Strategies for quantifying GeneChip expression for large studies

Darlene Goldstein
Institut de mathématiques
École Polytechnique Fédérale
Lausanne, SWITZERLAND

MBI Workshop: Analysis of Gene Expression Data
12 October 2004

Outline

- Expression measures for Affymetrix GeneChips
- Advantages and disadvantages for large studies
- Subset strategies
- Resampling strategy
- Comparison of strategies
- Remaining issues/conclusions

What is the problem?

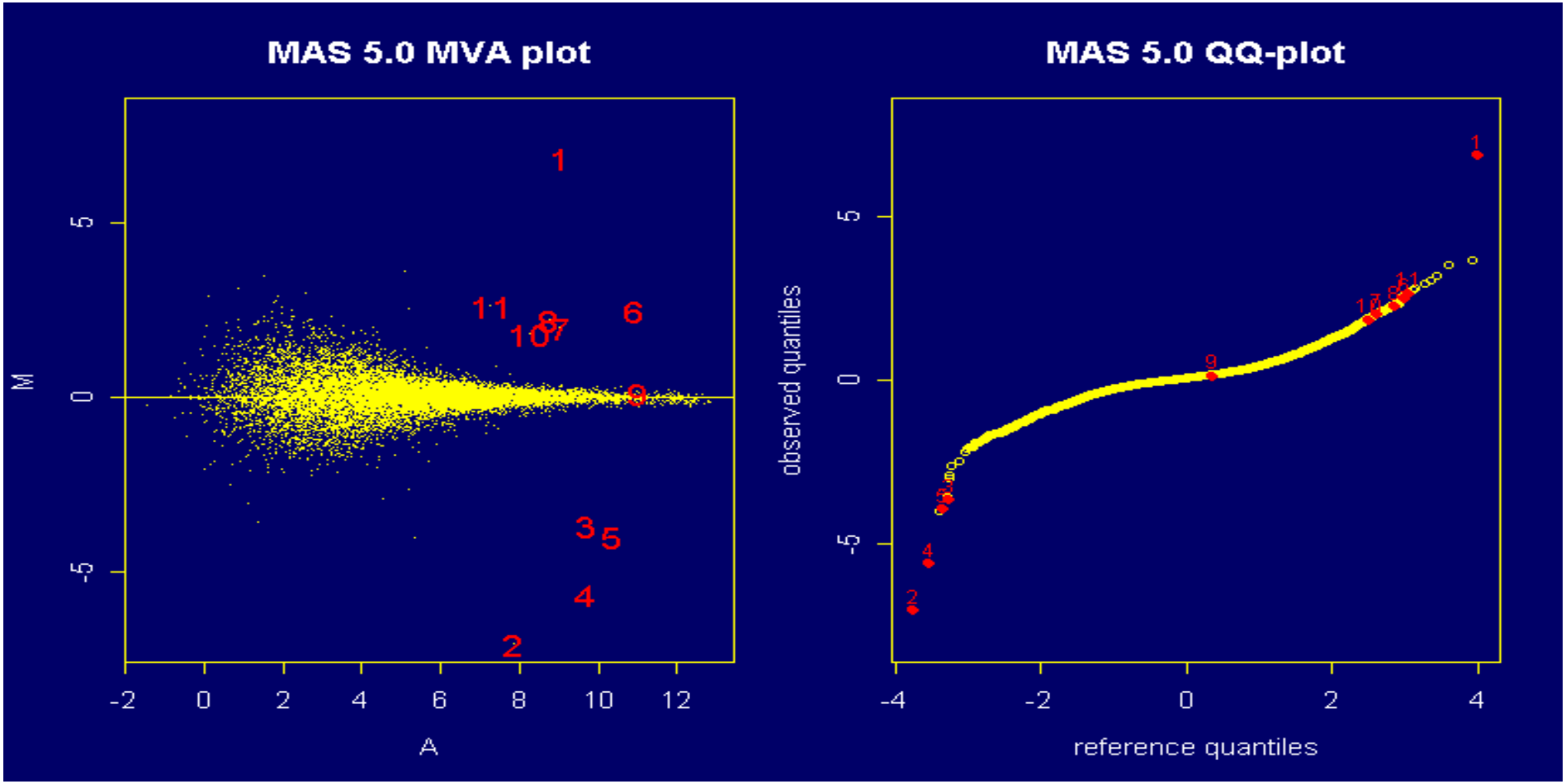
- *The good news:* biological and medical investigators are taking our advice ☺, so that microarray studies are now *becoming larger*

- *The less good news:* limitations in computing capabilities can make quantifying expression *more difficult* ☹

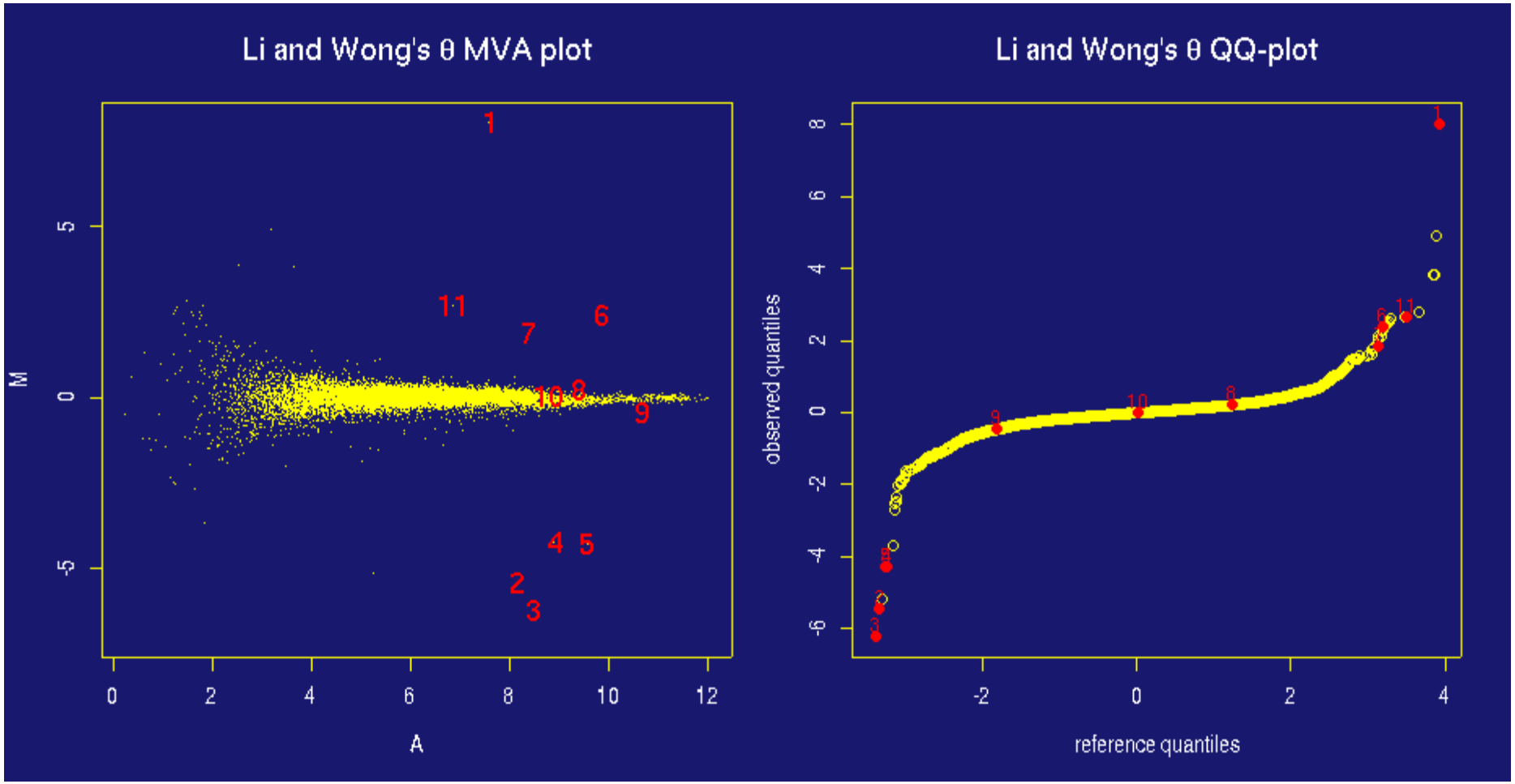
Expression measures

- *MAS 5.0* – Affymetrix
- *Model Based Expression Index* (MBEI) –
  - Li-Wong method; windows executable **dChip**
- *Robust Multichip Analysis* (RMA) –
  - Irizarry *et al.*, Bolstad *et al.*; R pkg **affy**
- Other methods include:
  - plier, plier+16 (Hubbell, new Affymetrix)
  - *vsn* (Huber *et al.*, Rocke)
  - *gcrma* (Wu *et al.*)
- Visit http://affycomp.biostat.jhsph.edu/

