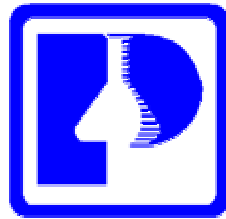


# Topological adjustments to the Genechip expression values.

Andrey Ptitsyn



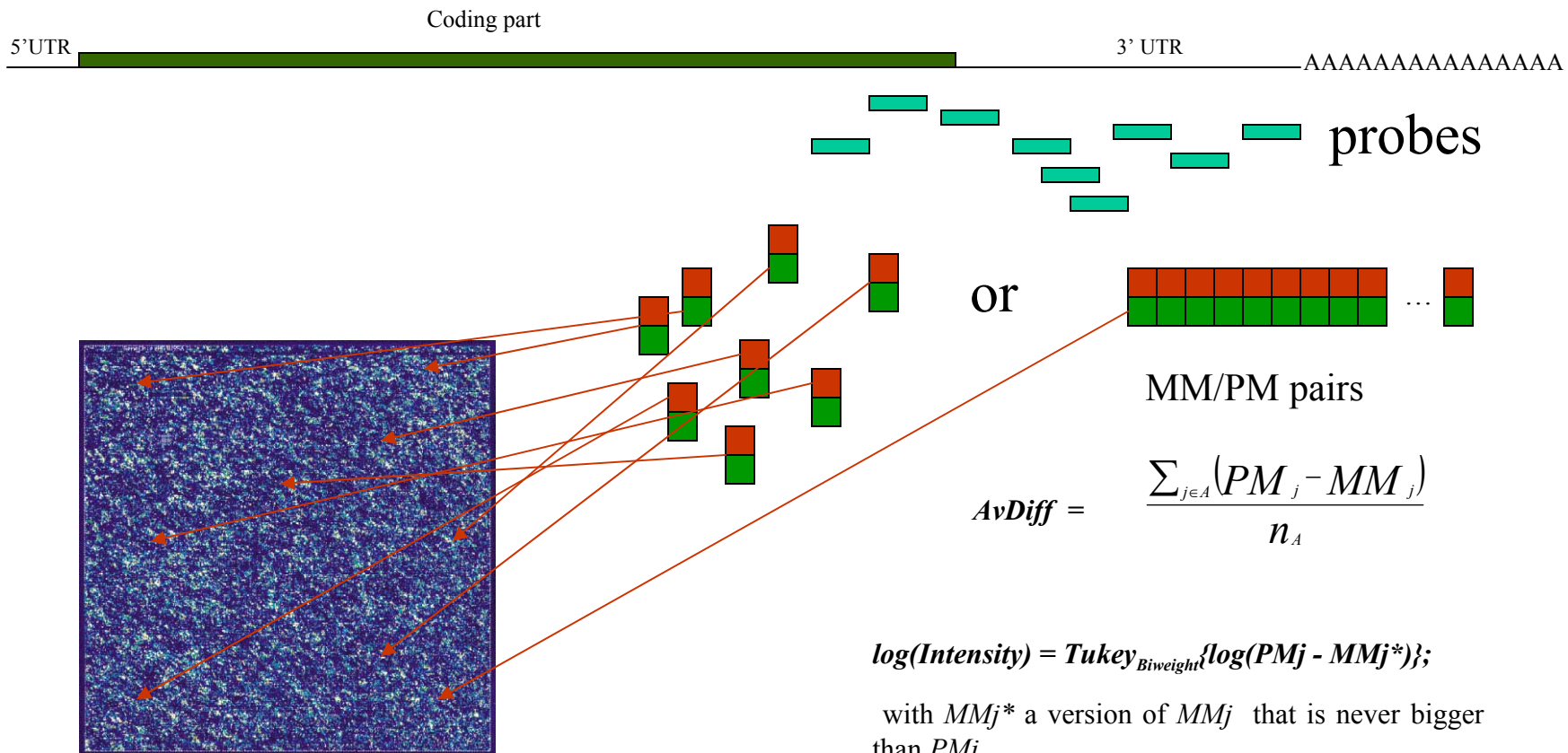
Pennington Biomedical Research Center  
Louisiana State University  
Baton Rouge

