# Testing for the Differential Expression of Genes at the Probe Level of Affymetrix Microarray data

T. Khamiakova

Interuniversity Institute for Biostatistics and Statistical bioinformatics


Universiteit Hasselt

29 September 2010

# Overview

# Advantage of Probe Level Analysis



Standard practice:

- summarization (with previous pre-processing)
- analysis of differential expression

Drawback of Summarization:

- Loss of information

Alternative:

- Differential expression on probe level

## Probes:

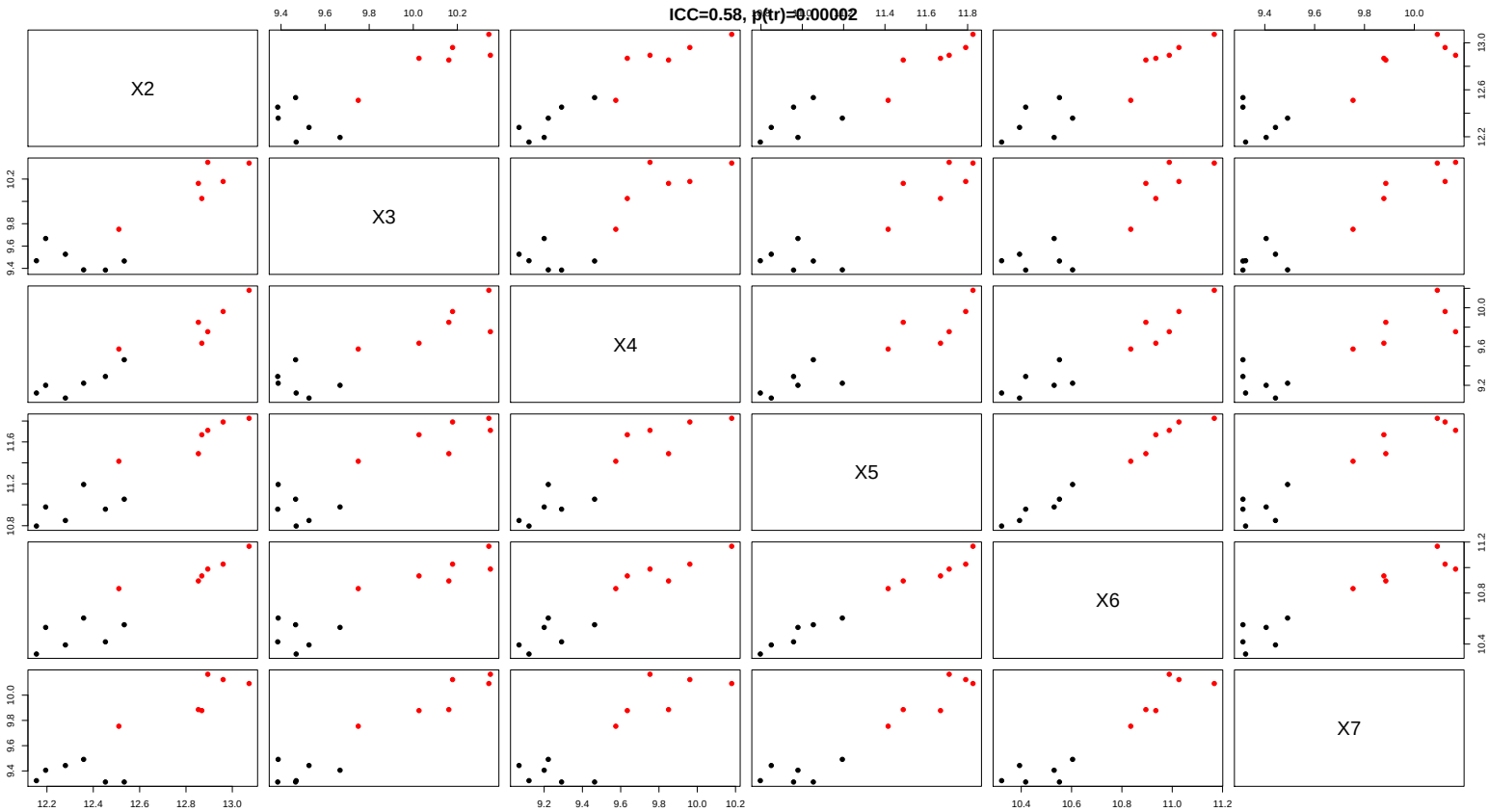
- Probeset - multivariate measurement of gene expression
- Probe = random variable measuring expression of gene, has a specific mean $\mu_j$

