University of Essex

Identification of spatial biases in Affymetrix oligonucleotide microarrays

Jose Manuel Arteaga-Salas, Graham J. G. Upton, William B. Langdon and Andrew P. Harrison

University of Essex, U. K.

Agenda

1. Introduction
2. **Identification** of spatial biases **with replicates** .
3. **Reduction** of spatial biases **with replicates**.
4. **Identification** of spatial biases **without replicates**.
5. **Reduction** of spatial biases **without replicates**.
6. Conclusions

1. Introduction

- Microarrays are popular tools to measure gene expression.
- Several laboratories invest important resources on this technology.
- Affymetrix Oligonucleotide Microarrays **contain** spatial biases in their hybridizations (Suarez-Fariñas *et. al*. (2005); Langdon *et. al.* (2008)). The problem is **independent** of chip-type.
- Some methods have been proposed to **identify** and **reduce** these biases **for replicated arrays**.
- **No methods available** for experiments **without** replication.

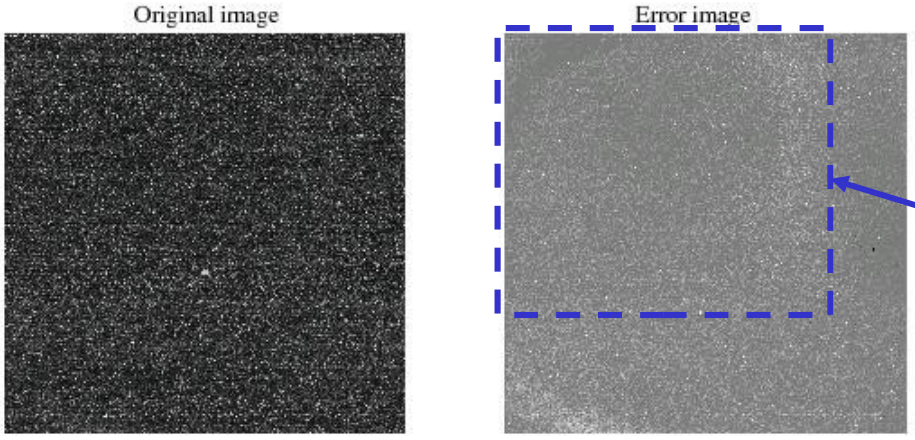
2. Identification of spatial flaws w/replicates

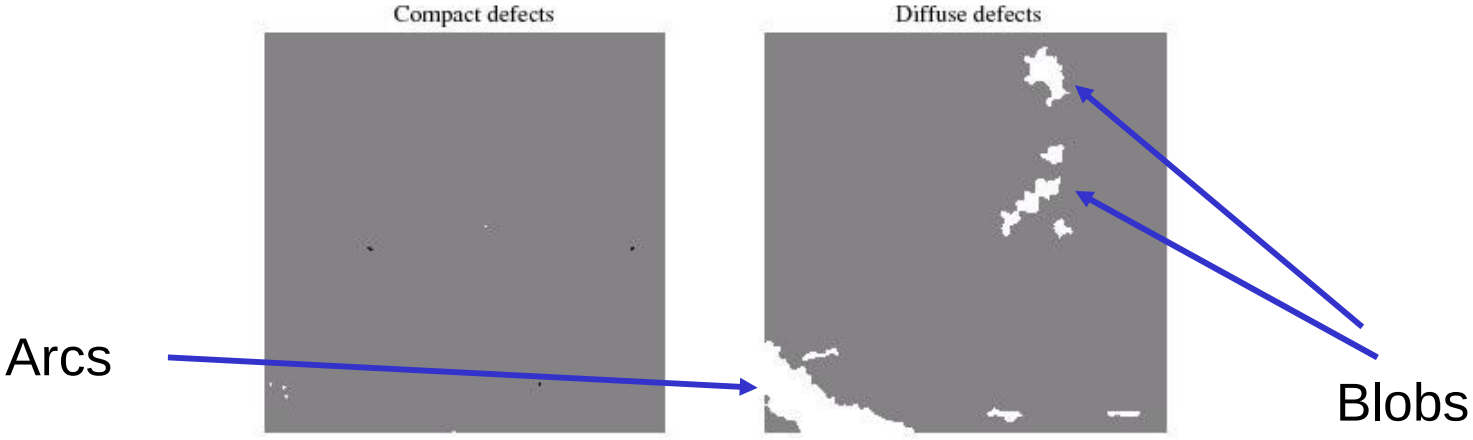
- Suarez-Fariñas *et. al.* (2005) developed the "**Harshlight**" package (available in Bioconductor).

- Harshlight uses statistical and image processing methods to identify spatial defects.

- After identification of flawed locations in the array the user can correct by **substituting with the median** value of all the available arrays at each location, **or with "N/A"**.

- Disadvantage: ONLY works in the presence of replicate arrays.





Chip summary:

Extended defects: the variance of the Error Image explained by the background is 11.64

	compact	diffuse
Number of clusters found:	9	9
Percent of the surface covered by the defects:	0.02	5.34

Harshlight report for 3 replicates of the GSE4217 experiment available at GEO (arrays GSM96262-4)

2.1 Another method

- Arteaga-Salas *et. al.* (2008) developed an independent method to identify spatial biases using replicate arrays.

- For location *(i,j)* and replicate *r* calculate d_{ijr}

$$d_{ijr} = \frac{L_{ijr} - \alpha_{ij}}{\beta_{ij}}$$

Where L_{ijr} is the logarithm of the observed intensity values, α_{ij} is the median of the L_{ijr} values and β_{ij} is the standard deviation of the L_{ij} values.

- Select locations where **abs(d_{ijr})>25%** (say).

- The selected locations represent "unusually high" or "unusually low" values, in comparison with a **reference set** (in this case, the reference set is the **median** of all replicates).

