Quality assessment for high-dimensional data:

- Dead pixels in CT scans for 3D printed objects
- Molecular biology in cancer treatment decisions

Dr Julia Brettschneider

Department of Statistics, University of Warwick

Warwick Statistics

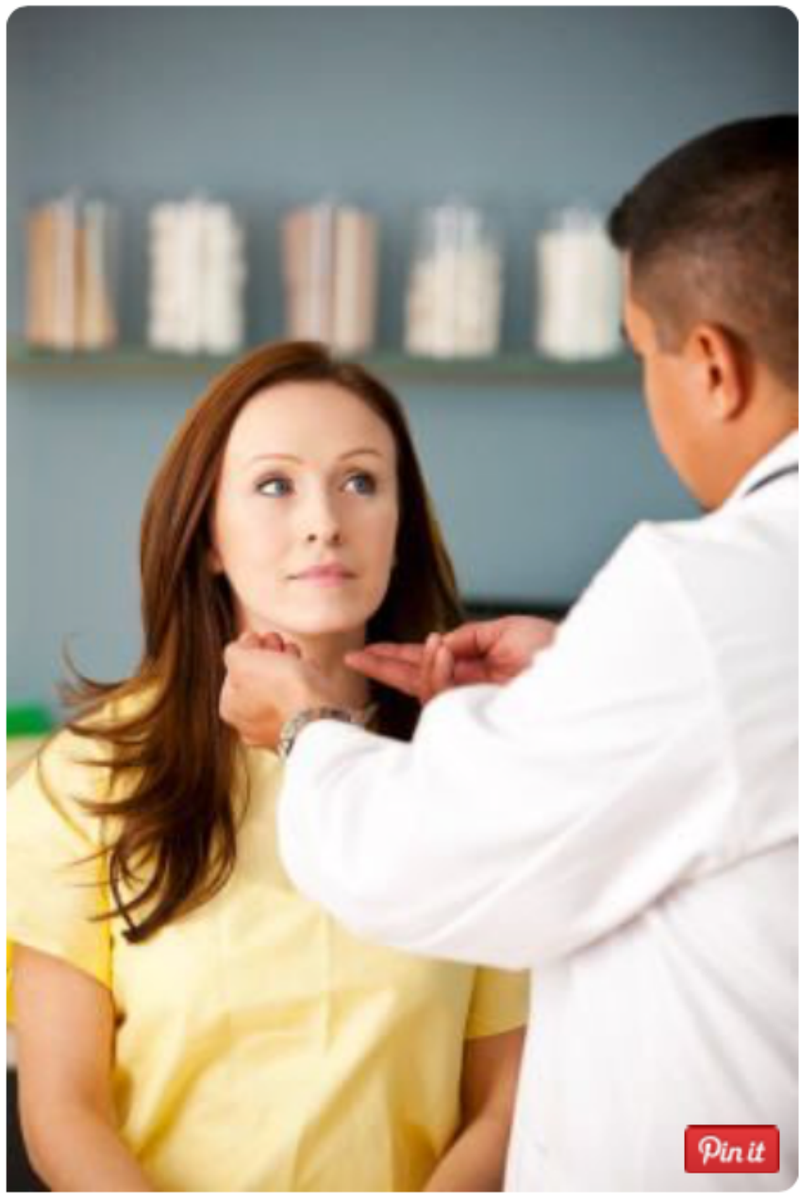
AS & RU
March 27, 2017

THE UNIVERSITY OF WARWICK

Application of statistical genomics: Cancer prognosis

Afirma Thyroid Analysis May Help Patients Avoid Surgery

Nodules No Longer Classifed as Inconclusive or Indeterminate



The Afirma Thyroid Analysis may help patients with an indeterminate nodule avoid unnecessary thyroid surgery. istockphoto

By **Mary Shomon** - Reviewed by a **board-certified** physician.
Updated July 16, 2016

Thyroid cancer is the fastest-growing cancer in the United States. There were an estimated 44,670 new cases in 2010, according to the American Cancer Society. Along with the increased awareness of thyroid cancer comes increased scrutiny of thyroid nodules. Thanks to more vigilant monitoring, ultrasounds, and x-rays, more thyroid nodules are being detected and evaluated.

When a thyroid nodule is considered suspicious -- meaning that it has characteristics that may suggest thyroid cancer -- the key evaluation is a fine needle aspiration (FNA) biopsy.

The FNA biopsy helps determine whether the nodule is malignant (thyroid cancer) or benign.

Application of spatial statistics: Dead pixels

Technology

Nintendo Switch owners complain about dead pixels

By Jane Wakefield
Technology reporter

7 March 2017 | Technology Share

Some users have reported "annoying" screen glitches

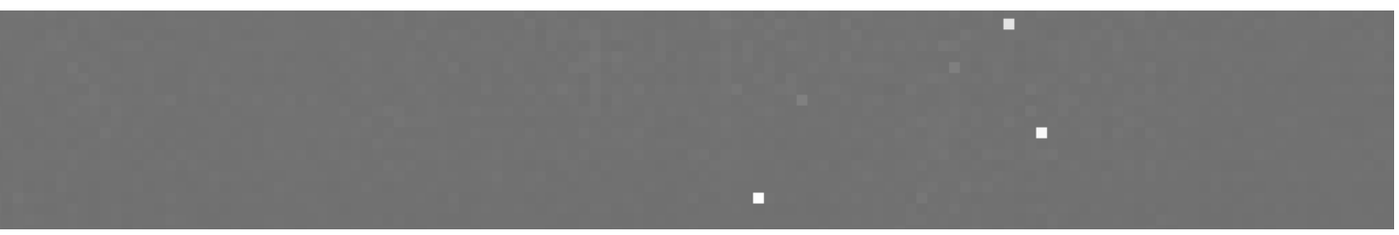
Thousands of owners of Nintendo's new console, Switch, have complained about dead or stuck pixels creating distracting and annoying dark squares on their screens.

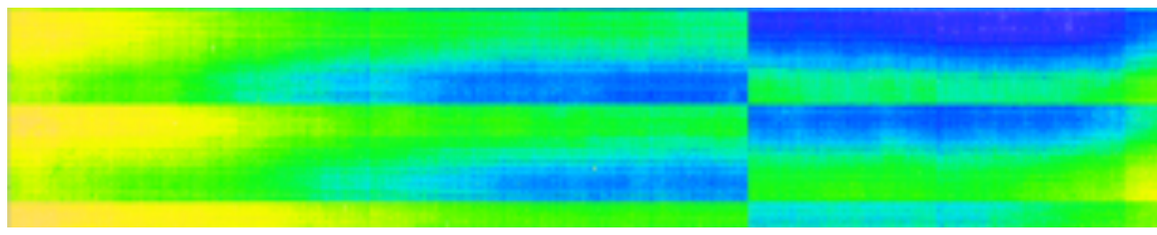
Application of spatial statistics:

Dead pixels in detectors of computed tomography machines

Part of quality control for 3D printed objects

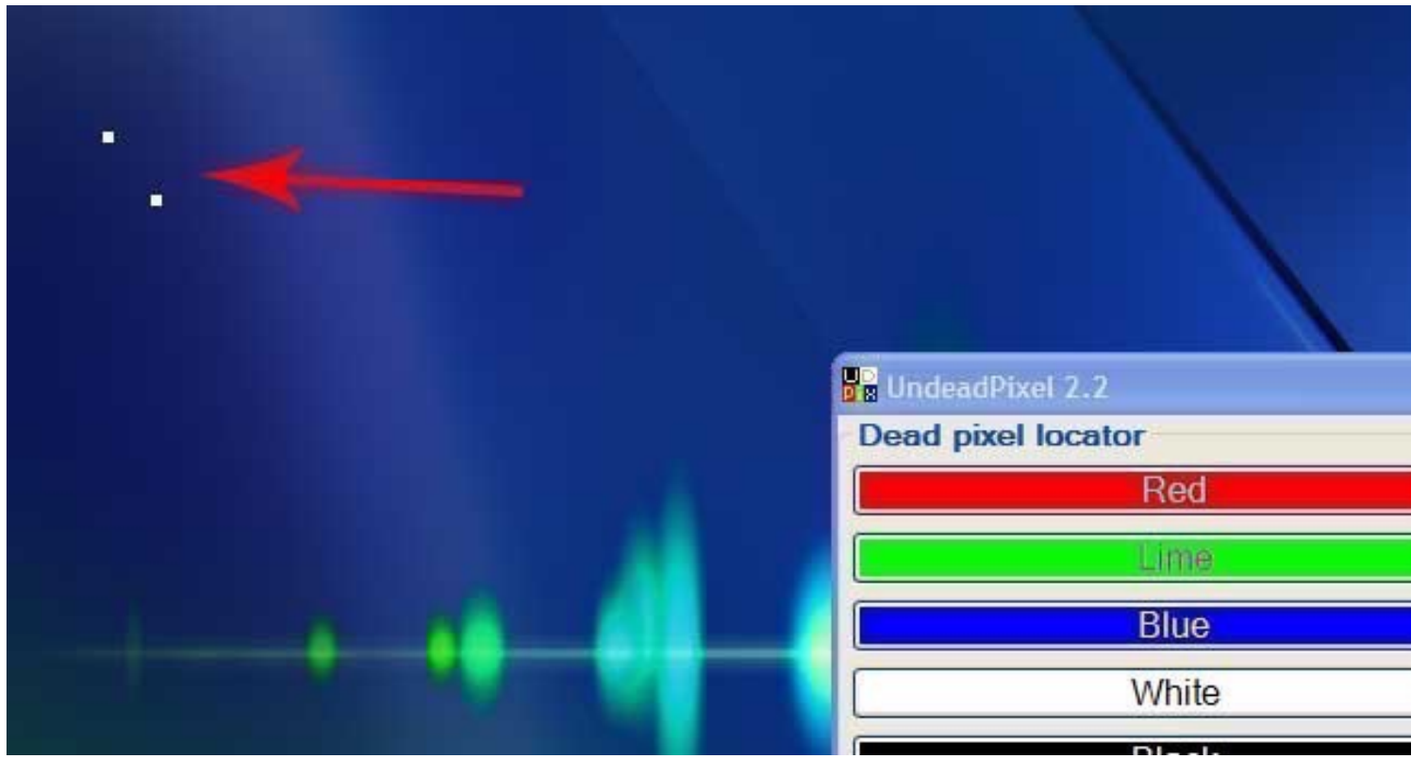
joint project with Warwick Manufacturing Group



