downwards Trend

- $\mathcal{T} = \{L \geq M1 \geq M2 \geq K\}$
- $\overline{y}_g(t_i) = m^{down}(t_i) + \epsilon_i$
- Isotonic Regression for downwards trend:



Statistic

Definition of E2 Statistic

E2 (Barlow [1],Robertson et al. [3]):

$$\overline{E}_{01}^{2up} = 1 - \frac{\sum_{kj}(y_{kj} - \hat{m}^{up}(t_i))^2}{\sum_{kj}(y_{kj} - \overline{y})^2}, \tag{4}$$

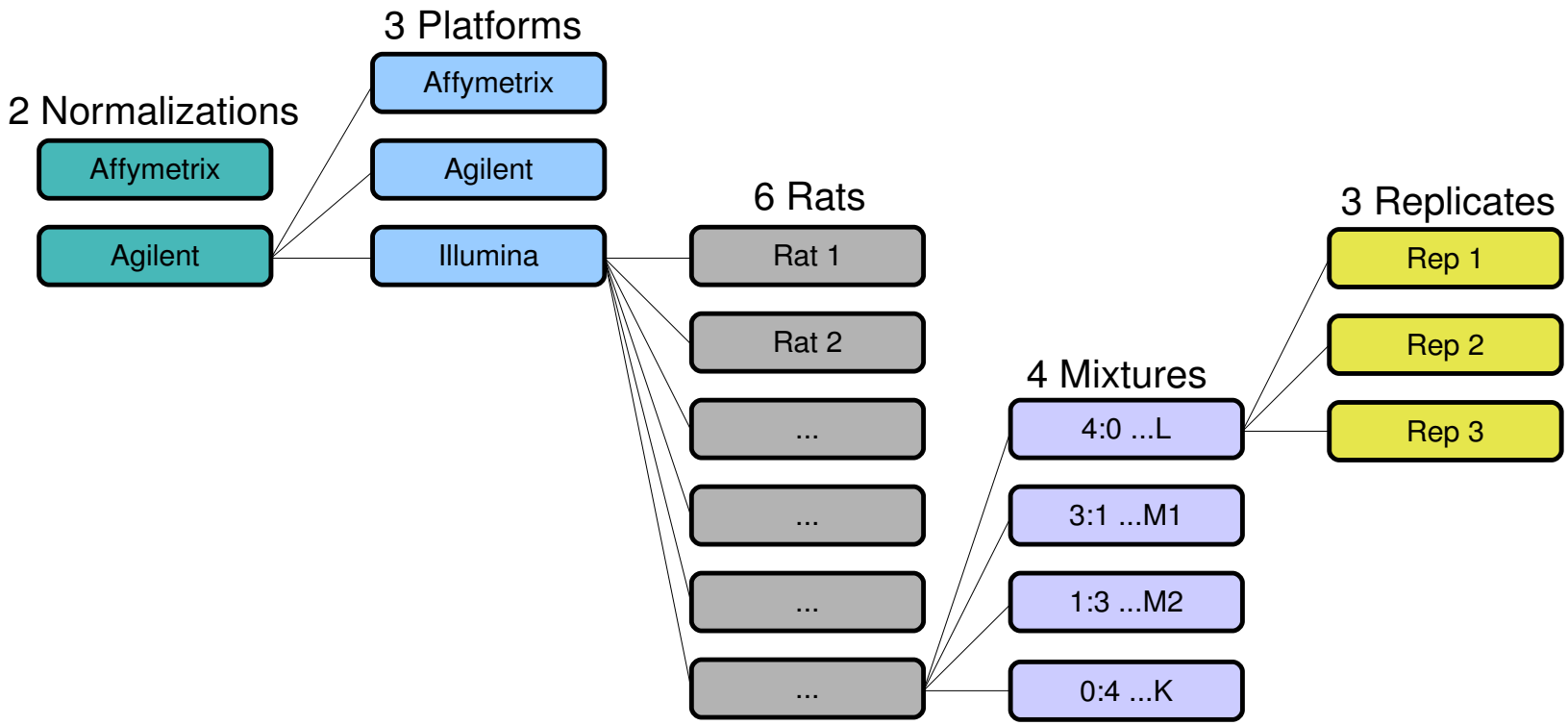
- Likelihood-ratio:

$$\overline{E}_{01}^{2up} = 1 - \frac{ESS}{TSS}$$

p-Value Combination

Capturing the Hierarchical Variance Structure

- Revisit the design hierarchy
- Now we add a new level: Normalization



Normalizations

Baseline vs. Quantile Normalization

- Both widely used

Baseline Normalization

Align per array medians

1. From each array remove array-wise median
2. To each array add overall median

Removes systematic location shifts

Quantile Normalization

Align order statistics

1. Per array - reduce expressions to ranks
2. Per array - reassign ranks to quantiles from mean distribution (means of order statistics)

Removes any systematic disturbance that keeps the order

Normalizations

Baseline vs. Quantile Normalization

- Both widely used

Baseline Normalization

Align per array medians

1. From each array remove array-wise median
2. To each array add overall median

Removes systematic location shifts

Quantile Normalization

Align order statistics

1. Per array - reduce expressions to ranks
2. Per array - reassign ranks to quantiles from mean distribution (means of order statistics)

Removes any systematic disturbance that keeps the order

Normalizations

Baseline vs. Quantile Normalization

- Both widely used

Baseline Normalization

Align per array medians

1. From each array remove array-wise median
2. To each array add overall median

Removes systematic location shifts

Quantile Normalization

Align order statistics

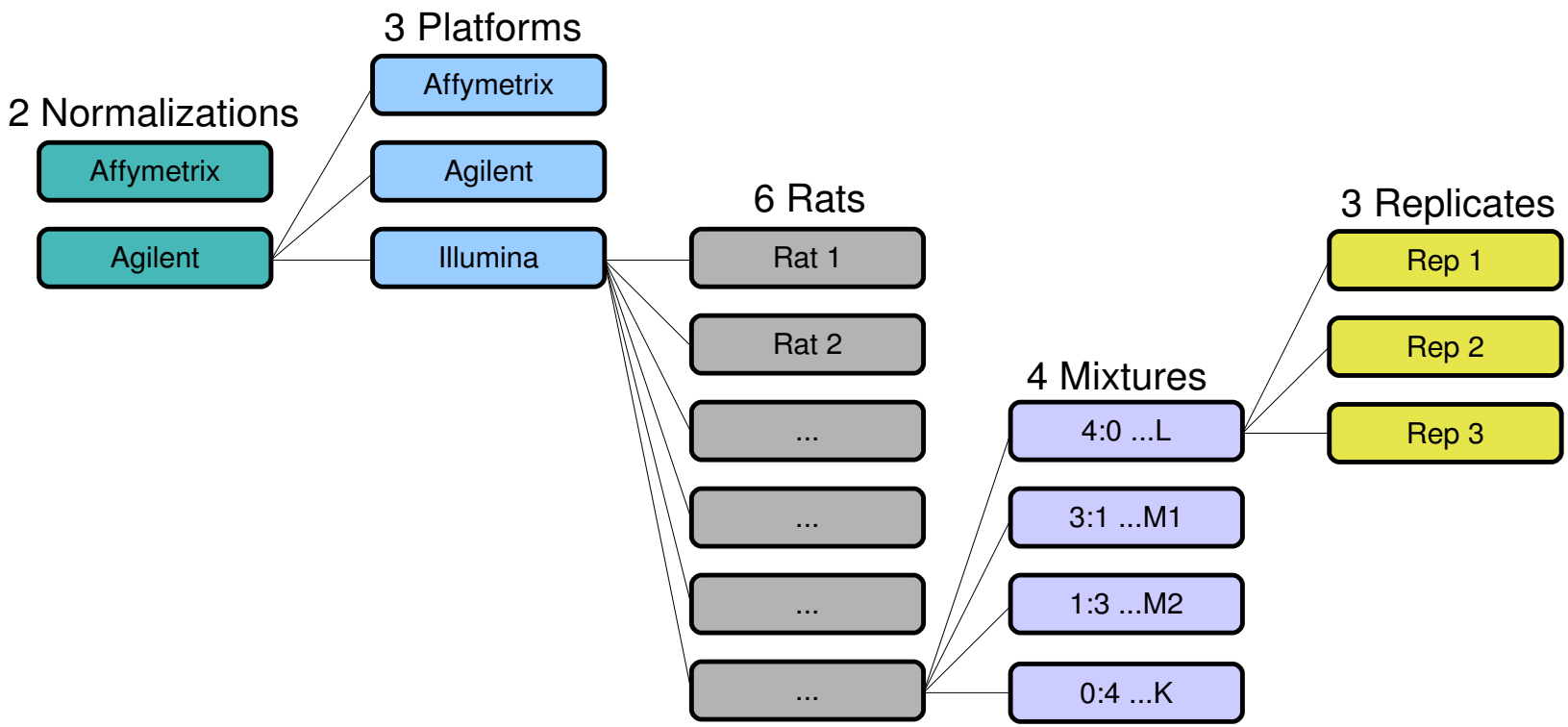
1. Per array - reduce expressions to ranks
2. Per array - reassign ranks to quantiles from mean distribution (means of order statistics)

