Weighting of microarrays to improve quality of inference

-an empirical Bayes approach

Anders Sjögren, Erik Kristiansson, Mats Rudemo, Olle Nerman
Department of Mathematical Statistic, Chalmers University of Technology, Sweden

Overview (1/2)

- Microarray experiments explained
- Quality of different steps in microarray experiments varies between arrays
- Currently - outlier or non-outlier array
- We propose modelling of array specific variance components

Overview (2/2)

- Gene specific variance components with prior distribution, empirical Bayes
- A statistic is produced with known distribution
- Performance is evaluated on simulated data

Biological question

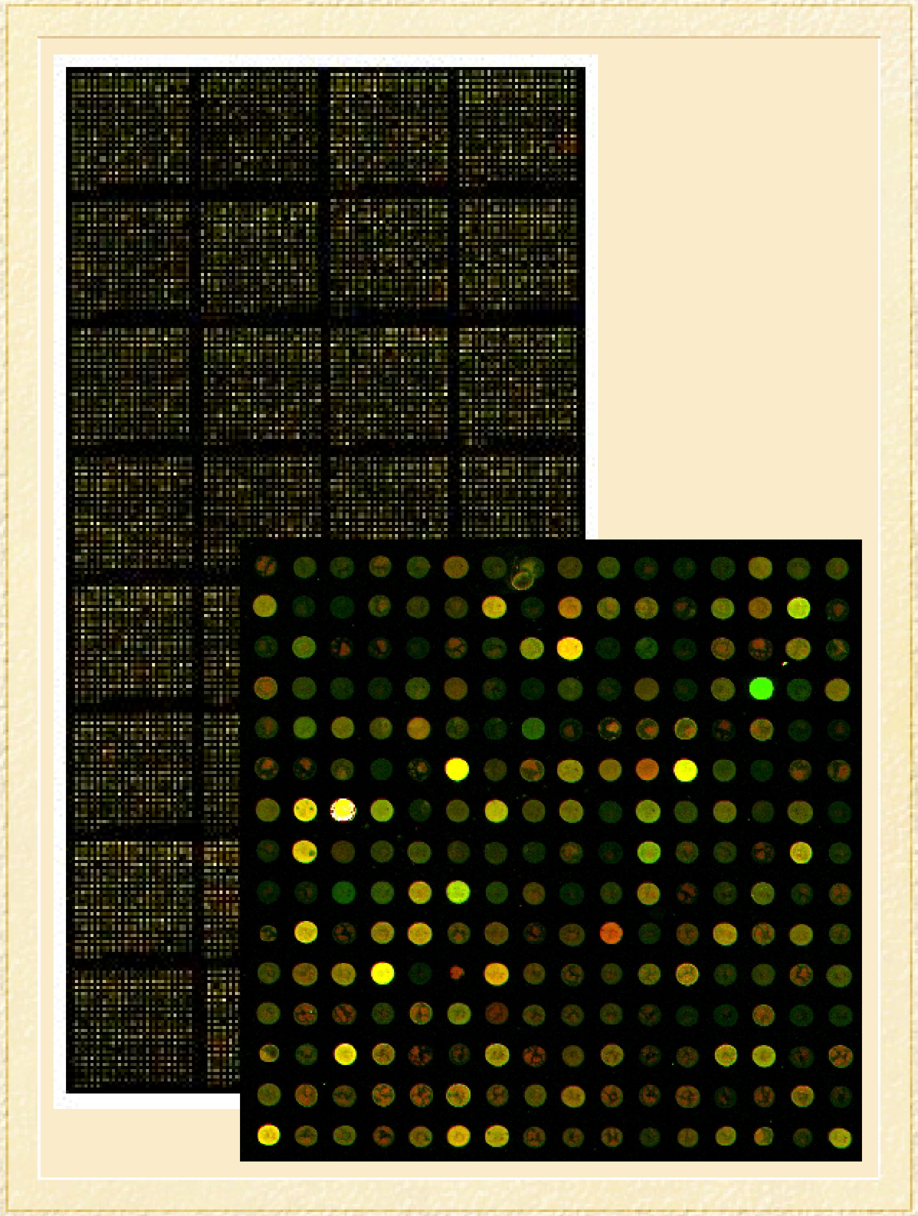
- What genes are differentially expressed between two (paired) conditions?

Differentially expressed?

- Central dogma of molecular biology:
  DNA – RNA – Protein
- Microarrays measure RNA levels.
- Two main subtechnologies:
  Two-color spotted cDNA microarrays
  Oligonucleotide microarrays (Affymetrix)