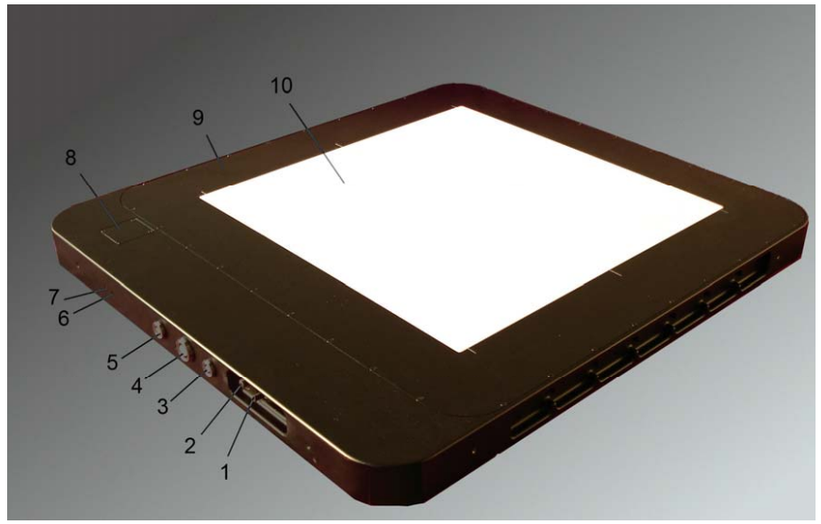


Dead pixels

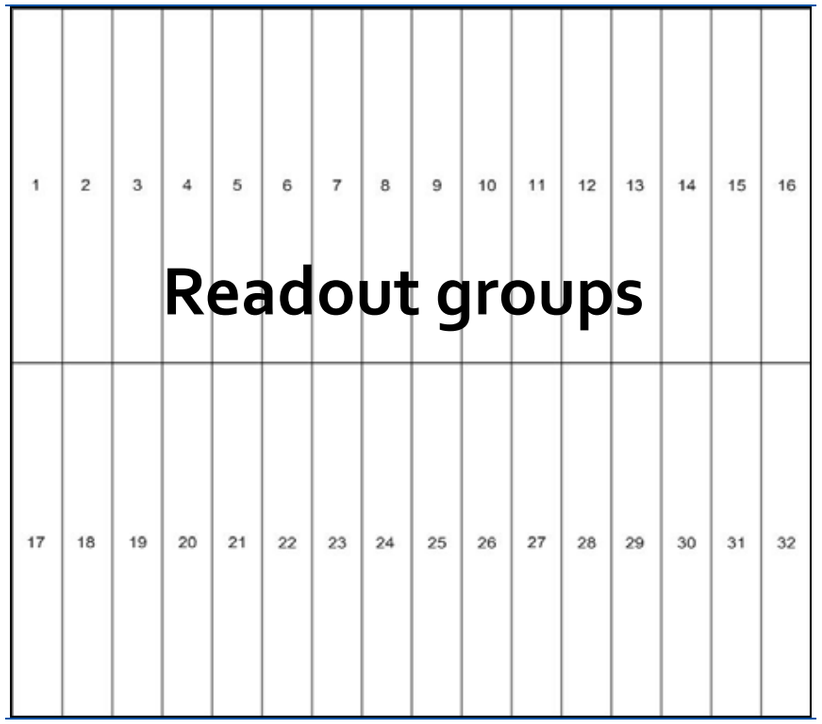
- Occur on detectors of LCD screens, digital cameras, CT scanners…
- Quantify damage
- Describe characteristics
- Reasons for damage
- Speed of decay

