Reality checks can help against drowning in data and algorithms.

Willem Talloen

Senior Manager, Janssen Pharmaceutica
Guest Professor, University of Hasselt

Biomina lunch meeting
30/03/2018

What to do with all this data?

We are drowning in information but starved for knowledge.
John Naisbitt

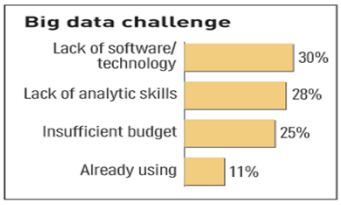
Big data is big market & Big business

"today: look at Facebook, Google, Twitter: their value is not (just) in owning the data but in aggregating, mining, filtering and repurposing it."

http://www.mediafuturist.com/data-is-the-new-oil/

Challenges in Handling Big Data

- The Bottleneck is in technology
  - New architecture, algorithms, techniques are needed
- Also in technical skills
  - Experts in using the new technology and dealing with big data



"Looks like you've got all the data —what's the holdup?"

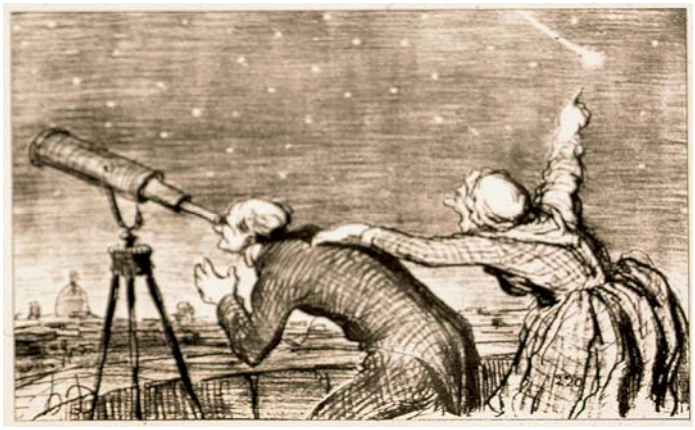
Public data

Many of the omics data generated through public funding find their way into database resources that are available online. Although the data are publicly available, there is a growing concern that much of these data sit in these databases without being used or fully analyzed. Concerns about the **growing disparity between data generation and the in-depth analysis of those data** are becoming more frequently voiced, including a perceptive comment by Harvey Blanch, professor of biochemical engineering at the University of California, Berkeley, who characterized massive omics data generation efforts as allowing us to "*know less, faster*."

Nature Chemical Biology 6, 787–789 (2010)

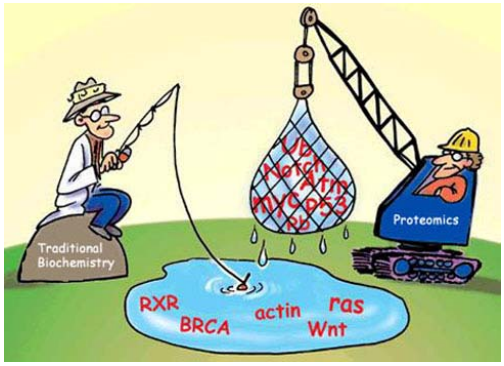
Information: The more, the better ?

- Why do we want to measure as much as possible?
  → Focusing on 'usual suspects' may miss obvious signals.



High-content biotechnologies

- Why do we want to measure as much as possible?
  - → Because we can…
  - Mass-Spectrometry, Microarrays, Sequencing, …
  - -omics data



A microarray simultaneously measures the expression of the whole (human) genome

Microarrays/NGS

- Best thing about microarrays/NGS:
- Analyse 10,000s of genes simultaneously
- Won't miss anything

- Worst thing about microarrays/NGS:
- Analyse 10,000s of genes simultaneously
- Can end up missing the interesting results in a mass of false positives

Slide from Chris Harbron

Typical workflow of high-dimensional data analysis

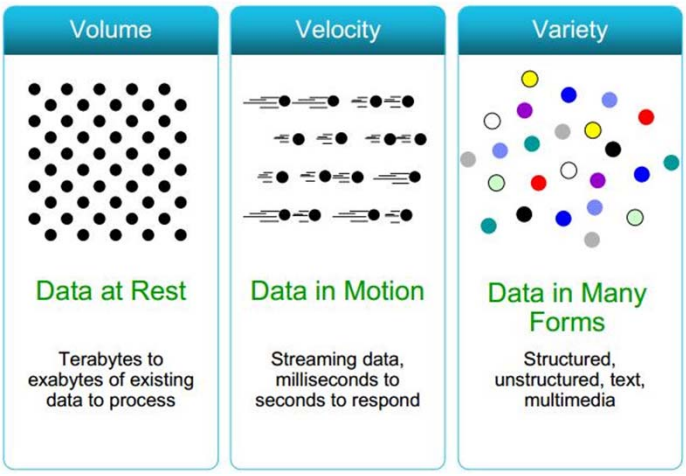
Typical workflow of high-dimensional data analysis

What is Big Data?

"Big data are high volume, high velocity, and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Gartner 2012)

Volume - exceeds limits of traditional column and row relational DB - constantly growing	Requires	**Vertical scalability** - ability to grow storage to accommodate new 'records'
Velocity - arrives rapidly, often in real time	Requires	**Data streaming** - real time processing, analysis and transformation
Variety - does not have a standard structure, e.g. text, images	Requires	**Horizontal scalability** - ability to add additional data structures

Big data



Microarrays

Problems with microarray data analysis mainly due to

1. the high dimensionality of data
2. the technology (noise)
3. the biology (most genes are non-informative)

- Signal-to-noise ratio
  - SNR compares the level of a desired signal to the level of background noise

Gene filtering

The procedure of removing genes that have no chance of being differentially expressed or predictive, regardless of the hypothesis or prediction problem that would be addressed.

- Very fruitful
  - It increases the sensitivity of the analyses*
- No common practice
  - clear guidelines are lacking
  - rather dangerous as it bears the risk to exclude some potentially relevant genes.

*Calza 2007, Talloen 2007

Gene filtering

- Gene filtering ≠ Gene selection !!
  - Non-specific – Specific
  - Unsupervised – Supervised
  - e.g. detection limit – e.g. fold change, significance

The logic behind gene filtering

Most genes are not relevant for the experiment.

Microarrays/NGS can measure entire genomes at once. For a given experiment, focusing on a certain tissue in a few conditions, it is however a certainty that not the entire genome is of relevance.

1. Not all genes are expected to be expressed.
   Most tissues express only around 10,000 - 15,000 genes*.

2. Not all expressed genes vary across samples.
   Many 'house-keeping' genes are expressed at a certain level for the maintenance of the cell.

*Su2002, Jongeneel2003

The advantage of gene filtering

It increases the sensitivity of the analyses
by reducing the dimensionality of the data set.
Less false positive results.