## Data structure

- each array can add extra variability to GE measurements
- in more complex designs fixed effects also introduce variability

# Illustration





# Adjustment for covariates

## LMM-2:

$$log(PM_{ijk}) = \mu_j + \alpha \cdot T_{ik} + b_i + \epsilon_{ijk}$$

- $\mu_j$ - probe-specific mean
- $b_i \sim N(0, \sigma_b^2)$ - array-to-array variability
- $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ - measurement error
- $\alpha$ - covariate effect
- $T_{ik}$ - covariate indicator

# Example Dataset



## Sialin DataSet (Janssen Pharmaceutica):

- Two groups of mice: wild-type (WT) and knock-out (KO)
- Expression measured on Day 18
- Big changes in gene expression are expected
- Sample size: 6 WT, 6 KO
- $\approx$ 16000 genes

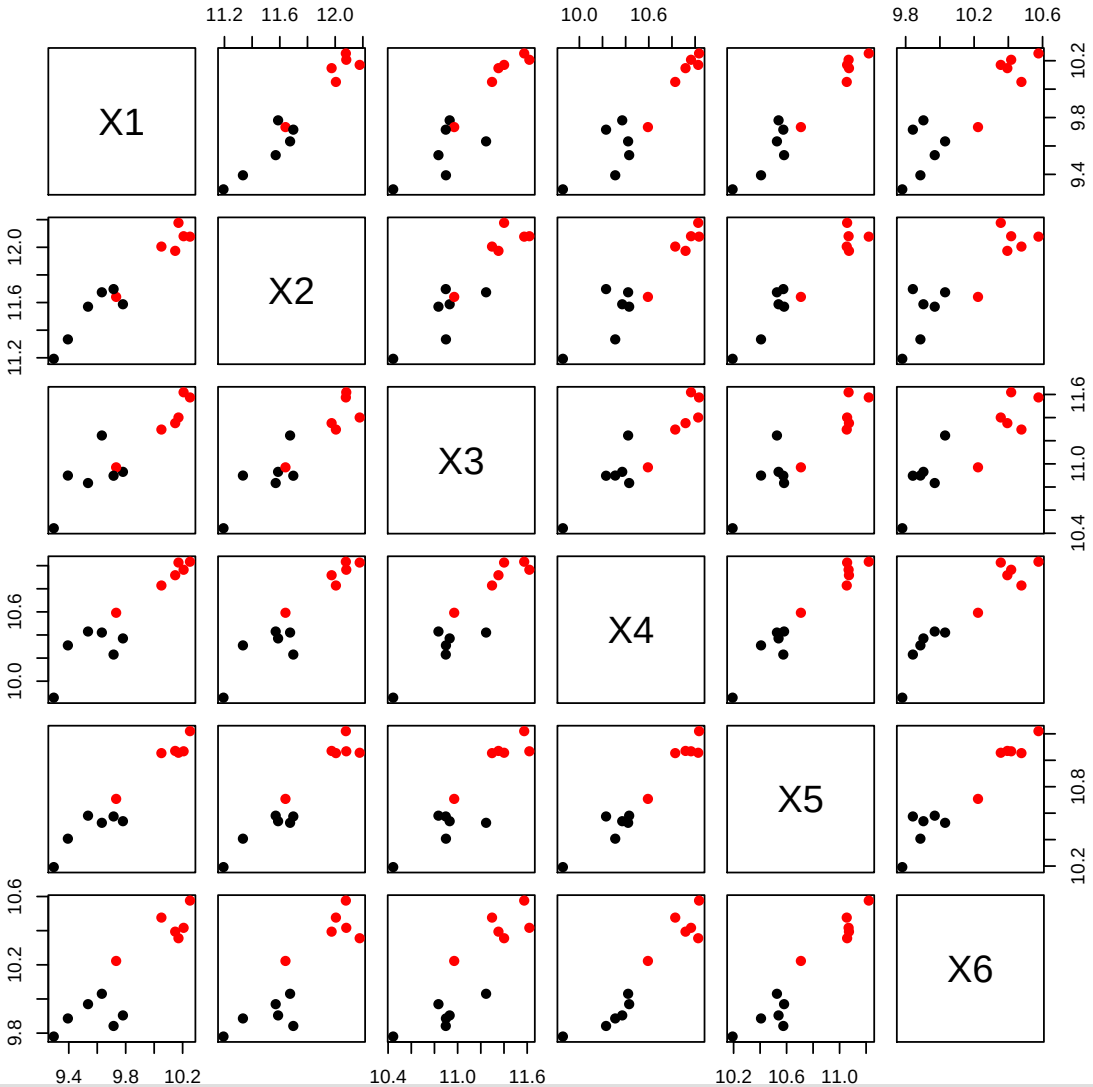
## Number of significant tests

Testing for variance component and treatment effect

| | | $\sigma_b$ | | |
|---|---|---|---|---|
| | | NS | S | Total |
| Trt | NS | 14910 | 869 | 15779 |
| | S | 572 | 44 | 616 |
| | Total | 15482 | 913 | 16395 |