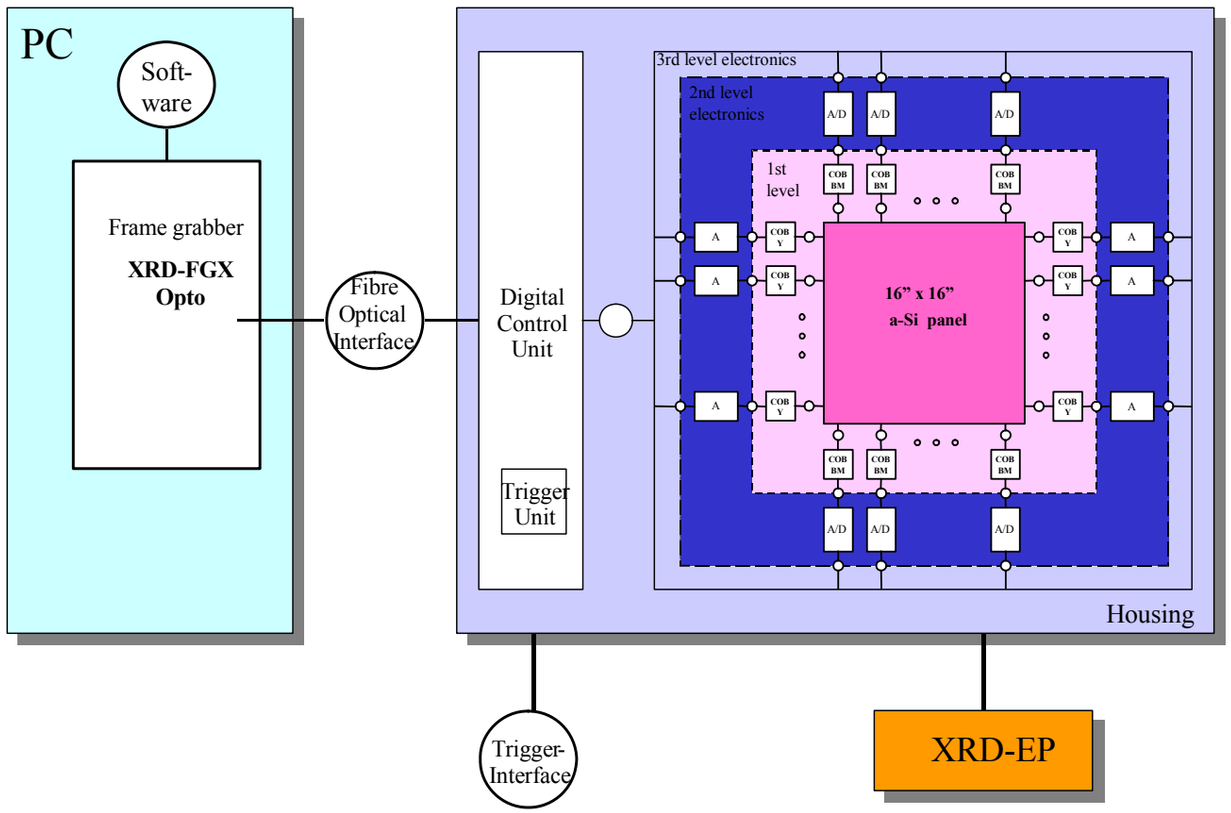


X-ray detectors and bad pixel maps



Perkin Elmer
XRD 1621

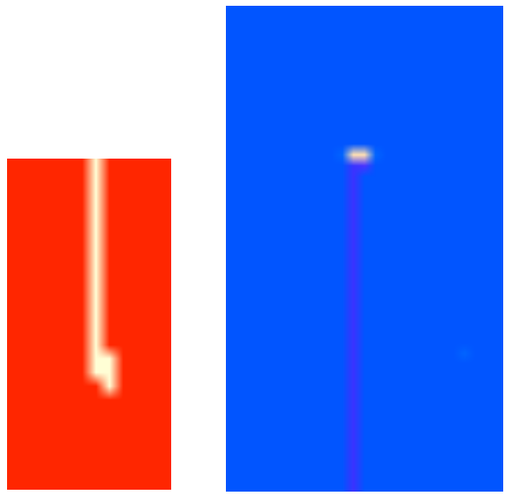


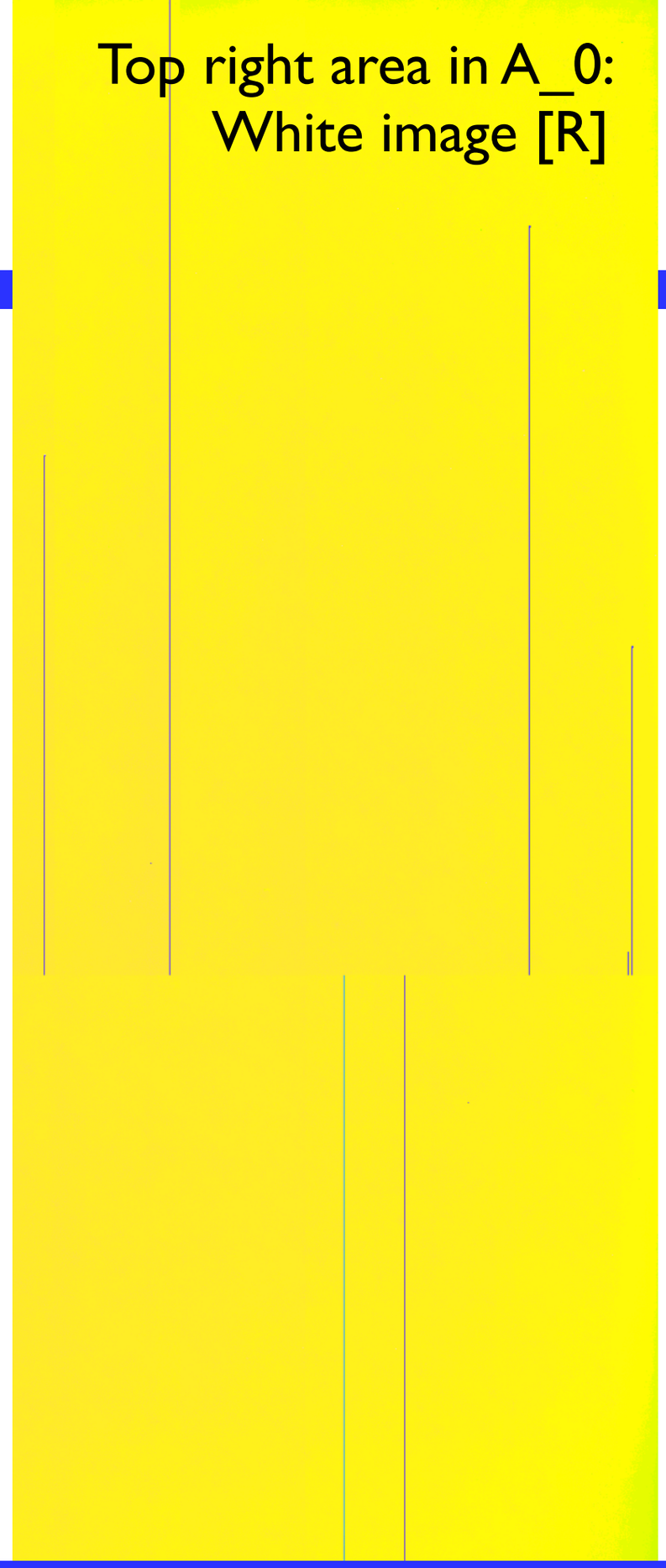


- "Underperforming" (sensitivity, noise, uniformity)
- Bad pixel map with coordinates

Local defects: Dead lines

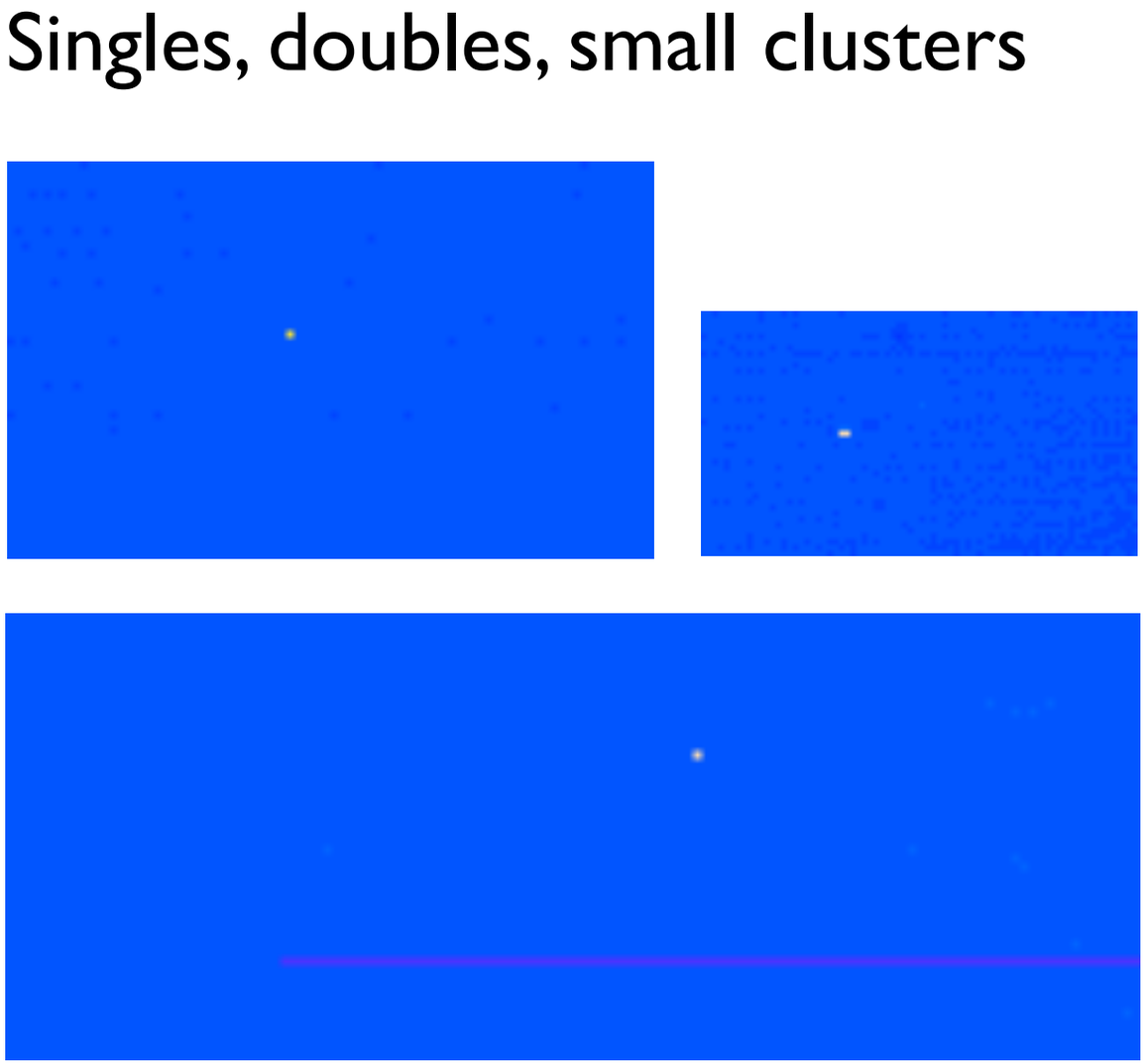
- Lines on bad pixel images
- From centre horizontal line outwards
- Clusters at the end



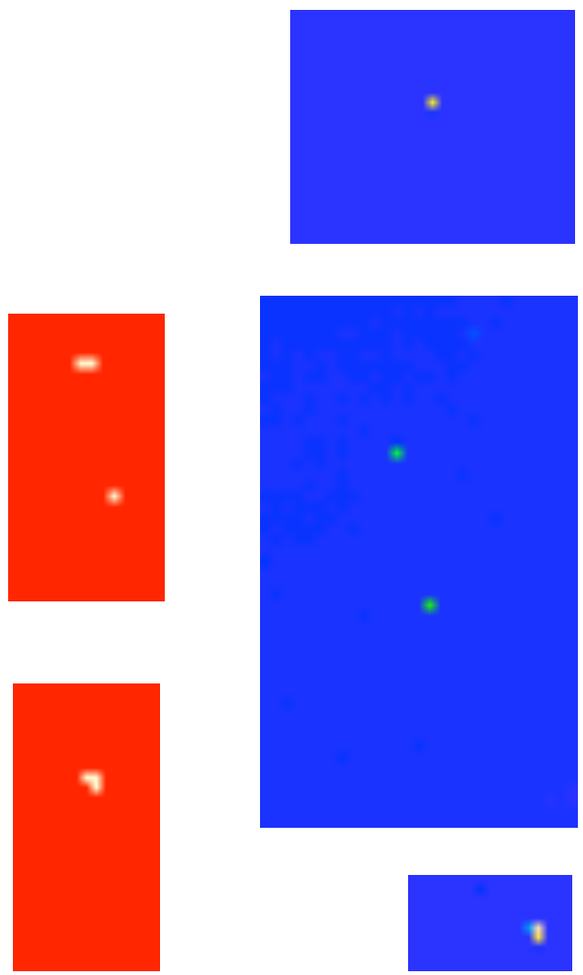


Local defects: Isolated dead pixels

Singles, doubles, small clusters



A_0: Grey image [R]