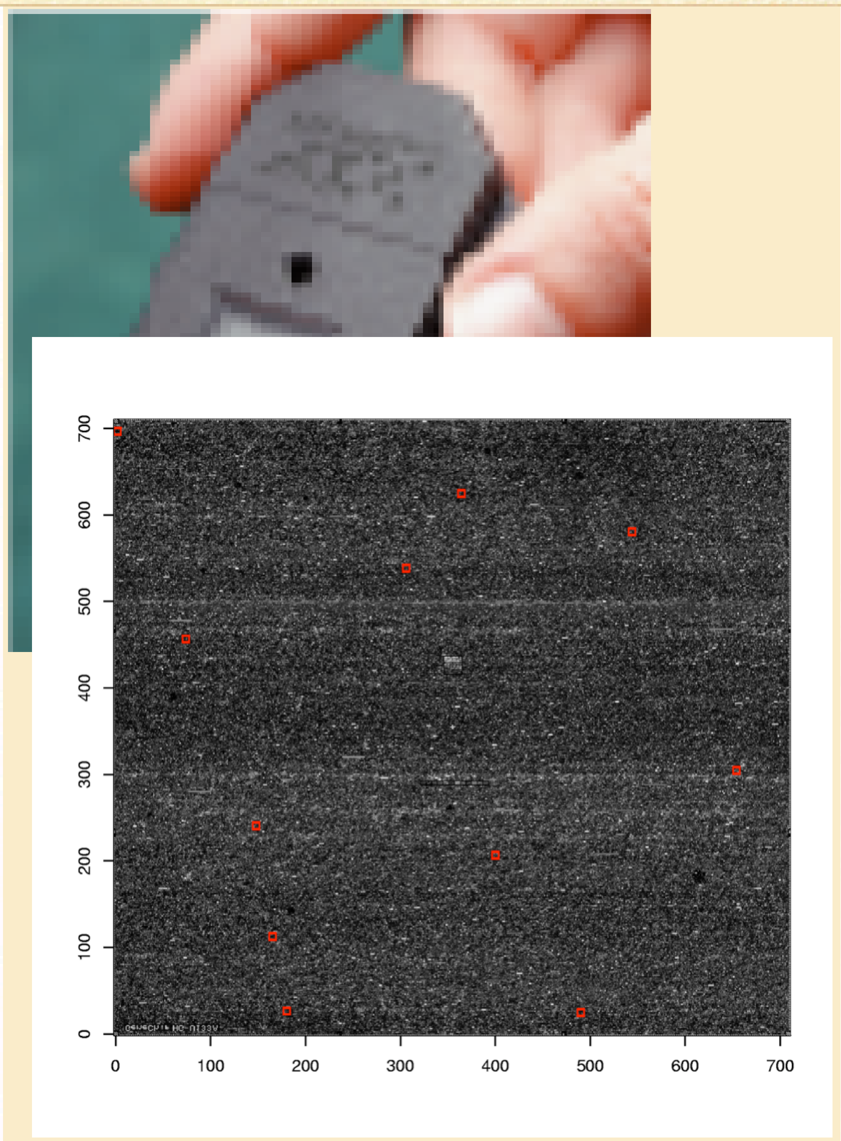
Two-color spotted cDNA microarrays

- Manufacturing – cDNA probes from reverse-transcribed mRNA
- Two colors (red and green) for different samples
- Comparative analysis



Oligonucleotide microarrays (Affymetrix)

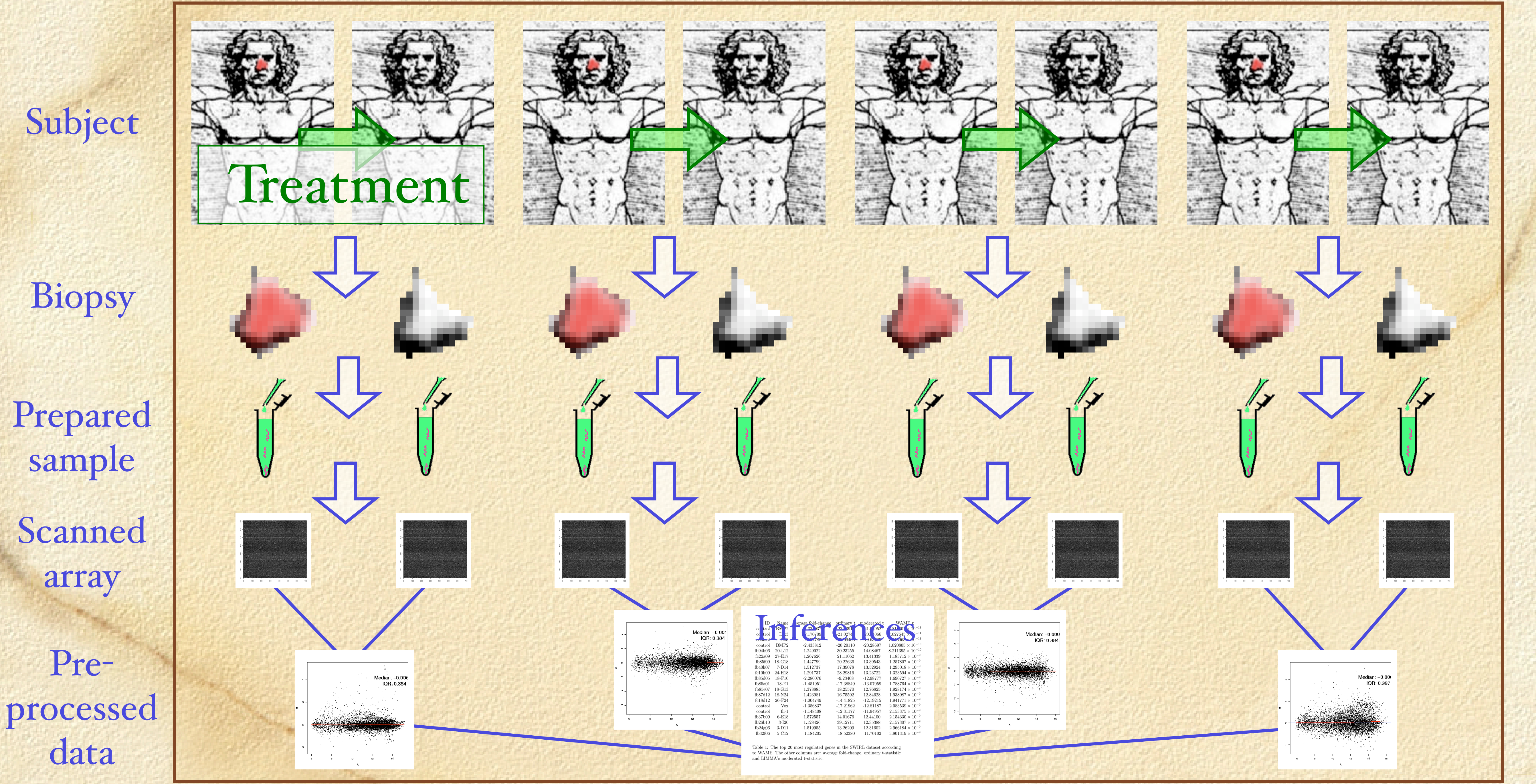
- Manufacturing – Litographic process
- One color per array
- Direct analysis
- Every gene represented by several (11-20) probes