- It reduces the problem of
  - overfitting
  - multiple testing
    - decrease the proportion of false positives in the top gene lists
    - diminish the impact of multiple testing corrections

- Correction techniques serve as a cure, not a prevention

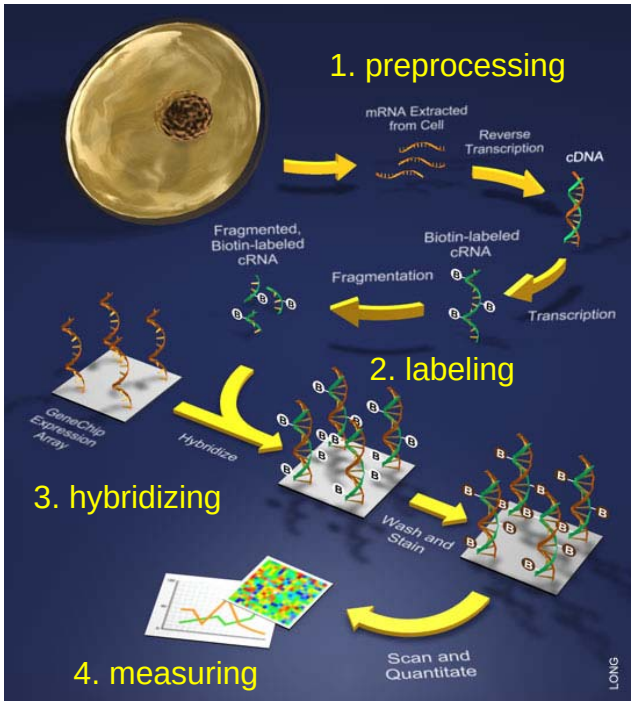
Filtering approaches

1. Filtering by signal
   - remove genes with a signal close to background level
   - unfortunately, there is no clear detection limit
2. Filtering by variation across samples
   - remove genes that do not change in a given experiment
   - the choice of what is 'too low' is very arbitrary
3. MAS5 absent/present (A/P) calls*
   - identifies whether the target transcript was detected or not by the probe set
     - A gene is called present when the PM probe intensities are statistically higher than the MM probe intensities
4. I/NI calls

*Liu 2002

I/NI calls

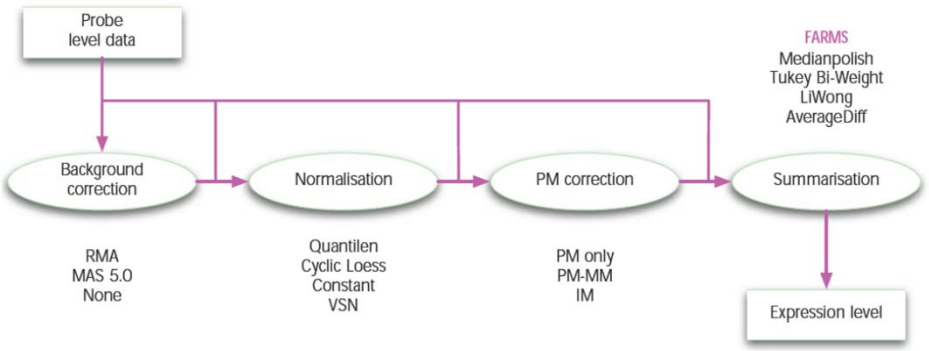
- Makes use of the multiple probes in a probeset
  - Multiple probes can be regarded as repeated measures of the same signal
- Informative probeset : the same pattern should be seen by most probes within the probeset
  - array-to-array variation is similar for the repeated probes
  - array-to-array variation > probe-to-probe variation within an array
  - biological signal > technical noise
- Bayesian estimate of a signal to noise ratio
  - Binary classification



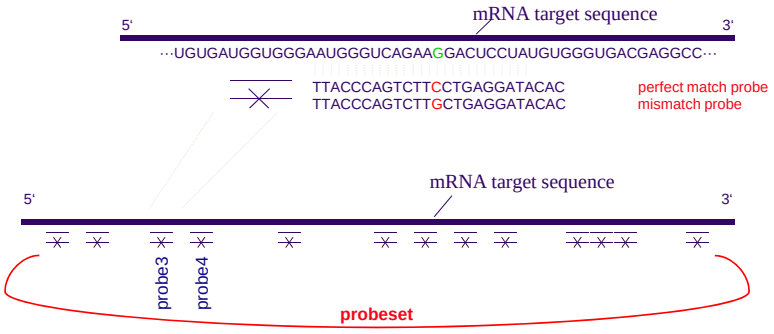
Affymetrix technology

Image Courtesy of Bioteach

Preprocessing chain



Microarray design

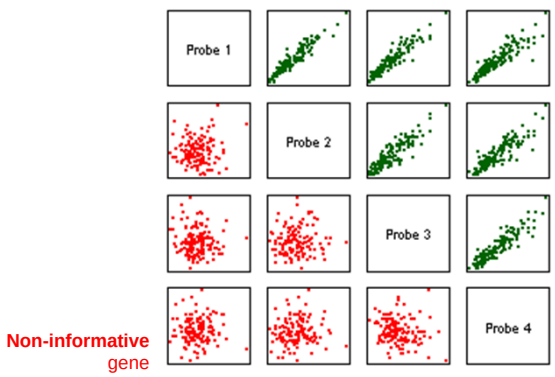


- A gene is a probeset = a set of multiple probes

  → Use the probes as repeated measurements!

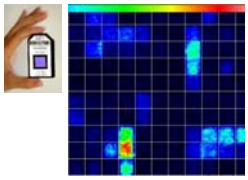
I/NI calls

- Calls genes informative or non-informative
  - No correlation amongst probes → gene is non-informative
  - High correlation amongst probes → gene is informative

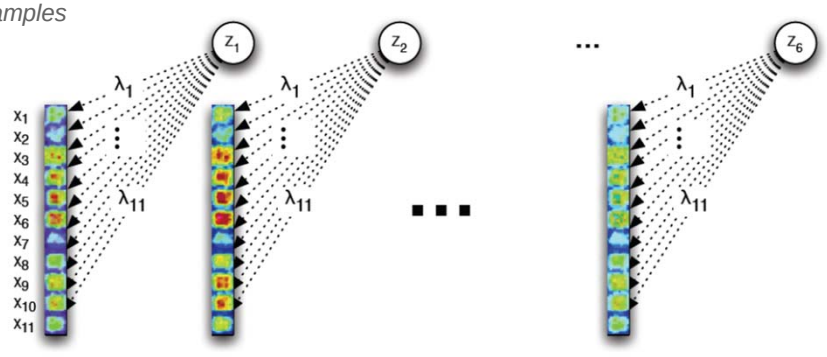


Factor analysis

$$\boldsymbol{x} = \boldsymbol{\lambda} z + \epsilon$$



11 probes
6 samples



Hochreiter et al. 2006 *Bioinformatics*

Factor analysis

$$x = \lambda z + \epsilon$$

- factor : $z \sim N(0,1)$
- noise: $\varepsilon \sim N(0, \psi)$
- data: $x \sim N(0, \lambda\lambda^T + \psi)$