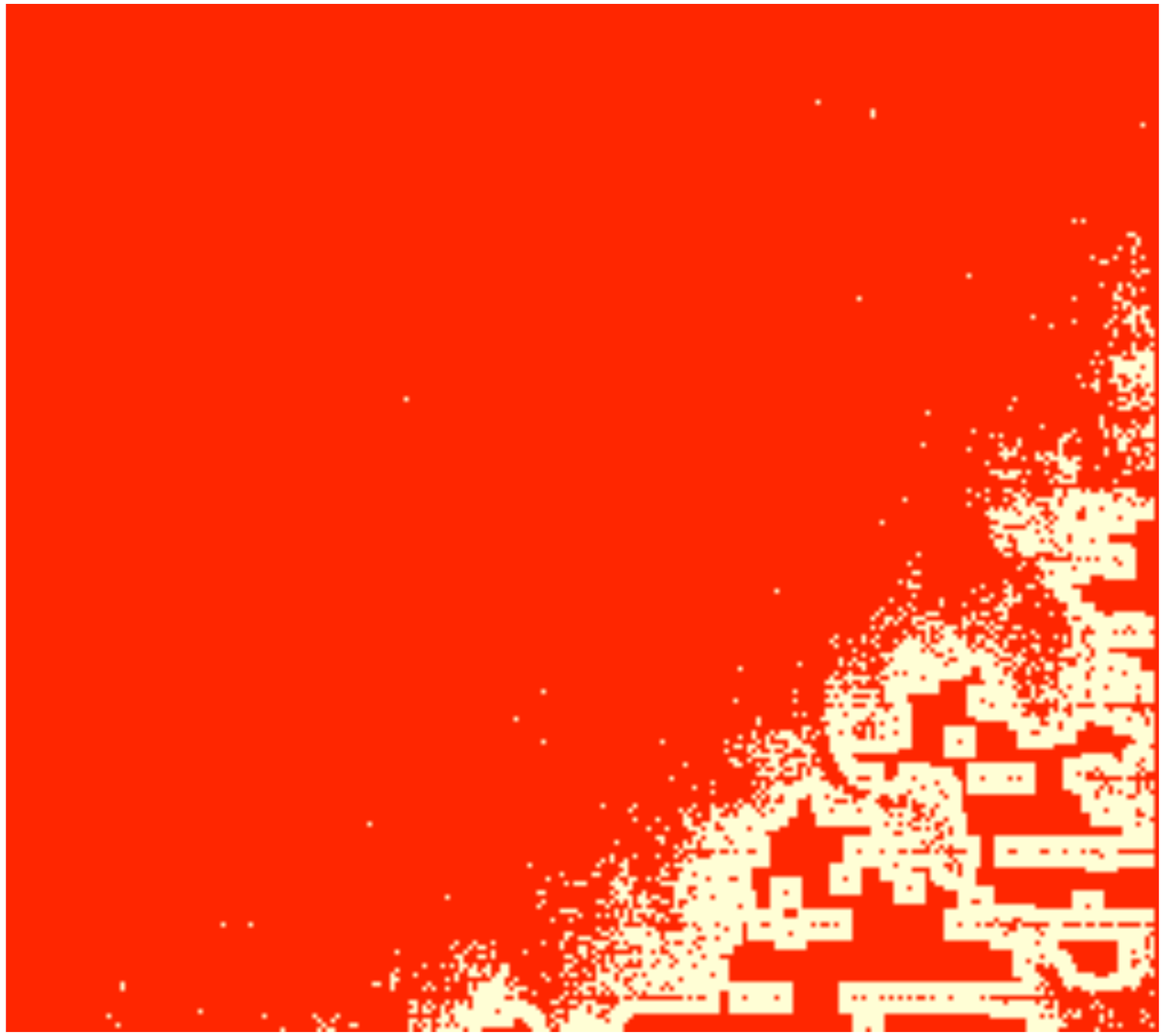


A_0:
bp binary
image [R]

A_0: Black
image [R]

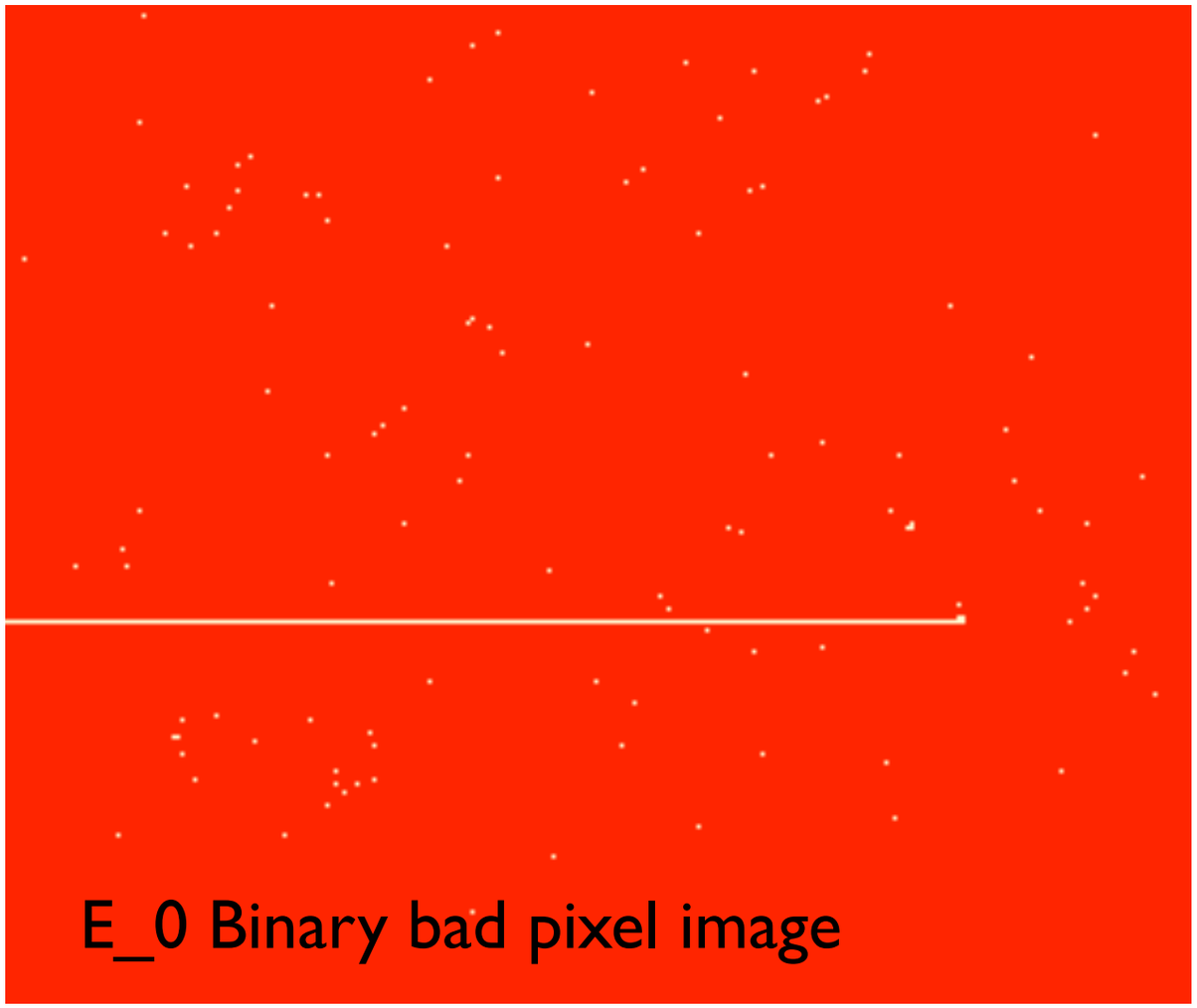
Local defects: Corners

B_0: Binary bad pixel image [R]