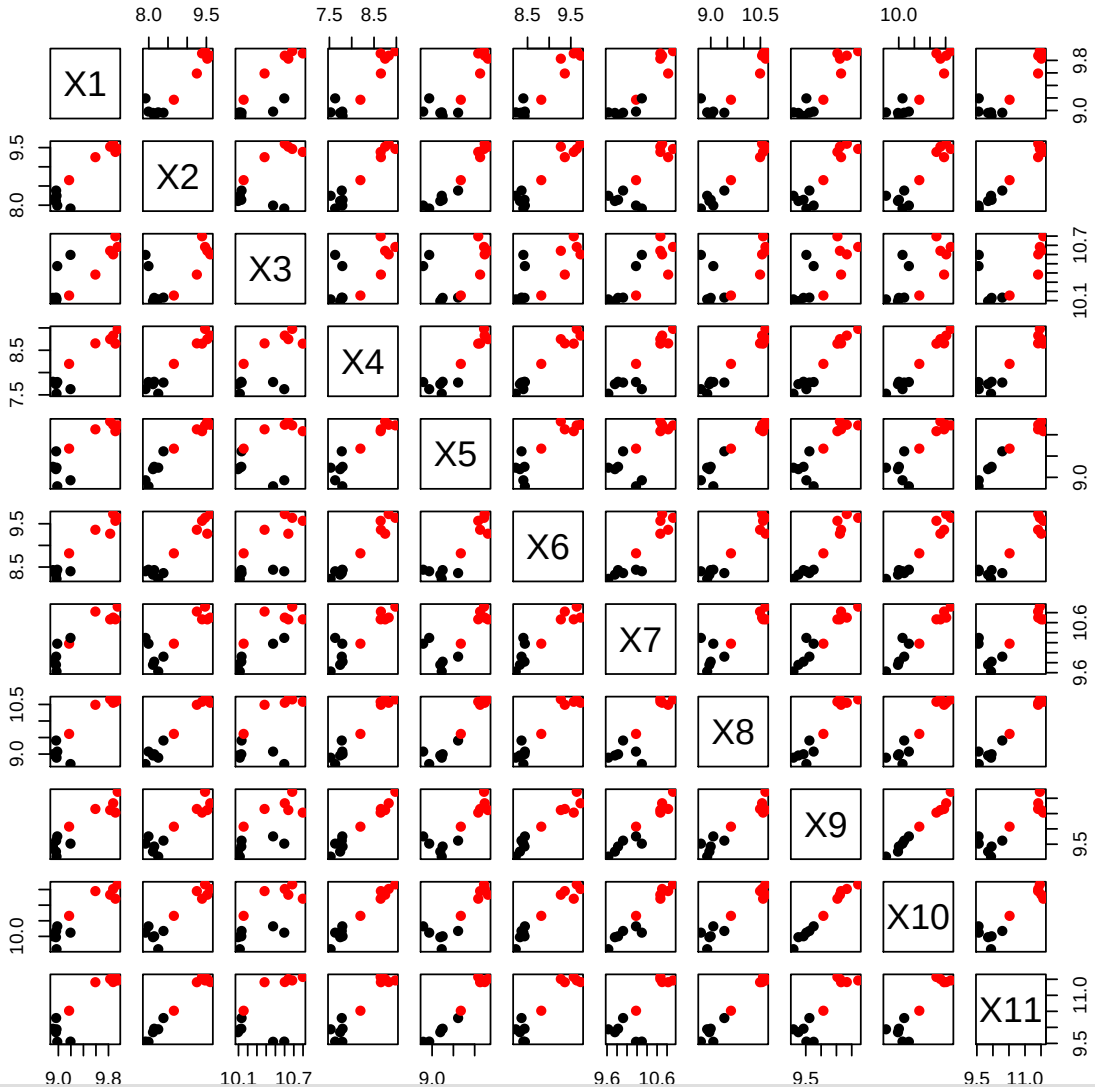
# Significant Treatment and ICC