Removes any systematic disturbance that keeps the order

p-Value Combination

Capturing the Hierarchical Variance Structure

- Revisit the design hierarchy
- We want p



p-Value Combination

Inverse Normal Method

- Combine one-sided p-values:

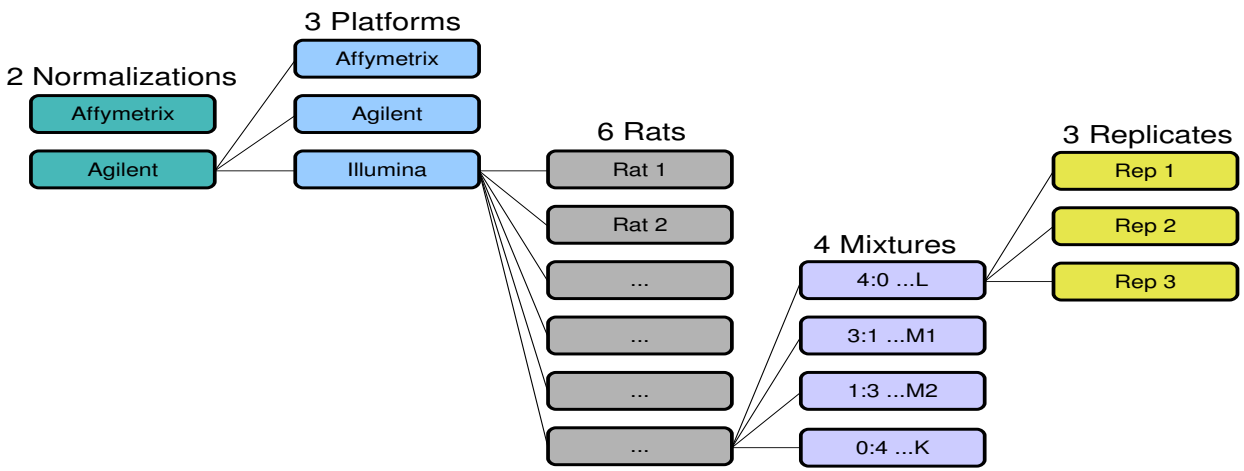
$$p_g^{C,up} = 1 - \Phi\left(\frac{1}{\sqrt{N}} \sum_i \Phi^{-1}(1 - p_{ig}^{up})\right), \tag{5}$$

- $p_g^{C,down}$ analogue
- uniformly distritibuted conservative one-sided p-values
- Bonferroni correct directional decision:
  $p_g^C = 2\min(p_g^{C,up}, p_g^{C,down})$.

p-Value Combination

Per Animal p-Values

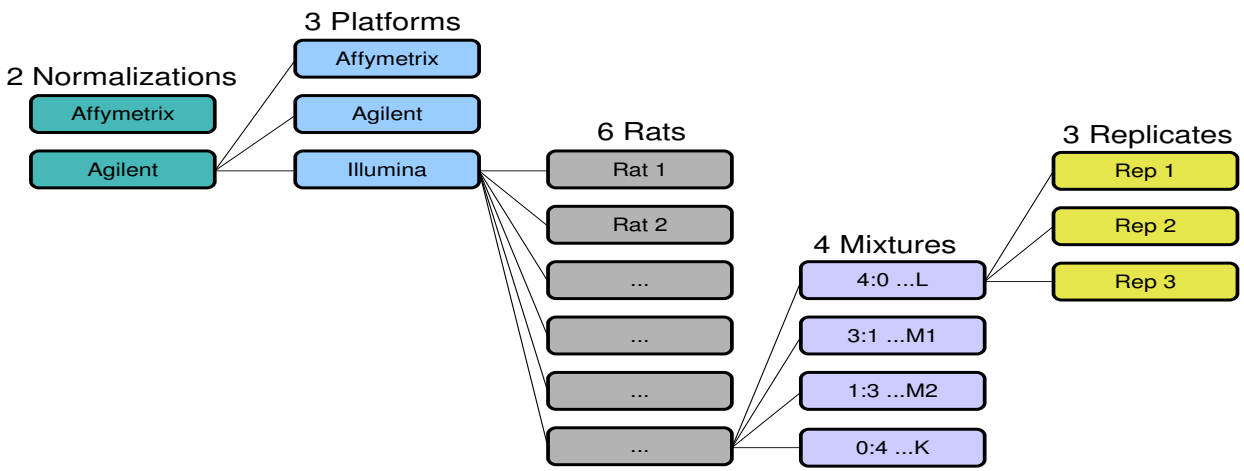
- 6 Animals $\times$ 3 Platforms $\times$ 2 Normalizations $\rightarrow$ 36 times $p^{up}_{Norm,Plat,ig}$, $p^{down}_{Norm,Plat,ig}$, $p_{Norm,Plat,ig}$
- Combine the 6 $\times$ 6 $p^{up}_{Norm,Plat,ig}$, $p^{down}_{Norm,Plat,ig}$ to get get 6: $p^{C^{Plat},up}_{Norm,g}$, $p^{C^{Plat},down}_{Norm,g}$, and $p^{C^{Plat}}_{Norm,g}$
- Combine the 3 $p^{C^{Plat},up}_{Norm,g}$, $p^{C^{Plat},down}_{Norm,g}$ to get 2: $p^{C^{Norm},up}_{g}$, $p^{C^{Norm},down}_{g}$