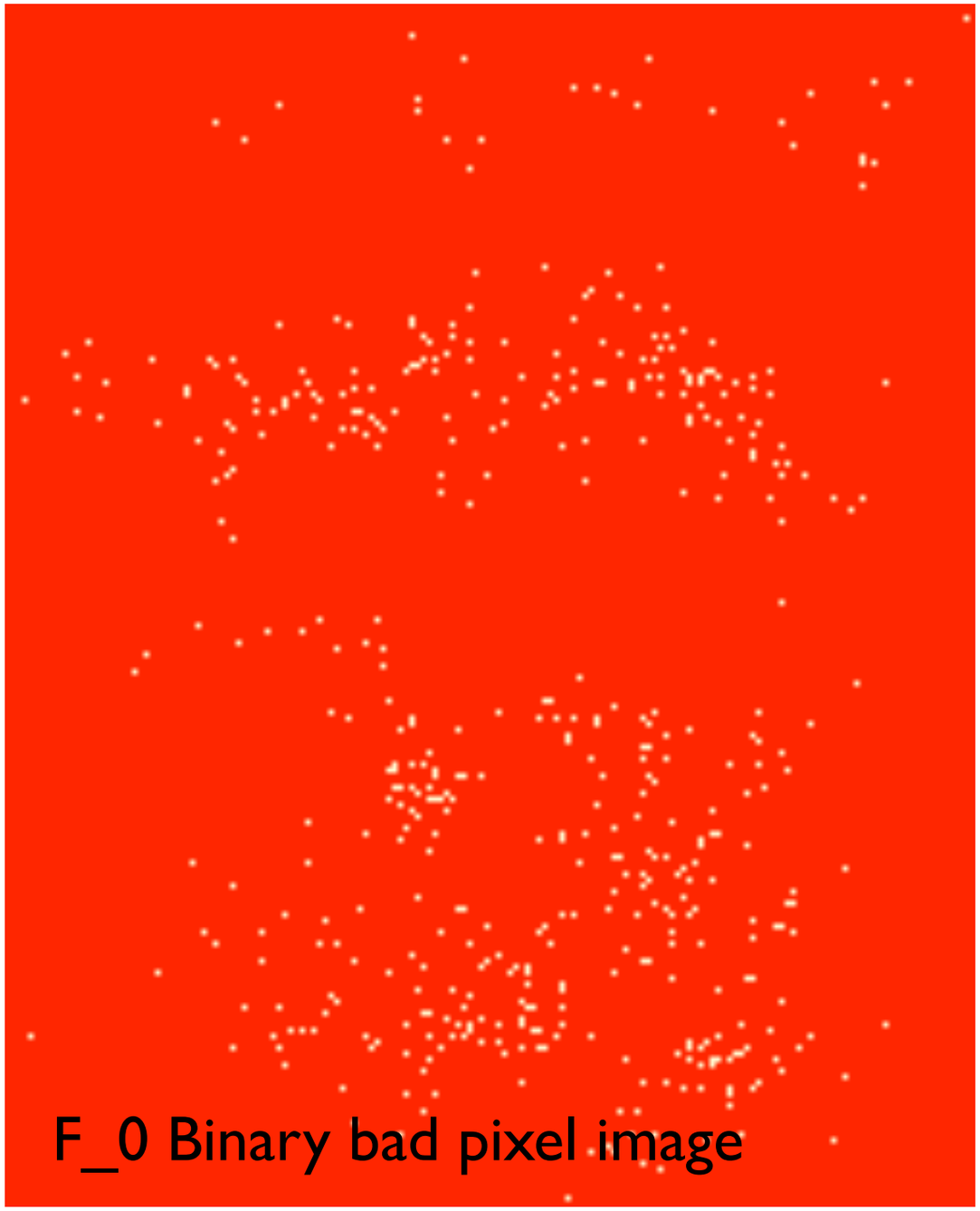


Local defects: Patches

- Areas with high density area of bad pixels



E_0 Binary bad pixel image



F_0 Binary bad pixel image

Point processes and spatial statistics

Mathematical model:
Interpret dead pixels a spatial point process

What is its distribution?
E.g. are there clusters?

Point processes and spatial statistics

Mathematical model:
Interpret dead pixels a spatial point process

What is its distribution? Clusters? Repulsion?

K-function: for h>0, K(h) is the **expected number of extra points in circle of radius h, rescaled by density**

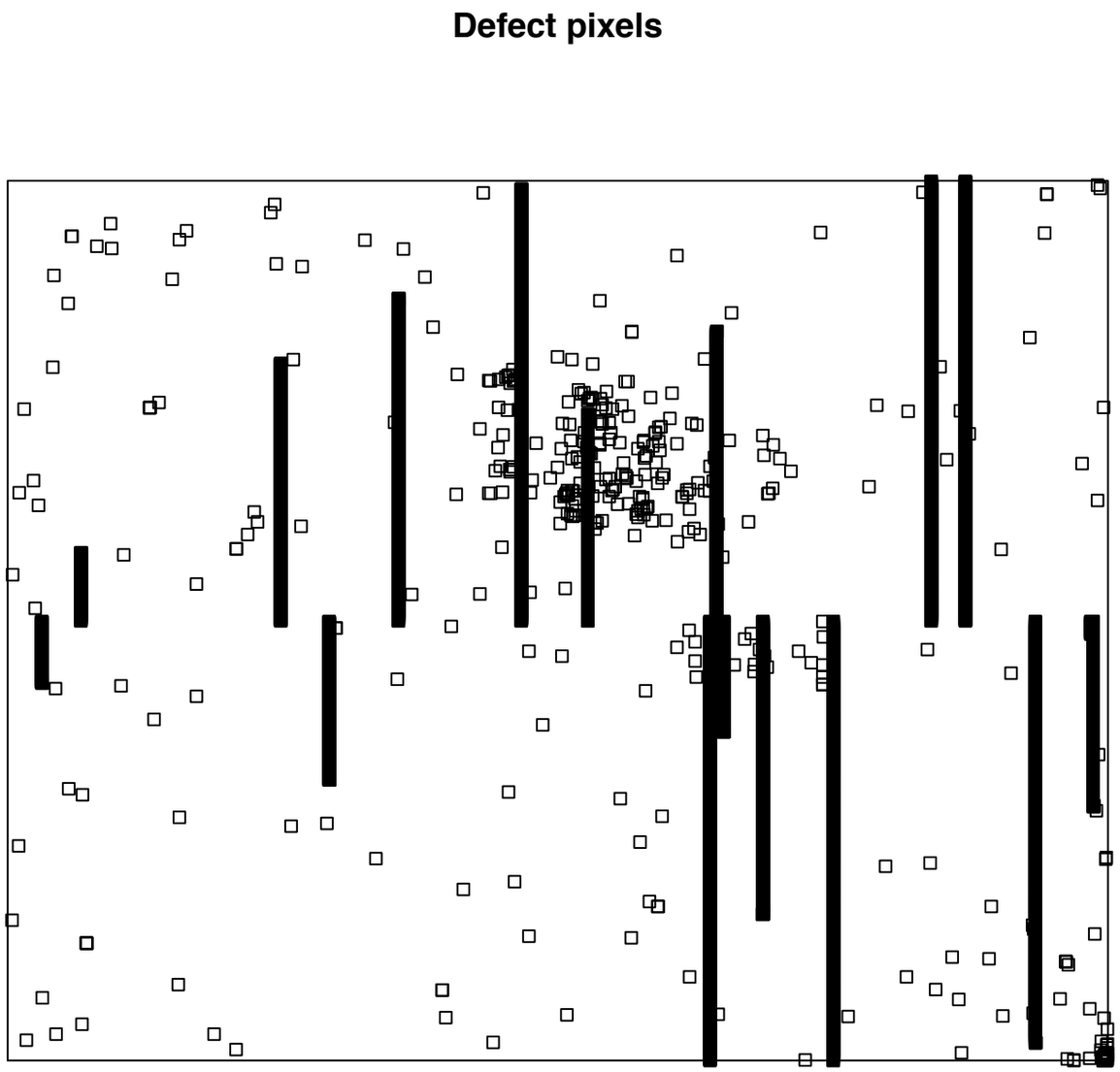
$K(h) = \frac{1}{\lambda} E[N(C_h - \{s\}) \mid N(s) = 1]$ (s location with point)

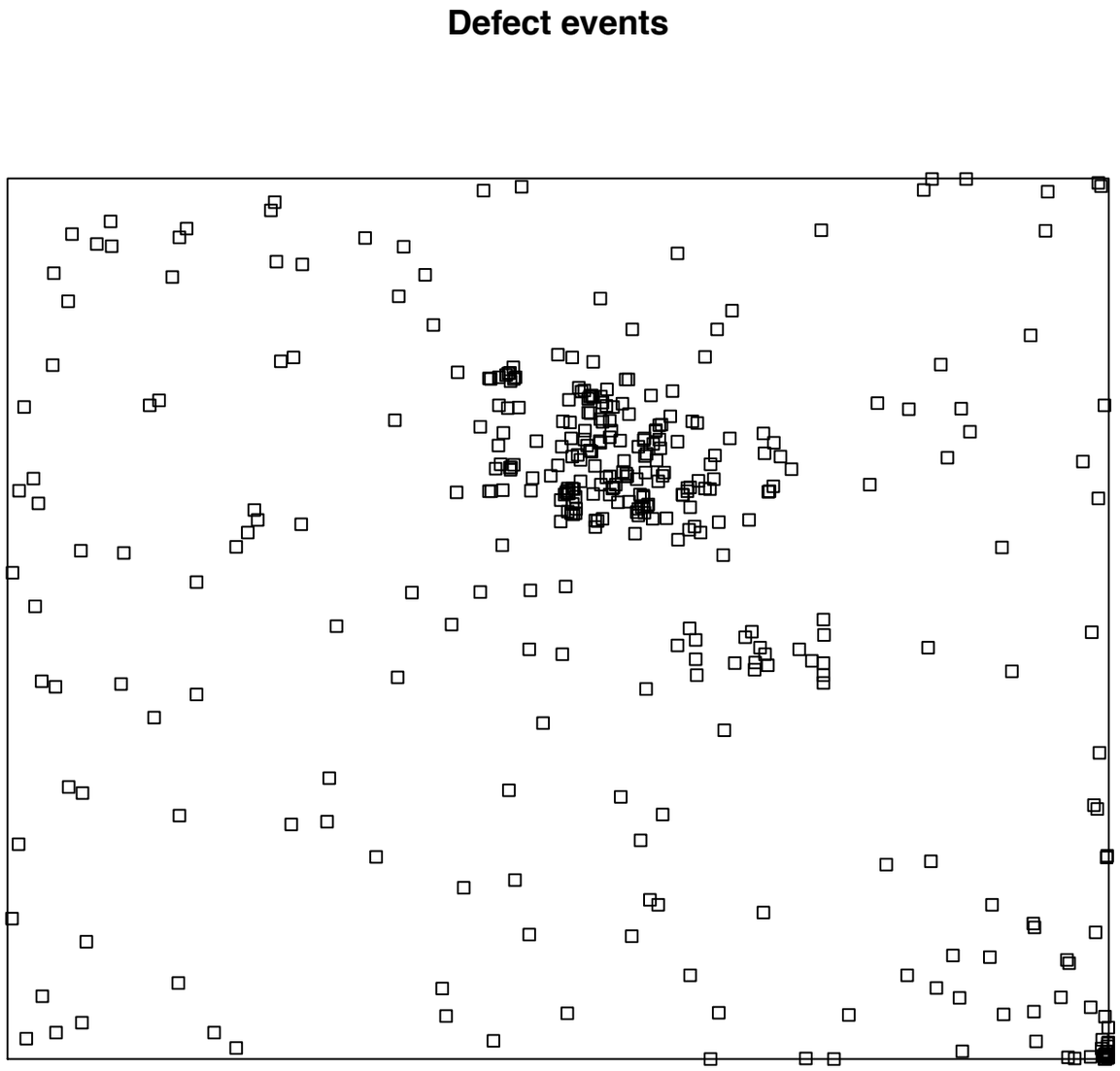
For stationary processes: Proportional to its area $K(h) = \lambda a(C_h)$