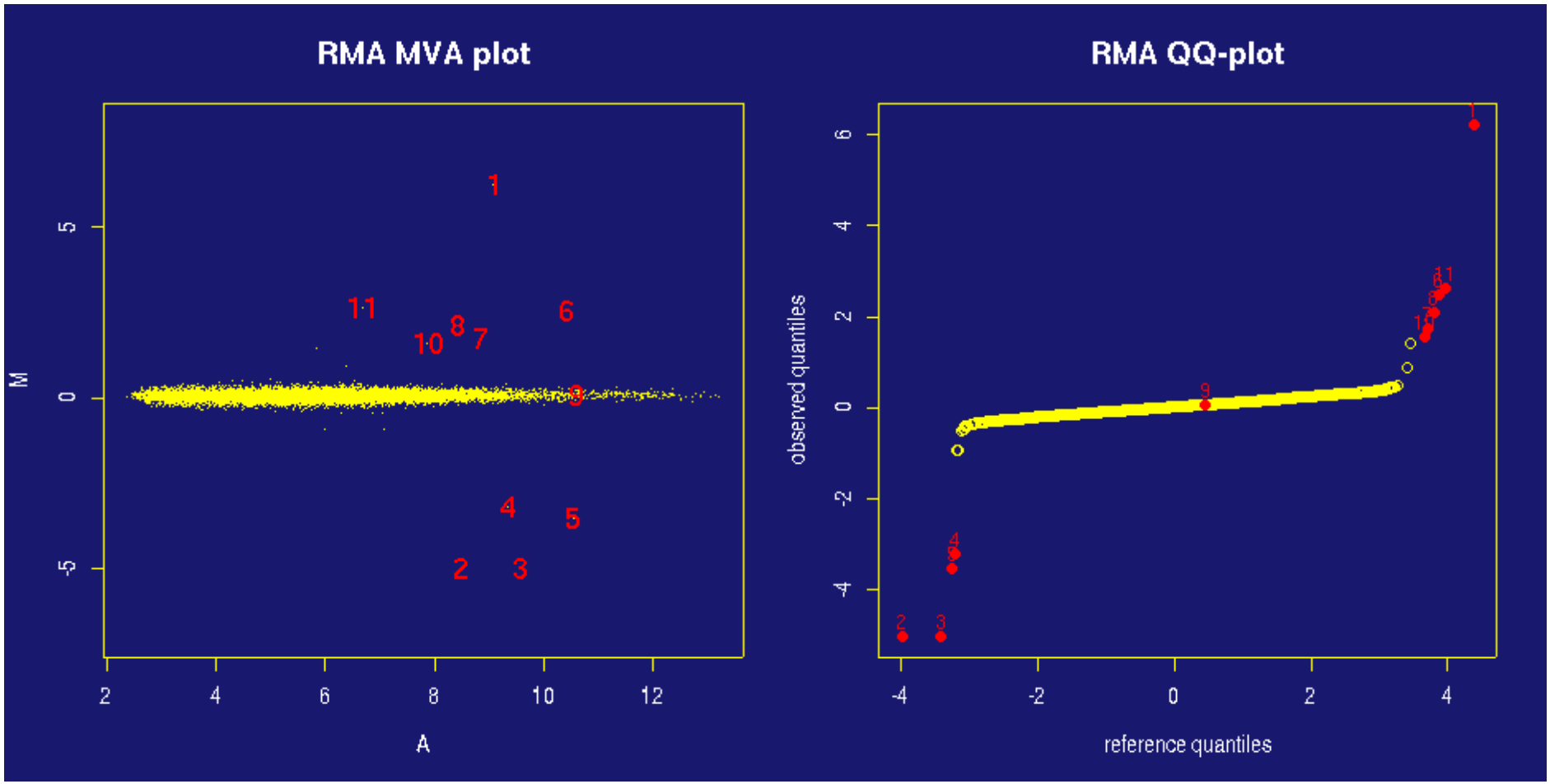
Differential expression: MAS 5.0



Differential expression: Li-Wong



Differential expression: RMA



Advantages and drawbacks (I)

- *MAS 5.0* – Affymetrix
  - quick when scale each separately to target
  - problem – variance of lower expression, get many false positives (new algorithm plier+16 might improve this; using vsn on top of MAS 5.0 also improves)
- *Model Based Expression Index* (MBEI)
  - model improves on MAS 5.0
  - can fit with many chips (up to ~ 400)
  - still room for improvement of expression quantification

Advantages and drawbacks (II)

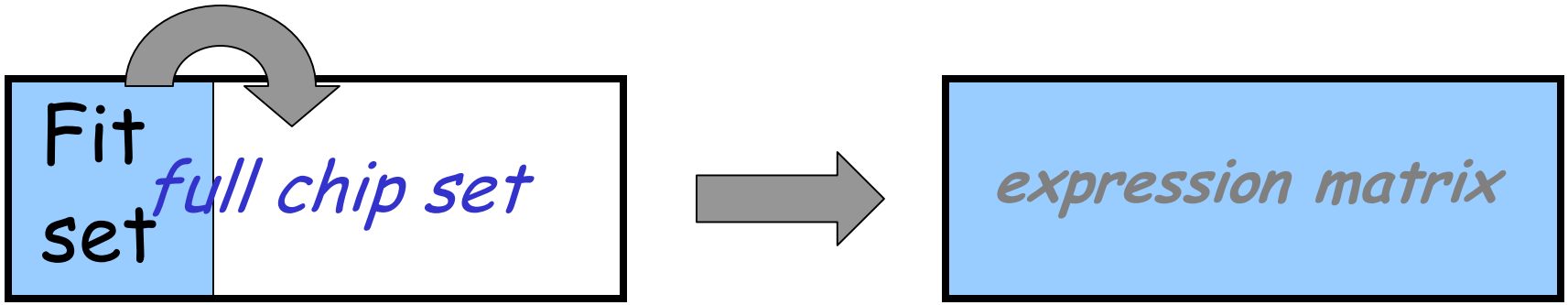
- *Robust Multichip Analysis* (RMA)
  - background correction, quantile norm, chip + probe model (median polish fitting)
  - performs well on calibration data sets
  - computational improvements (*e.g.* justRMA)
  - can still have computational problems with very large studies

  => Subset strategies

Subset strategy: Extrapolation

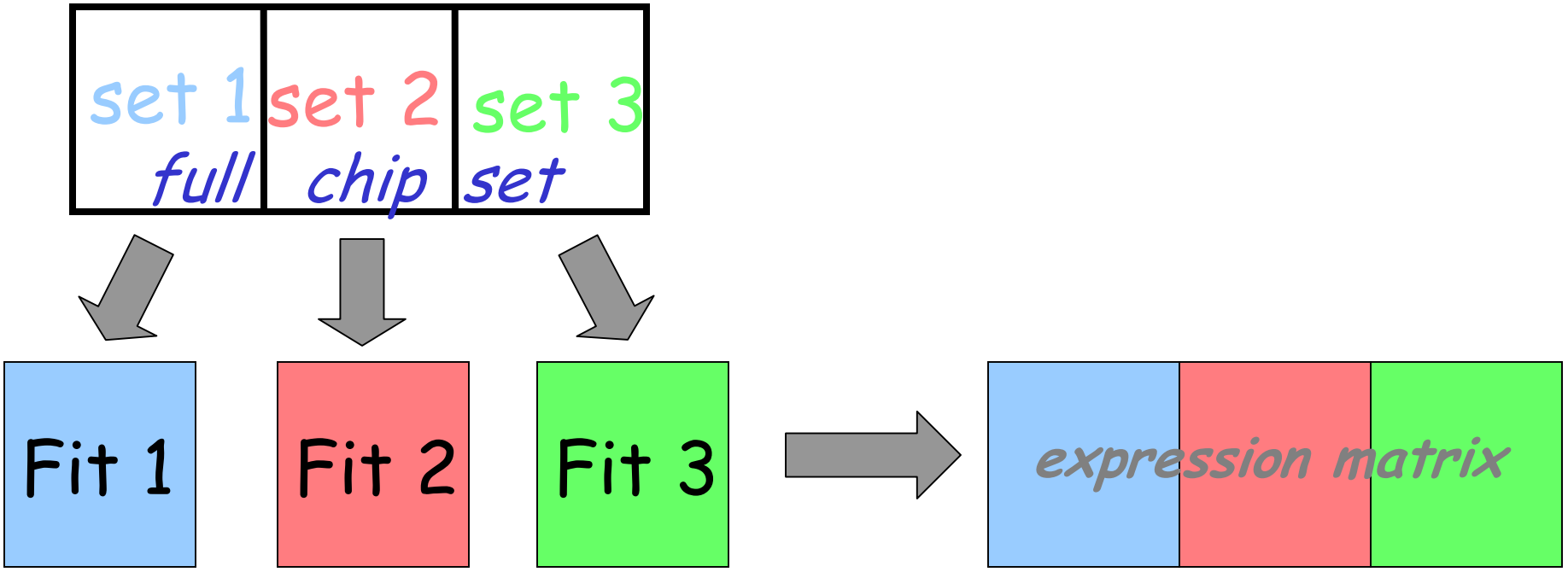
- Fit model on *only a subset of chips*
- *Apply model* to remaining chips