p-Value Combination

Per Animal p-Values

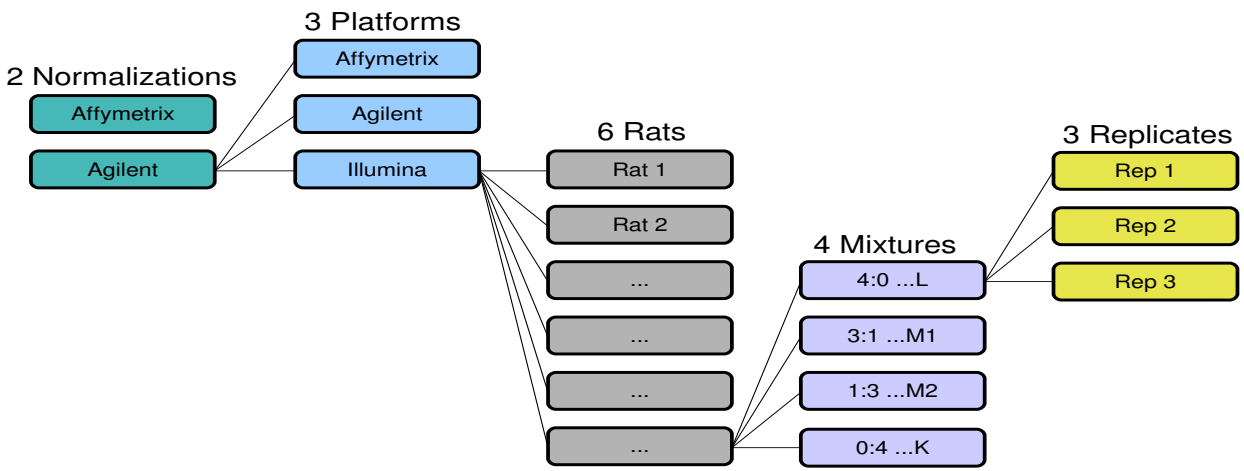
- 6 Animals $\times$ 3 Platforms $\times$ 2 Normalizations $\longrightarrow$ 36 times $p^{up}_{Norm,Plat,ig}$, $p^{down}_{Norm,Plat,ig}$, $p_{Norm,Plat,ig}$
- Combine the 6 $\times$ 6 $p^{up}_{Norm,Plat,ig}$, $p^{down}_{Norm,Plat,ig}$ to get get 6: $p^{C^{Plat},up}_{Norm,g}$, $p^{C^{Plat},down}_{Norm,g}$, and $p^{C^{Plat}}_{Norm,g}$
- Combine the 3 $p^{C^{Plat},up}_{Norm,g}$, $p^{C^{Plat},down}_{Norm,g}$ to get 2: $p^{C^{Norm},up}_{g}$, $p^{C^{Norm},down}_{g}$



p-Value Combination

Per Animal p-Values

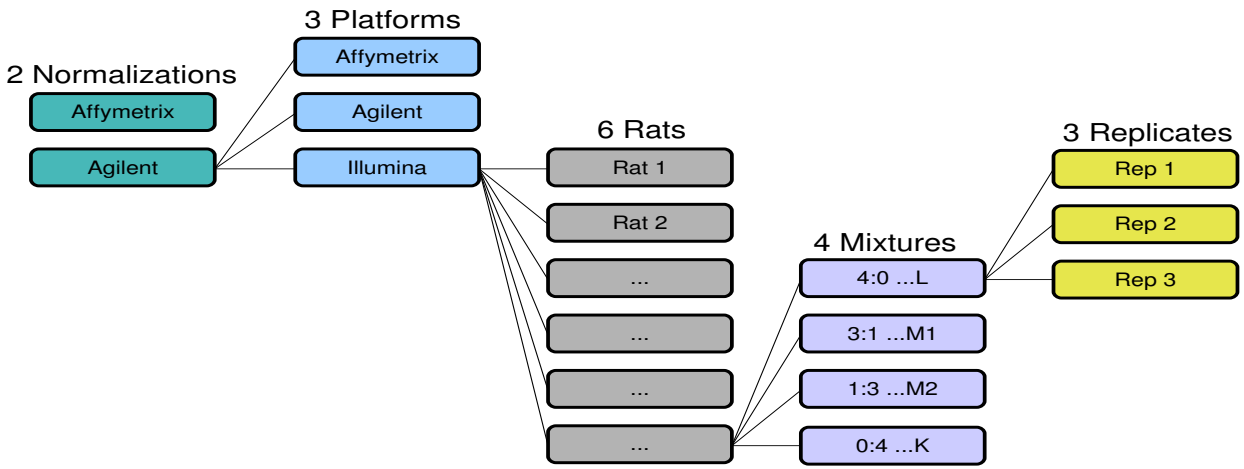
- 6 Animals $\times$ 3 Platforms $\times$ 2 Normalizations $\longrightarrow$ 36 times $p^{up}_{Norm,Plat,ig}$, $p^{down}_{Norm,Plat,ig}$, $p_{Norm,Plat,ig}$
- Combine the 6 $\times$ 6 $p^{up}_{Norm,Plat,ig}$, $p^{down}_{Norm,Plat,ig}$ to get get 6: $p^{C^{Plat},up}_{Norm,g}$, $p^{C^{Plat},down}_{Norm,g}$, and $p^{C^{Plat}}_{Norm,g}$
- Combine the 3 $p^{C^{Plat},up}_{Norm,g}$, $p^{C^{Plat},down}_{Norm,g}$ to get 2: $p_g^{C^{Norm},up}$, $p_g^{C^{Norm},down}$



p-Value Combination

Per Animal p-Values

- 6 Animals $\times$ 3 Platforms $\times$ 2 Normalizations $\longrightarrow$ 36 times $p^{up}_{Norm,Plat,ig}$, $p^{down}_{Norm,Plat,ig}$, $p_{Norm,Plat,ig}$
- Combine the 6 $\times$ 6 $p^{up}_{Norm,Plat,ig}$, $p^{down}_{Norm,Plat,ig}$ to get get 6: $p^{C^{Plat},up}_{Norm,g}$, $p^{C^{Plat},down}_{Norm,g}$, and $p^{C^{Plat}}_{Norm,g}$
- Combine the 3 $p^{C^{Plat},up}_{Norm,g}$, $p^{C^{Plat},down}_{Norm,g}$ to get 2: $p^{C^{Norm},up}_{g}$, $p^{C^{Norm},down}_{g}$



p-Value Combination

Summary

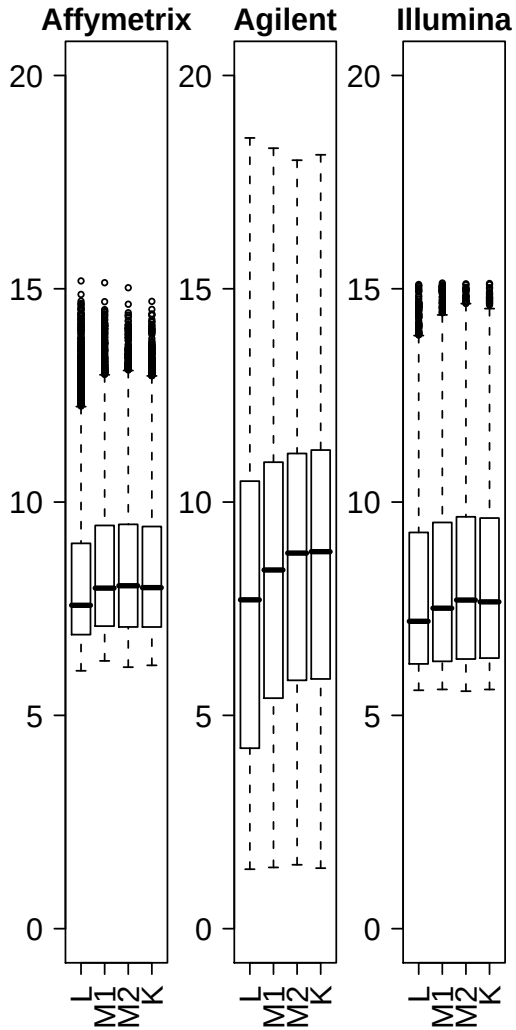
- Comptute one sided permutation test p-values for each animal, on each platform seperately with Quantile - and Baseline - normalized data.
- Combine per animal tests from each plaform.
- Combine per platform tests from each normalization.