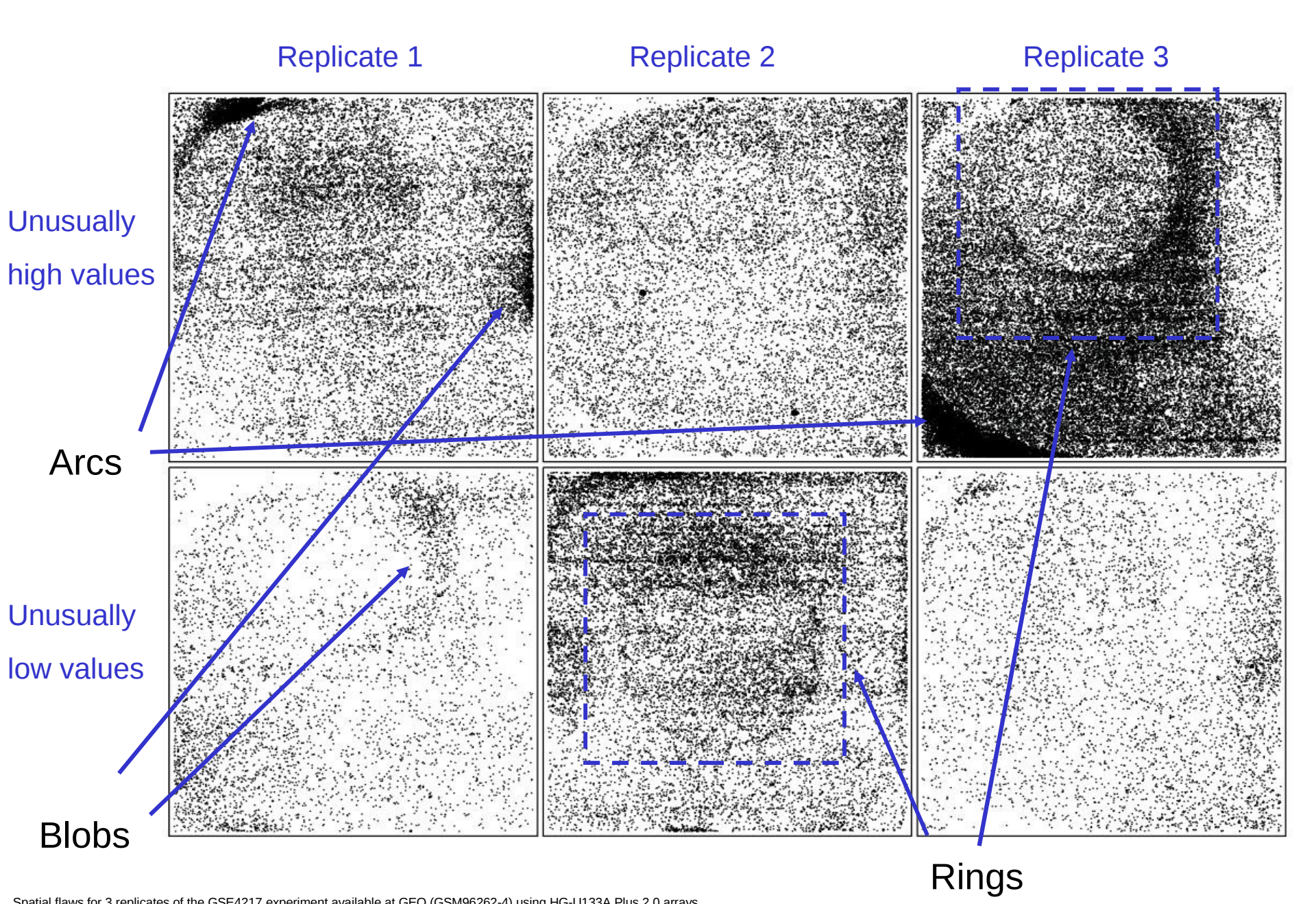
- Disadvantage: ONLY works in the presence of replicate arrays.

- **Next**:

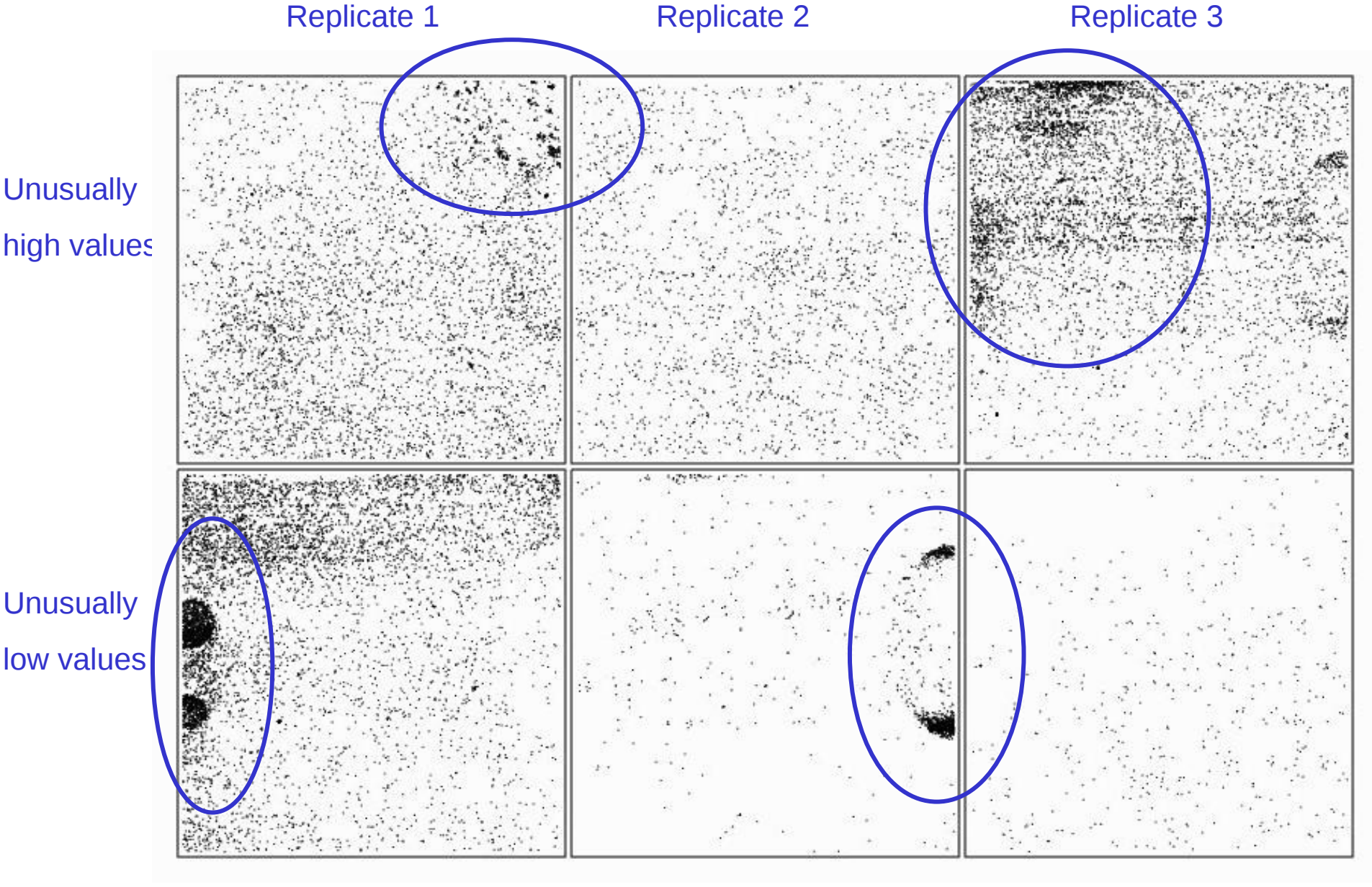
Example 1: Three **HG-U133 Plus 2.0** replicates (from GEO).