Nature of microarray data

- Many dimensions (genes), typically 5000-44000
- Few replicates (arrays), typically 3-100

Experiment overview (affy)



Subject

Biopsy

Prepared sample

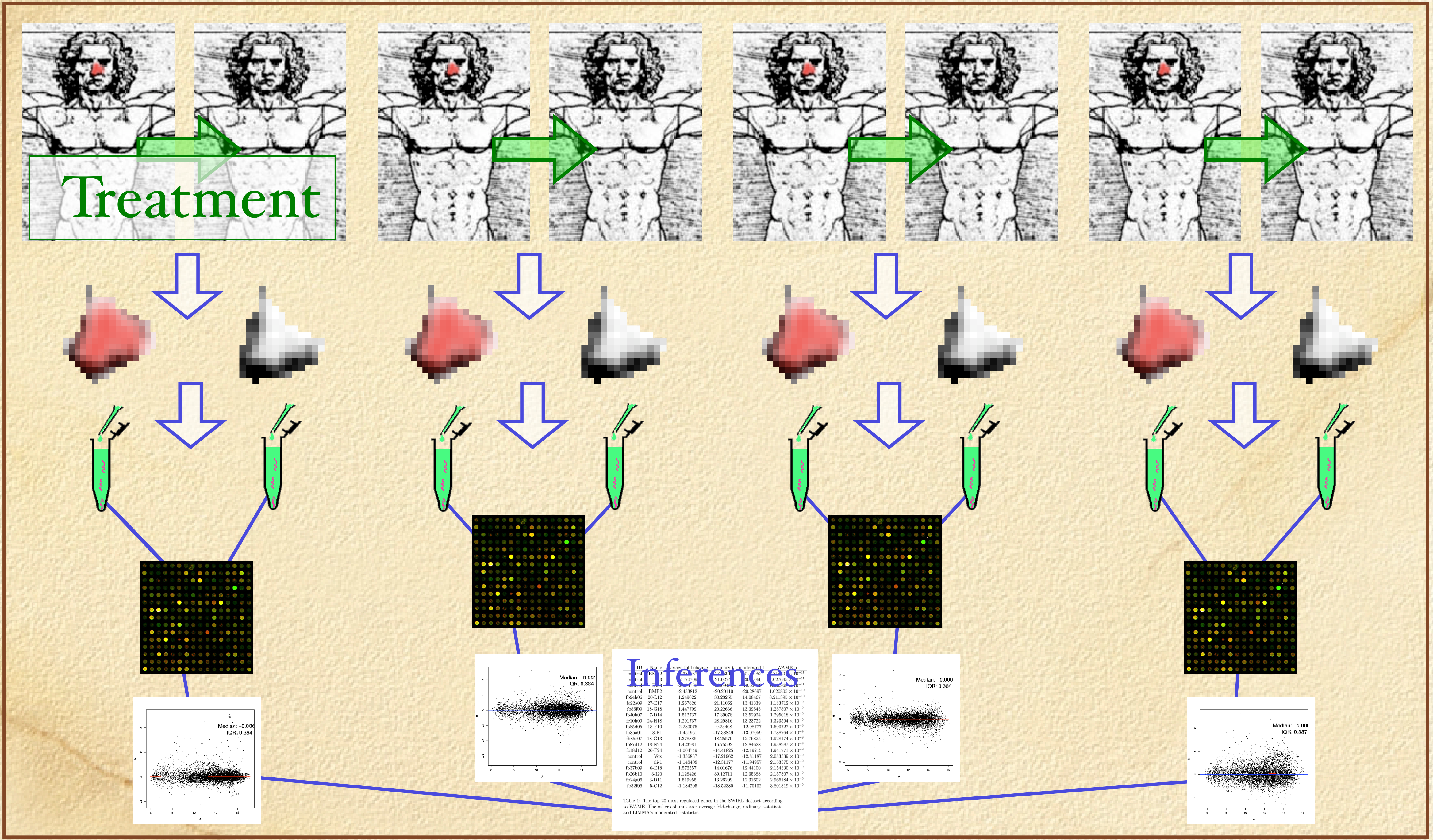
Scanned array

Pre-processed data

ID	Name	average fold-change	ordinary t	moderated t	WAME p
control	BMP2	[illegible]	[illegible]	[illegible]	[illegible] $\times 10^{-11}$
control	[illegible]	[illegible]	-21.0274[illegible]	[illegible]	7.027645 $\times 10^{-11}$
control	Dlx3	[illegible]	[illegible]	[illegible]	[illegible] $\times 10^{-11}$
control	BMP2	-2.433812	-20.20110	-20.28697	1.020805 $\times 10^{-10}$
fb94h06	20-L12	1.249022	30.23255	14.08467	8.211395 $\times 10^{-10}$
fc22a09	27-E17	1.267626	21.11062	13.41339	1.183712 $\times 10^{-9}$
fb85f09	18-G18	1.447799	20.22636	13.39543	1.257807 $\times 10^{-9}$
fb40h07	7-D14	1.512737	17.39078	13.52924	1.295018 $\times 10^{-9}$
fc10h09	24-H18	1.291737	28.29816	13.23722	1.323594 $\times 10^{-9}$
fb85d05	18-F10	-2.280076	-9.23408	-12.98777	1.690727 $\times 10^{-9}$
fb85a01	18-E1	-1.451951	-17.38849	-13.07059	1.788764 $\times 10^{-9}$
fb85e07	18-G13	1.378885	18.25570	12.76825	1.928174 $\times 10^{-9}$
fb87d12	18-N24	1.423981	16.75592	12.84628	1.938987 $\times 10^{-9}$
fc18d12	26-F24	-1.004749	-14.41825	-12.19215	1.941771 $\times 10^{-9}$
control	Vox	-1.356837	-17.21962	-12.81187	2.083539 $\times 10^{-9}$
control	fli-1	-1.148408	-12.31177	-11.94957	2.153375 $\times 10^{-9}$
fb37b09	6-E18	1.572557	14.01676	12.44100	2.154330 $\times 10^{-9}$
fb26b10	3-I20	1.128426	39.12711	12.35388	2.157307 $\times 10^{-9}$
fb24g06	3-D11	1.519955	13.26209	12.31602	2.966184 $\times 10^{-9}$
fb32f06	5-C12	-1.184205	-18.52380	-11.70102	3.801319 $\times 10^{-9}$

Table 1: The top 20 most regulated genes in the SWIRL dataset according to WAME. The other columns are: average fold-change, ordinary t-statistic and LIMMA's moderated t-statistic.

Experiment overview (cDNA)