Results

Finally!

Exploratory Analysis

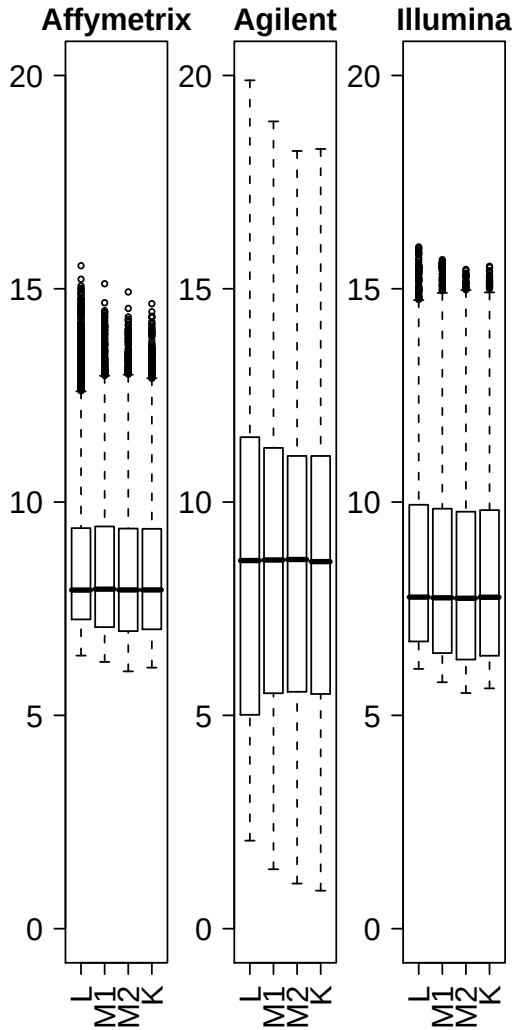
Distribution of Group Means on Raw Data



- Location-shift
- Higher messenger-RNA content in kidney?
- Both normalization methods remove any visible trends in location
- Baseline
- Quantile - also in scale

Exploratory Analysis

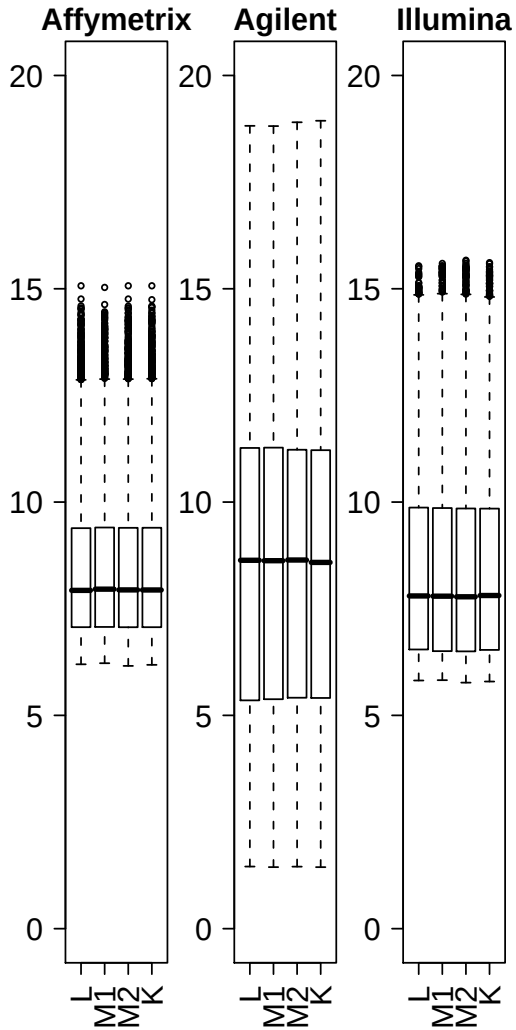
Distribution of Group Means on Raw Data



- Location-shift
- Higher messenger-RNA content in kidney?
- Both normalization methods remove any visible trends in location
- Baseline
- Quantile - also in scale

Exploratory Analysis

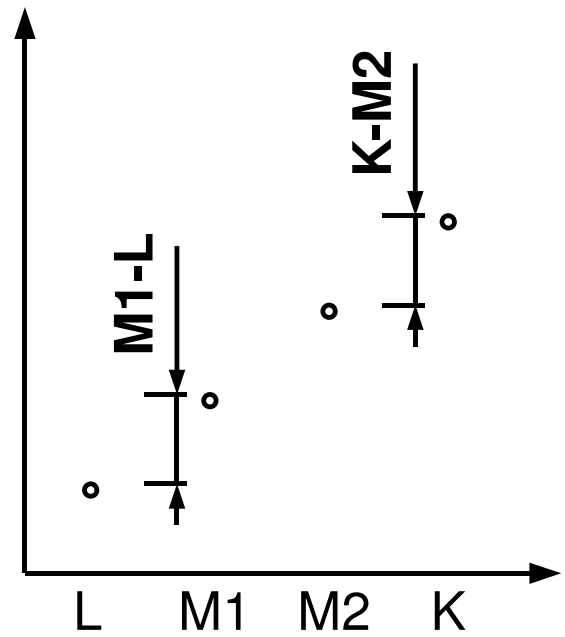
Distribution of Group Means on Raw Data



- Location-shift
- Higher messenger-RNA content in kidney?
- Both normalization methods remove any visible trends in location
- Baseline
- Quantile - also in scale

Exploration of Trend

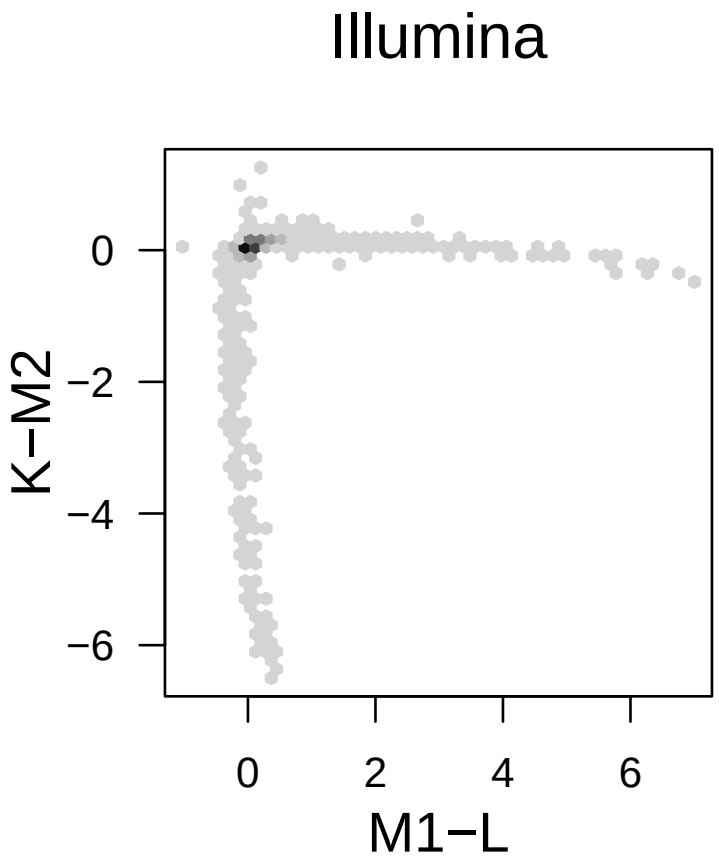
Relationship between Increases



- Relationship between first/second increase
- Scatterplot - Illumina: Trends not linear; When first increase large then last increase small and vice versa
- Scatterplot - Agilent
- Scatterplot - Affymetrix
- Rightmost point
- Lowest point
- Saturation?