  => get gene expression measure for each gene



Subset strategy: Partition

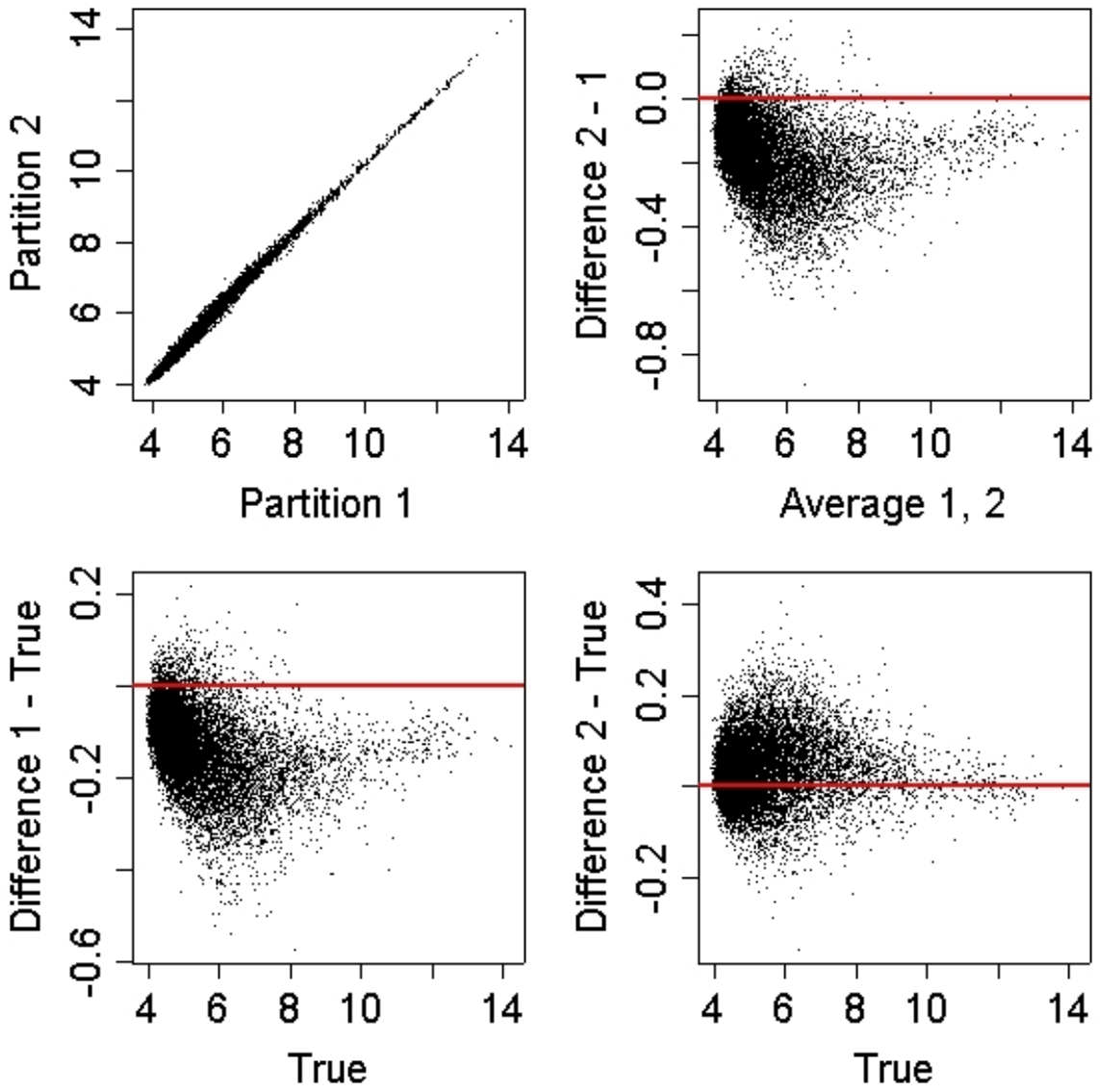
- *Partition* chips into subsets
- Fit *separate models* within each subset
- Combine to get full set



Problems

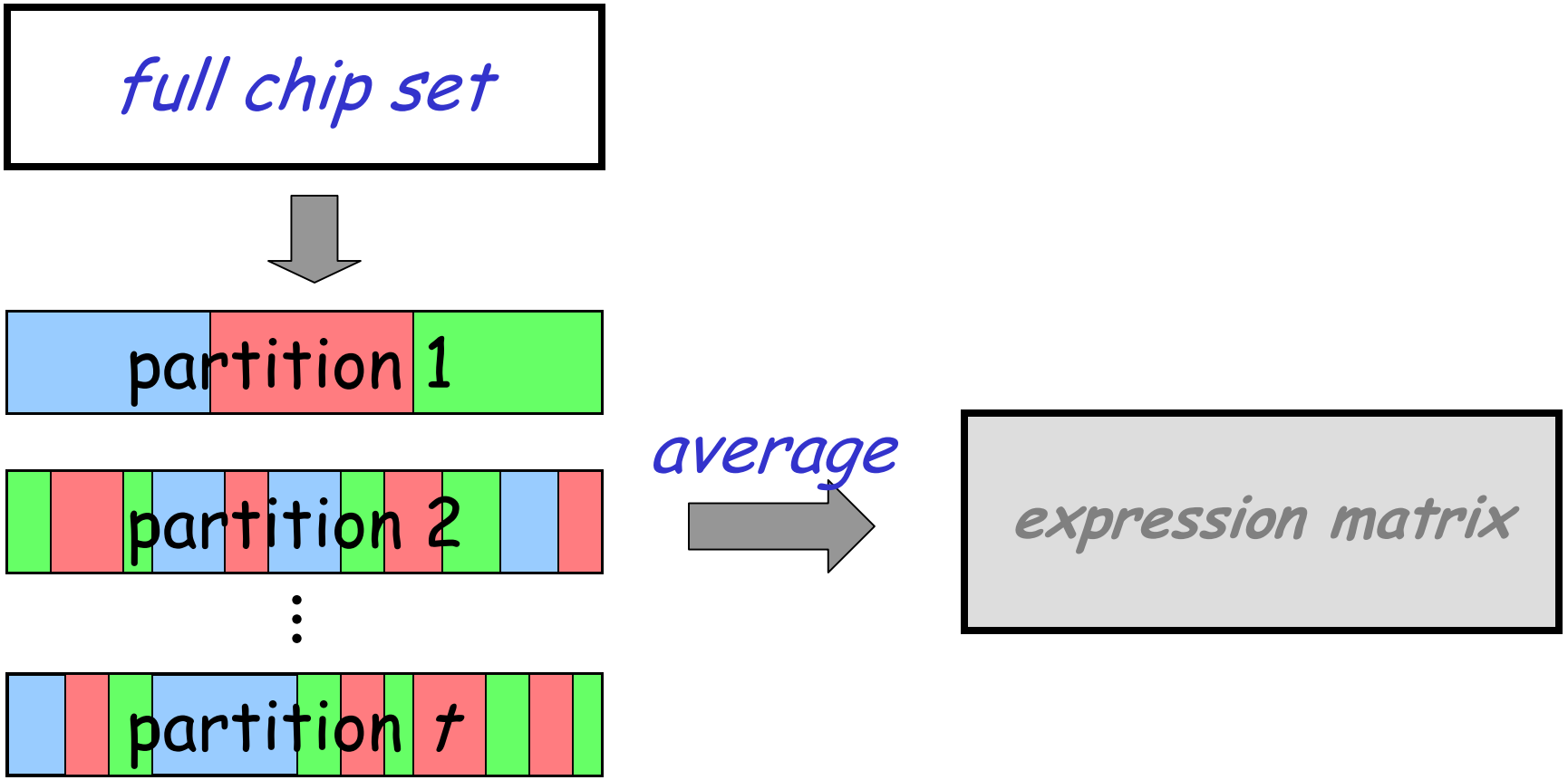
- Extrapolation
  - fitting set characteristics '*locked in*'
  - what if fitting set is 'bad' in some way?
- Partitioning has this problem as well, although to a lesser degree
- Both strategies exhibit some *variability*; perhaps more than we would like to see ...

