Statistics and Microarrays

Magnus Åstrand

Jan 2008

Outline

- 45 sec introduction to microarrays
- Methods for finding differentially expressed genes
- The WAME model
- Affymetrix type arrays
- Locally moderated t-tests
- Probe-level analysis

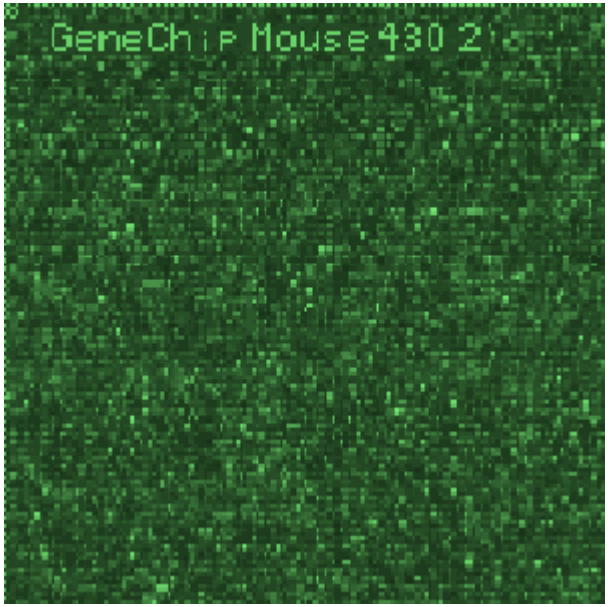
Microarrays

- Measures gene expression:
  ”Which genes are active”
  Species,Tissue and time dependent

- Many variants/platforms:
  Ex: Affymetrix

- Many ”genes” (20-40K)

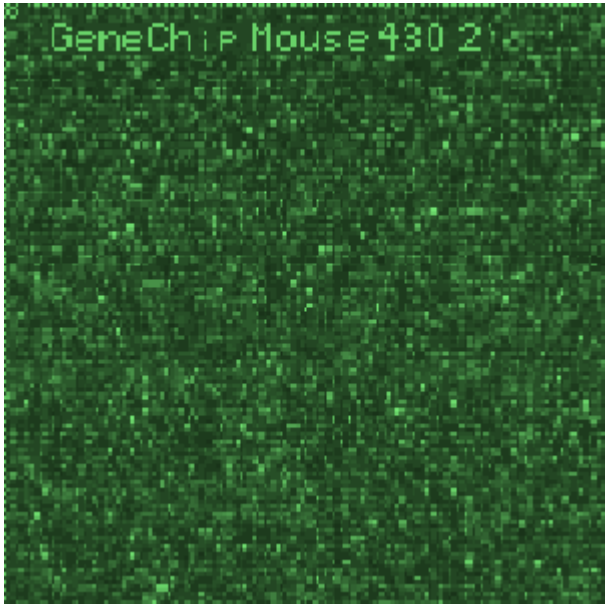
- Few observations (5-10 replicates)



Microarrays

- Measures gene expression:
  ”Which genes are active”
  Species,Tissue and time d...

Which genes are differentially expressed between two conditions?



- Many ”genes” (20-40K)

- Few observations (5-10 replicates)

Many solutions!

- Empirical Bayes methods & penalized t-test
  - Baldi & Long (2001), Lönnstedt & Speed (2002), Smyth (2004),....
  - Adjust gene-specific sample-variance (or std) towards a global estimate.
- Variance as a function of intensity
  - Eaves et al (2002), Jain et al (2003), Comander et al (2004),....
- Fully baysian methods
  - BGX (Hein et at 2005)
  - multi-mgMOS (Liu et al 2006)