Exploration of Trend

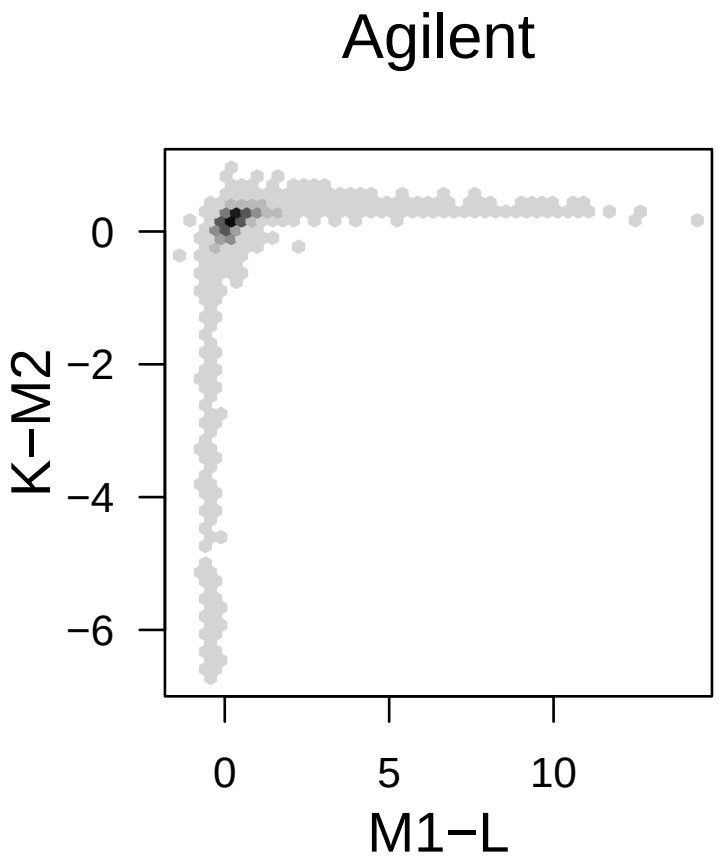
Relationship between Increases



- Relationship between first/second increase
- Scatterplot - Illumina: Trends not linear; When first increase large then last increase small and vice versa
- Scatterplot - Agilent
- Scatterplot - Affymetrix
- Rightmost point
- Lowest point
- Saturation?

Exploration of Trend

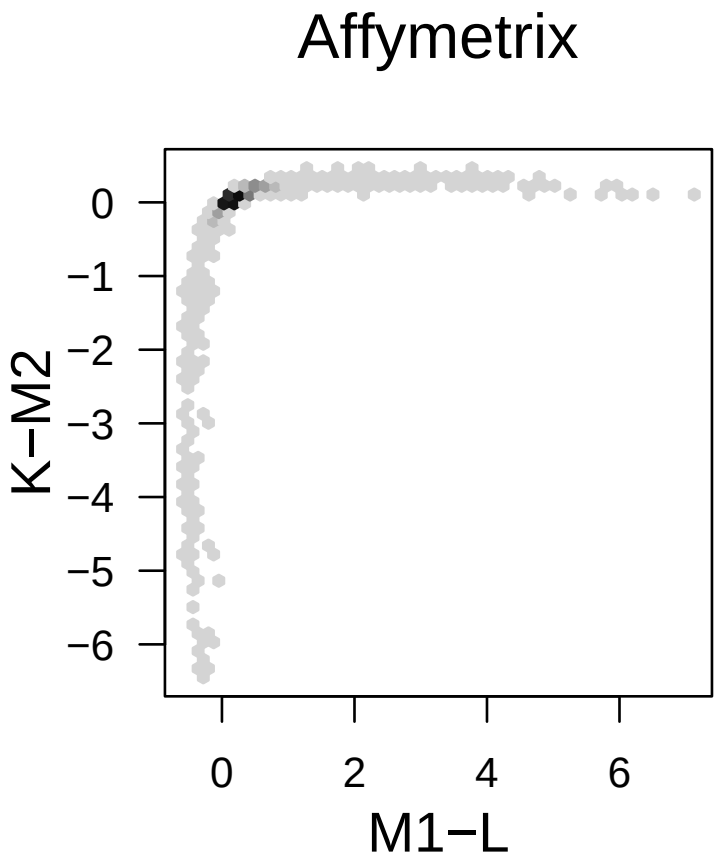
Relationship between Increases



- Relationship between first/second increase
- Scatterplot - Illumina: Trends not linear; When first increase large then last increase small and vice versa
- Scatterplot - Agilent
- Scatterplot - Affymetrix
- Rightmost point
- Lowest point
- Saturation?

Exploration of Trend

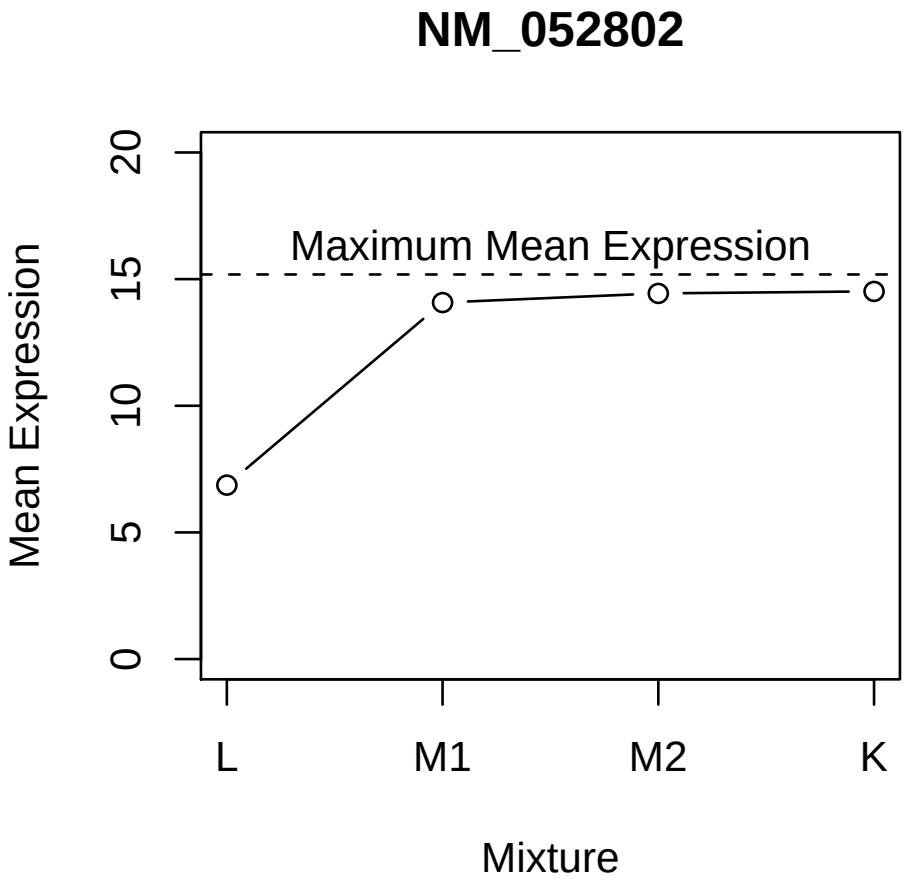
Relationship between Increases



- Relationship between first/second increase
- Scatterplot - Illumina: Trends not linear; When first increase large then last increase small and vice versa
- Scatterplot - Agilent
- Scatterplot - Affymetrix
- Rightmost point
- Lowest point
- Saturation?