Subject

Biopsy

Prepared sample

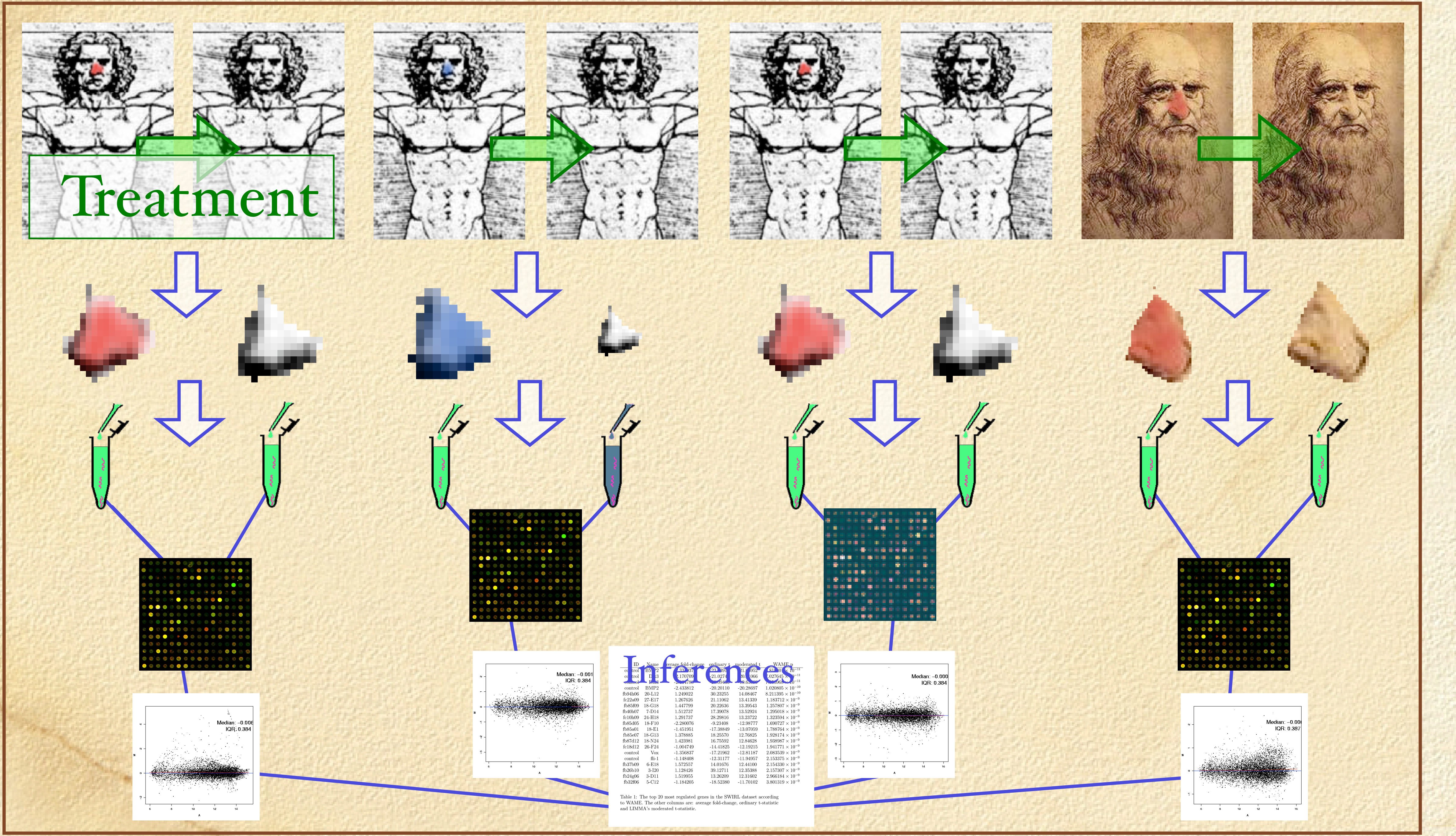
Scanned array

Pre-processed data

ID	Name	average fold-change	ordinary t	moderated t	WAME p
co[illegible]trol	BM[illegible]2	[illegible]33[illegible]37	-23.[illegible]391	[illegible]21.[illegible]952	[illegible]81[illegible]30[illegible] $\times 10^{-11}$
co[illegible]trol	D[illegible]x3	[illegible]2.170709	-21.0274[illegible]	[illegible]20.[illegible]066	[illegible]7.027645 $\times$ [illegible]
[illegible]	[illegible]	-2.[illegible]4788	[illegible].09436	-19.63[illegible]	7.[illegible]5696[illegible] $\times 10^{-11}$
control	BMP2	-2.433812	-20.20110	-20.28697	1.020805×10^{-10}
fb94h06	20-L12	1.249022	30.23255	14.08467	8.211395×10^{-10}
fc22a09	27-E17	1.267626	21.11062	13.41339	1.183712×10^{-9}
fb85f09	18-G18	1.447799	20.22636	13.39543	1.257807×10^{-9}
fb40h07	7-D14	1.512737	17.39078	13.52924	1.295018×10^{-9}
fc10h09	24-H18	1.291737	28.29816	13.23722	1.323594×10^{-9}
fb85d05	18-F10	-2.280076	-9.23408	-12.98777	1.690727×10^{-9}
fb85a01	18-E1	-1.451951	-17.38849	-13.07059	1.788764×10^{-9}
fb85e07	18-G13	1.378885	18.25570	12.76825	1.928174×10^{-9}
fb87d12	18-N24	1.423981	16.75592	12.84628	1.938987×10^{-9}
fc18d12	26-F24	-1.004749	-14.41825	-12.19215	1.941771×10^{-9}
control	Vox	-1.356837	-17.21962	-12.81187	2.083539×10^{-9}
control	fli-1	-1.148408	-12.31177	-11.94957	2.153375×10^{-9}
fb37b09	6-E18	1.572557	14.01676	12.44100	2.154330×10^{-9}
fb26b10	3-I20	1.128426	39.12711	12.35388	2.157307×10^{-9}
fb24g06	3-D11	1.519955	13.26209	12.31602	2.966184×10^{-9}
fb32f06	5-C12	-1.184205	-18.52380	-11.70102	3.801319×10^{-9}

Table 1: The top 20 most regulated genes in the SWIRL dataset according to WAME. The other columns are: average fold-change, ordinary t-statistic and LIMMA's moderated t-statistic.

Experiment overview - in reality