- λz models the correlation between the data elements
- ε accounts for the independent noise in the data

Hochreiter et al. 2006 *Bioinformatics*

I/NI calls $x = \lambda z + \epsilon$

Variance of z given x

$$\mathrm{var}(z \mid x) = (1 + \lambda^T \Psi^{-1} \lambda)^{-1}$$

- value between [0-1]
- reflects how much variation in the probe set data *x* is explained by the factor *z*
  - informative : large λ and small ψ $(1+\infty)^{-1}$ → $var(z|x) = 0$
  - non-informative : small λ and large ψ $(1+0)^{-1}$ → $var(z|x) = 1$
  - signal-to-noise-ratio of 1 : → $var(z|x) = 0.5$

Talloen et al. 2007 *Bioinformatics*

Prior knowledge

- Observed variance in the data is often low
  → high values of λ are unlikely
- Most genes from a (whole-genome) chip are non-relevant
  → most genes with a $\lambda \approx$ zero
- Increasing mRNA conc leads to a larger signals
  → negative values of λ are not plausible

Extension of Factor analysis

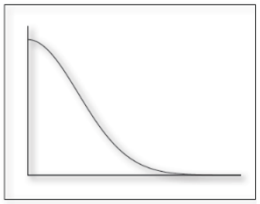
- bayesian posterior:

$$p(\boldsymbol{\lambda}, \boldsymbol{\Psi} \mid \{\boldsymbol{x}\}) \propto p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi})\, p(\boldsymbol{\lambda}, \boldsymbol{\Psi})$$

- prior assumption:

$$p(\boldsymbol{\lambda}, \boldsymbol{\Psi}) = p(\boldsymbol{\lambda})$$
$$\lambda_j = \max\{y_j, 0\}$$
$$y_j \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda)$$

  - σ_λ mean of probe variance



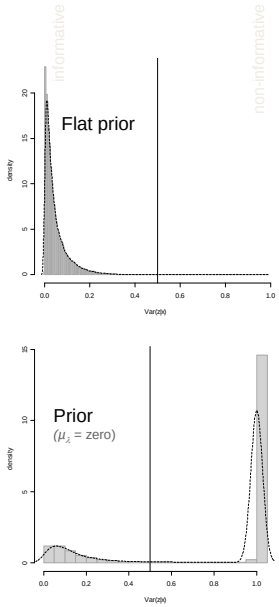
Hochreiter et al. 2006 *Bioinformatics*

I/NI calls

μ_λ = zero

- local minimum of zero for the loadings λ.
- the λ of non-informative probe sets are shrinked to zero, resulting in a *var(z|x)* close to 1

→ clear bimodal distribution of *var(z|x)* with distinct modes for non-inf. and inf. genes



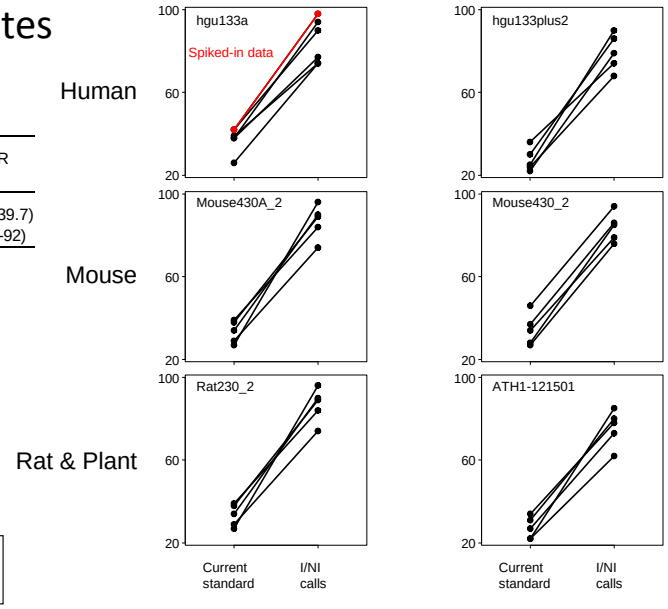
Application on real-life and spiked-in data

- Application on 30 real-life studies
  - 6 of the most commonly used microarray chips
    - Human (2)
    - Mouse (2)
    - Rat
    - Plant (Arabidopsis)
- Validation on a human spiked in data set
  - hgu133a
  - spiked in data set

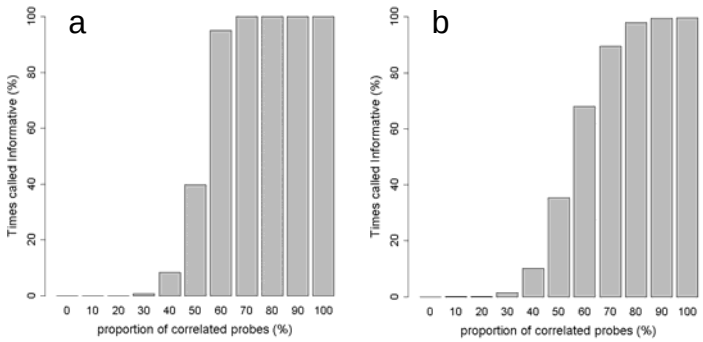
Exclusion rates

	median	IQR
A/P filtering	**33.2**	(28.8-39.7)
I/NI filtering	**83.8**	(81.9-92)

The lines represent datasets of the 31 microarray experiments



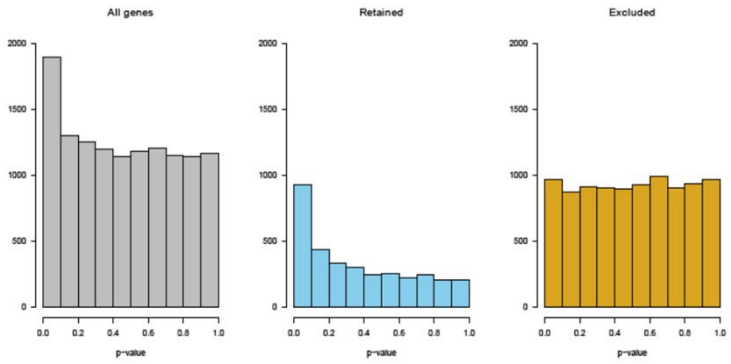
Proportion of correlated probes necessary to call a probe set informative by I/NI calls.