Example 2: Three **HG-U133A** replicates (from Affymetrix).

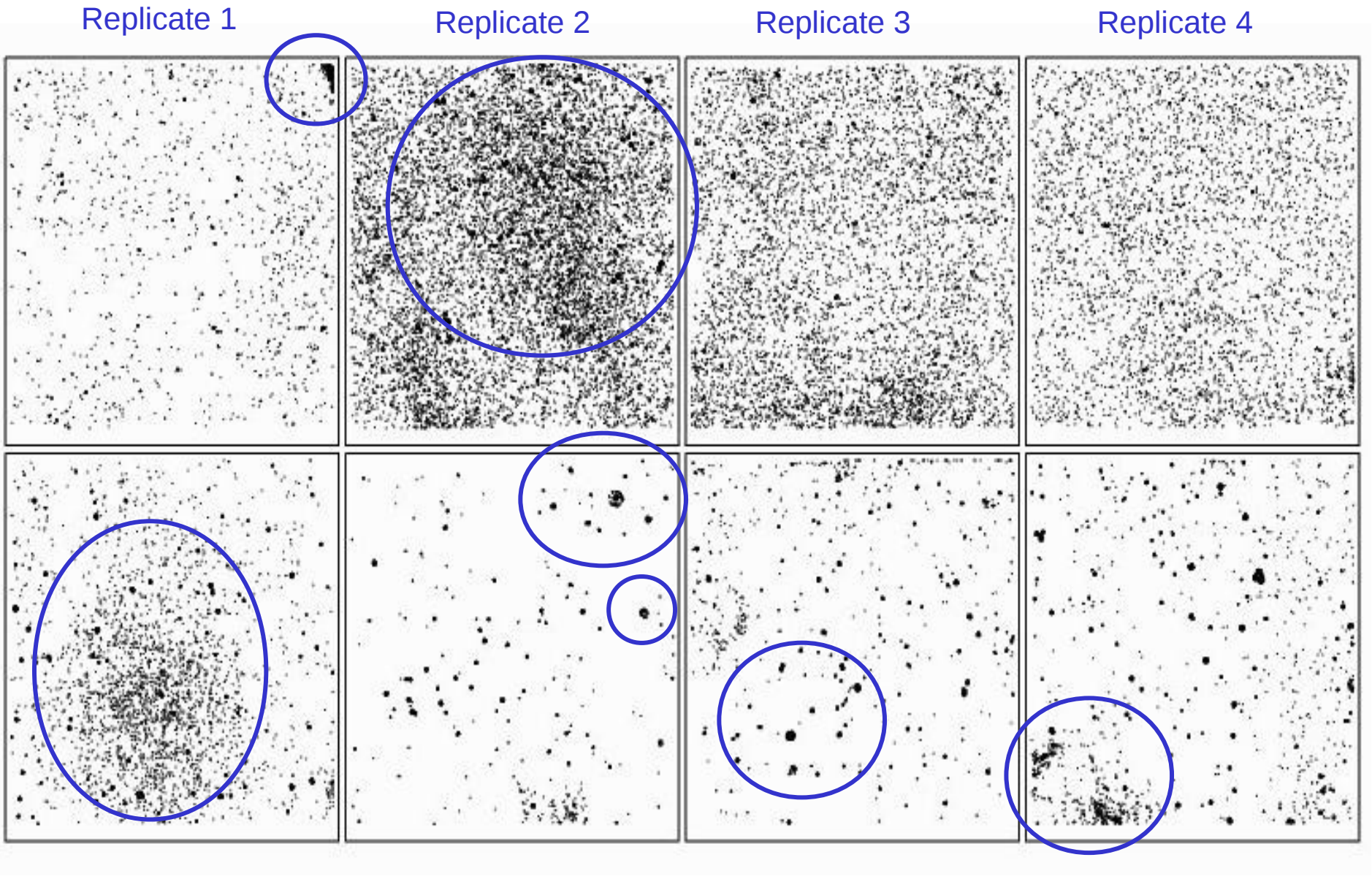
Example 3: Four **DrosGenome1** replicates (from GEO).