Exploration of Trend

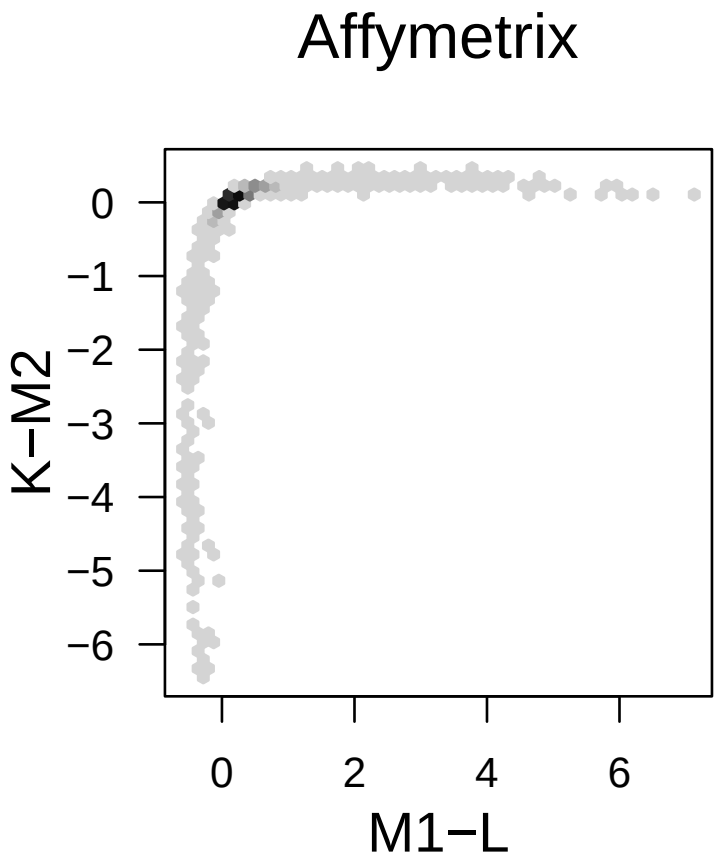
Relationship between Increases



- Relationship between first/second increase
- Scatterplot - Illumina: Trends not linear; When first increase large then last increase small and vice versa
- Scatterplot - Agilent
- Scatterplot - Affymetrix
- Rightmost point
- Lowest point
- Saturation?

Exploration of Trend

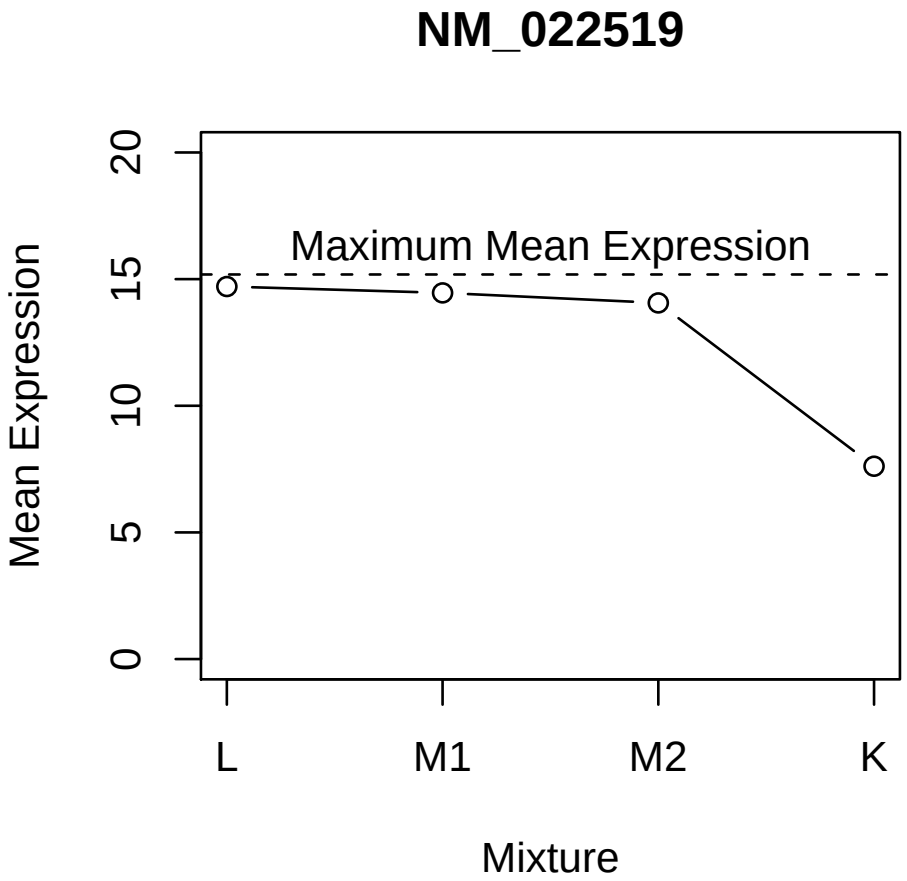
Relationship between Increases



- Relationship between first/second increase
- Scatterplot - Illumina: Trends not linear; When first increase large then last increase small and vice versa
- Scatterplot - Agilent
- Scatterplot - Affymetrix
- Rightmost point
- Lowest point
- Saturation?

Exploration of Trend

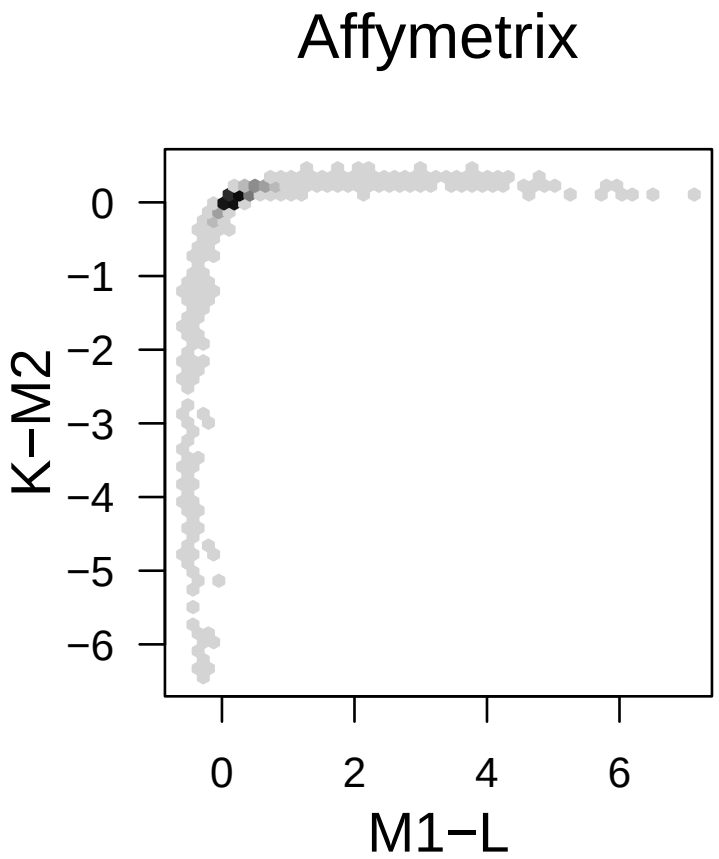
Relationship between Increases



- Relationship between first/second increase
- Scatterplot - Illumina: Trends not linear; When first increase large then last increase small and vice versa
- Scatterplot - Agilent
- Scatterplot - Affymetrix
- Rightmost point
- Lowest point
- Saturation?