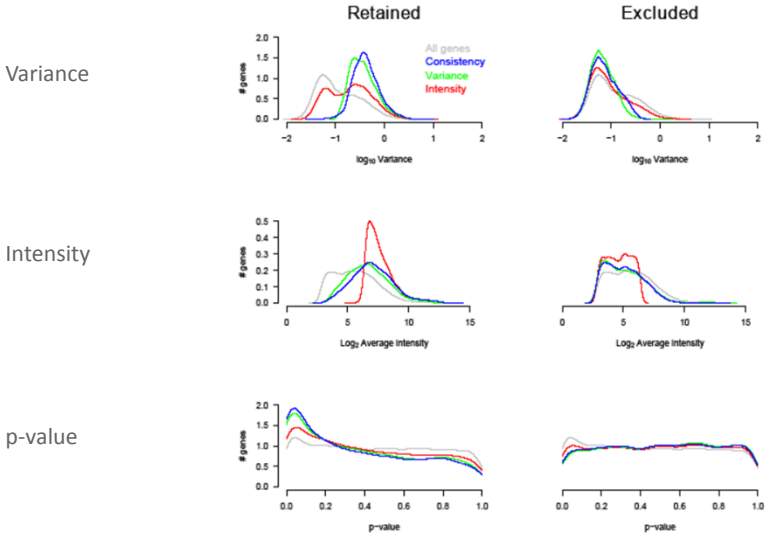
a) Simulation starting from 1000 randomly generated probe sets, so that there is a complete absence of any correlation structure between the 11 probes (i.e. 0% of the probes correlated), and making two probes perfectly correlated (i.e. 10% of the probes being correlated), then three probes (20%) until all 11 probes are all correlated (i.e. 100%).

b) Simulation starting from 1000 informative probe sets (with 11 probes) from a real data set (E-MEXP-714 from ArrayExpress), so that there is a correlation structure in the data. For simplicity, we assume that in this starting situation all 11 probes (i.e., 100%) are correlated. In reality, this is not always true as some probe sets were called informative because many –but not all– probes were correlated. Then, the intensity values are scrambled across arrays for one probe, creating a situation where at most 90% of the probes are correlated. And so forth until all probes are scrambled creating a complete lack of any correlation structure in the data (i.e., 0%).

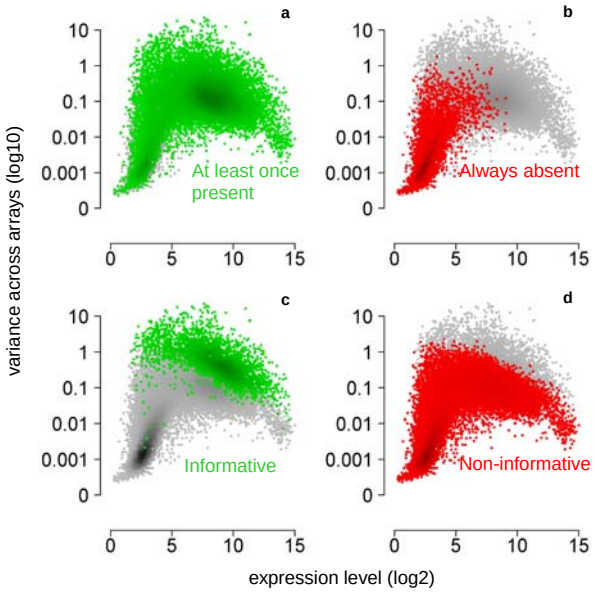
Effects of I/NI calls on statistical tests



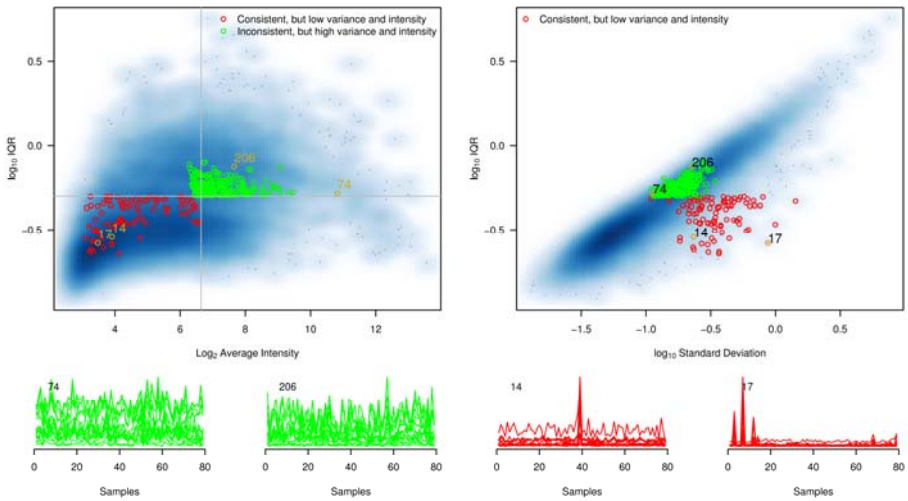
Distribution differences between gene filtering techniques



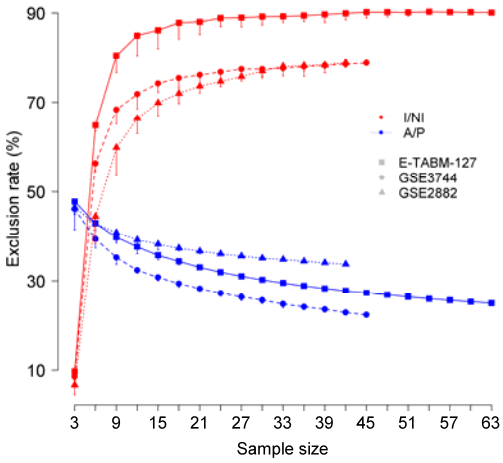
Distributional properties of I/NI and AP calls



Distributional properties of I/NI calls and var/intensity



Relation between sample size and exclusion rates for I/NI calls and A/P calls.



This result was obtained by starting from three complete data sets; E-TABM-127 (n=63), GSE3744 (n=45) and GSE2882 (n=41), and by sequentially deleting three randomly chosen arrays untill only three arrays remained. This sequential elimination of three arrays was repeated five times for every data set. The variation across these 5 simulations are represented by error bars.

Spiked Data Set

- Spiked data
  - Artificially entered 42 genes in a background mixture
  - Known concentrations

- 42 spiked-in genes among 22300 in sample

- A/P calls excludes 42.2% (keeps 12869 genes)
- I/NI calls excludes 99.5% (keeps 113 genes)
  - These 113 genes contain all 42 spike-in genes

Effects of I/NI calls on statistical tests

- I/NI calls results in
  - Smaller gene lists with fewer false positives
  - After multiple testing correction: larger gene lists

Number of significant genes:

	total
A/P filtering	12869
I/NI filtering	133

Effects of I/NI calls on statistical tests

- I/NI calls results in
  - Smaller gene lists with fewer false positives
  - After multiple testing correction: larger gene lists

Number of significant genes:

	before multiple testing correction		after correction		total
	spiked-in	not spiked-in	spiked-in	not spiked-in	
A/P filtering	29	711	7	2	12869
I/NI filtering	29	13	29	5	133

Conclusions

I/NI calls offers a critical solution to the curse of high-dimensionality in the analysis of microarray data

- It filters informative genes in a statistically sound and objective manner
- Reduces the dimensionality of the data
  - The smaller gene set contains less false positives
  - A smaller set of genes eventually needs to be tested
    - Multiple testing and overfitting less problematic

Carefully filter genes

Be always cautious when filtering genes.