# Data sources:



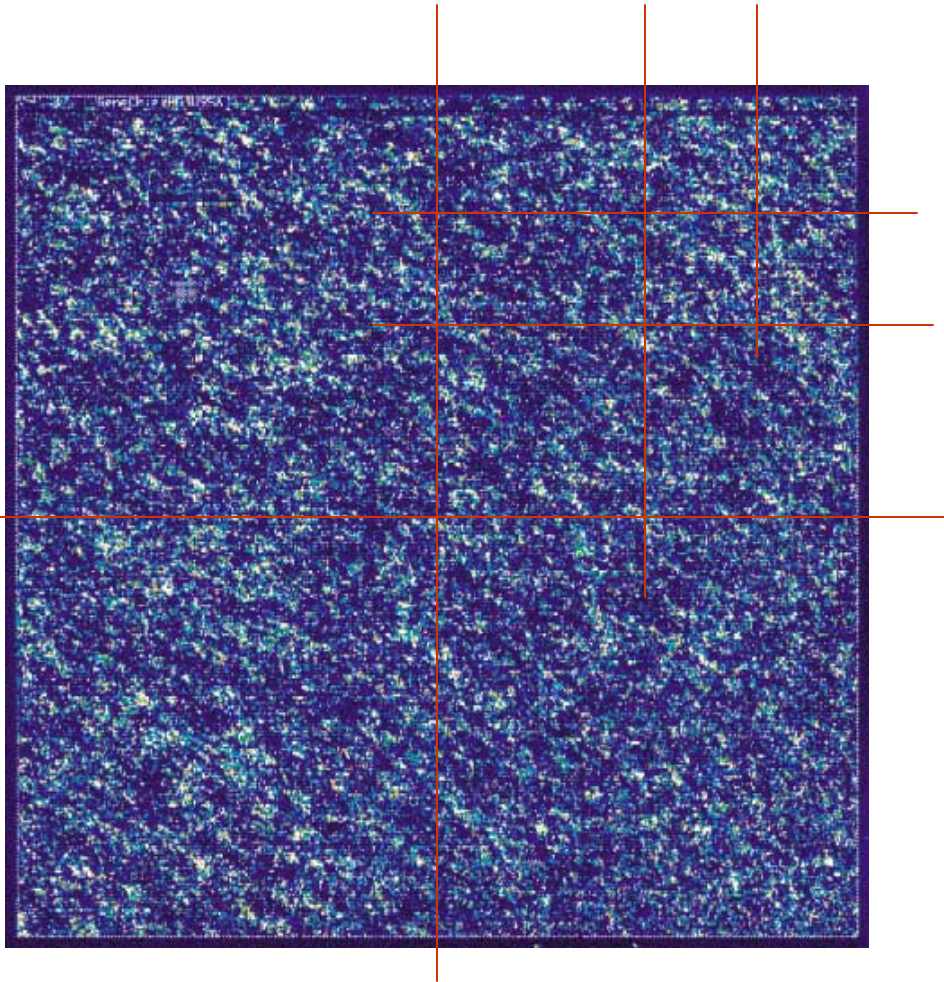
- All data used is courtesy of Affymetrix
- The same data which was used for the development of the Affymetrix algorithms
- The data was made available at the Affymetrix website
- The same data set was used for the CAMDA02 challenge

# From probe intensities to estimation of gene expression



Data from an experiment showing the expression of thousands of genes on a single GeneChip® probe array. Image courtesy of Affymetrix.

# Is it really as uniform as it looks?



T-test of mean intensities

$$t_{AB} = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\sigma_{AB}^2 \left( \frac{1}{n_A} + \frac{1}{n_B} \right)}}$$

Data from an experiment showing the expression of thousands of genes on a single GeneChip® probe array. Image courtesy of Affymetrix.

# The algorithm to determine the appropriate grid

## Begin

calculate *overall\_mean*;

**for** *grid* from 2 to *maximum\_grid*

**for** each local area of the *grid*

    compare *local\_mean* to the *overall\_mean*;

**if** *local\_mean* not equal to the *overall\_mean* **store** *grid*;

**end\_for**

**end\_for**

shuffle all probes; //stir, not shake ☺

**for** *uf\_grid* from 2 to *maximum\_grid*{

**for** each local area of the *uf\_grid*

    compare *local\_mean* to the *overall\_mean*;

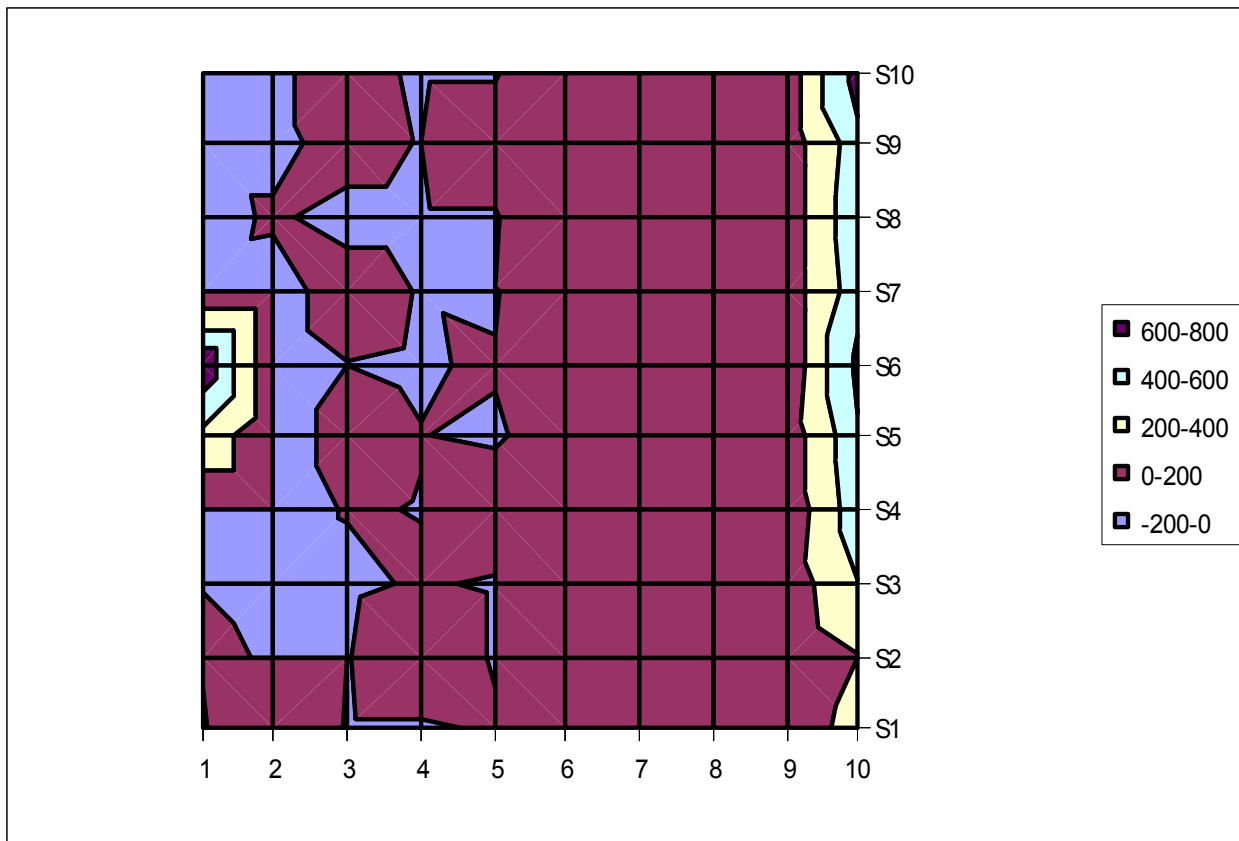
**if** *local\_mean* not equal to the *overall\_mean* **store** *uf\_grid*;