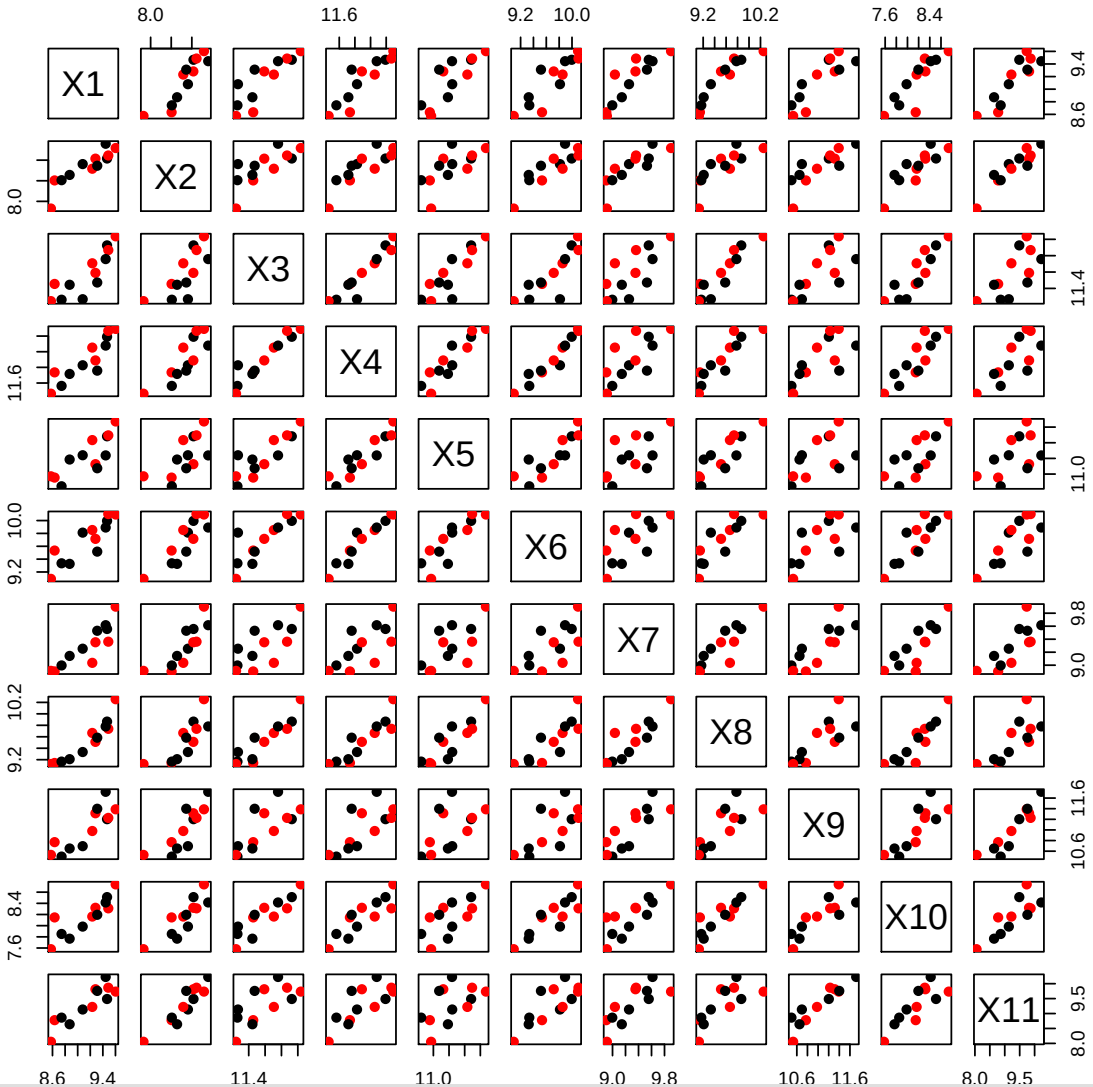
# Significant Treatment, Non-Significant ICC