Spatial flaws for 3 replicates of the GSE4217 experiment available at GEO (GSM96262-4) using HG-U133A Plus 2.0 arrays



Spatial flaws for 3 replicates of the HG-U133A SpikeIn Experiment -- Affymetrix



Spatial flaws for 4 replicates of the GSE6515 experiment available at GEO (GSM149276-9) using DrosGenome1 arrays

3. Reducing spatial biases w/replicates

- Harshlight proposes to **substitute** flawed locations **with the median (HMS)** of all the arrays at each location or with "**N/A**".

- Arteaga-Salas *et. al.* (2008) introduced **two procedures** to assist with flaw removal:

CPP (complementary probe pair) adjustment, suitable **only** for **replicated** arrays.

LPE (local probe effect) adjustment, suitable for **replicate or non-replicate** arrays.

- CPP and LPE can be used **separately** or **in sequence**.

3.1 Local Probe Effect (LPE) adjustment

- LPE can be used whenever *R* (*R>2)* **arrays** are available.
- It uses the spatial structure in a **5 x 5 window** centred at location *(i,j)* to decide whether adjustment should take place.
- For array *r* we **first** calculate the values d_{ijr} given by,

$$d_{ijr} = \frac{L_{ijr} - \alpha_{ij}}{\beta_{ij}}$$

Where L_{ijr} is the logarithm of the observed value, α_{ij} is the median of the L_{ijr} values and β_{ij} is the standard deviation of the L_{ij} values.

- Now, define I_{ij} and G_{ij} as follows:

I_{ij} – The identifier of the array where d_{ijr} has largest absolute value.

G_{ij} – Is 1 if the *d*-value with largest magnitude is positive, otherwise is equal to -1.

- Using these two values calculate E_{ij} with,

$$E_{ij} = I_{ij} \times G_{ij}$$

So, with *R* arrays, E_{ij} takes **one** of the values { *-R,-(R-1),…,-2,-1, 1,2,… (R-1),R* }

- **An example**,

Cell at location *(i,j)*

	r=1	*r*=2	*r*=3
Original	45	38.8	34952
L_{ijr}	3.807	3.658	10.462
d_{ijr}	-0.558	-0.596	**1.154**

α_{ij} =	5.976
β_{ij} =	3.886

I_{ij} =	3
G_{ij} =	1
E_{ij} =	**3**

5 x 5 window centered at *(i,j)*

-1	-1	**3**	1	-2
-1	**3**	**3**	**3**	**3**
3	**3**	**3**	**3**	1
3	**3**	**3**	**3**	**3**
-1	3	**3**	**3**	-2

17 cases where *E=3*

- If the **5 x 5 window** contains a **majority** of **informative locations** (PM or MM only) with the **same *E*-code**, then a **spatial bias** is present.

We **adjust** the value in cell *(i,j,r).*

- Let Δ be the set of ***N*** informative locations within the window (in the example, ***N=17***).

- For each location in Δ we calculate the ***d***-values for array ***r*** in need of correction, and let $\bar{d}$ be their average.

- The adjusted value L^a_{ijr} is given by,

$$L^a_{ijr} = L_{ijr} - \beta_{ij}\,\bar{d}$$

3.2 Results

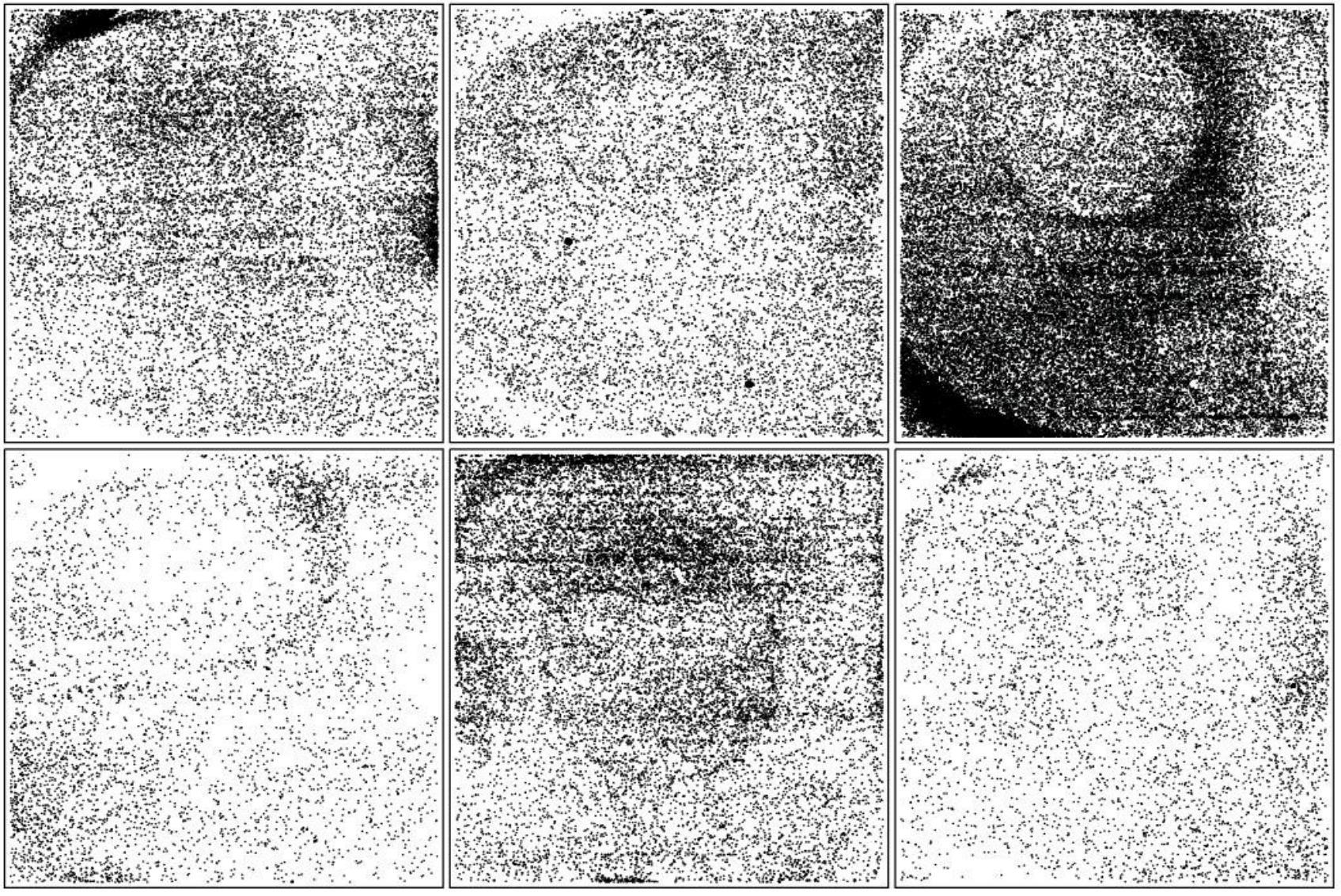
- We apply LPE+CPP and Harshlight Median Substitution (HMS) to