**end\_for**

**end\_for**

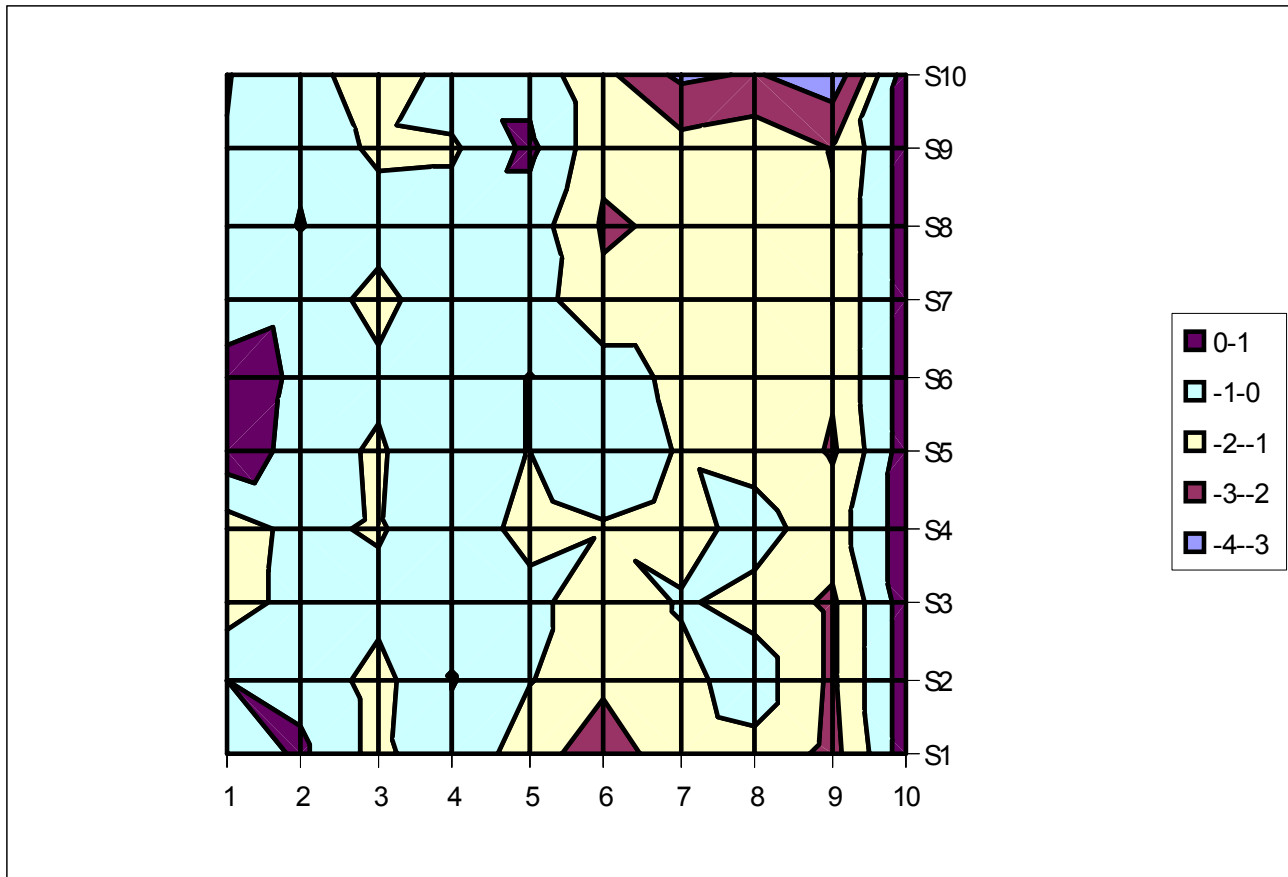
**end**

# Genechip topological maps:mean



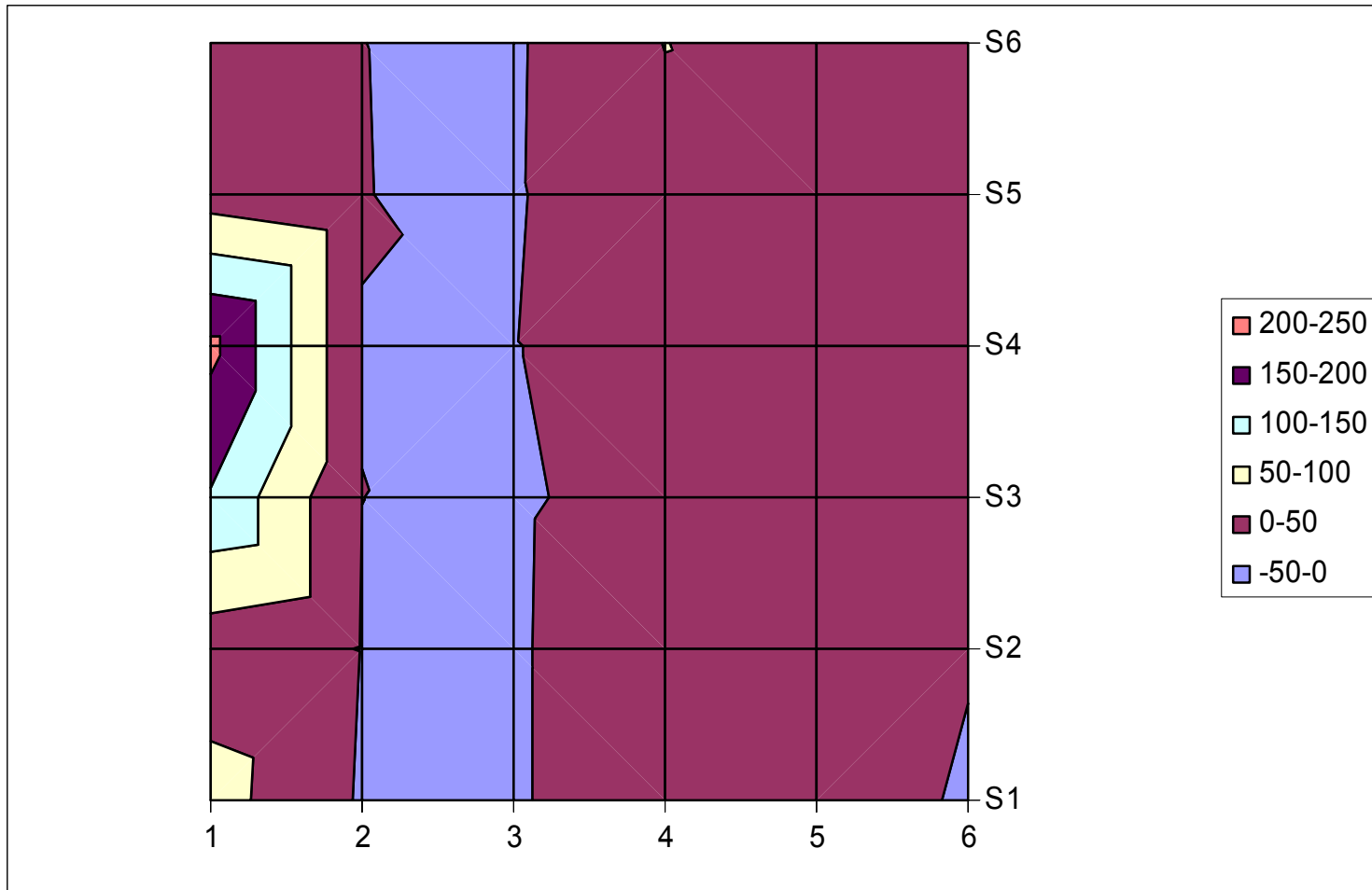
Heat map representation of intensity pattern in Affymetrix *E.coli* chip, 000825\_bgd\_Ds1.CEL from CAMDA02 data set. The grid is 6, which is the maximal grid for which shuffled probes' mean intensity in every square is equal to the overall mean intensity by t-test. For raw, non-shuffled probes the grid is 4, i.e. mean intensity of at least one of 16 square sub-regions is different from the overall mean for this chip. The color represents the difference of  $t$  value for each area of the grid from tabulated  $F$  distribution value with 0.05 confidence level and  $\infty$  degrees of freedom. In blue areas (below 0) the means are equal, while other areas show significant deviation.

# Genechip topological map: standard deviation



Heat map of significance of difference between local grid and overall standard deviations. With conf. 0.05 only a few areas differ from the overall standard deviation, a spot in the middle of the left side and the narrow area along all right side