ID	Name	average fold-change	ordinary t	moderated t	WAME p
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible] $\times 10^{-11}$
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
control	BMP2	-2.433812	-20.20110	-20.28697	1.020805×10^{-10}
fb94h06	20-L12	1.249022	30.23255	14.08467	8.211395×10^{-10}
fc22a09	27-E17	1.267626	21.11062	13.41339	1.183712×10^{-9}
fb85f09	18-G18	1.447799	20.22636	13.39543	1.257807×10^{-9}
fb40h07	7-D14	1.512737	17.39078	13.52924	1.295018×10^{-9}
fc10h09	24-H18	1.291737	28.29816	13.23722	1.323594×10^{-9}
fb85d05	18-F10	-2.280076	-9.23408	-12.98777	1.690727×10^{-9}
fb85a01	18-E1	-1.451951	-17.38849	-13.07059	1.788764×10^{-9}
fb85e07	18-G13	1.378885	18.25570	12.76825	1.928174×10^{-9}
fb87d12	18-N24	1.423981	16.75592	12.84628	1.938987×10^{-9}
fc18d12	26-F24	-1.004749	-14.41825	-12.19215	1.941771×10^{-9}
control	Vox	-1.356837	-17.21962	-12.81187	2.083539×10^{-9}
control	fli-1	-1.148408	-12.31177	-11.94957	2.153375×10^{-9}
fb37b09	6-E18	1.572557	14.01676	12.44100	2.154330×10^{-9}
fb26b10	3-I20	1.128426	39.12711	12.35388	2.157307×10^{-9}
fb24g06	3-D11	1.519955	13.26209	12.31602	2.966184×10^{-9}
fb32f06	5-C12	-1.184205	-18.52380	-11.70102	3.801319×10^{-9}

Table 1: The top 20 most regulated genes in the SWIRL dataset according to WAME. The other columns are: average fold-change, ordinary t-statistic and LIMMA's moderated t-statistic.

Nature of microarray data - part II

□ Quality of data from arrays differ, technically and biologically (differing variances)

□ Many dimensions (genes), typically 5000-44000

□ Few replicates (arrays), typically 3-100

□ Spurious significants problem (small s)

Established models dealing with spurious significance

Efron, et al (2001): $t^g_{\text{penalized}} = \frac{\bar{x}_g}{s_{90}+s_g}$, where s_{90} is the 90:th percentile of all s_g.