- The decision not to look at certain genes generally has a dramatic impact on the final conclusions. The more stringent the filtering is applied, the larger this impact will be.
- This large effect of the filtering can however be both beneficial or detrimental, depending on whether the biologically relevant genes were kept or excluded.
- Be mindful when the signal is small (for example when only few outlying samples in a large manel may be of interest)

Literature

Talloen et al. 2007 *Bioinformatics*

Original papers

I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data.

Willem Talloen[1*§], Djork-Arné Clevert[2,3§], Sepp Hochreiter[2], Dhammika Amaratunga[4], Luc Bijnens[1], Stefan Kass[1] and Hinrich W.H. Göhlmann[1]

[1] Johnson & Johnson Pharmaceutical Research & Development, a division of Janssen Pharmaceutica n.v., Beerse, Belgium

[2] Institute of Bioinformatics, Johannes Kepler Universität Linz 4040 Linz, Austria

[3] Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin

[4] Johnson & Johnson Pharmaceutical Research & Development, Raritan, US

ABSTRACT

Motivation: DNA microarray technology typically generates many measurements of which only a relatively small subset is informative for the interpretation of the experiment. To avoid false positive results, it is therefore critical to select the informative genes from the large noisy data before the actual analysis. Most currently available filtering techniques are supervised and therefore suffer from a potential risk of overfitting. The unsupervised filtering techniques, on the other hand, are either not very efficient or too stringent as they may mix up signal with noise. We propose to use the multiple probes measuring the same target mRNA as repeated measures to

2003) and increases the risk of overfitting in classification methods (Bellman, 1961). Ideally, the high-dimensionality of microarray data should be reduced before the actual analysis by excluding all the non-informative genes. This need for suitable data reduction approaches resulted in the development of many feature selection methods to separate signal from noise, i.e. the informative from the non-informative genes. Most selection algorithms are supervised like the various methods implemented within classification algorithms (Vapnik, 2000), and the ranking of genes on fold changes or test-statistics. As supervised feature selection approaches often suffer from overfitting (Varshavsky et al., 2006) and selection bias

DEVELOPMENT OF GENE SIGNATURES

Gene signatures

A condition's gene signature is the group of genes in a type of cell whose combined expression pattern is uniquely characteristic of that condition.

- Examples of conditions
  - Response to therapy
  - Disease
  - Prognosis

- Generated by
  - Omics technologies
    - generating high-dimensional data
  - Classification algorithms/ feature selection methods
    - Extracting molecular signatures from the high-dimensional data
    - select a small number of features (genes) with high predictive accuracy (Diaz-Uriarte2006).

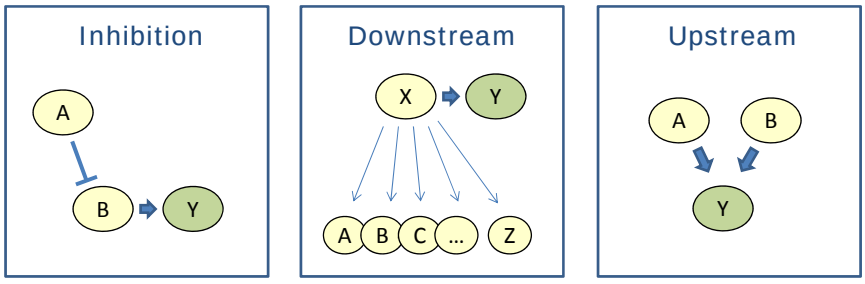
Currently implemented "Signatures"

- A number of gene expression signatures have been developed to help identify those patients at highest risk for recurrent disease
- This may avoid adjuvant chemotherapy administration to patients at low risk
  - Intrinsic breast cancer subtypes
- Currently under development
  - 70-gene Signature
  - 21-gene Signature
  - 2-gene Signature
  - 50-gene signature

Slide from Jane Chawla

	70-gene Signature	21-gene Signature	2-Gene Ratio	Intrinsic Subtypes
Analysis Approach	Supervised	Supervised	Supervised	Unsupervised
Tissue Type	Fresh or Frozen	Formalin-Fixed, Parafin-embedded	Formalin-Fixed, Parafin-embedded	Formalin-Fixed, Parafin-embedded
Technique	DNA microarrays	Q-RT-PCR	Q-RT-PCR	Q-RT-PCR
Prognostic	Untreated pts age<60, T1-2, LN-	Untreated & TAM-treated ER+/LN-	TAM-treated, ER+/LN- untreated	TAM-treated
Predictive	NO	Benefit to TAM +/- CMF/MF	Response to TAM	NO
Validation	Retrospective	Retrospective	Retrospective	Retrospective
Prospective Trials	MINDACT	TAILORX	NONE	NONE

Why a multiple marker signature?



$\frac{B}{A}$ $mean(A, B, ..., Z)$ $either\ A\ or\ B$

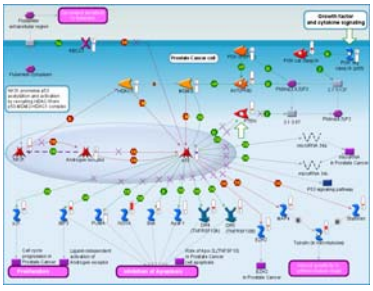
When do multiple markers outperform single markers?

1. Inhibition/catalyzation
   - The genes interact in such a way that their relative proportion is marking the endpoint of interest.
2. Downstream effects
   - Borrowing strength of markers within a pathway. The genes are coregulated and belong to the same pathway that marks the endpoint of interest. Combining the expression levels of multiple genes improves the robustness and the predictive accuracy of the biomarker.
3. Upstream effects
   - There may be multiple causes for a sample to show the phenotype. These different upstream causes should be integrated into a signature to make it more general.