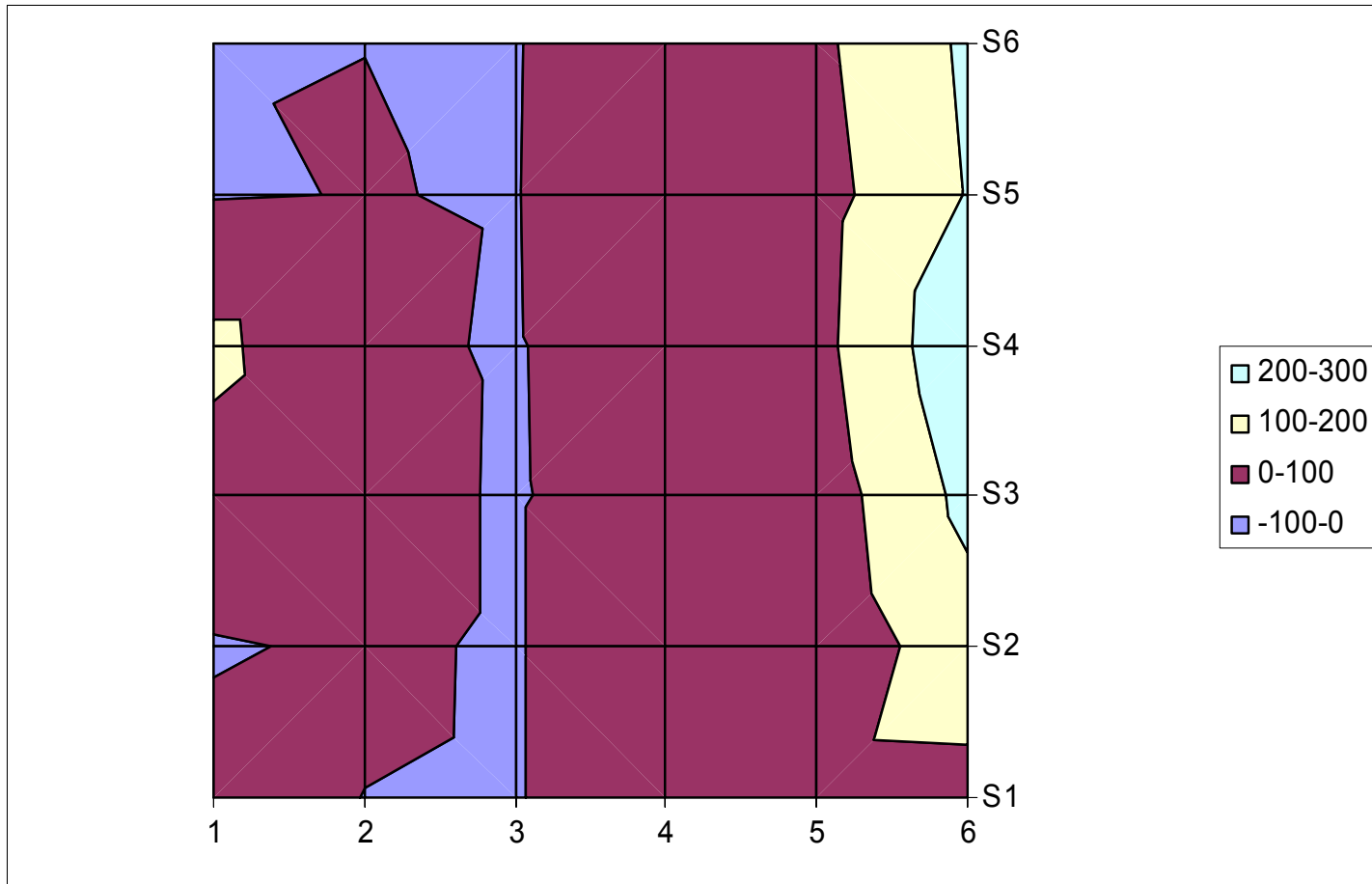
# Same chips, different flavors



The same 000825\_bgd\_Ds1.CEL chip as on previous slides, only MM probes taken into account