Example 1 to illustrate the **reduction** of the spatial biases:

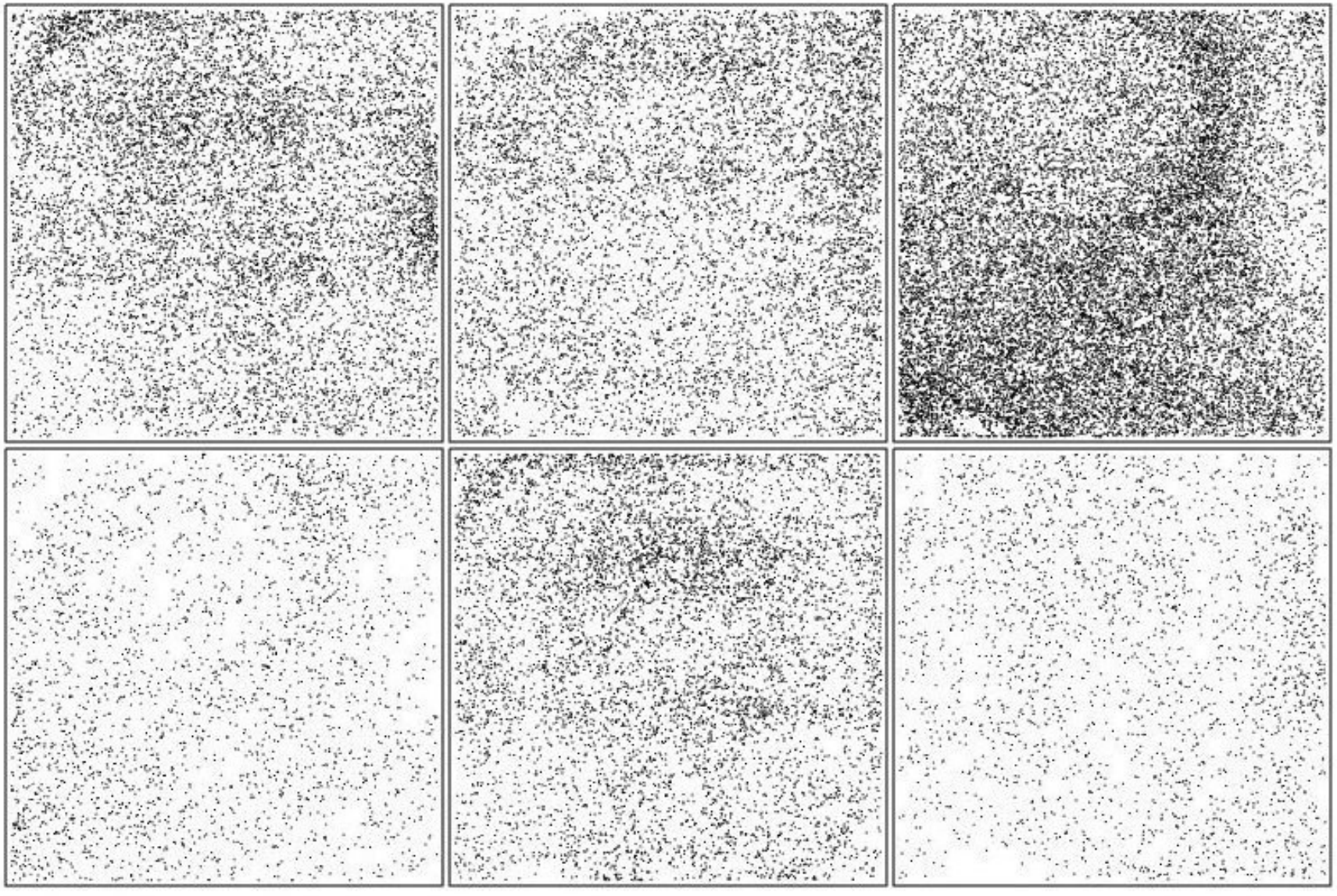
Total % of defects

	replicate 1	replicate 2	replicate 3
original	→ **6.3**	→ **7.9**	→ **8.9**
HMS (once)	1.7	3.0	3.3
HMS (twice)	0.8	2.2	2.3
CPP	0.9	0.9	1.8
LPE	3.8	5.3	5.2
CPP+LPE	0.8	0.9	1.8
LPE+CPP	→ **0.6**	→ **0.6**	→ **1.7**

Example 1 (three HG-U133 Plus 2.0 replicates)