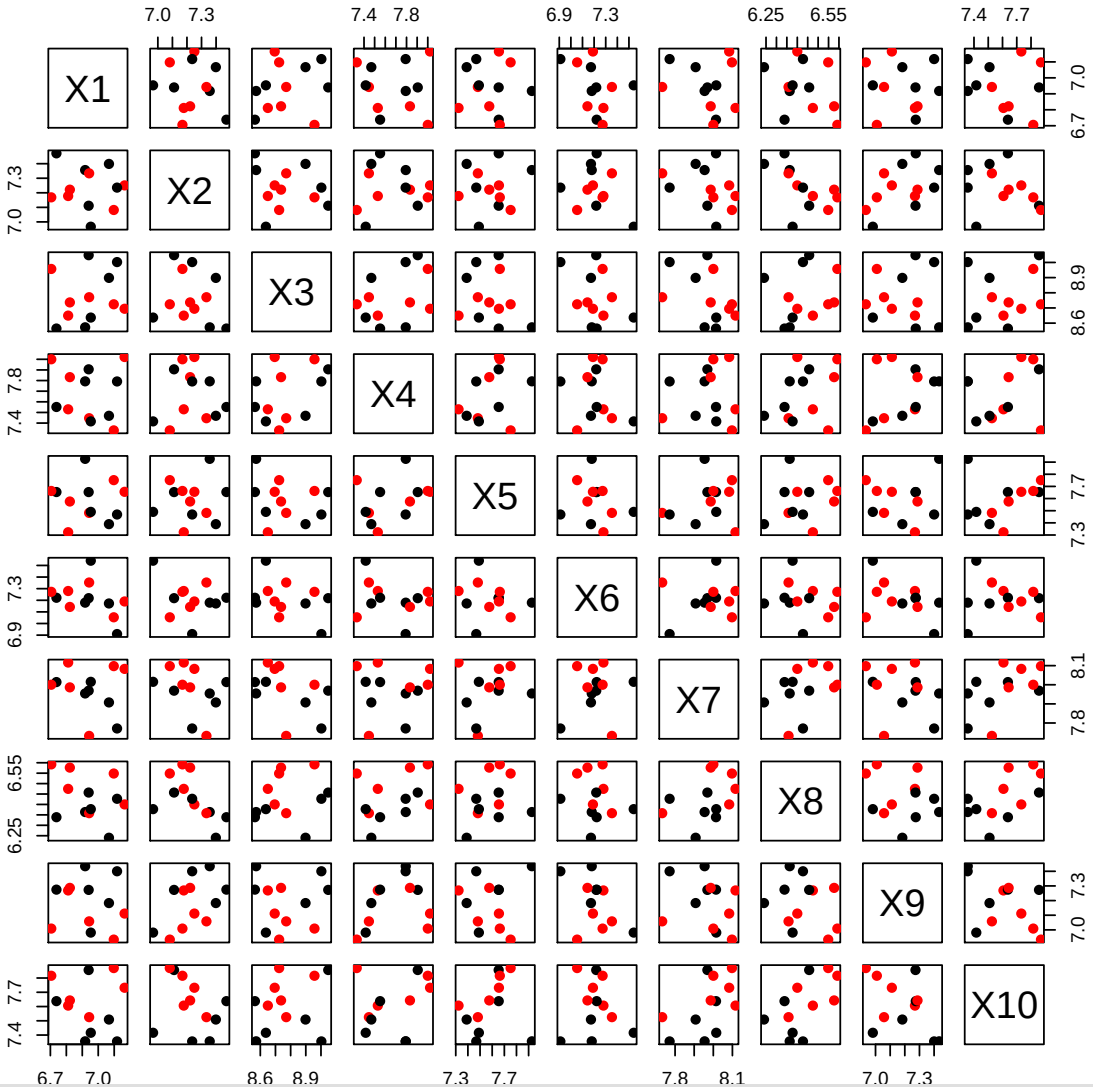
# Non-Significant Treatment, Significant ICC