Weighted & Moderated t-tests

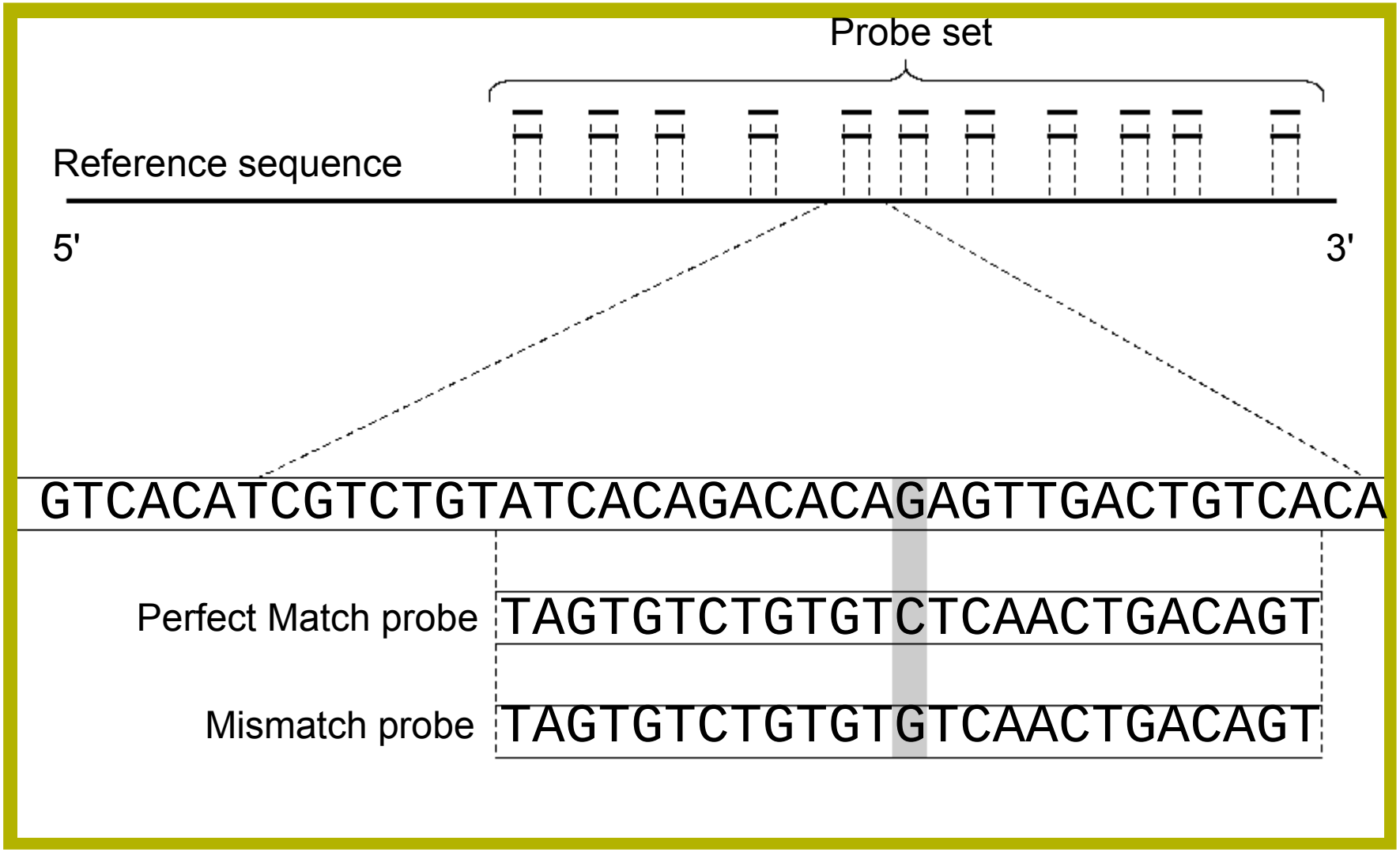
$$\boldsymbol{x}_{\boldsymbol{g}} | \sigma_g^2 \sim N_n(\mu_g, \sigma_g^2 \Sigma)$$
$$\sigma_g^2 \sim \Gamma^{-1}(\tfrac{1}{2} m, \tfrac{1}{2} m \nu)$$