Under CSR: $K(h) = \frac{1}{\lambda}(\lambda \pi h^2) = \pi h^2$

Higher level defect model (Step 1)

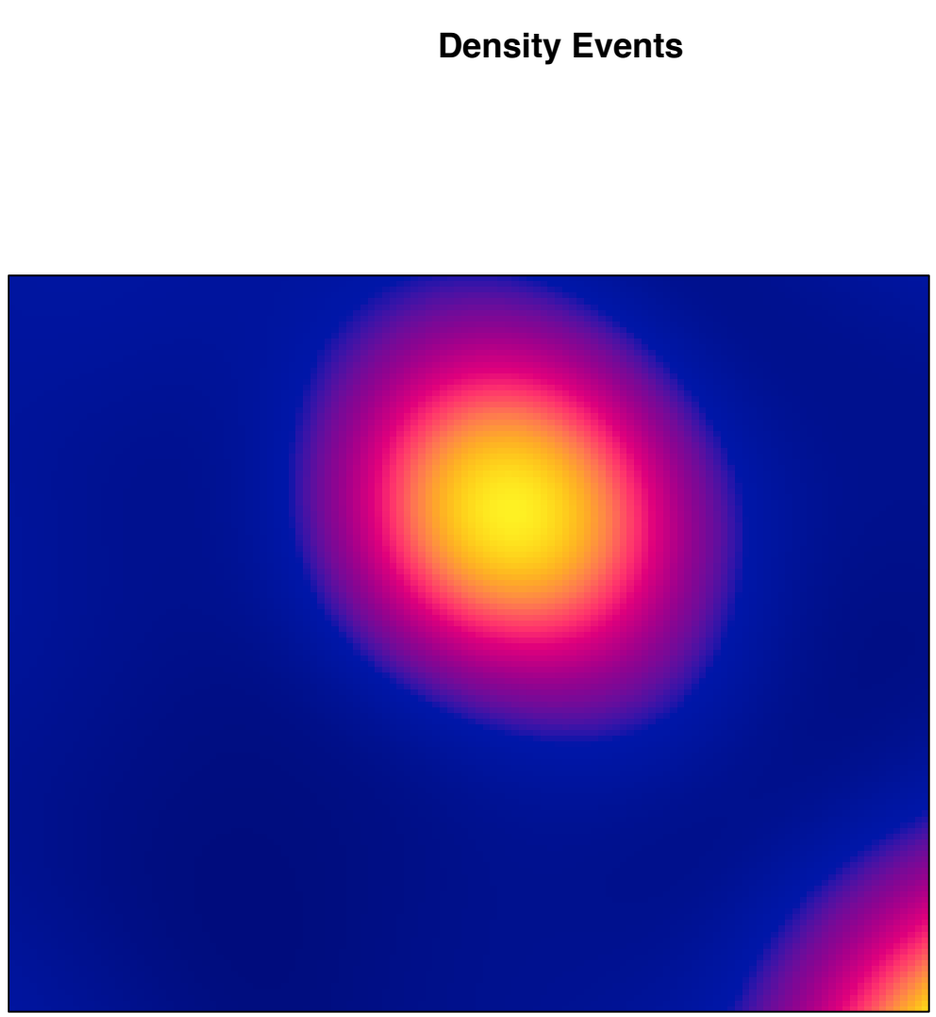
Conversion of point process to *event process*

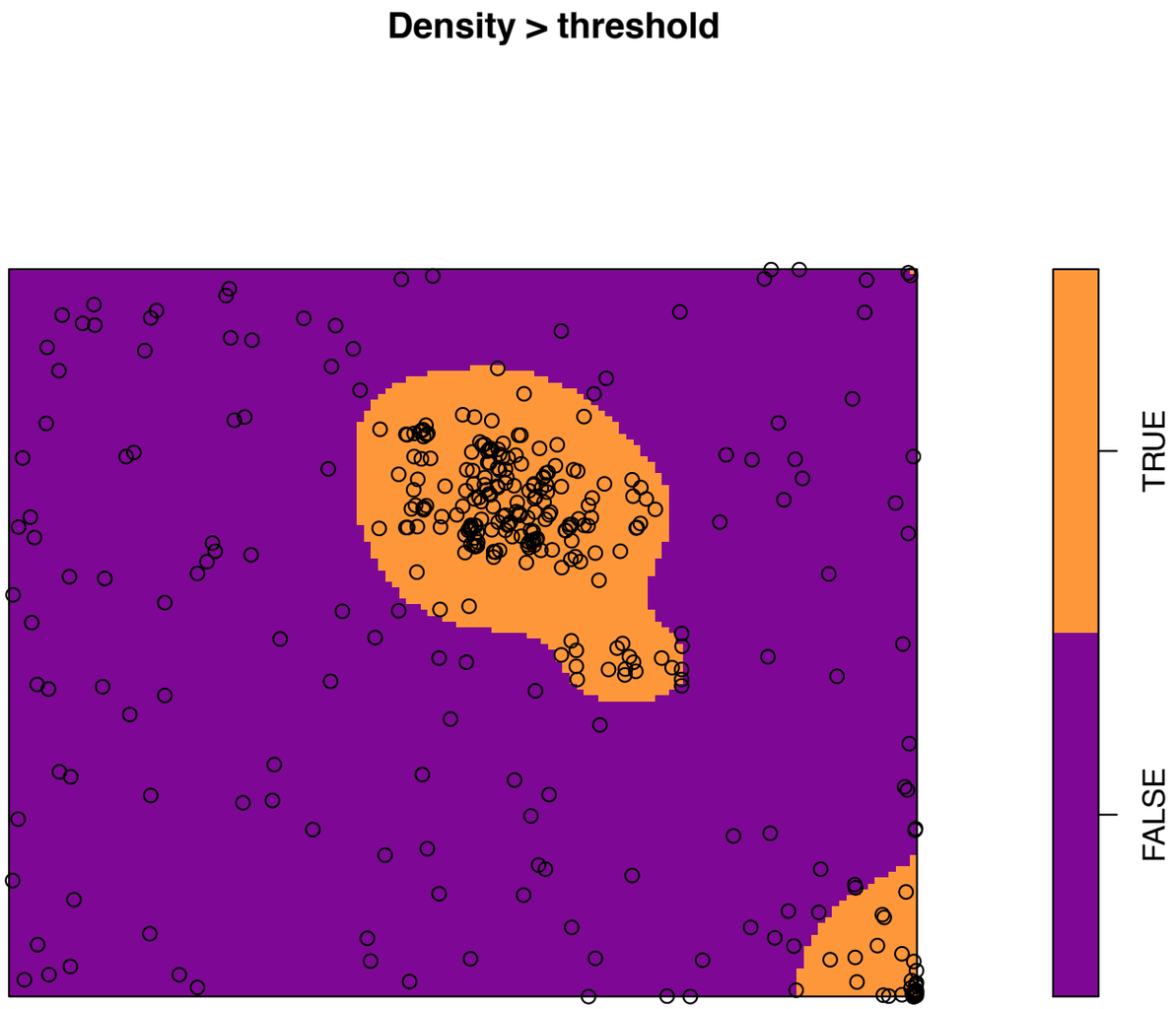




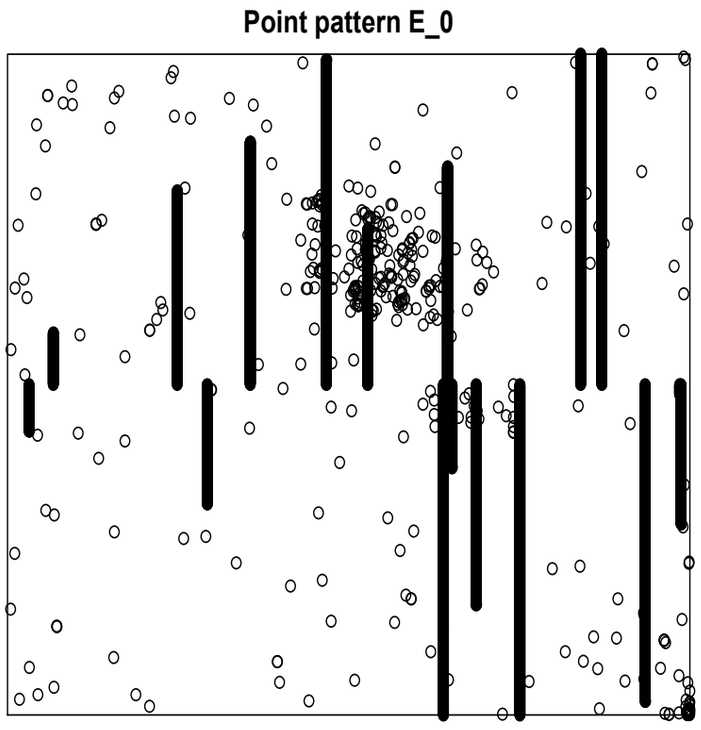
Density based thresholding (Step 2)