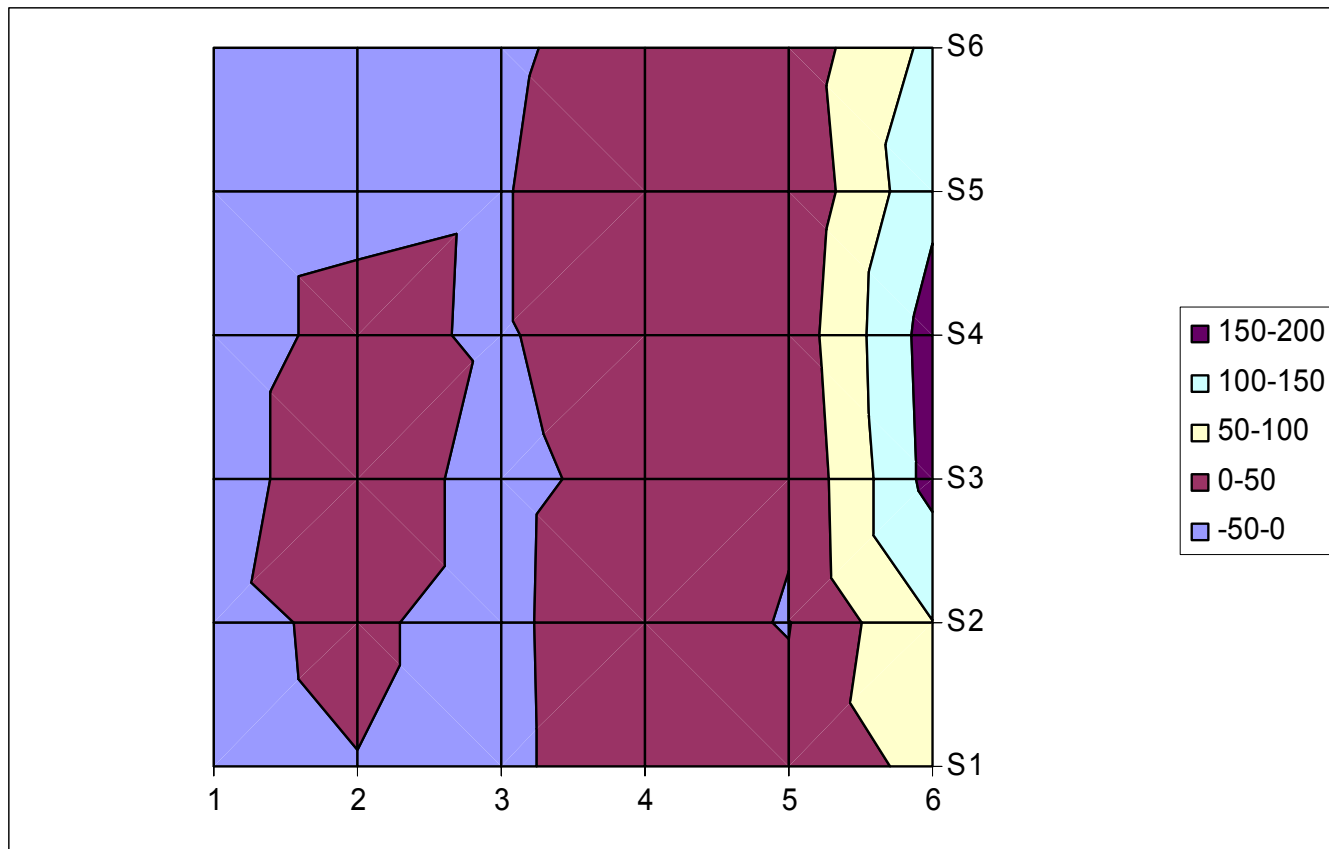


# Same chips, different flavors



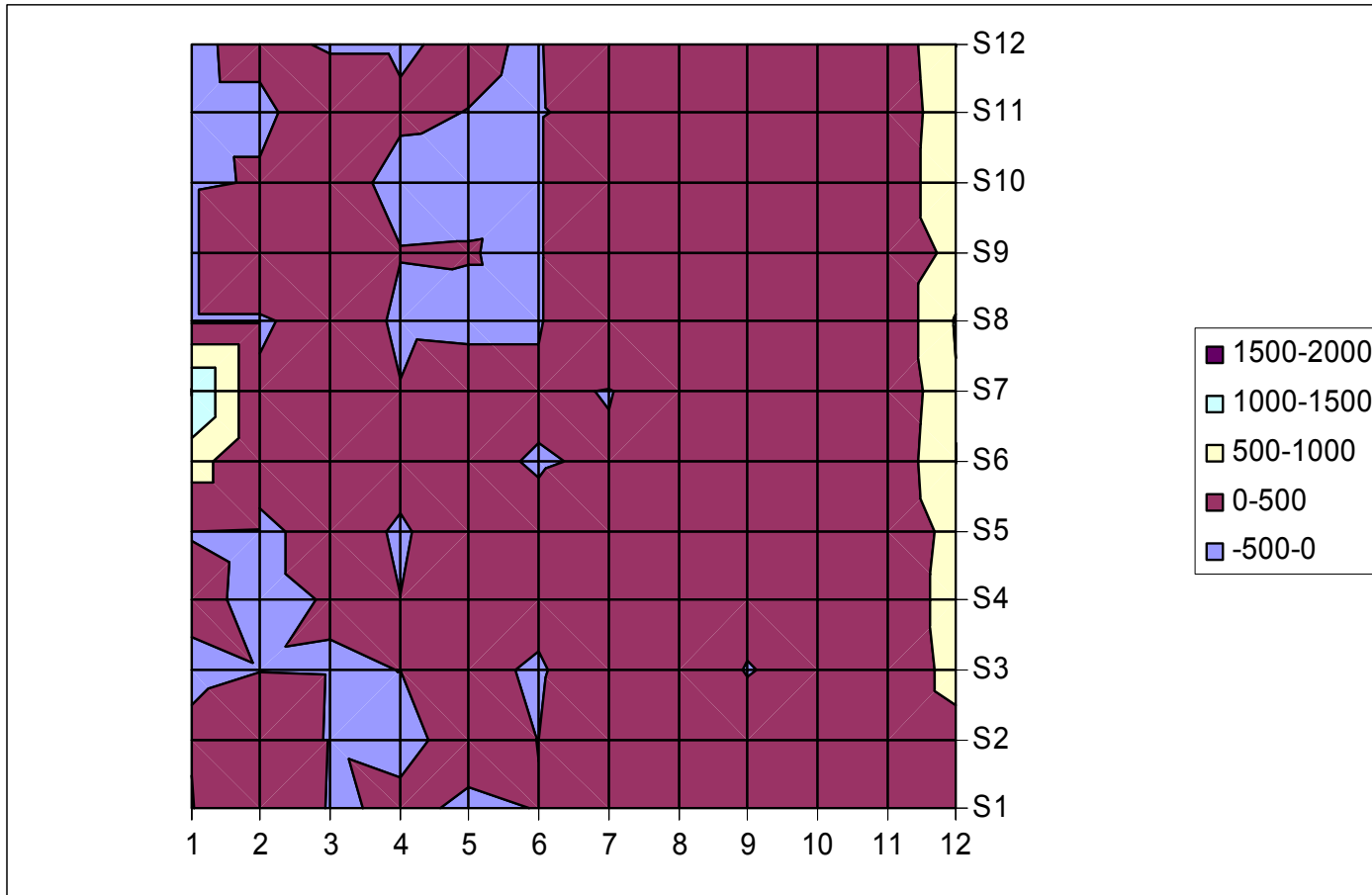
The same 000825\_bgd\_Ds1.CEL chip as on previous slides, only PM values taken into account