Variability of partitioning: expression



Resampling strategy

- Apply subset strategy many times on *different subsets* (generated randomly)



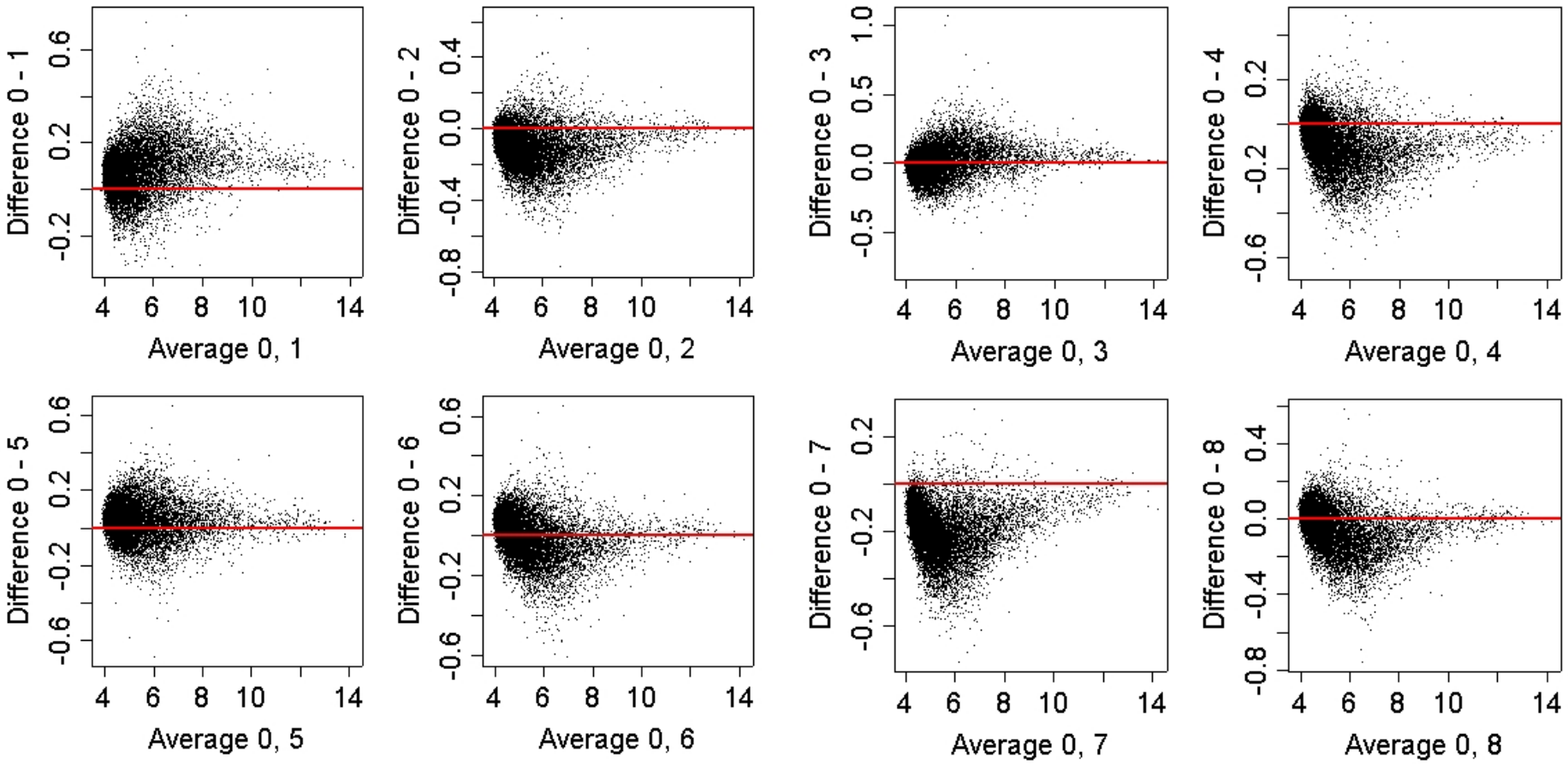
Strategy comparison study

- *Main ingredient:*

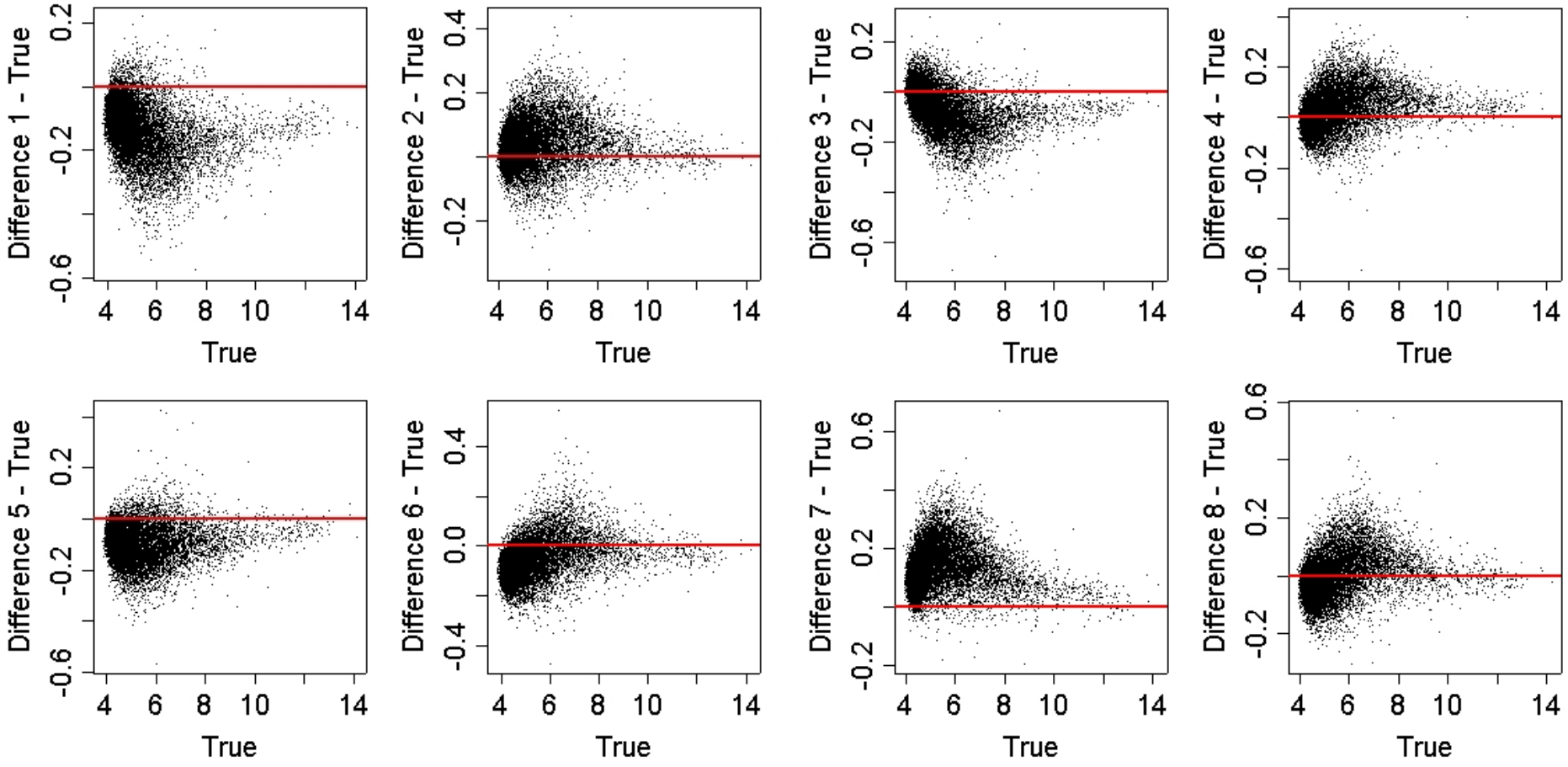
  compare *expression measures* and *test statistics* from a large (full) data set to those from subset, resampling strategies

- Many times for subsets of given sizes
- Data sets:
  - ALL (St. Jude Children's Hospital); 335 chips, publicly available
  - HD (international collaboration); about 70 individuals, 3 tissues per individual