Specific examples

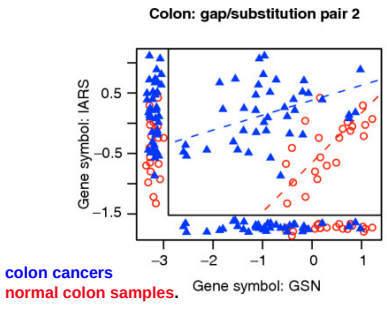
Downstream

- P53 signalling
- KRAS signalling
- IL6 signalling
- AR signalling

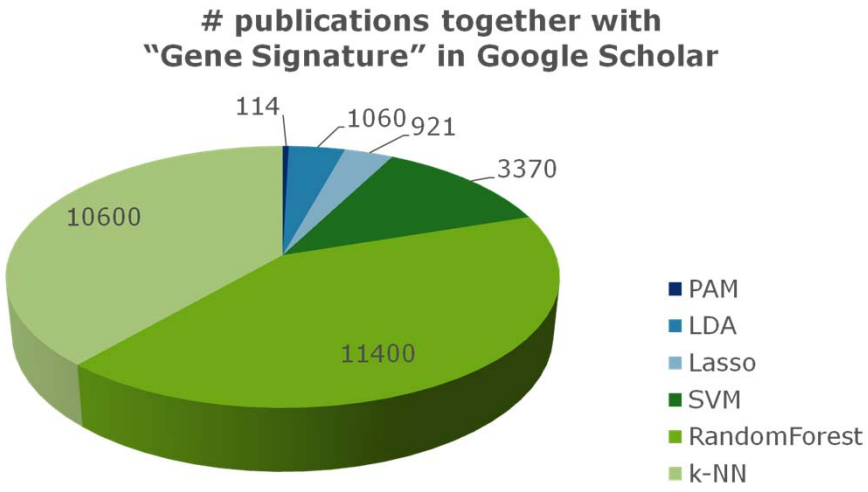


Inhibition

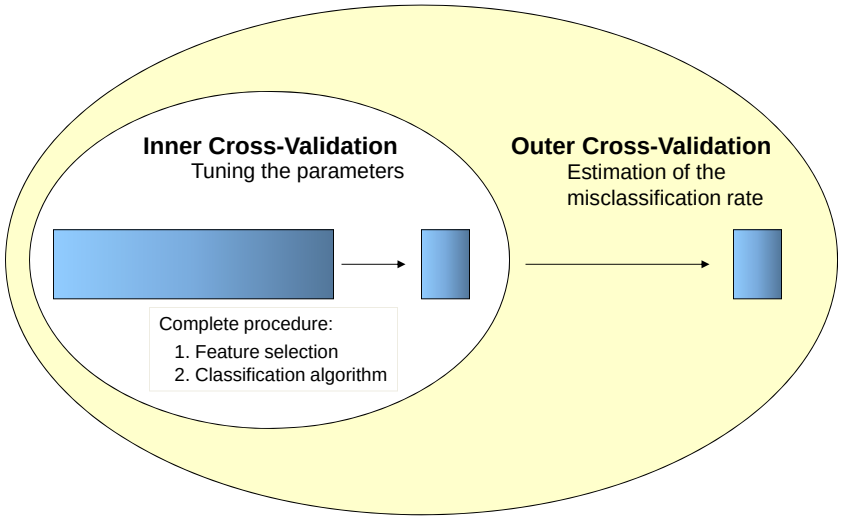
- Protein/Creatinine Ratio (PCR).
  - The 2005 UK Chronic Kidney Disease guidelines states that PCR is a better test than 24 hour urinary protein measurement.



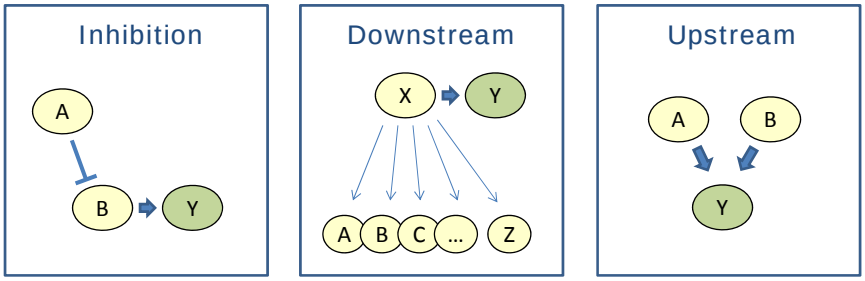
Used classification algorithms



Cross validation



Why a multiple marker signature?



Univariate :	fails	works	fails
Multivariate :	works	'average'	works

Hypothesis-driven classification

- Downstream
  1. Feature selection: gene by gene analysis
  2. Composite index (average) of top genes
- Inhibition
  - Linear model
- Upstream
  - Classification tree

Why rarely hypothesis-driven?

- Biological hypothesis formulation rare in Omics experiments
  - Exploratory searches
  - Pathway knowledge is far from comprehensive
- Omics data properties imposed new and interesting statistical challenges.

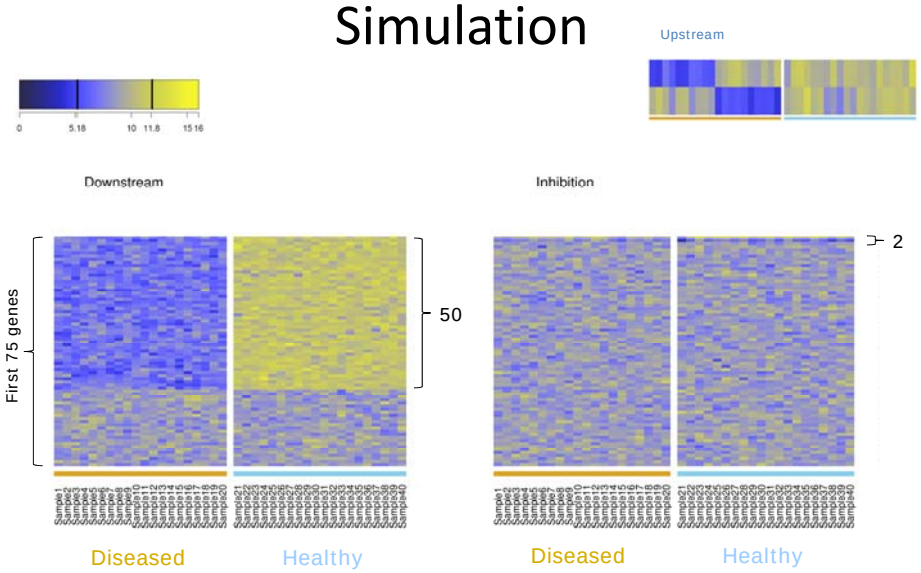
→This enthusiasm made many researchers forget to think about the biological relevance of these developed classification algorithms.