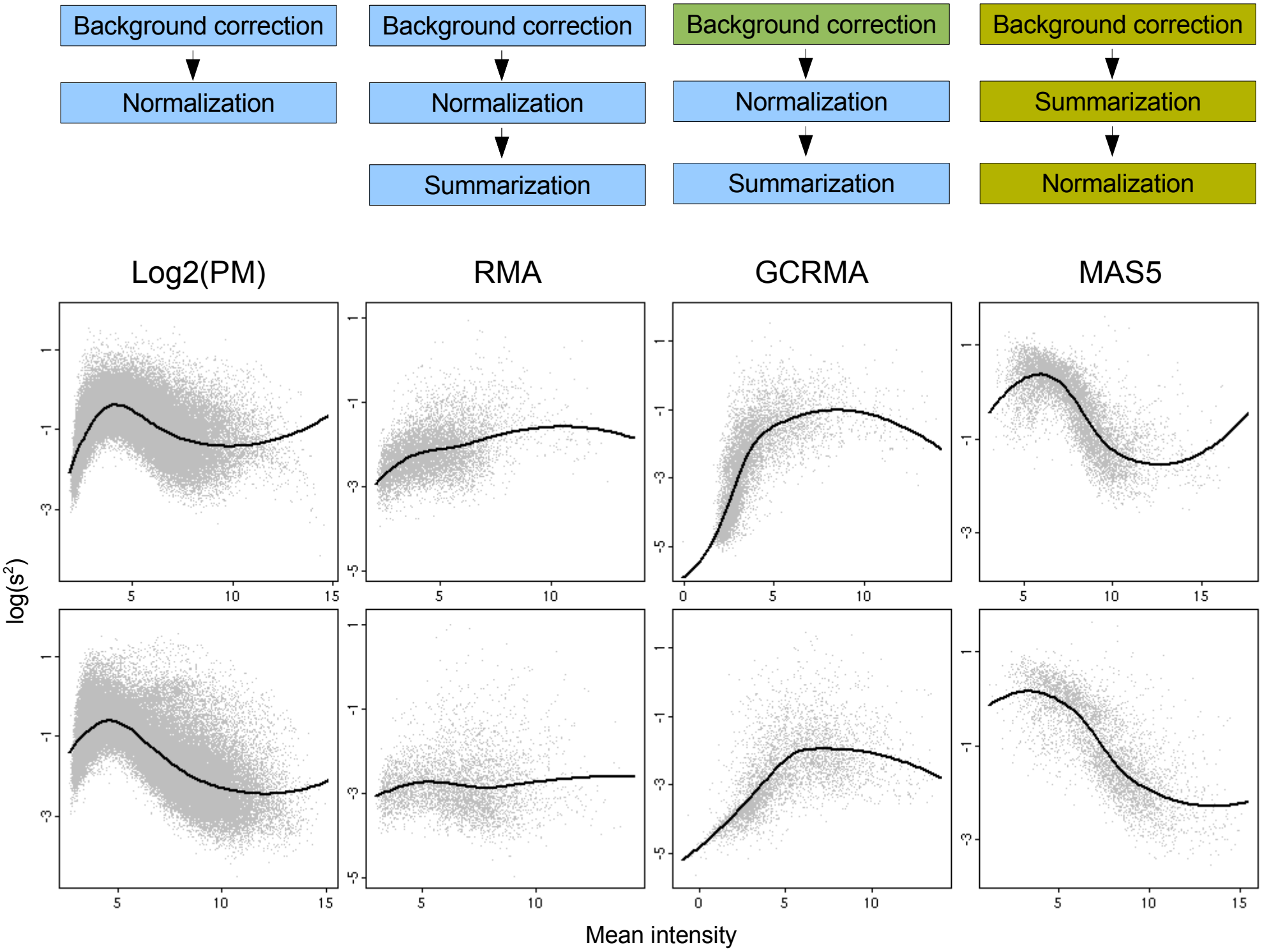
!!

WAME, Kristiansson & Sjögren et al (2005,2006,2007)