Exploration of Trend

Relationship between Increases



- Relationship between first/second increase
- Scatterplot - Illumina: Trends not linear; When first increase large then last increase small and vice versa
- Scatterplot - Agilent
- Scatterplot - Affymetrix
- Rightmost point
- Lowest point
- Saturation?

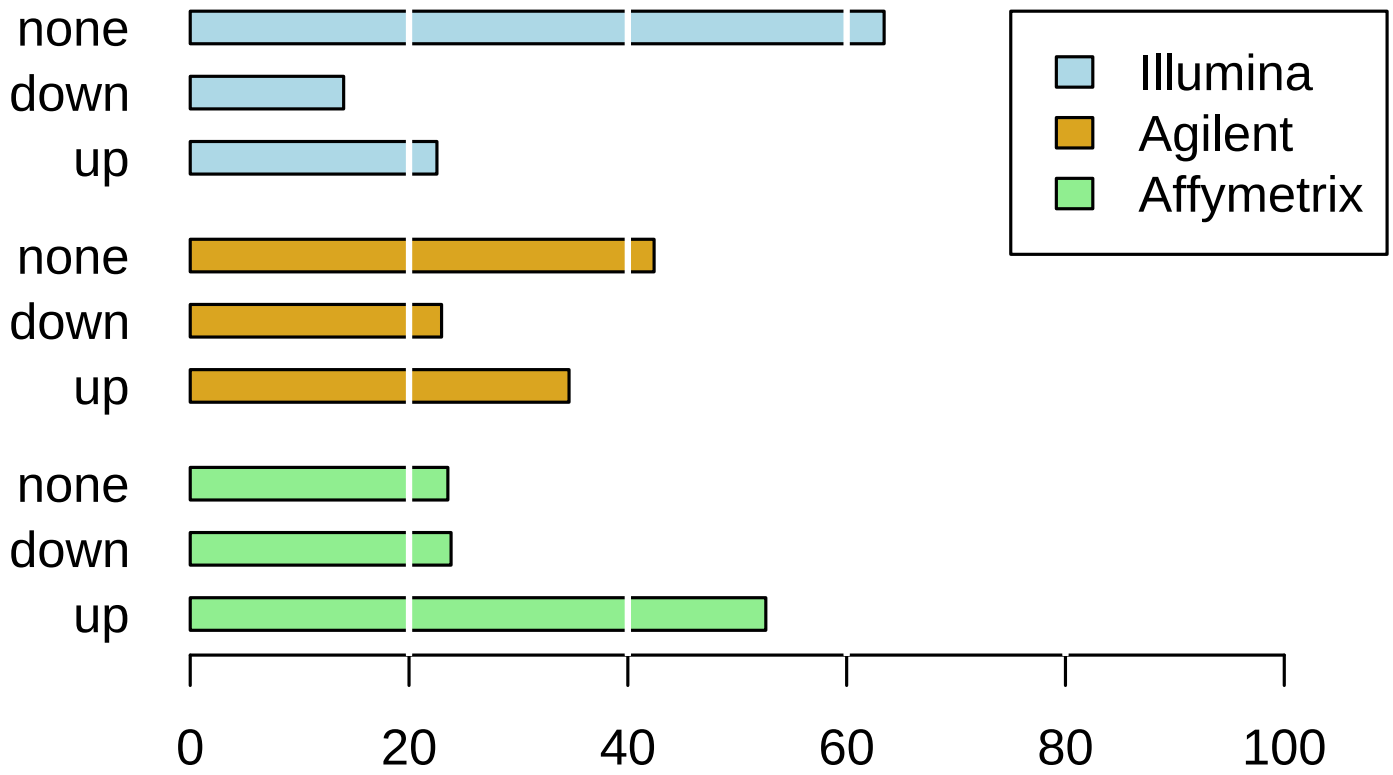
Test Setup

Settings

- *R* package `IsoGene` provided by Lin et al.
- 20000 permutations (1 week on Cluster)
- 2 Normalization Methods $\times$ 3 Platforms $\times$ 6 Animals
- 6111 well annotated genes available on all platforms
- remove one animal from Illumina data
- Family Wise Error: Bonferoni-Holm