Lönnstedt & Speed (2002); Smyth(2004): Empirical Bayes:

$\bar{X}_g|\mu_g, \sigma_g^2 \sim N(\mu_g, \sigma_g^2/N_I)$,

$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{N_I-1}\chi^2_{N_I-1}$ and $\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi^2_{d_0}$

Gives: $\tilde{t}_g = \frac{\bar{x}_g\sqrt{N_I}}{\sqrt{\mathbb{E}(\sigma_g^2|s_g^2)}} = \sqrt{\frac{d_0+(N_I-1)}{d_0 s_o^2+(N_I-1)s_g^2}}\bar{x}_g\sqrt{N_I}$

The proposed model

$$X_{ig}|c_g, \mu_g, \sigma_i^2 \sim N(\mu_g, c_g \cdot \boxed{\sigma_i^2})$$

$$c_g \sim \Gamma^{-1}(\boxed{\alpha}, \boxed{\beta})$$

Empirical Bayes
-estimating parameters from data

$$H_0 : \mu_g = 0, \;\; H_A : \mu_g \neq 0$$

Estimation strategy

□ ML estimate $\sigma_i^2, \alpha, \beta$ using the information of all genes. Estimated with high precision.

□ Build the statistic for μ_g with $\sigma_i^2, \alpha, \beta$ treated as known.

Estimating σ_i^2

Define $Y_j = X_{j+1} - X_1$, then $\mathbb{E}[Y] = 0$ and $\mathrm{Cov}[Y] = c_g \Sigma$, where $\Sigma = \mathrm{diag}(\sigma_2^2, \ldots, \sigma_n^2) + \sigma_1^2 \mathbf{1}_{(n-1)\times(n-1)}$.

Estimate ratios $r_i = \frac{\sigma_{i+1}^2}{\sigma_1^2}$ through transformation:

$v = (\frac{Y_1}{Y_1}, \cdots, \frac{Y_{N_I - 1}}{Y_1})$.

Numerical maximum likelihood estimation:

$l(r_1, \ldots, r_{N_I - 1} | \{X_{g,i}\}) =$

$C'' - \frac{N_G}{2} \log \left(|\tilde{\Sigma}| \right) - \frac{N_I - 1}{2} \sum_{g=1}^{N_g} \log \left(v_g' \tilde{\Sigma}^{-1} v_g \right)$, where

$\tilde{\Sigma} = \Sigma / \sigma_1^2 = \mathrm{diag}(r_1, \ldots, r_{N_I - 1}) + \mathbf{1}_{(n-1)\times(n-1)}$.

Estimating α, β

Treat σ_i^2 as known. Define $Y_j = X_{j+1} - X_1$.
Then $\mathbb{E}[Y] = 0$ and $\mathrm{Cov}[Y] = c_g\Sigma$, where
$\Sigma = \mathrm{diag}(\sigma_2^2, \ldots, \sigma_n^2) + \sigma_1^2 \mathbf{1}_{(n-1)\times(n-1)}$.

Define $S_g = Y_g'\Sigma^{-1}Y_g$, making $S_g \sim c_g\chi^2_{N_I-1}$.
Now, $S_g|\alpha, \beta \sim \Gamma^{-1}(\alpha, \beta) \cdot \chi^2_{N_I-1} = \frac{\Gamma((N_I-1)/2,1/2)}{\Gamma(\alpha,\beta)} =$
$2\beta \cdot \beta'(\frac{N_I-1}{2}, \alpha)$, where β' is the β'-distribution.

Finally, α, β are numerically ML estimated:
$l(\alpha, \beta|\{S_g\}) = C - \left(\alpha + \frac{N_I-1}{2}\right) \sum_{g=1}^{N_G} \log(s_g/2 + \beta) +$
$N_G \left[\alpha \log(\beta) + \log\Gamma\left(\alpha + \frac{N_I-1}{2}\right) - \log\Gamma(\alpha)\right]$

The statistic for μ_g

Treating $\sigma_i^2, \alpha, \beta$ as known, the unbiased statistic with minimal variance given c_g is :

$\bar{X}_g^w = (\sum_{j=1}^{N_I} 1/\sigma_j^2)^{-1} \sum_{j=1}^{N_I} \frac{1}{\sigma_j^2} X_{g,j}$,

$\bar{X}_g^w | c_g \sim N(\mu_g, \frac{c_g}{\sum_{j=1}^{N_I} \frac{1}{\sigma_j^2}})$.

Conditioning on S_g,

$f_{\bar{X}_g^w|S_g}(x|s) = \int f_{\bar{X}_g^w|c_g,S_g}(x|c,s) f_{c_g|S_g}(c|s)\, dc$, yields:

$\bar{X}_g^w|S_g \sim \mu_g + Z^{\text{st}}_{\alpha+\frac{N_i-1}{2}} \cdot \sqrt{\frac{2\beta+s_g}{\sum_{i=1}^{N_I} 1/\sigma_i^2}}$,

where Z_a^{st} is the student-Z distribution with a degrees of freedom.