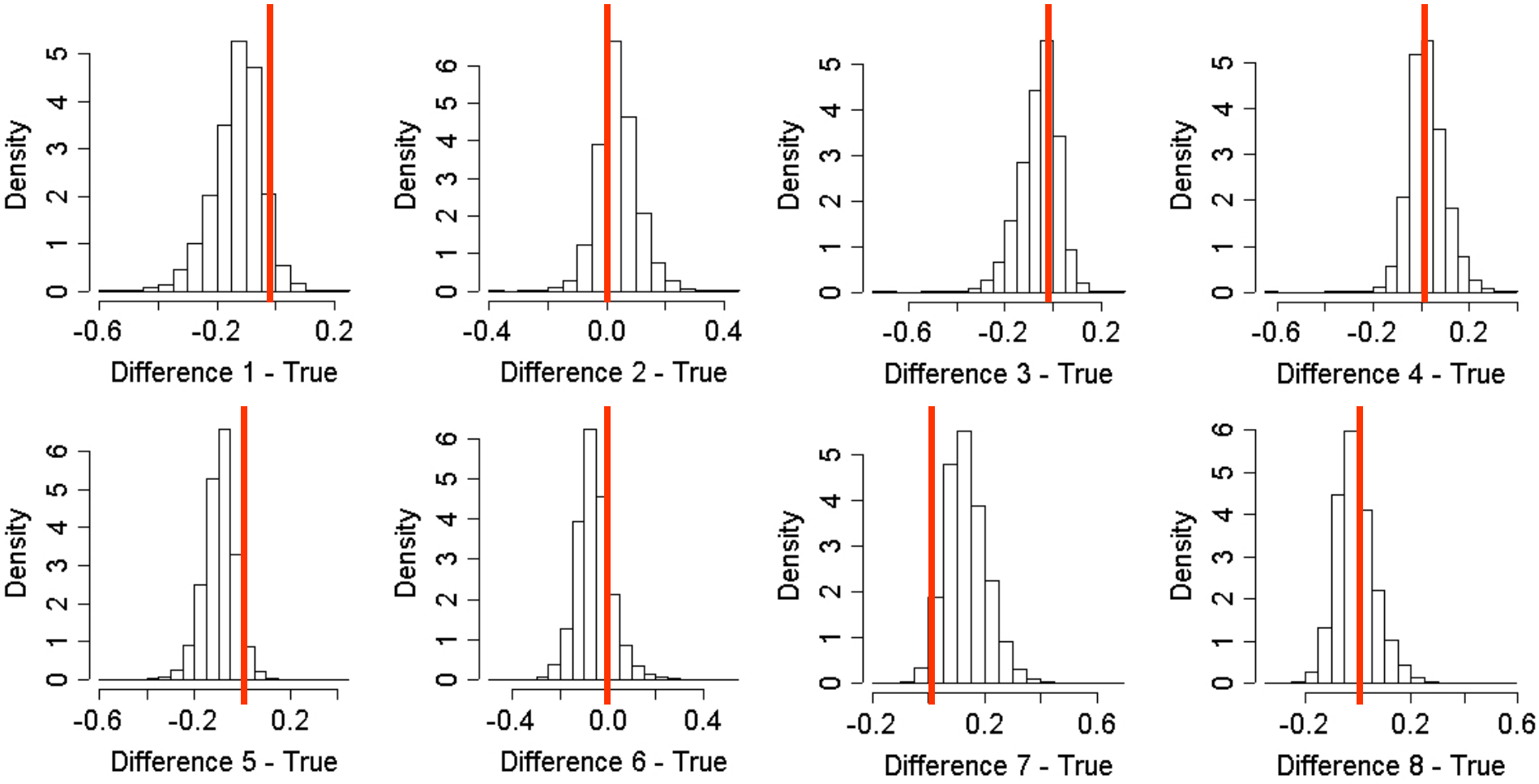
Partition replicates of one ALL chip



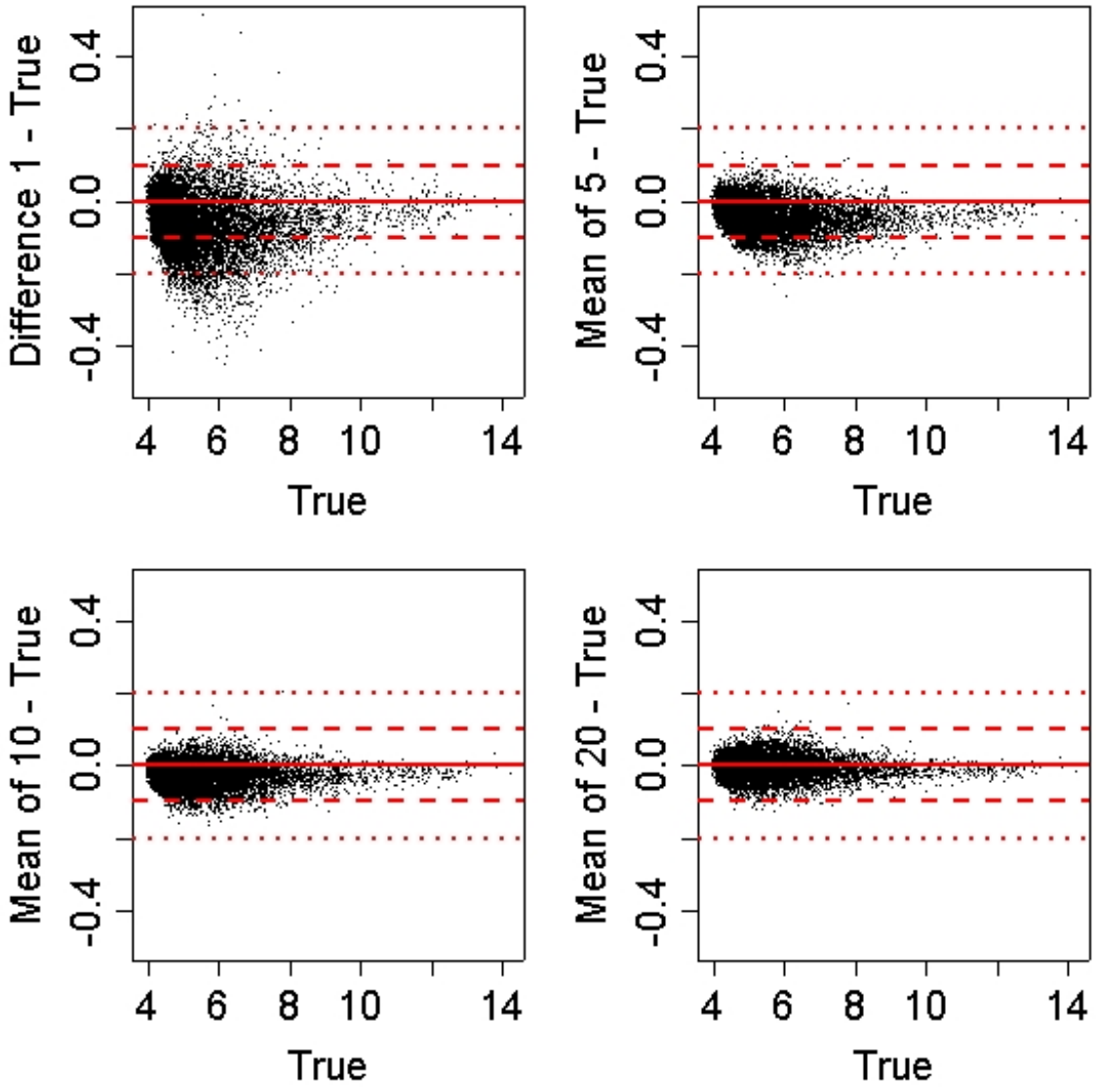
Partition replicates (1 chip) vs. true



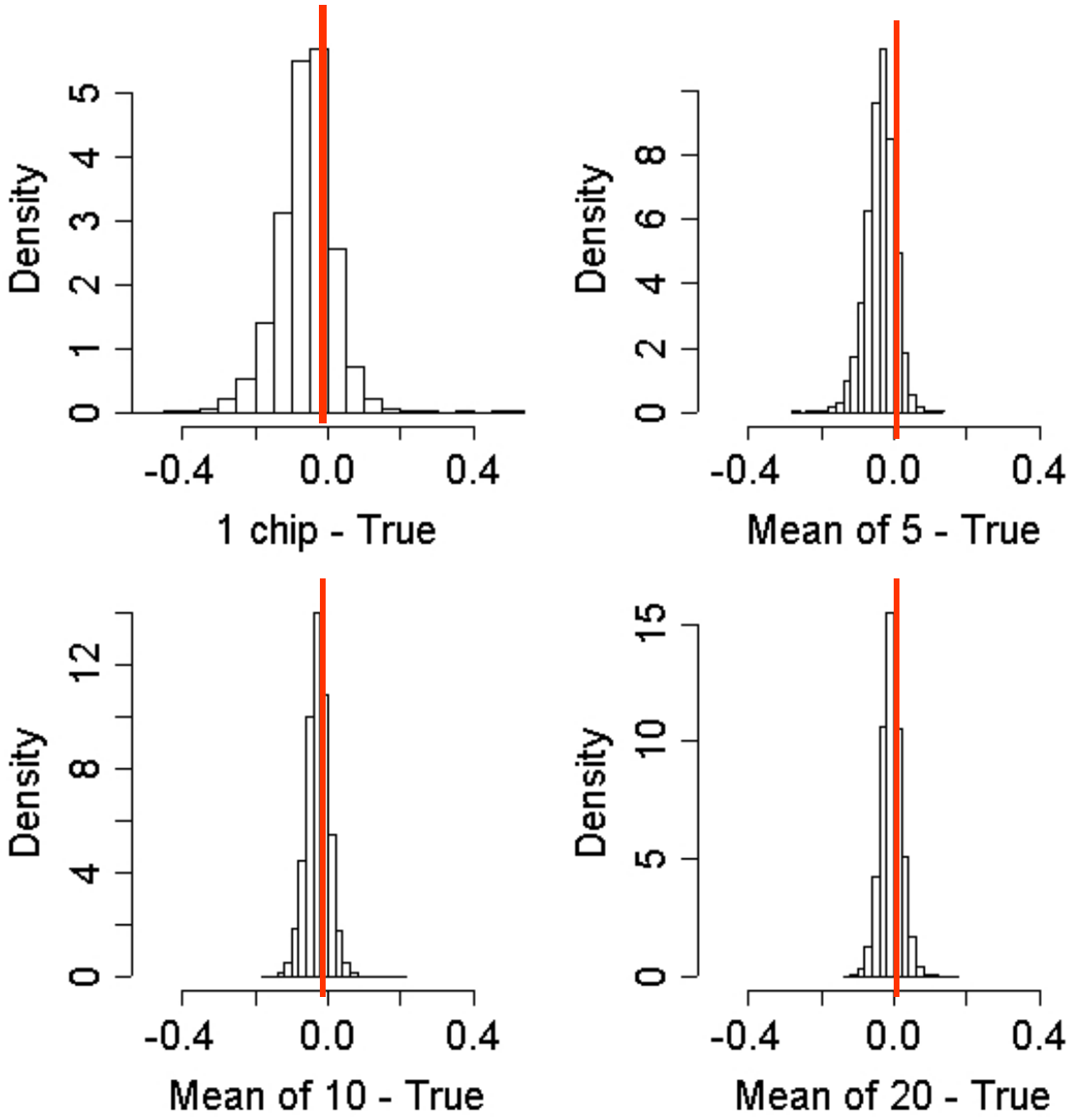
Bias of single chips



Differences mean vs. true



Bias of the mean

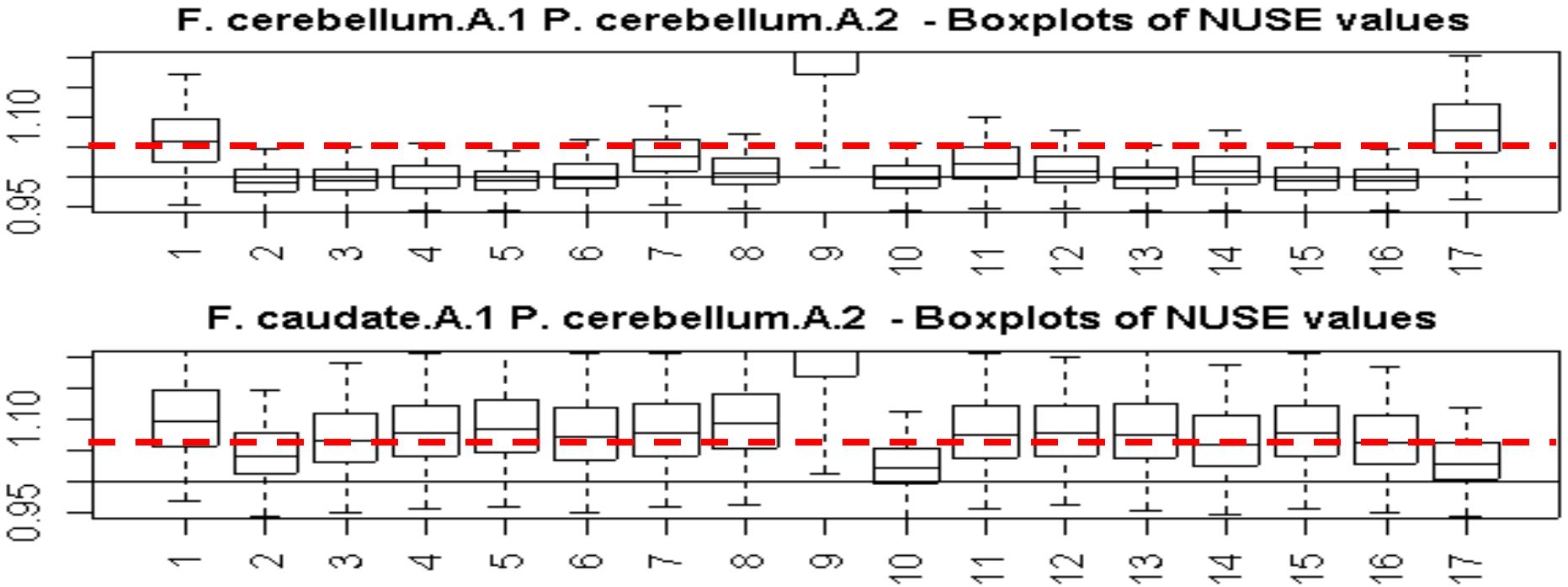


Effect on decisions

- Choice of the fitting set *can be problematic*
  - time trend
  - multi-center studies