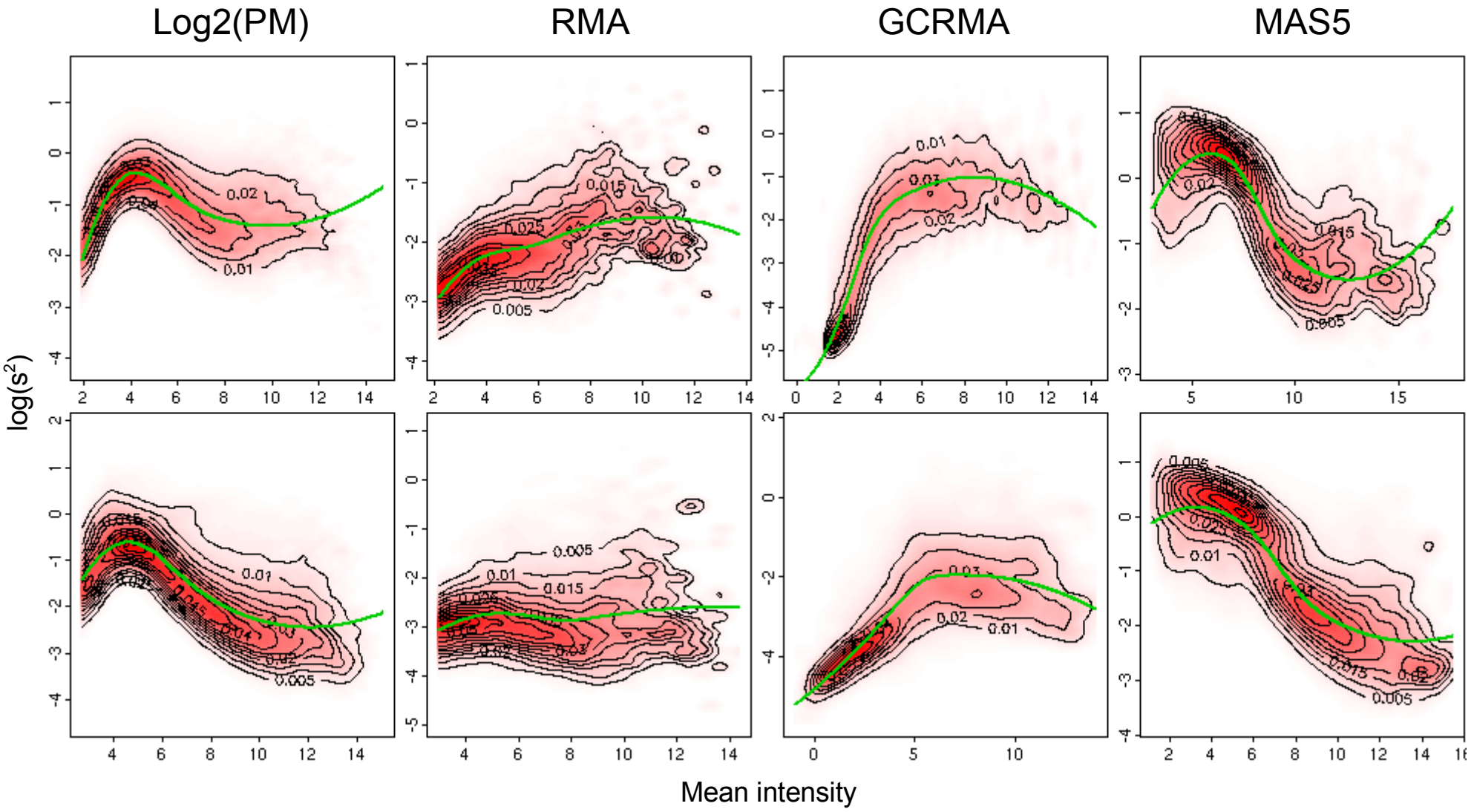
- Unequal variances
  - Poor quality on the RNA or on the array it self
- Covariances
  - Different mixtures of cell-types across the arrays
  - Unobserved covariates that influence our data (e.g. QC parameters)

Affymetrix type arrays




Background correction
Normalization
Background correction
Normalization
Summarization
Background correction
Normalization
Summarization
Background correction
Summarization
Normalization
Log2(PM)
RMA
GCRMA
MAS5
log(s2)
Mean intensity

Affymetrix type arrays



Extending the WAME-model

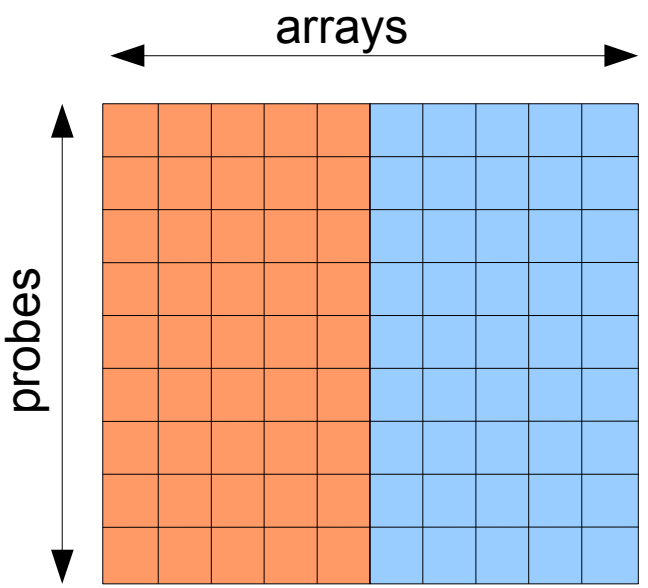
$$\boldsymbol{x}_g | \sigma_g^2 \sim N_n(\mu_g, \sigma_g^2 \Sigma)$$
$$\sigma_g^2 \sim \Gamma^{-1}(\tfrac{1}{2} m, \tfrac{1}{2} m \nu(\bar{\mu}_g))$$

!!