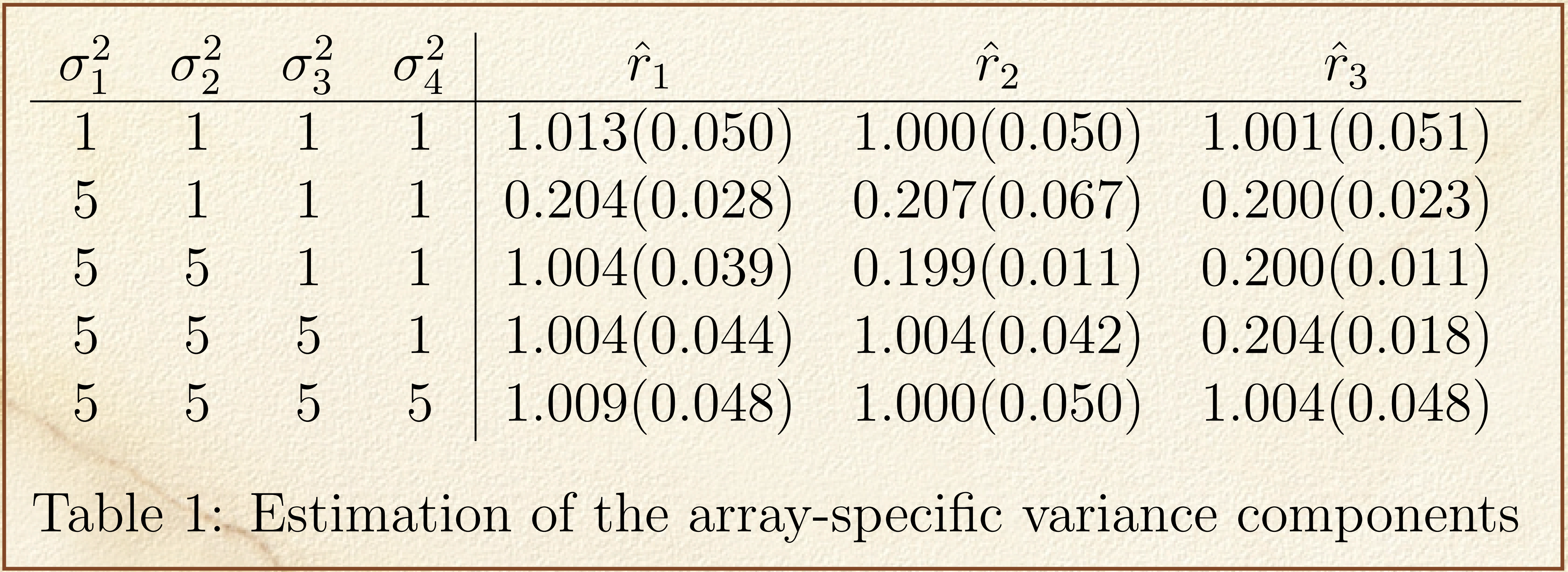
Evaluation of performance

- Simulated data according to model
- Public datasets [future]: e.g. swirl (mutated zebra-fish)
- Comparing with established statistics:
  fold change, ordinary t,
  penalized t (Efron et al),
  moderated t (Smyth; LIMMA)

Simulated data (1/3)

- 5000 genes, 4 arrays, alpha=1.5, beta=0.5
  150 regulated genes with expected value ±1

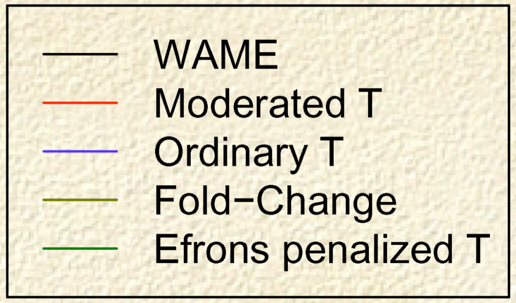
- 4 different array specific variance situations