- How are subsequent *decisions* (*e.g.* on DE, choice of genes for followup) affected
  - can't compare true/false positives, because *don't have a 'known' result* (for HD we might obtain qpcr data on some 'interesting' genes)

Decisions can depend on fitting set

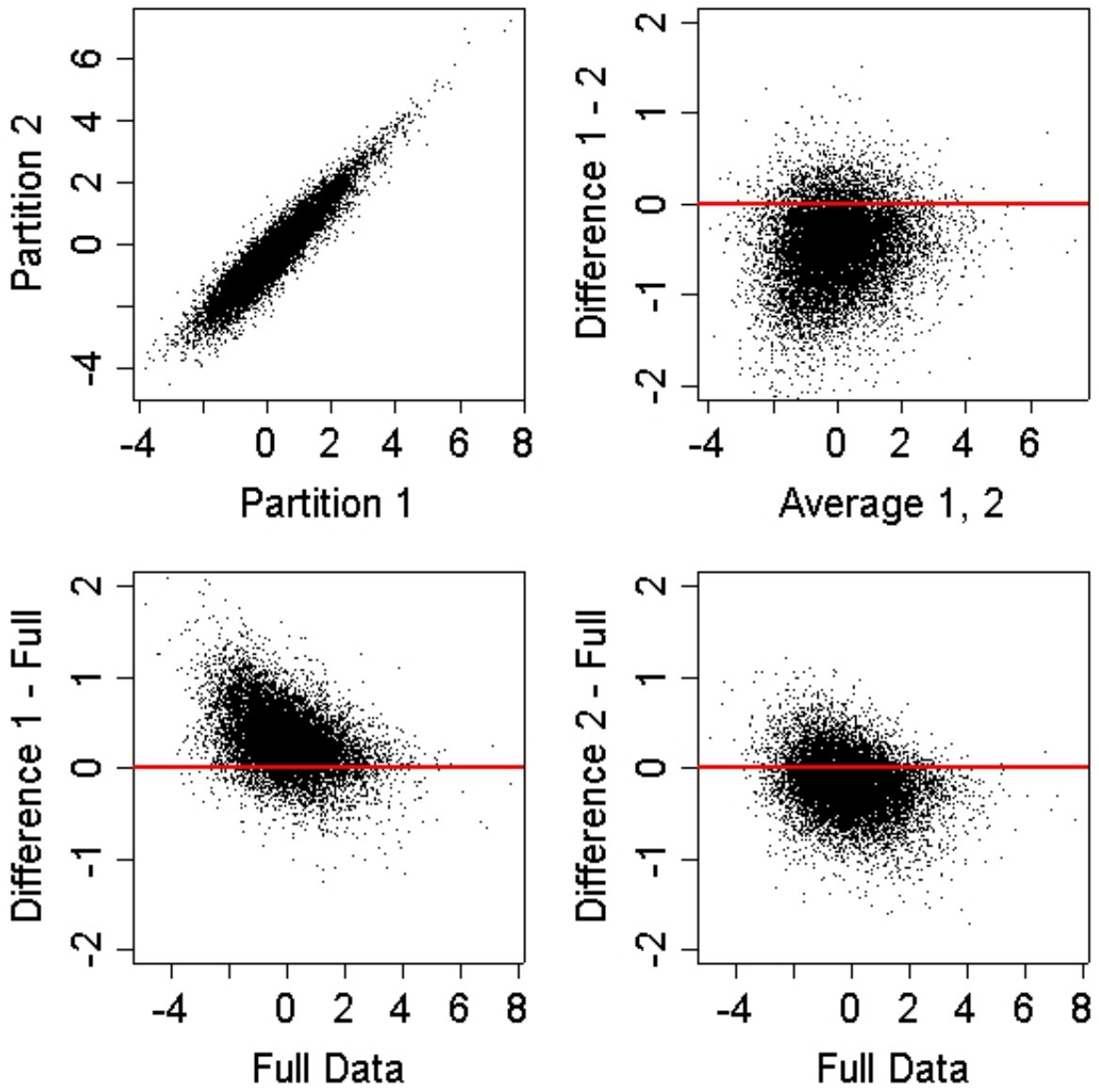
- Used *two different fitting sets* to estimate expression on *same 'left out' set* (extrapolation)
- Used the resulting expression values in calculations of *other quantities* (*e.g.* NUSE)