Simulation illustration

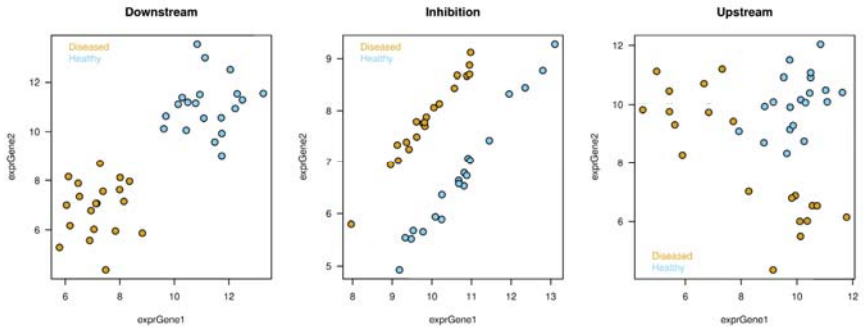
- Random data
  - 40 samples (2 groups x 20)
  - 1000 genes
- Downstream:
  - 50 differentiating genes
- Inhibition:
  - 2 genes which ratio differentiates
- Upstream:
  - Samples are differentiated either by gene 1 or by gene 2

```
# downstream
downstr <- simulateData2(nEffectRows = 50, betweenClassDifference = 4,
        nNoEffectCols = 0, withinClassSd = 1)
```

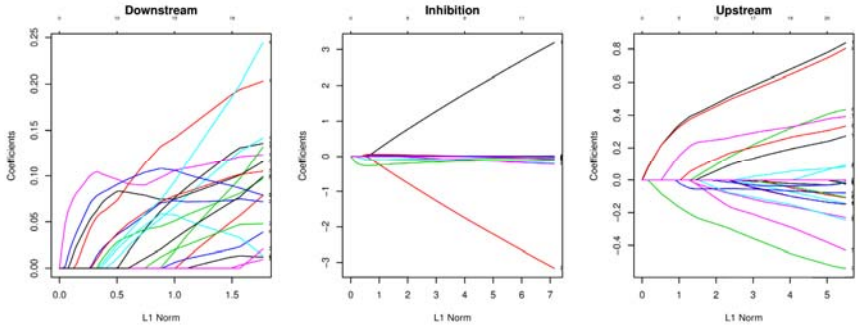
Simulation



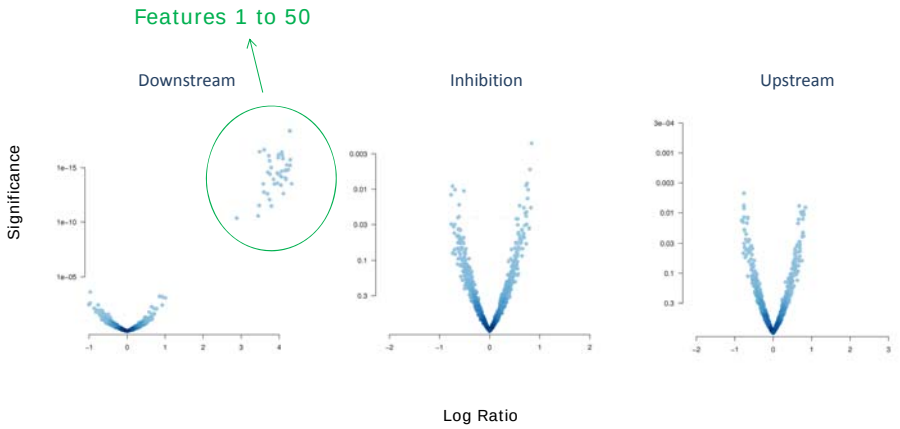
Plot of top two genes



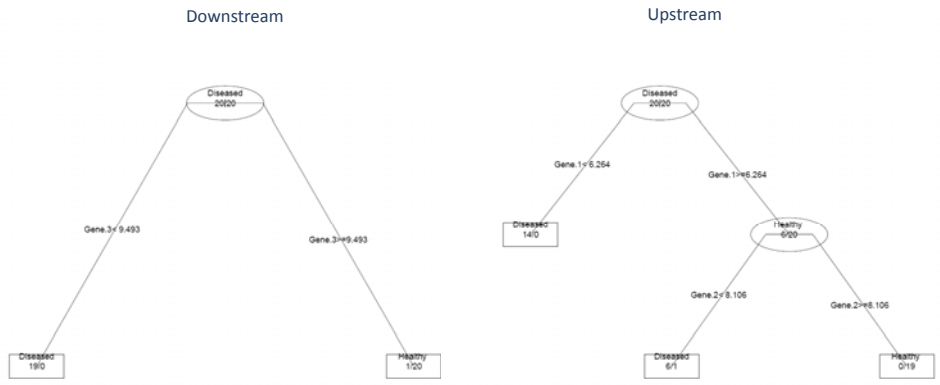
Lasso



gene by gene t-tests

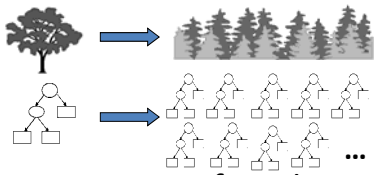


Recursive partioning



From tree to forest

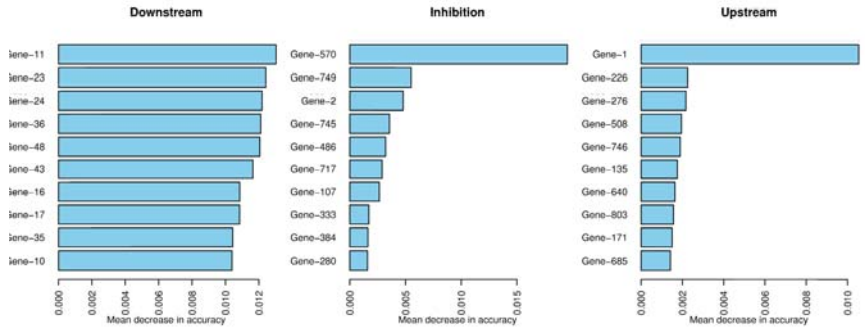
- Random forest: Construct a collection of trees and combine results of individual trees.



- For each tree two different sources of randomness:
  - random training set (bootstrap)
  - random gene selection

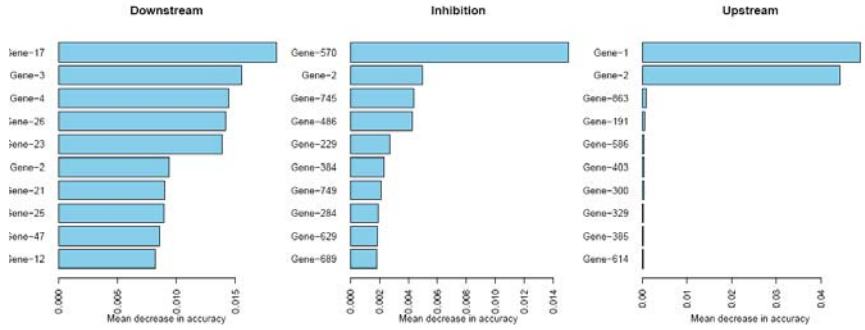
Random Forest

N = 40 samples



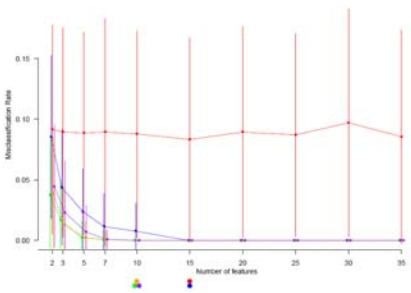
Random Forest

N = 100 samples

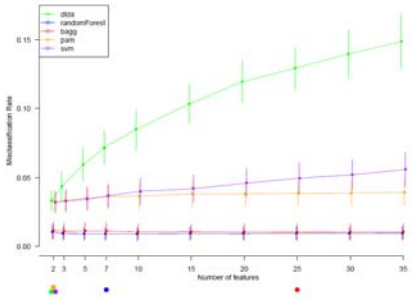


Classification

downstream



upstream



Next steps

- Investigate effects of
  - Sample size
  - Signal
  - Unbalancedness
- Asses methods based on predictive accuracy