- EM-algorithm for fitting the model
- Probe level Locally moderated Weighted median-t (PLW)
  - Applies model on Perfect Match probe data
- Locally Moderated Weighted-t (LMW)
  - Applies model on expression indexes

Affymetrix type arrays

Perfect match intensity-matrix
(background corrected, normalized and logged)



RMA
(median polish)

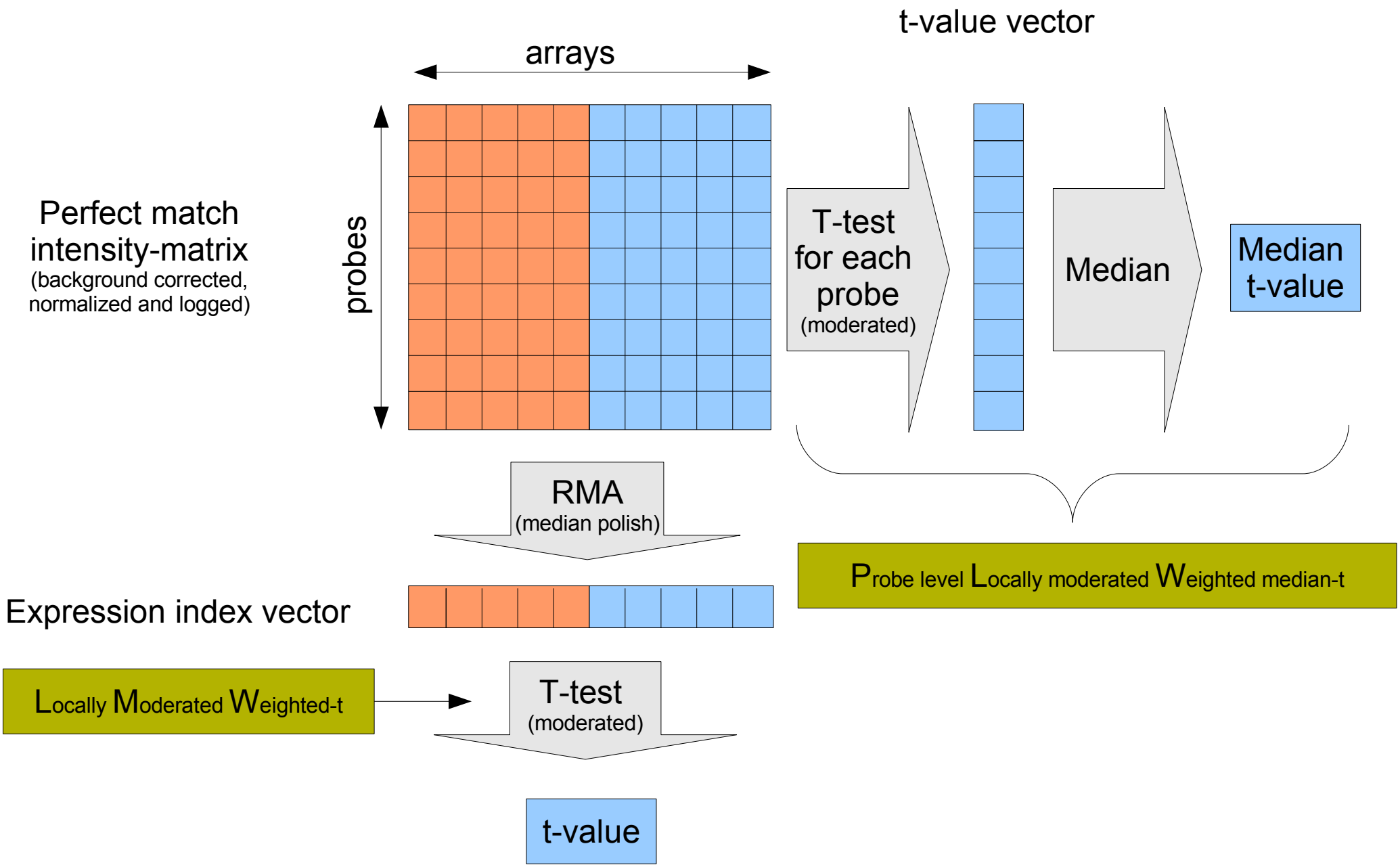
Expression index vector

Locally Moderated Weighted-t

T-test
(moderated)