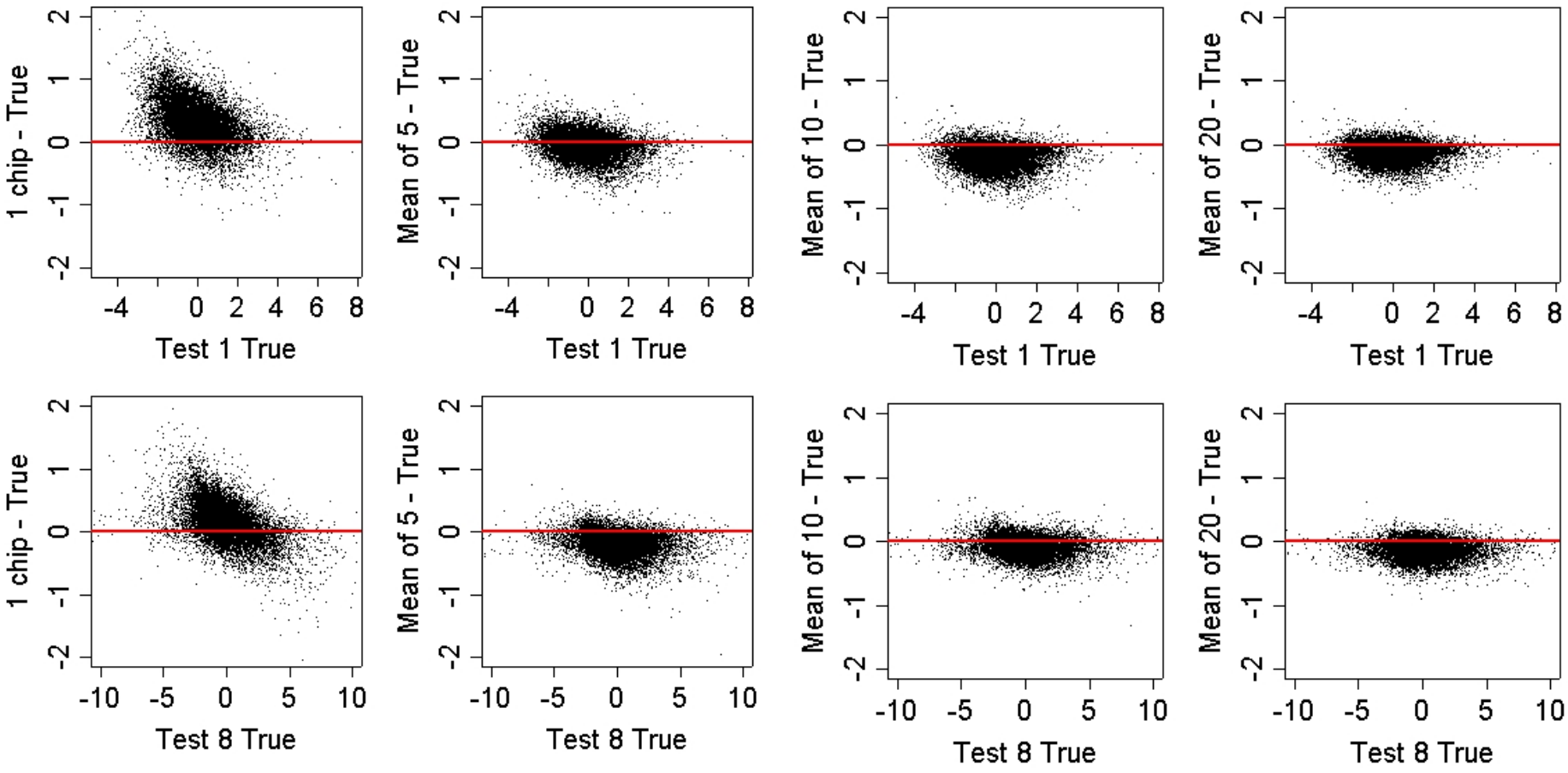
Example: *t*-tests between ALL types

- 12 *t*-tests:
  - type vs. normal (9 types)
  - 3 other tests with different sample sizes (large vs. large, small vs. small, large vs. small)
- Did not use shrinkage (moderated t), since sample sizes are not too small
- Compute on: full data, each partition, and using mean *expression* across partitions