# Non-Significant Treatment, Non-Significant ICC





## Simple Analysis

- on probe level data (estimated treatment effect from LMM)
- t-test on summarized data (FARMS and RMA summarized)
- FDR correction at 0.05

**Number of significant genes with treatment effect**

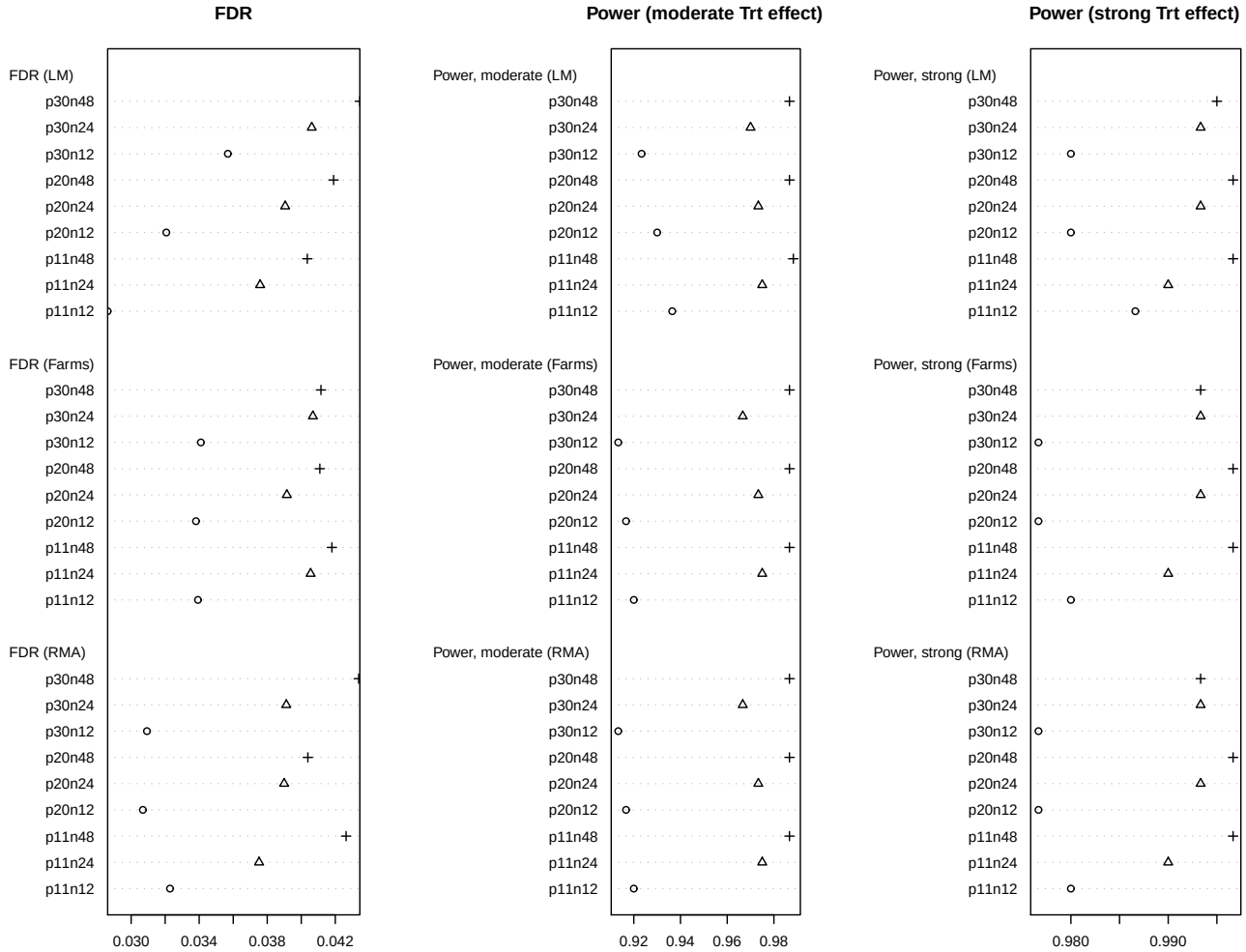
| | Probe Level | FARMS | RMA |
|---|---|---|---|
| LM, t-test | 616 | 553 | 371 |