# Same chips, different flavors



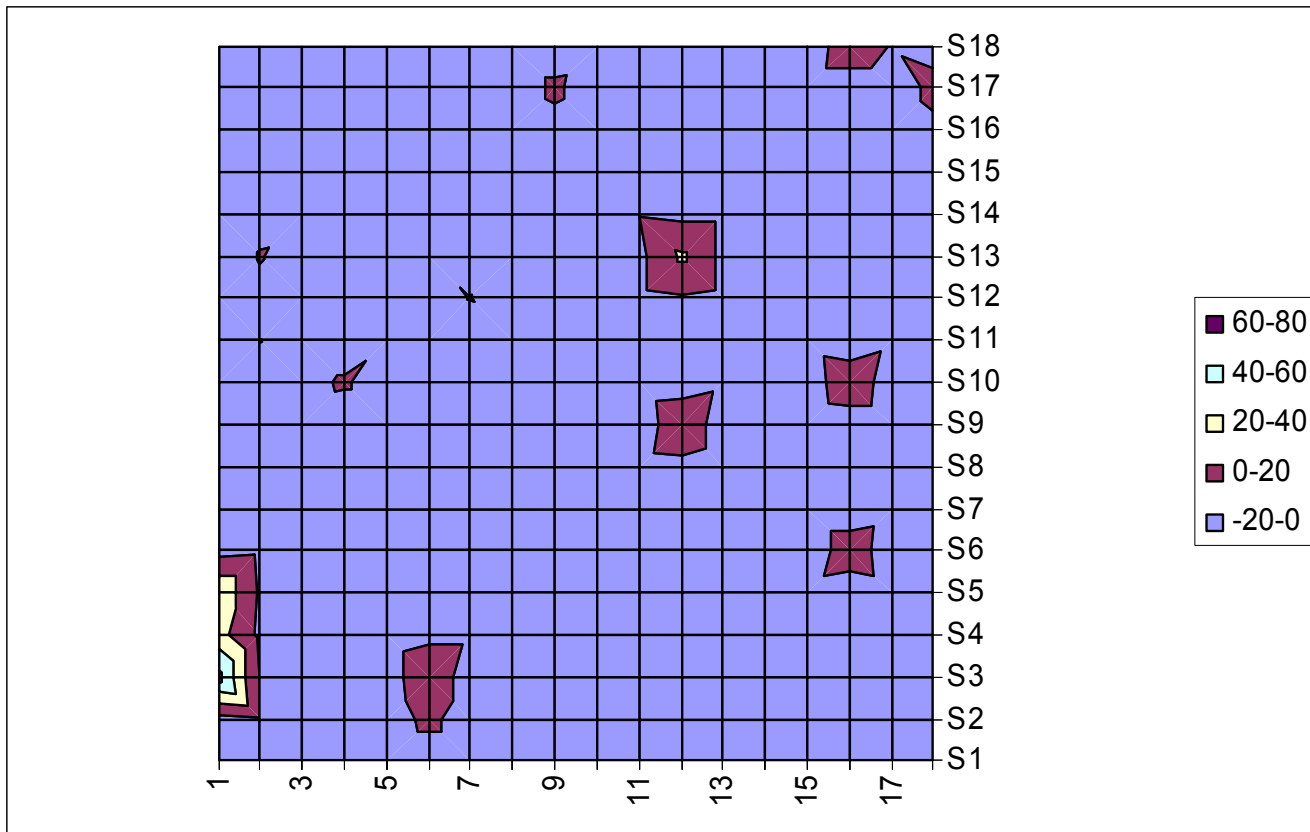
The same 000825\_bgd\_Ds1.CEL chip as on previous slides, pair-wise PM-MM values

# Same chips, more ketchup



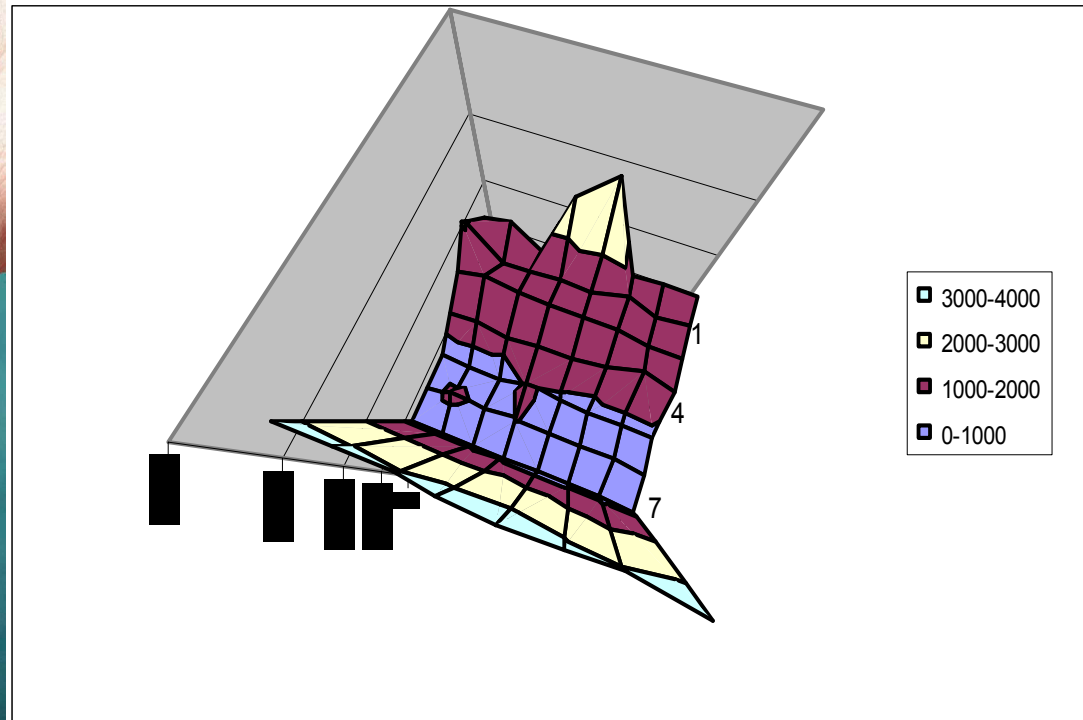
The same 000825\_bgd\_Ds1.CEL as before, logarithmic (ln) values of pair-wise PM/MM differences

# More chips



A similar topological map can be made for a human U95 chip (2353s99hpp\_av08.CEL in this example).

# Hypothesis: the origin of the Affy's belly button



A different representation of the same 000825\_bgd\_Ds1.CEL chip. This surface shows only raw mean intensity for each square of the grid. Comparing it to the map of significance we can say how meaningful deviation from the overall mean value is for each part of the chip

Affymetrix GeneChip® probe array. Image courtesy of Affymetrix.