t-value

Affymetrix type arrays



Does it work?

5 spike-in datasets

PLW & LMW compared with 11 other methods.

Based on RMA processed data.

Results summarized by ROC-curve AUC

Does it work?

5 spike-in datasets

PLW & LMW compared with 11 other methods.

Based on RMA processed data.

Results summarized by ROC-curve AUC

	Affymetrix		Golden	Gene Logic	
Method	U95	133A	Spike	Tonsil	AML
PLW	96(1)	93(6)	40(1)	87(1)	86(1)
LMW	96(2)	94(1)	32(3)	84(3)	80(4)
LPE	94(5)	93(10)	38(2)	84(2)	85(2)
WAME	95(3)	94(2)	32(7)	81(5)	78(7)
Efron-t	94(6)	93(4)	32(5)	79(7)	79(5)
IBMT	95(4)	94(3)	32(8)	78(8)	76(8)
FC	92(11)	93(5)	31(10)	83(4)	85(3)
logit-T	94(9)	92(11)	-(-)	80(6)	79(6)
LIMMA	94(7)	93(7)	32(9)	76(9)	75(9)
SAM	94(8)	93(8)	32(5)	74(11)	74(10)
Shrink-t	94(10)	93(9)	32(4)	75(10)	73(11)
PPLR	88(12)	90(12)	-(-)	71(12)	69(12)
t-test	85(13)	86(13)	25(11)	57(13)	52(13)

Area under ROC curves up to 100 false positives => optimum is 100. Numbers within parenthesis are within data set ranks for the methods compared. Methods are ordered with respect to mean rank across the five data sets.

Does it work?

The same analysis but with

GCRMA

and

MAS5

replacing

RMA

		Affymetrix		Golden	Gene Logic	
	Method	U95	133A	Spike	Tonsil	AML
GCRMA	PLW	97(1)	92(8)	38(2)	87(1)	87(1)
	LMW	95(2)	93(1)	21(10)	84(2)	79(4)
	LPE	95(4)	91(10)	40(1)	82(4)	86(2)
	IBMT	95(3)	93(2)	27(4)	81(7)	76(6)
	Efron-t	94(5)	93(4)	26(5)	82(5)	79(5)
	WAME	94(8)	93(3)	26(6)	83(3)	75(8)
	LIMMA	94(7)	93(5)	27(3)	80(8)	73(9)
	FC	93(10)	93(7)	25(7)	81(6)	86(3)
	SAM	94(6)	93(6)	24(9)	79(9)	76(7)
	Shrink-t	94(9)	92(9)	25(8)	78(10)	70(10)
	t-test	86(11)	84(11)	15(11)	64(11)	53(11)
MAS5	LMW	89(1)	87(1)	54(1)	79(1)	70(2)
	IBMT	87(2)	87(2)	52(2)	77(3)	69(3)
	LPE	84(3)	84(3)	49(3)	78(2)	79(1)
	WAME	71(6)	81(5)	15(6)	69(4)	54(8)
	LIMMA	71(7)	81(6)	17(5)	67(6)	54(6)
	SAM	74(4)	81(4)	1(8)	67(5)	54(9)
	Shrink-t	71(8)	80(7)	10(7)	67(7)	54(7)
	t-test	73(5)	76(8)	38(4)	60(9)	47(10)
	Efron-t	65(9)	72(9)	1(9)	66(8)	57(4)
	FC	56(10)	61(10)	0(10)	58(10)	55(5)

Summary & Future work

- Excellent performance on 3+3 array
- What about larger groups?

- Only looked at ranking of genes.
- Explore how to control FDR.

- Look at performance on gene-sets analysis
  (identifying regulated groups of genes, where gene-groups are externally defined)