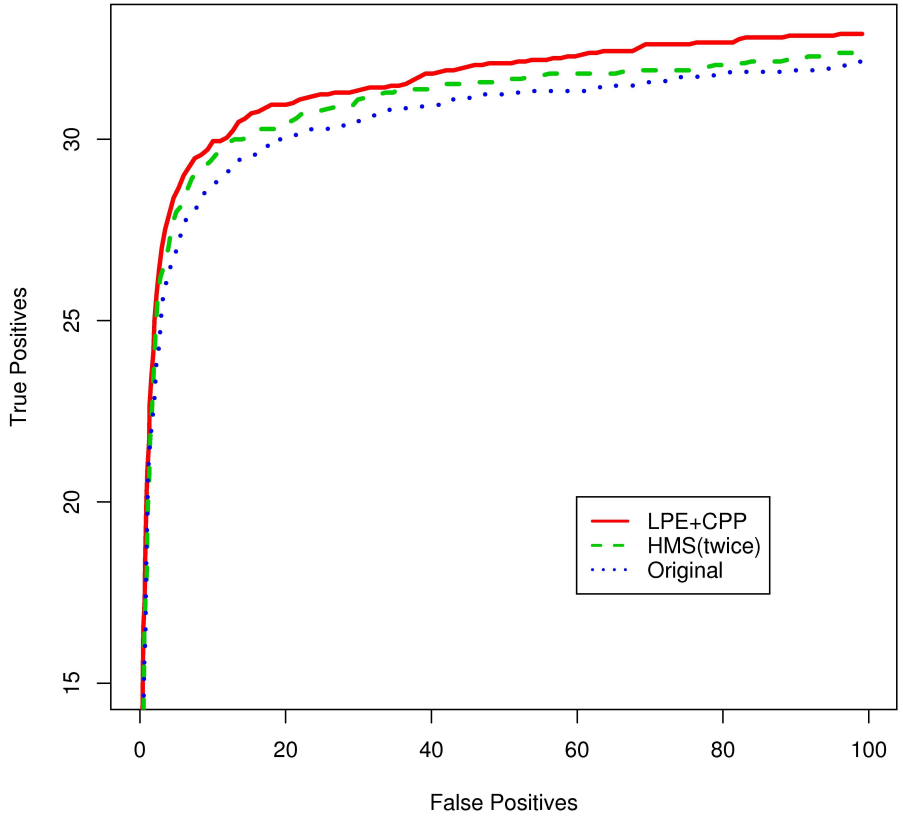


Example 1 after LPE+CPP



How do we know that these adjustments are the *appropriate* adjustments?

ROC curves to measure the rate of false/negative positives in the HG-U133A Spike-In Experiment (Affymetrix) before and after Spatial Flaws Reduction. Gene Expression summarized with RMA.



From Arteaga-Salas *et. al. (2008)* in "Statistical Applications in Genetics and Molecular Biology" (SAGMB).

4. Identification --- without replicates

- In the absence of replicates the two methods described before are not applicable to visualize spatial flaws.

- To identify spatial biases without replicates we need an **alternative reference set** to compare the values.

- Langdon *et. al.* (2008) calculated an "**Average GeneChip**" and a "**Variance GeneChip**" using Affymetrix Chips in the Gene Expression Omnibus (GEO) as available in February 2007.

- This was done separately by Chip type and organism.

4.1 The Average GeneChip

- To obtain the “Average GeneChip” the **arithmetic mean** of the **natural logarithm** of the observed probe values in each available chip was calculated.

- The **upper and lower 0.5%** of the values were discarded to avoid the effects of outliers.

- Using the same set of data the **variance** was calculated to obtain the “Variance GeneChip”.

4.2 Steps to visualize spatial biases

Let *A* be the Average GeneChip, *V* the Variance GeneChip and *L* the logarithm of the observed values.

1. For each location *(i,j)* in the array, calculate

$$h_{ij} = \frac{L_{ij} - A_{ij}}{\sqrt{V_{ij}}}$$

2. **Sort** h_{ij} by column *j*. For each sorted value assign a rank, and store them in **array *K***.

3. Define a "**sub-array**" centered at *(i,j).* A sub-array **size 11 x 11** includes enough spatial information in a neighbourhood.

4. The sub-array centered at K_{ij} contains information about PM/MM/other probes. To **avoid correlated values** we do not consider adjacent cells (only one probe in a PM,MM probe pair). In total we select **61 probes** from the total 121 available.