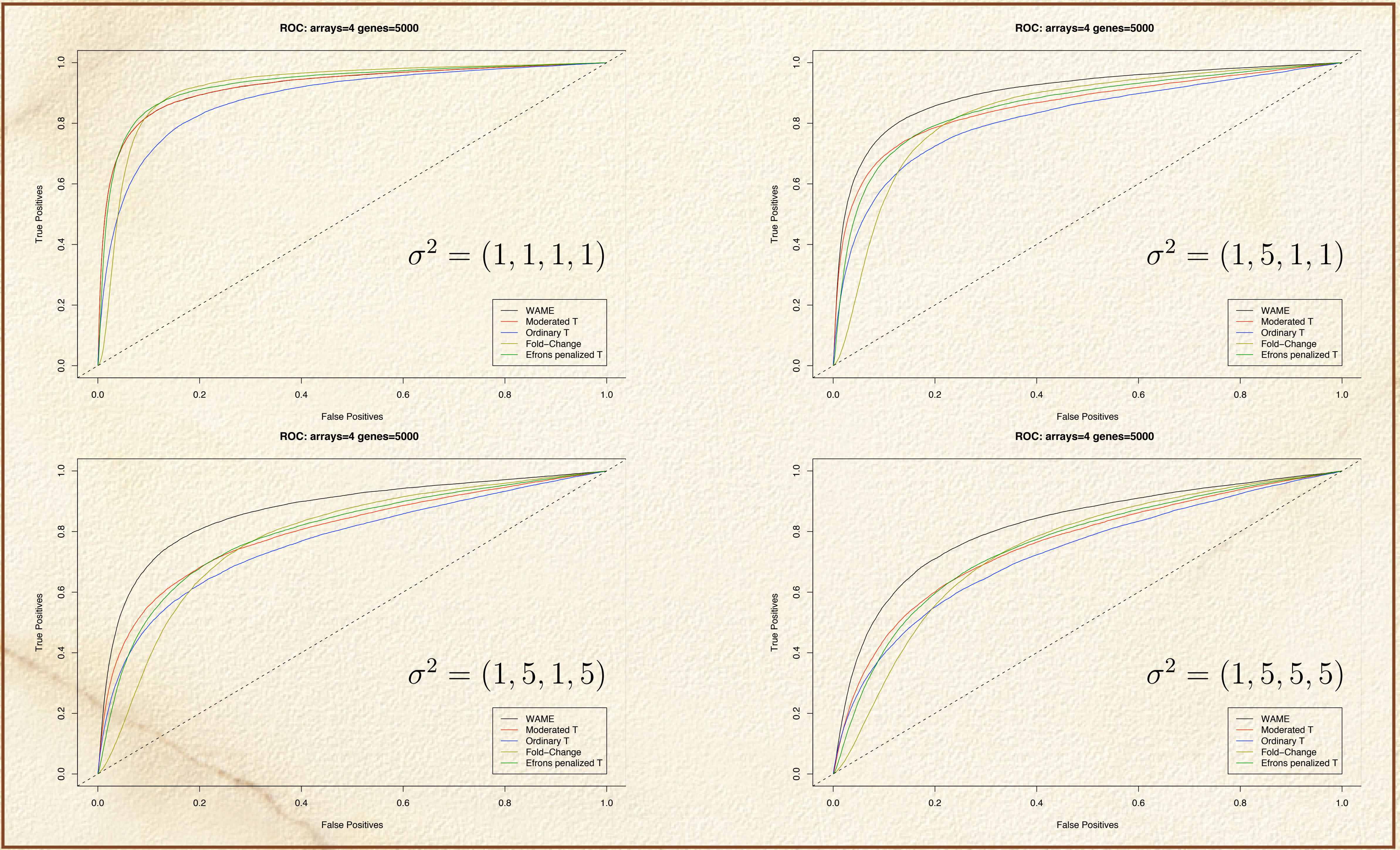


σ_1^2	σ_2^2	σ_3^2	σ_4^2	$\hat{r}_1$	$\hat{r}_2$	$\hat{r}_3$
1	1	1	1	1.013(0.050)	1.000(0.050)	1.001(0.051)
5	1	1	1	0.204(0.028)	0.207(0.067)	0.200(0.023)
5	5	1	1	1.004(0.039)	0.199(0.011)	0.200(0.011)
5	5	5	1	1.004(0.044)	1.004(0.042)	0.204(0.018)
5	5	5	5	1.009(0.048)	1.000(0.050)	1.004(0.048)

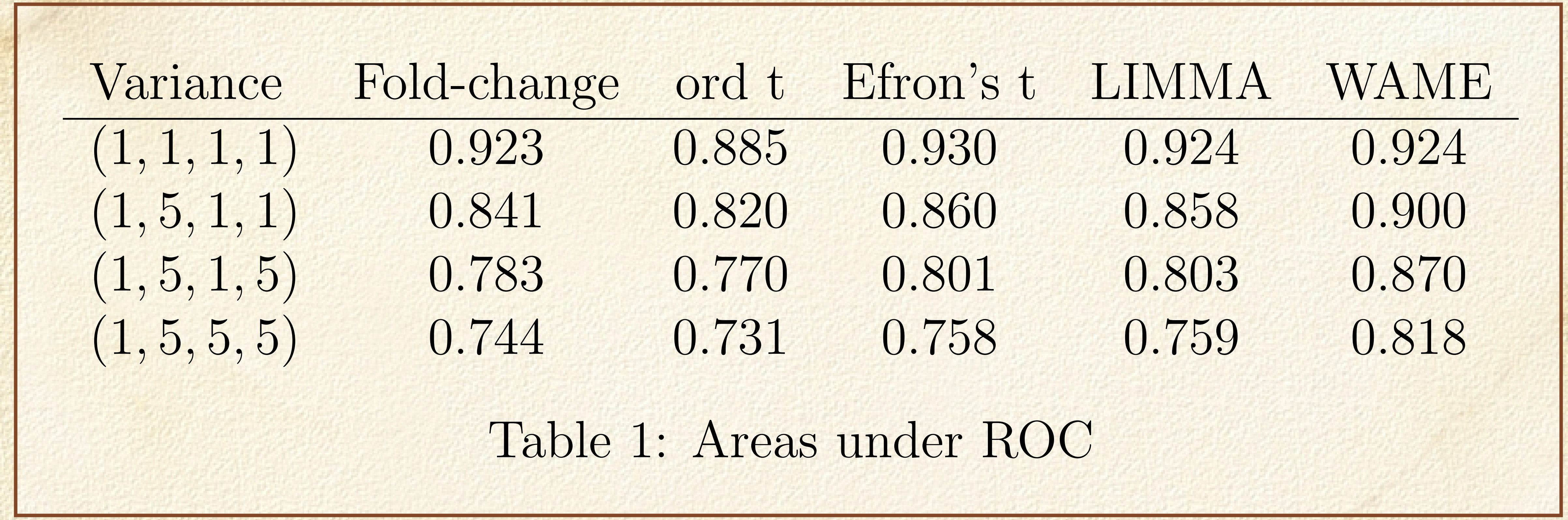
Table 1: Estimation of the array-specific variance components

Simulated data (2/3)





Simulated data (3/3)



Variance	Fold-change	ord t	Efron's t	LIMMA	WAME
(1, 1, 1, 1)	0.923	0.885	0.930	0.924	0.924
(1, 5, 1, 1)	0.841	0.820	0.860	0.858	0.900
(1, 5, 1, 5)	0.783	0.770	0.801	0.803	0.870
(1, 5, 5, 5)	0.744	0.731	0.758	0.759	0.818

Table 1: Areas under ROC

Summary

- Microarray data have differences in quality.
- The proposed method models those differences as differences in variance.
- Spurious significance must be taken care of.
- On simulated data, the proposed method performs well.

Questions?

- How to validate biologically?
- Alternative ideas for statistic?