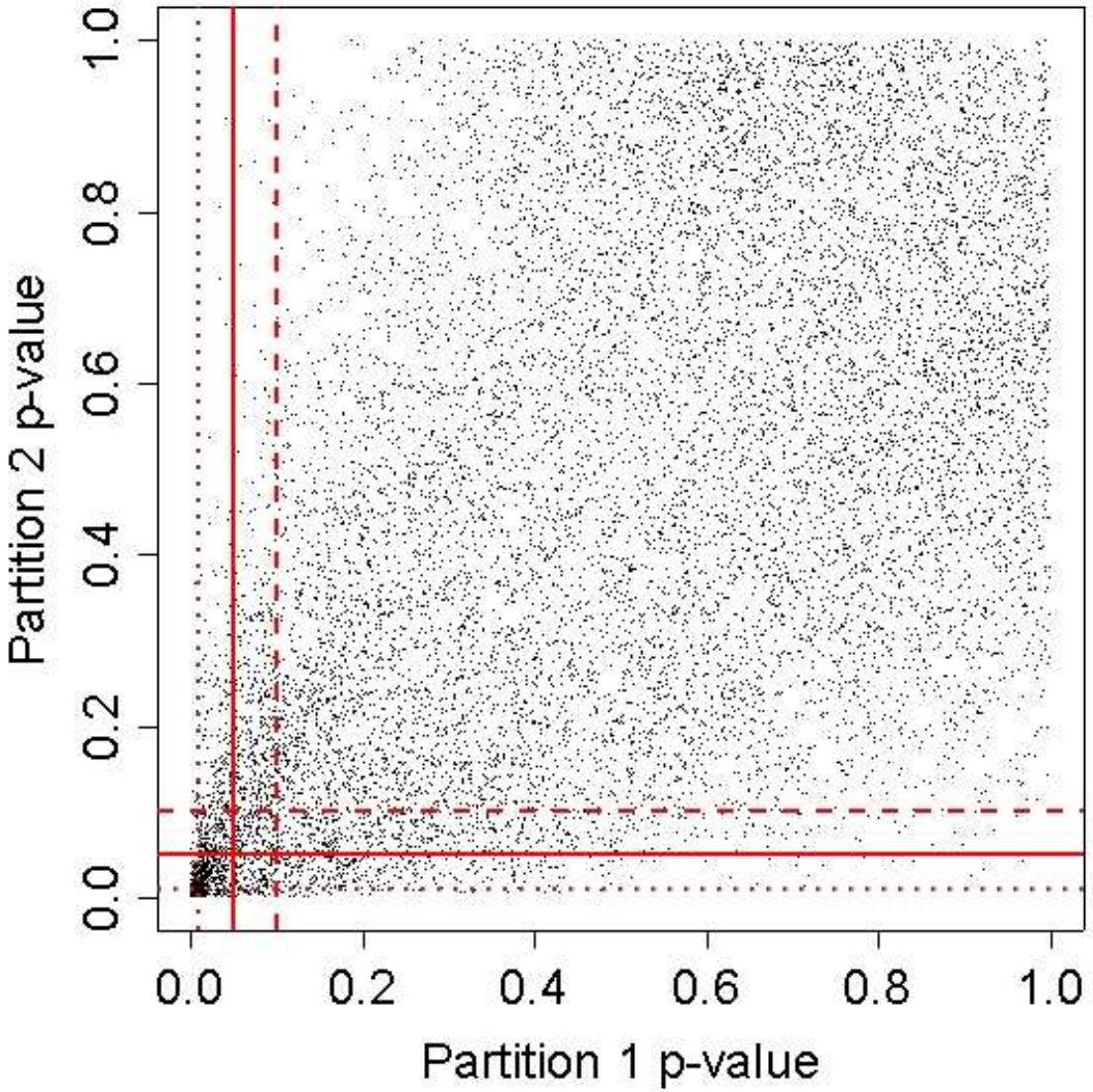
Single subset *t* compared to true



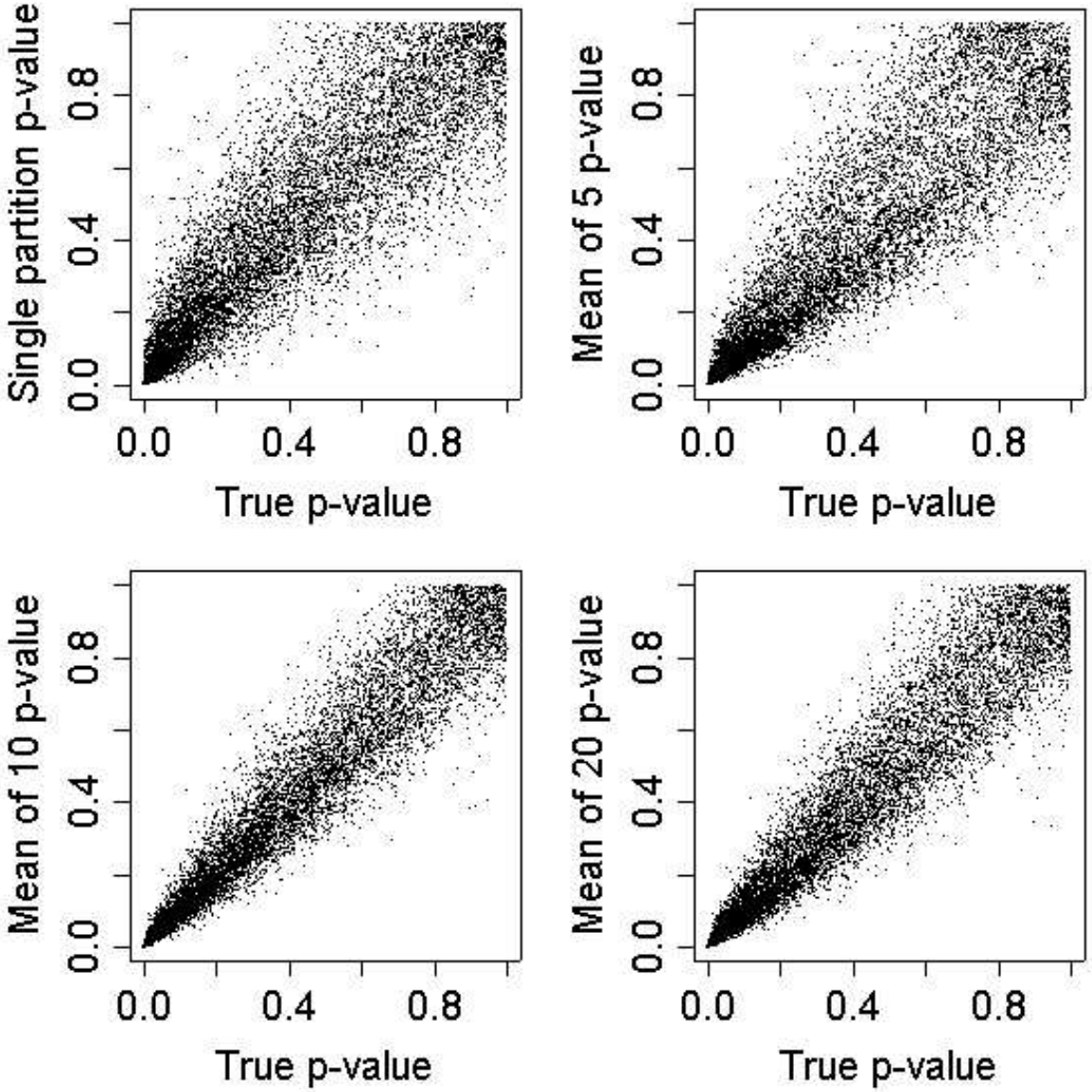
Mean expression *t* compared to true



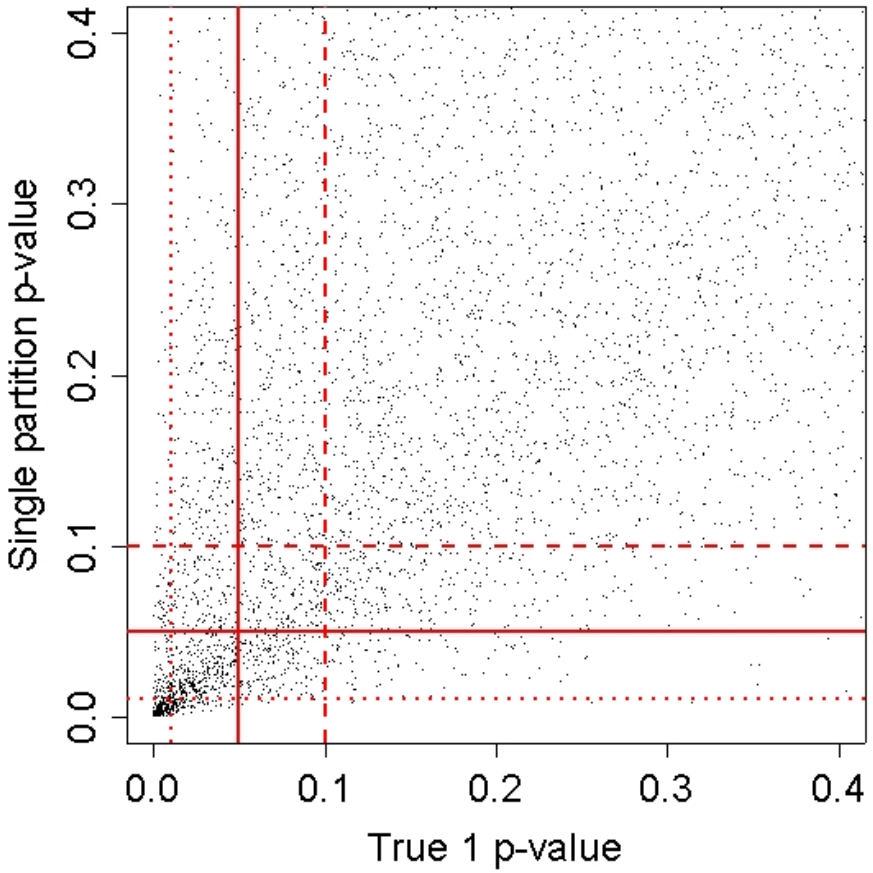
Variability of partitioning: p-values



Mean expression *p* compared to true



p-value agreement



#	*5%* (1172)	*1%* (381)
1	734/1012	262/358
5	991/1093	321/345
10	1020/1208	330/384
20	1064/1210	339/382

Possible computational improvements

- Ideally, the less the need for subset strategies, the better
- Improvements in computational feasibility would lessen need
- Wish list:
  - resolve memory management issues
  - potential for parallelization of some steps
- Vital-IT

Vital-IT

- Joint venture between academic and industrial partners (SIB-managed)
  - Universities of Lausanne, Geneva, Basel, EPFL, Ludwig Institute for Cancer Research
  - Hewlett-Packard, Intel
- High-performance computing center for life sciences
  - HP cluster of 32 servers, Itanium 2
  - Software development, optimization
  - Consulting for biology, medicine

Conclusion

- RMA for large studies not always possible – studies already in progress large enough to prohibit 'exact' RMA calculations
- Partition-resampling strategy seems 'safer' than using a single extrapolation or partition
- Here, differences are characterized and compared to full data RMA as 'truth'
- Ideally, it would be nice to be able to compare several strategies on large 'calibration' data sets ('known truth')

Acknowledgments

- *Ruth Luthi-Carter, EPFL*
- *Francois Collin, UCSF*
- *Ben Bolstad, SFSU*

- *Victor Jongeneel, LICR, SIB*
- *Li Long, SIB*

- *Terry Speed, UC Berkeley*
- *Rafael Irizarry, Johns Hopkins*

questions