   Calculate the scores Z_{ij},

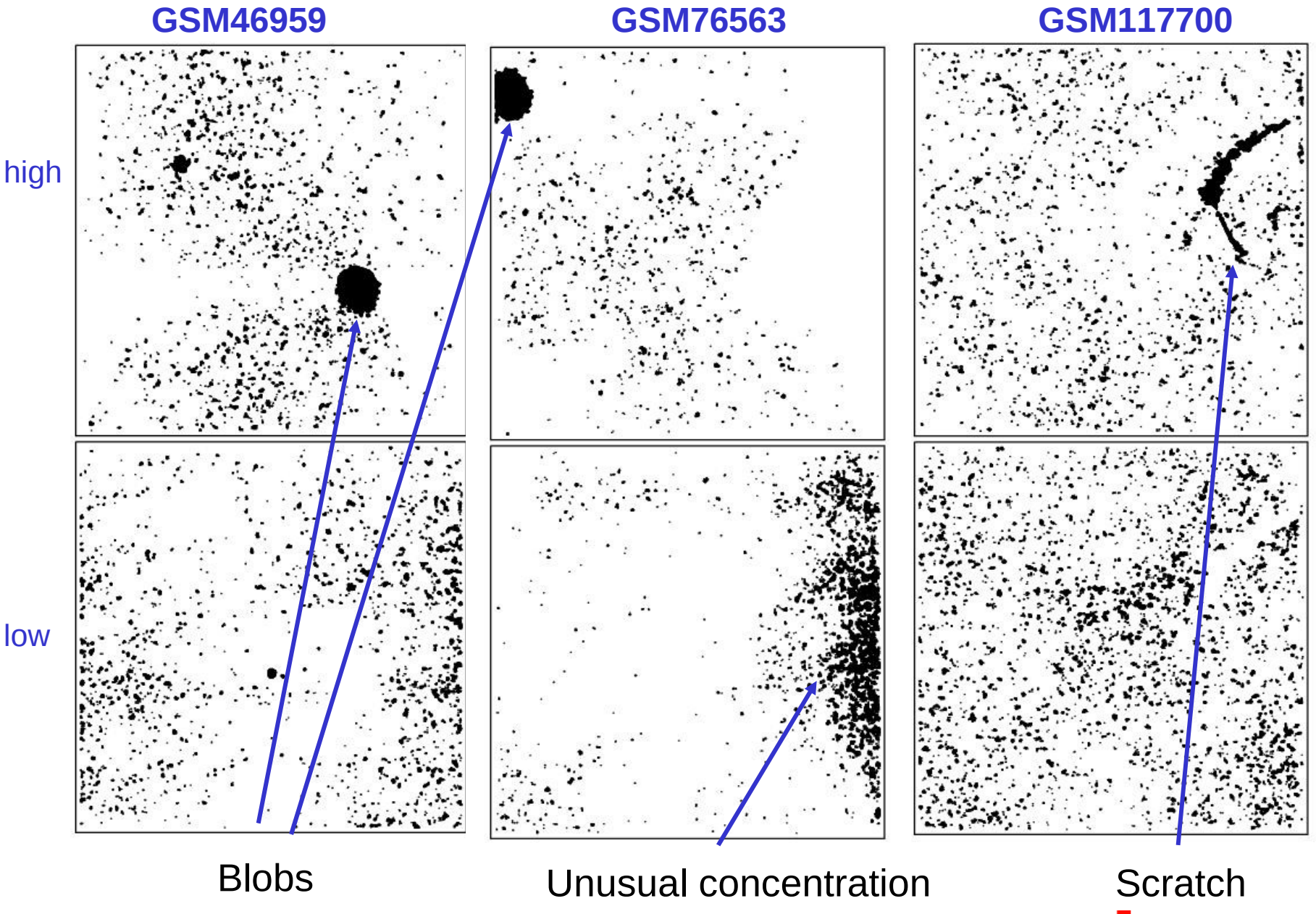
$$Z_{ij} = \frac{\sum_{n=1}^{61} K_n - 61 * \mu}{\sqrt{61 * \sigma^2}}$$

μ is the mean and σ^2 is the variance of a discrete uniform distribution (defined by the size of the chip).

The scores $Z \sim N(0,S^2)$. In the absence of spatial biases $S^2=1$.

5. **Plot** the locations where ***abs(Z)>= 2*S*** to identify neighbourhoods with unusually low or unusually high values.

Following these 5 steps we applied the procedure separately to **three HG-U133 Plus 2.0 arrays** from GEO (GSM46959, GSM76563 and GSM117700), from the accession number GSE2109.