## Results

- more significant genes from probe-level analysis
  - due to higher power (?)
  - due to higher error rate (?)
- Simulation Study
  - investigate power and FDR for probe level analysis
  - varied sample size and probeset size
  - varied treatment effect and intra-class correlation

# Results




FDR
Power (moderate Trt effect)
Power (strong Trt effect)
FDR (LM)
FDR (Farms)
FDR (RMA)
Power, moderate (LM)
Power, moderate (Farms)
Power, moderate (RMA)
Power, strong (LM)
Power, strong (Farms)
Power, strong (RMA)
p30n48
p30n24
p30n12
p20n48
p20n24
p20n12
p11n48
p11n24
p11n12
0.030
0.034
0.038
0.042
0.92
0.94
0.96
0.98
0.980
0.990

## Analysis of DE

- Probe-level analysis has higher power
- FDR is controlled for all methods

## Follow-Up

1. Power of other statistical tests on probe level (e.g. SAM)
2. Power of the probe level analysis compared to the filtered results
3. Extensions of the LMM

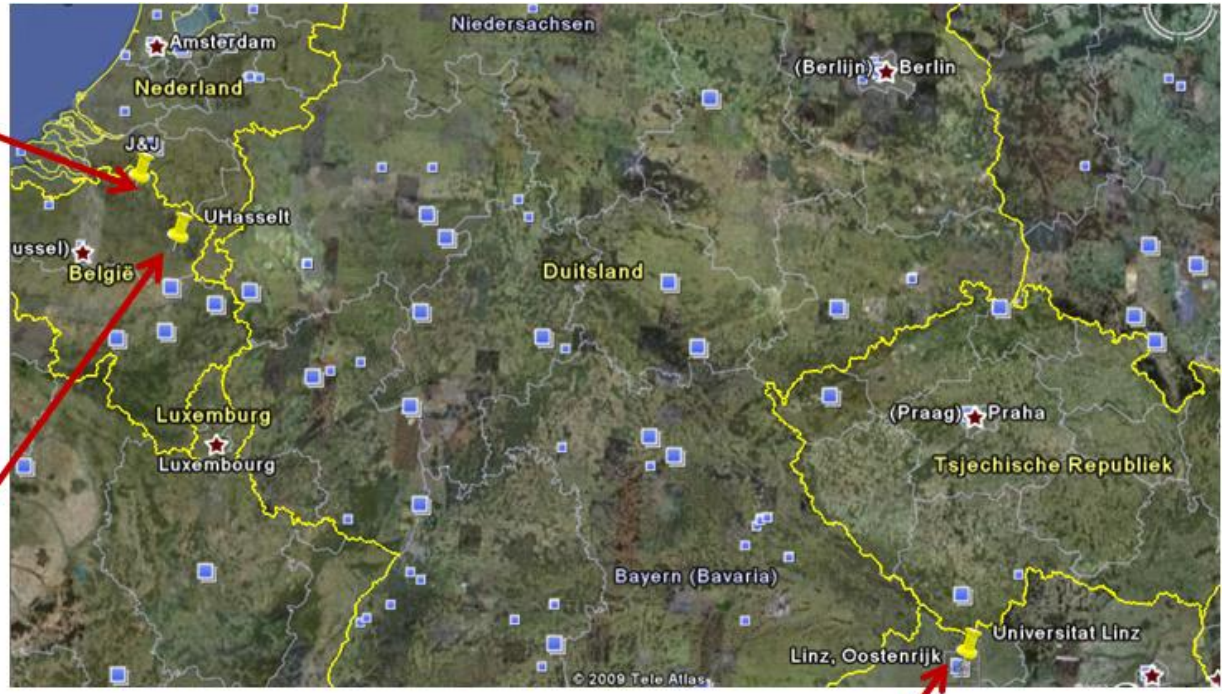
# Acknowledgments



## IWT project

**J&J PRD**
Hinrich Goehlmann
Willem Talloen
An De Bondt
Luc Bijnens

**I-Biostat:**
Adetayo Kasim
Ziv Shkedy
Dan Lin
Suzy Van Sanden




**Institute of Bioinformatics (Johannes Kepler University):**
Sepp Hochreiter
Djork-Arne Clevert