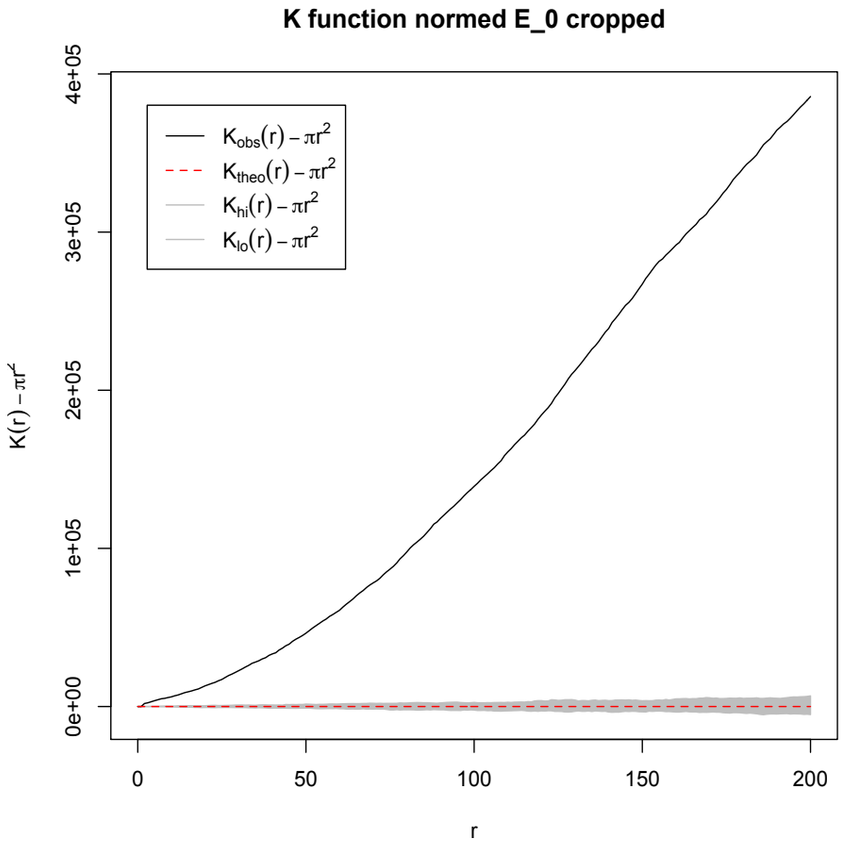
Remove areas with local density above threshold (medial +1.5 IQR)

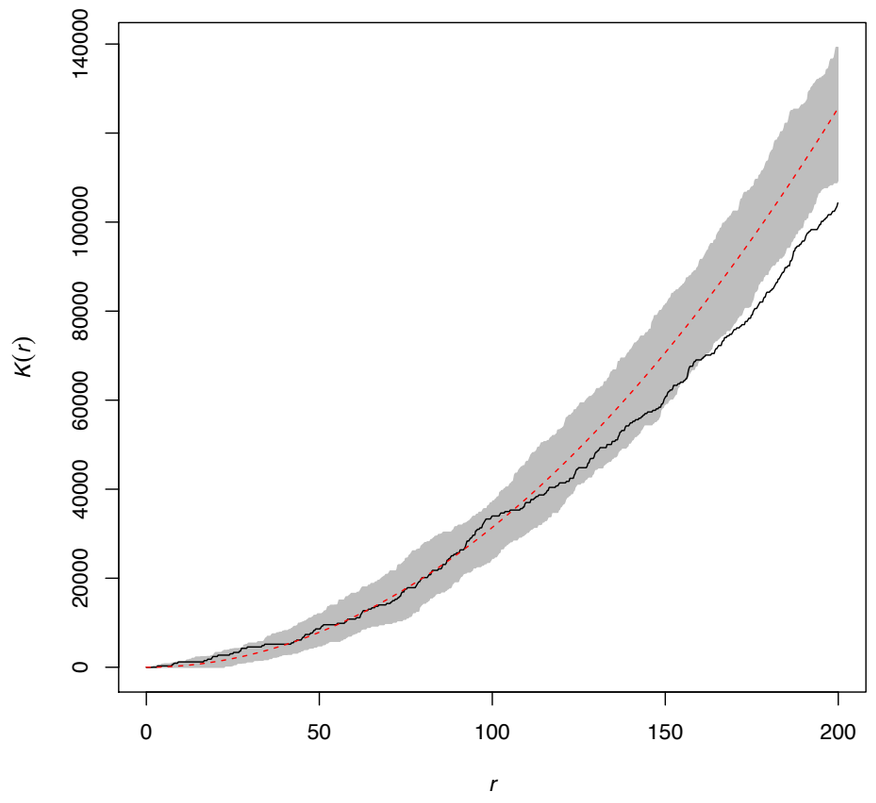




Point pattern and K-function modified







Dead pixels

Pixel process K function

Event process K function

Completely spatially at random

Spatial statistics for detector QA

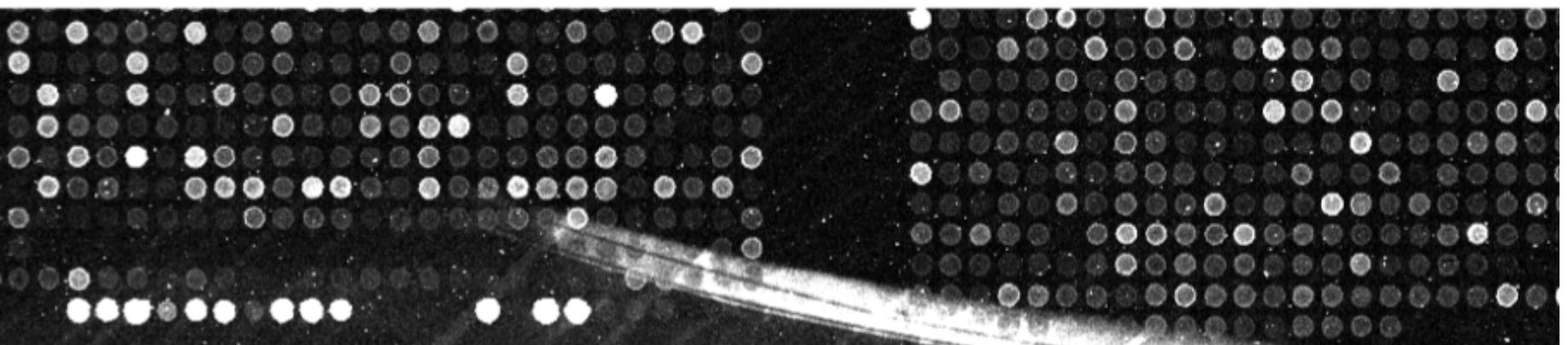
- Identify special causes of poor quality
- Remaining area CSR means general cause of poor quality
- Density in remaining area gives global quality score for the detector

Application of statistical genomics:

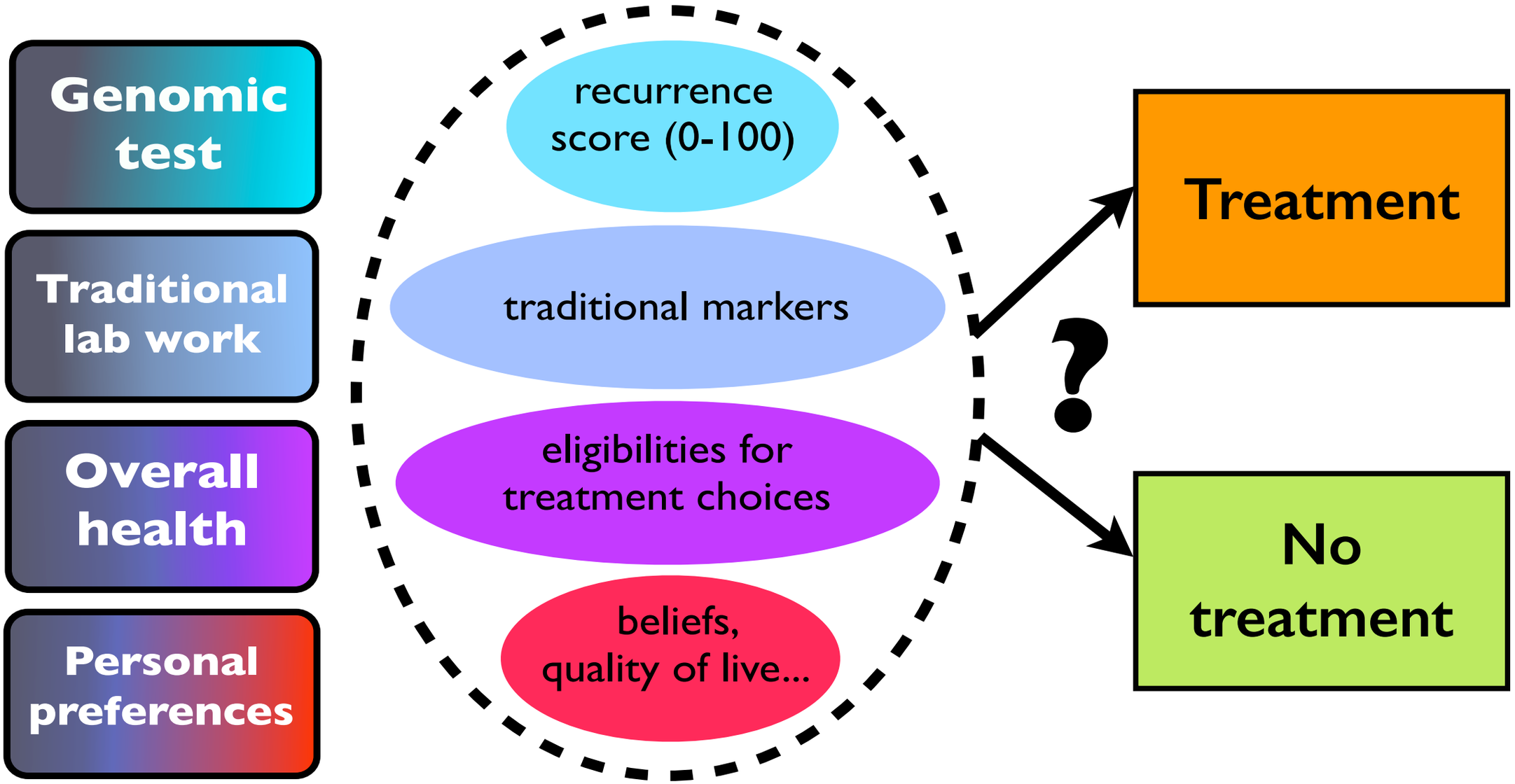
Data quality for cancer prognosis and treatment decision