GSM46959
GSM76563
GSM117700
high
low
Blobs
Unusual concentration
Scratch

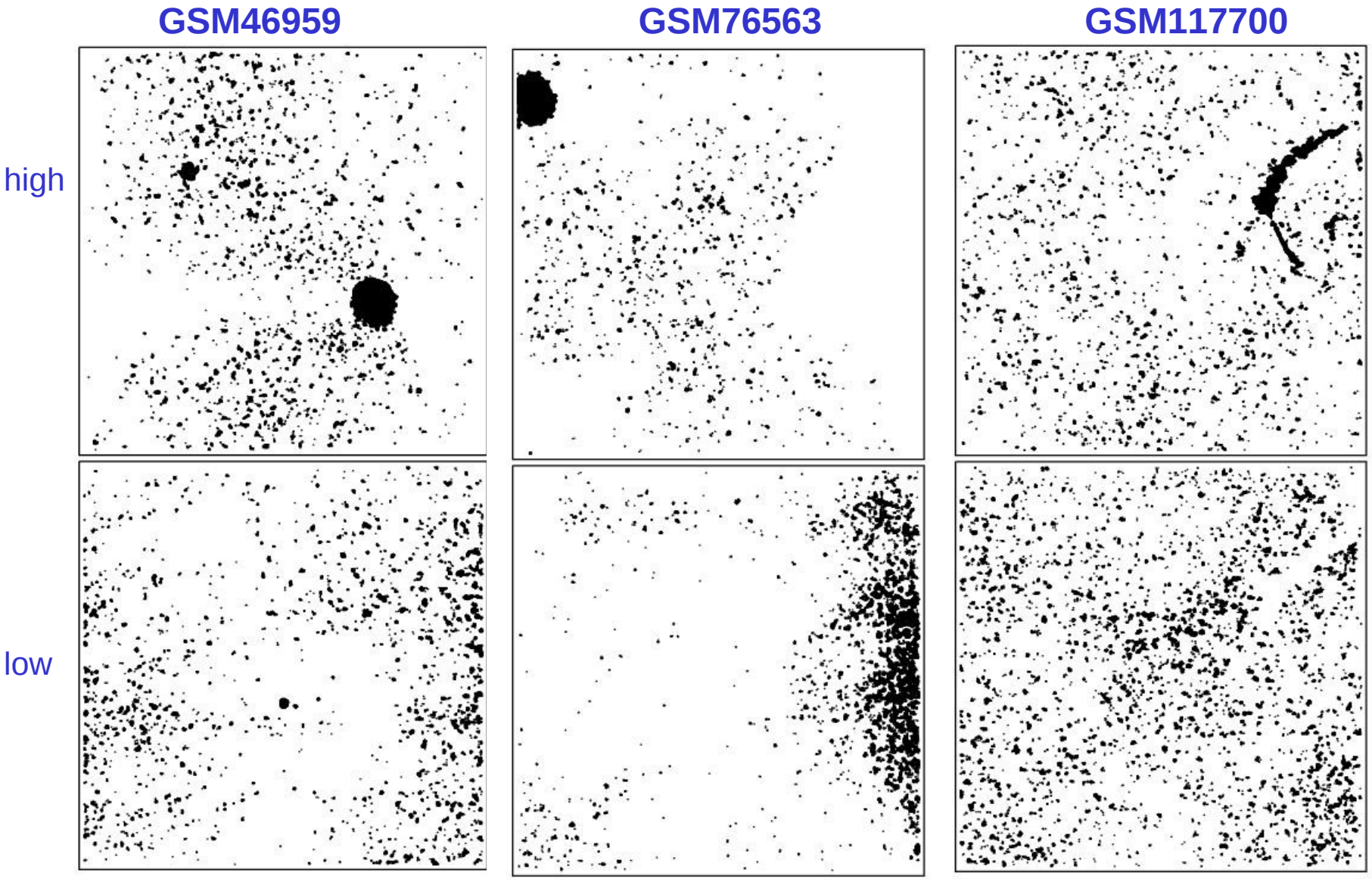
5. Reducing biases – without replicates

- **Problem**: In the absence of replicates, **two of the three** methods presented are not applicable (CPP and Harshlight are not, LPE is).
- Without replicates we don't know which are the "correct" values (we need some reference arrays).
- **Alternative:** We can **compare** a "contaminated" array **with other arrays** (at least two) **of the same type** where flaws have been **previously** reduced.
- In Section 4 we presented three HG-U133A Plus2.0 arrays "contaminated". In Section 3 we "cleaned" three replicate arrays of the same type.

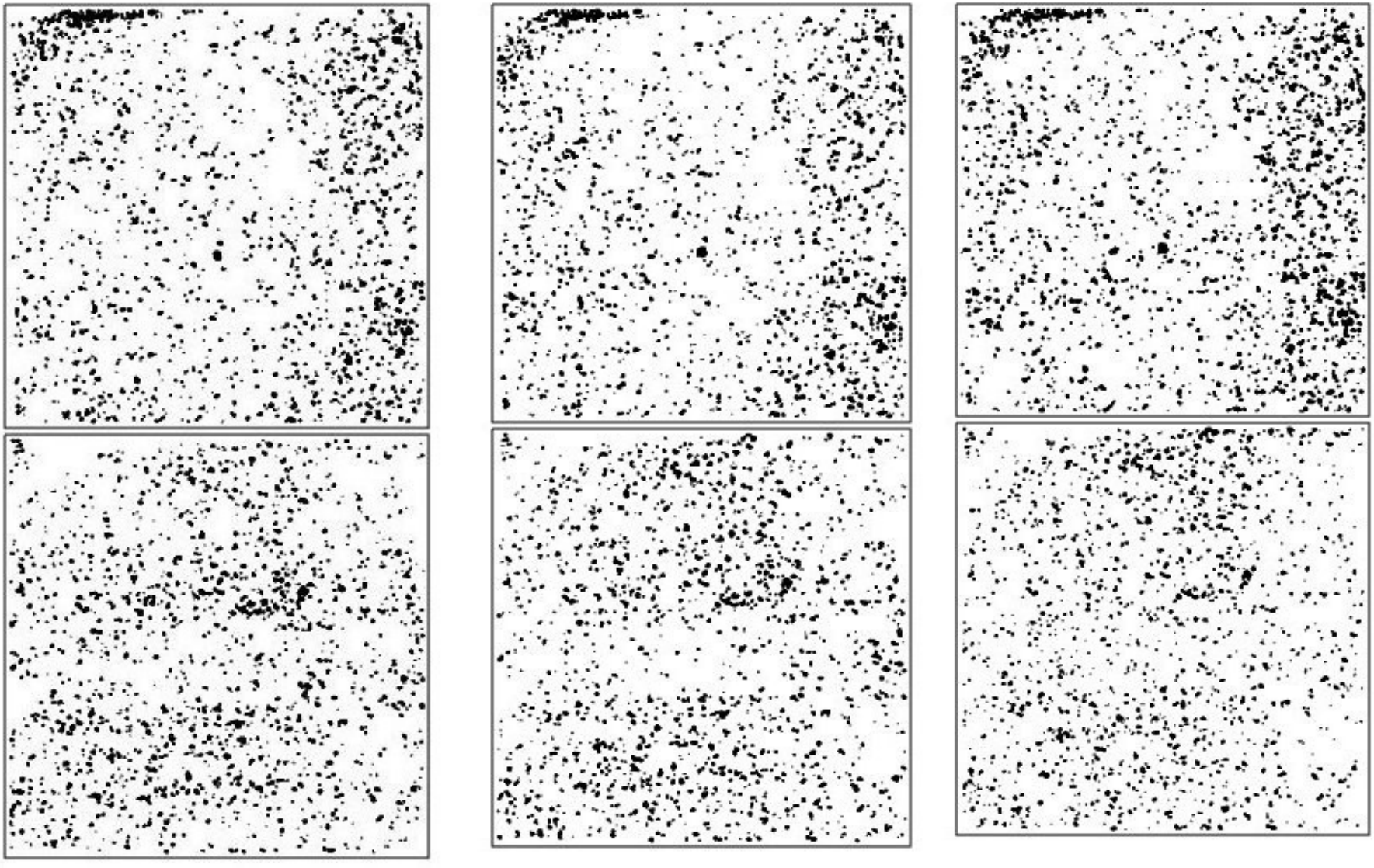
The "clean" arrays: **choose two** of the three replicates previously cleaned with LPE+CPP (let's choose the first and second replicates according to the Table).

The "contaminated" arrays: the three arrays presented in part 3.2 (the process is done **separately** for the three arrays).

- We now have **three arrays of the same type**.
- We can remove the flaws in the contaminated array using **LPE**.


GSM46959
GSM76563
GSM117700
high
low

Remaining flaws after LPE



5. Conclusions

- Oligonucleotide arrays **contain** spatial flaws in their hybridizations (they are usually manifested as “blobs”, “rings” or “scratches”).
- The problem **IS NOT** uncommon.
- Some methods to reduce flaws exist, but **not** for experiments **without replication**.
- Spatial biases **AFFECT** gene expression measurements.

THANK YOU!