Proportions of Significant Genes

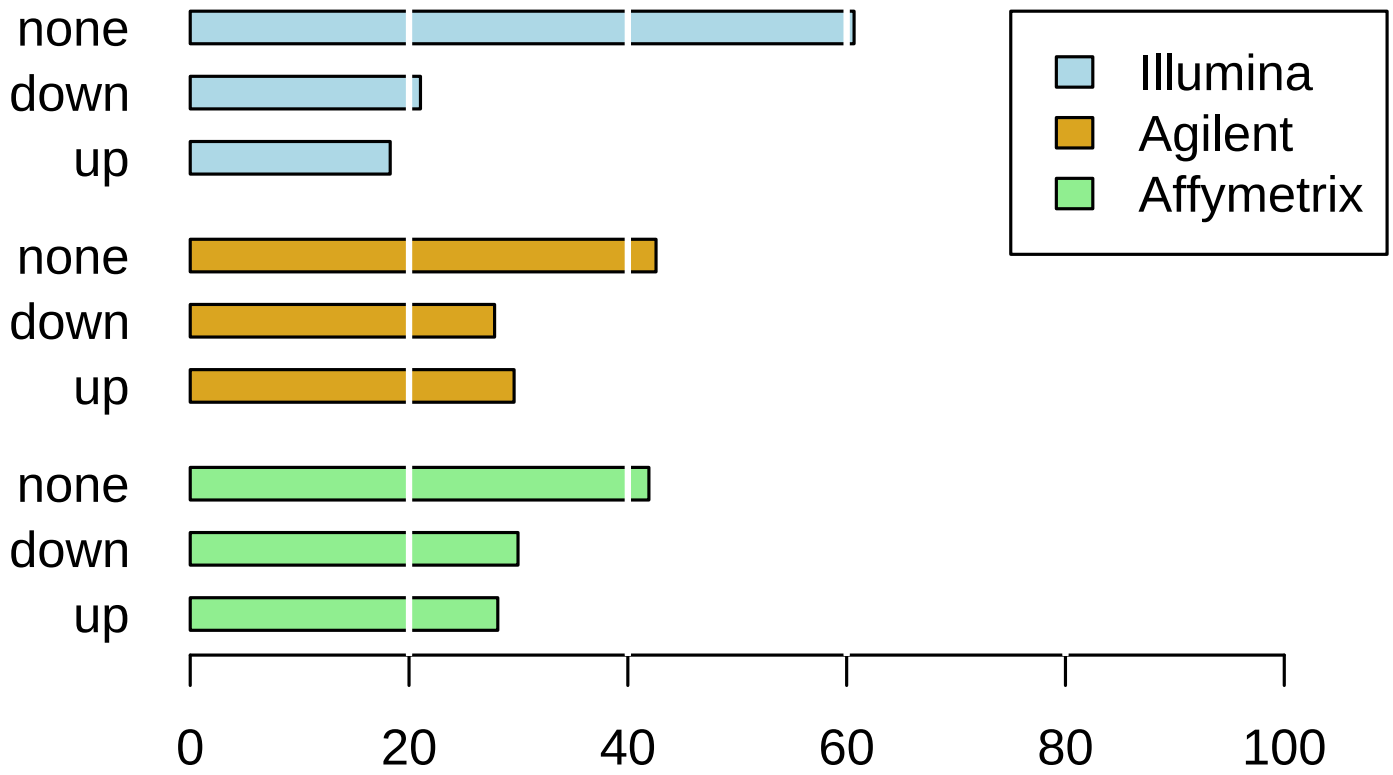
General Overview



- Baseline
- Quantile

Proportions of Significant Genes

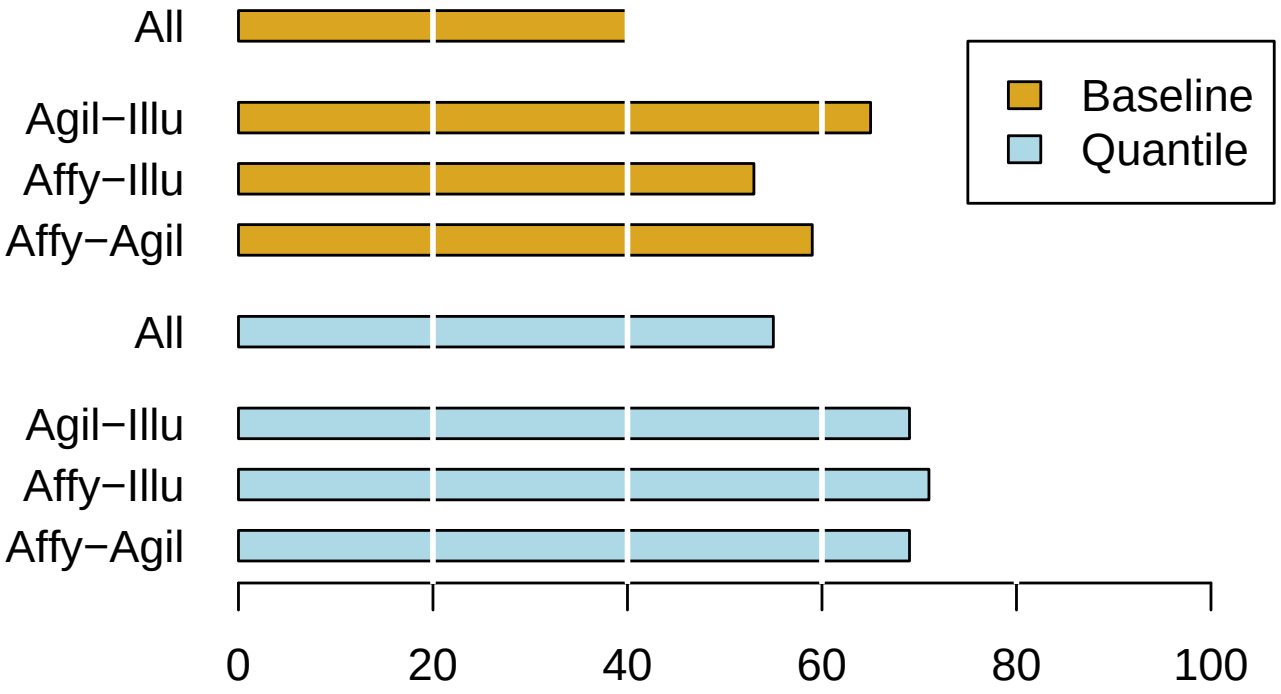
General Overview



- Baseline
- Quantile

Agreement Between Platforms

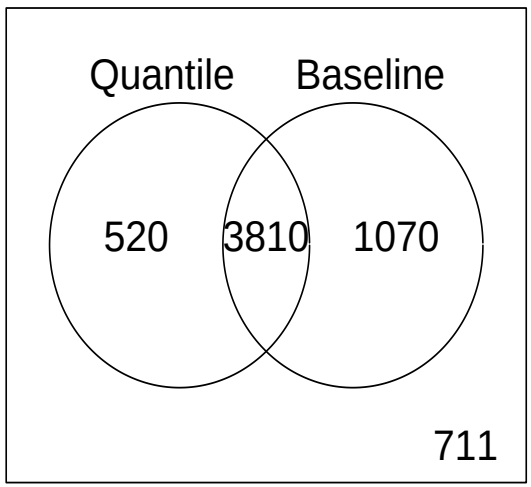
Number of Genes



- Fleiss' κ-coefficient - agreement across platforms using FWR adjusted combined p-Vaues
- Quantile Normalisation: .52
- Baseline Normalisation: .37

Agreement Between Normalizations

Number of Genes significant



Fleiss κ-coefficient: .57

- around 2 times more significant genes exclusive to baseline than to quantile normalized data
- more than 97% of genes exclusive to baseline normalized data are upregulated
- up-down in quantile exclusive genes 40:60

Summary

Results

Data

- Substantial number of genes show significant monotonicity
- Across platform agreement exceeds chance levels
- Agreement on baseline normalized data is worse
- Baseline noramlized data shows more upward trends - incomplete removal of total/messenger-RNA effect
- Genes exclusively significant in baseline data are mostly upward trends

Methods

- Isotonic regression as a means to detect monotonic trends
- p-Value combination as a means to compare results from differnt platforms.

Summary

Results

Data

- Substantial number of genes show significant monotonicity
- Across platform agreement exceeds chance levels
- Agreement on baseline normalized data is worse
- Baseline noramlized data shows more upward trends - incomplete removal of total/messenger-RNA effect
- Genes exclusively significant in baseline data are mostly upward trends

Methods

- Isotonic regression as a means to detect monotonic trends
- p-Value combination as a means to compare results from differnt platforms.

Summary

Results

Data

- Substantial number of genes show significant monotonicity
- Across platform agreement exceeds chance levels
- Agreement on baseline normalized data is worse
- Baseline noramlized data shows more upward trends - incomplete removal of total/messenger-RNA effect
- Genes exclusively significant in baseline data are mostly upward trends

Methods

- Isotonic regression as a means to detect monotonic trends
- p-Value combination as a means to compare results from differnt platforms.

Summary

Results

Data

- Substantial number of genes show significant monotonicity
- Across platform agreement exceeds chance levels
- Agreement on baseline normalized data is worse
- Baseline noramlized data shows more upward trends - incomplete removal of total/messenger-RNA effect
- Genes exclusively significant in baseline data are mostly upward trends

Methods

- Isotonic regression as a means to detect monotonic trends
- p-Value combination as a means to compare results from differnt platforms.

Summary

Results

Data

- Substantial number of genes show significant monotonicity
- Across platform agreement exceeds chance levels
- Agreement on baseline normalized data is worse
- Baseline noramlized data shows more upward trends - incomplete removal of total/messenger-RNA effect
- Genes exclusively significant in baseline data are mostly upward trends

Methods

- Isotonic regression as a means to detect monotonic trends
- p-Value combination as a means to compare results from differnt platforms.

Summary

Results

Data

- Substantial number of genes show significant monotonicity
- Across platform agreement exceeds chance levels
- Agreement on baseline normalized data is worse
- Baseline noramlized data shows more upward trends - incomplete removal of total/messenger-RNA effect
- Genes exclusively significant in baseline data are mostly upward trends

Methods

- Isotonic regression as a means to detect monotonic trends
- p-Value combination as a means to compare results from differnt platforms.

Thanks

- MSI - Martin Posch
- Statistic - Univie: Cluster

References

[1] Richard E. Barlow. *Statistical Inference Under Order Restrictions*. John Wiley and Sons Ltd, 1972.

[2] D. Lin, Z. Shkedy, D. Yekutieli, T Burzykowski, H. Gaehlmann, A. Bondt, T. Perera, T. Geerts, and L. Bijnens. Testing for trends in dose-response microarray experiments: a comparison of several testing procedures, multiplicity and resampling-based inference. *Statistical Applications in Genetics and Molecular Biology*, 2007.

[3] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. John Wiley & Sons Inc, 1988.

Thank you for your attention