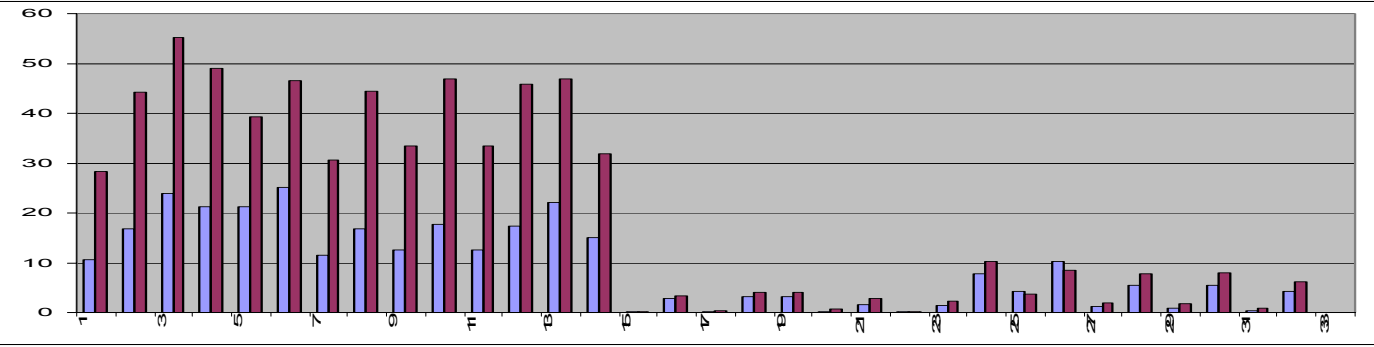
# What can we do about it?

- Fast quality control

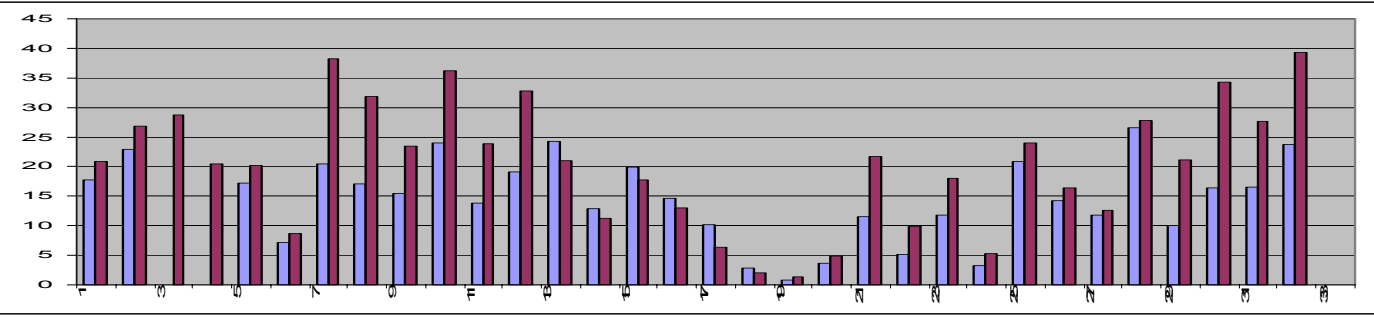
Chip 001003\_bgd\_Ds5 (*E.coli*) sticks out because of less uniform distribution of intensities. In this case difference between shuffled and non-shuffled grid is 6, while for other chips in this batch the difference is 4. The record in the table from the worksheet named “**Deviations**” within the excel file *Algo\_Dev\_Ecol.xls* reads “Left corner has large particle covering b3303\_rpsE & b3340\_c\_fusA”.

- Topological adjustment to the intensity values to compensate the local effects

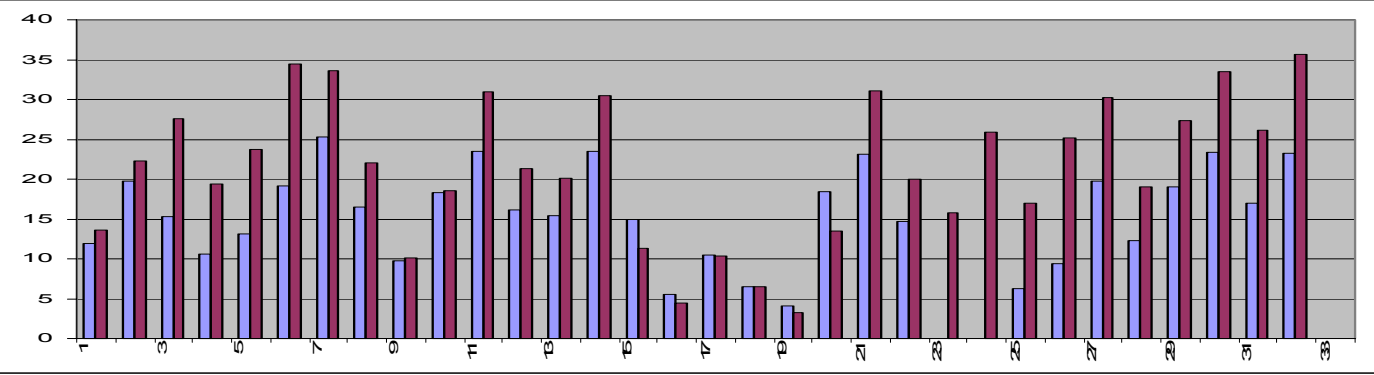
# Effect of topological adjustments on the probe intensities (U95 chip)



Unadjusted Z-score



Adjusted Z-score



# Acknowledgements:



- This work was partially supported by the BoR-HEF(2000-05)-08 grant