Consequence of correlated features

- When going for 'additive effects', only one feature will be selected from a pool of correlated features.
  - Strong correlations between genes lead to non-unique solutions, impeding the biological interpretation of the obtained signature
- Going for 'weighted average' takes advantage of the correlational structure to make the signature more robust

Conclusions

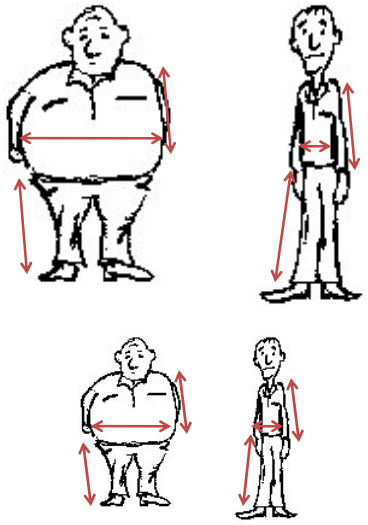
1. Statisticians should keep biology/hypotheses in mind when applying classification algorithms on Omics data
2. There are three main reasons why multiple markers outperform a single marker.
   - Downstream signalling
   - Inhibition/catalyzation
   - Different upstream causes
3. There are different statistical algorithms to adress each of these distinct hypotheses.

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

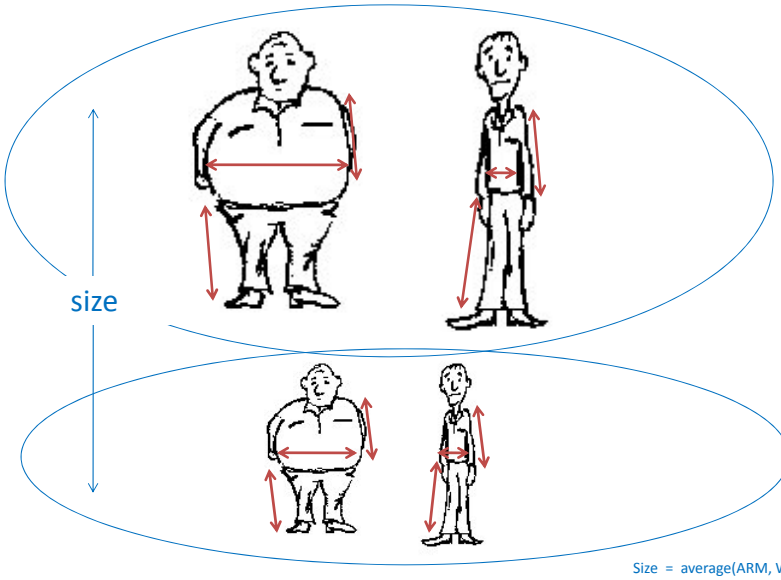
John Tukey (†2000)

Thank you

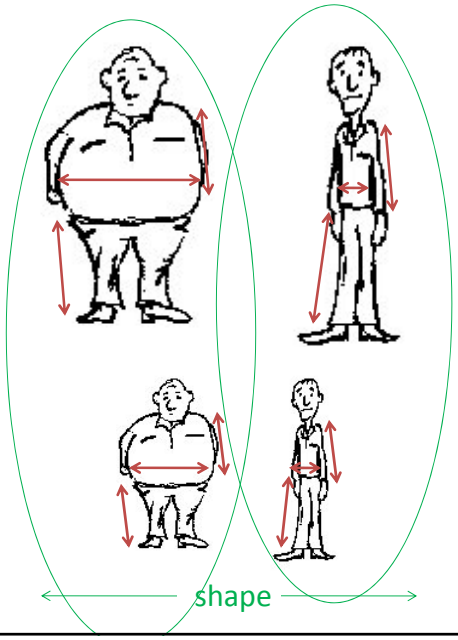
Analogy with Morphology: Differentiating people



Differentiating people

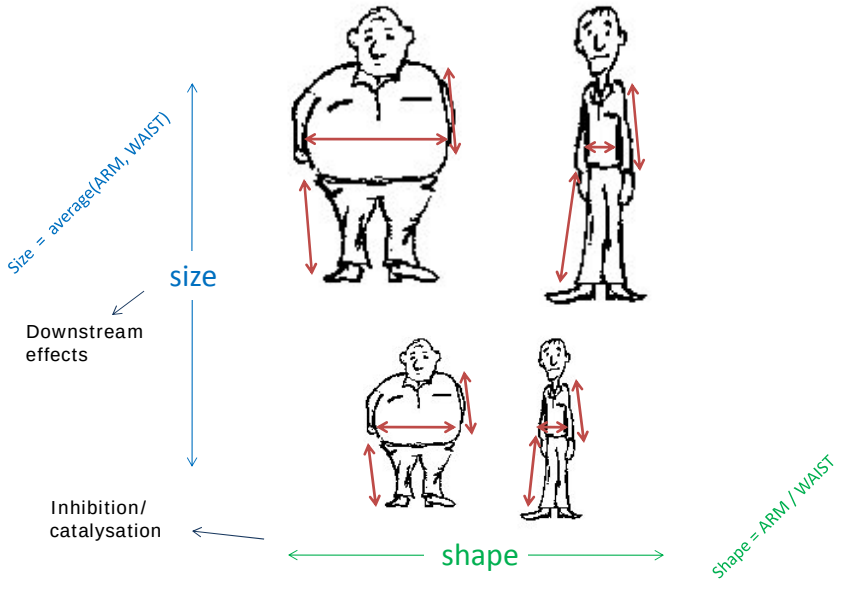


Differentiating people



Shape = ARM / WAIST

Differentiating people



How to combine morphological traits in an index

- Linear combinations after log transformation

$$\log\left(\frac{A}{B}\right) = \log(A) - \log(B)$$

$$Signature = \beta_1 * arm + \beta_2 * leg + \beta_3 * waist$$

Downstream effects

– Size: $all\ \beta_i > 0$

1st PC of PCA

Weighted average

Inhibition/ catalysation

– Shape: $\begin{cases} \beta_1 < 0\ and\ \beta_2 < 0 \\ \beta_3 > 0 \end{cases}$

LDA

Additive effects