Collaboration with Terry Speed's group in University of Berkeley, California

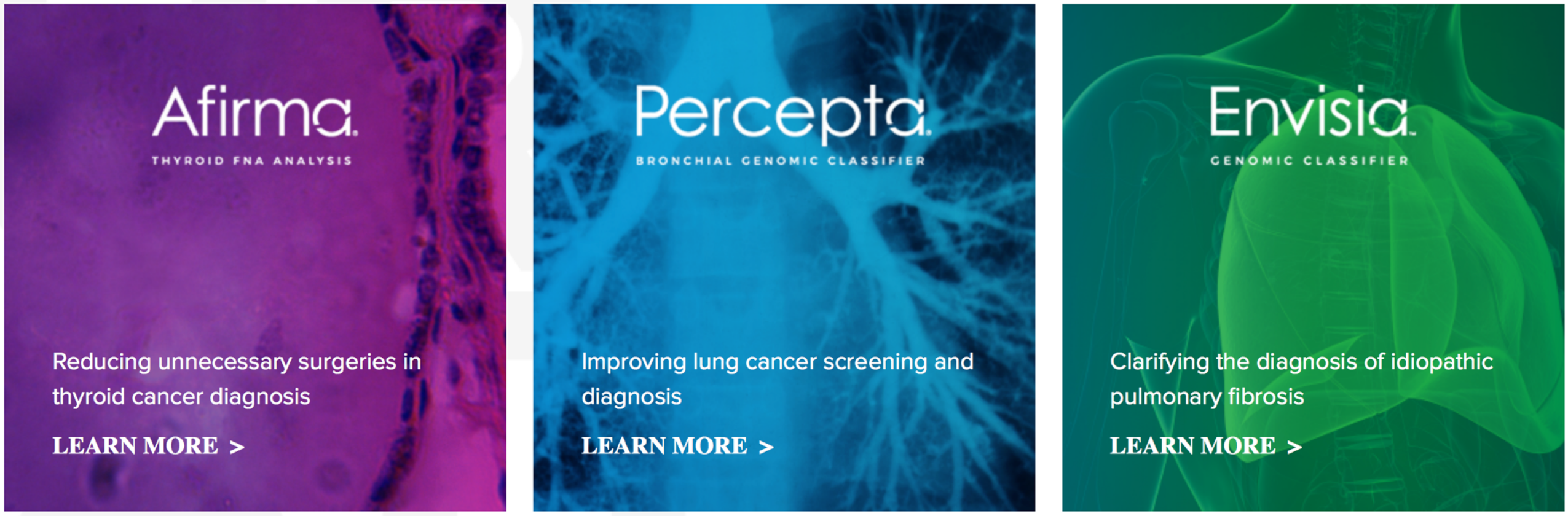
Results used by Bay Area Biotech companies



Decisions about invasive medical treatments

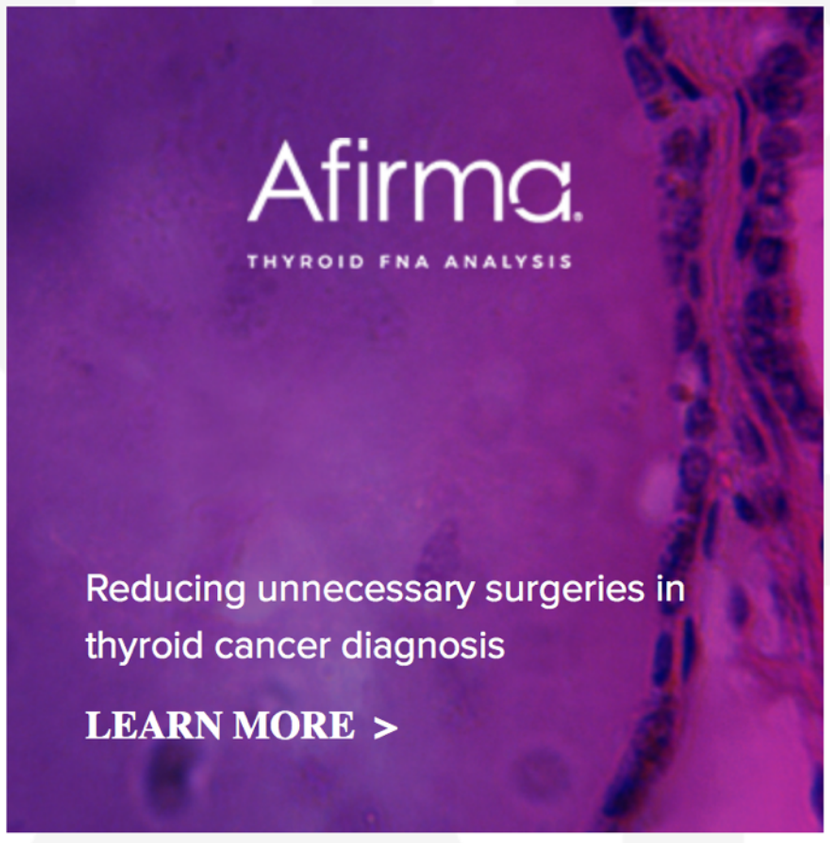


- Uncertainty, complex information (clinical tests e.g. OncotypeDX)
- Emotions interfering with judgement, multiple decision makers



Afirma (Veracyte) in practice:

- Traditional diagnosis in thyroid cancer 30% inconclusive and lead to surgery (plus life long treatment), but 80% turn out to be benign tutors
- Afirma avoid half of these surgeries (plus morbidity)
- Potential economics impact for US: $122 million savings

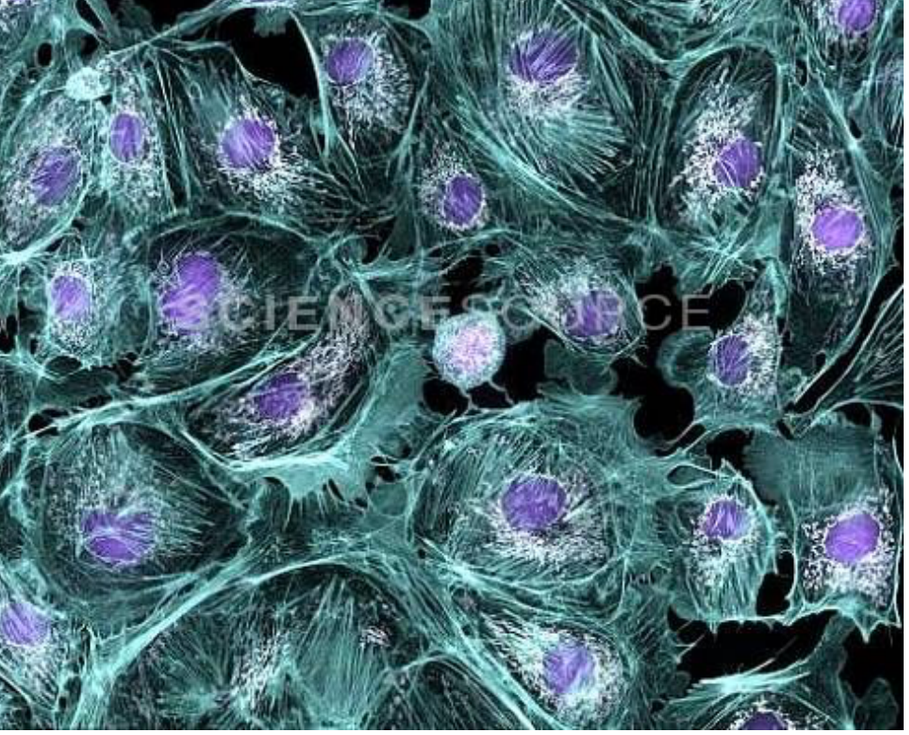


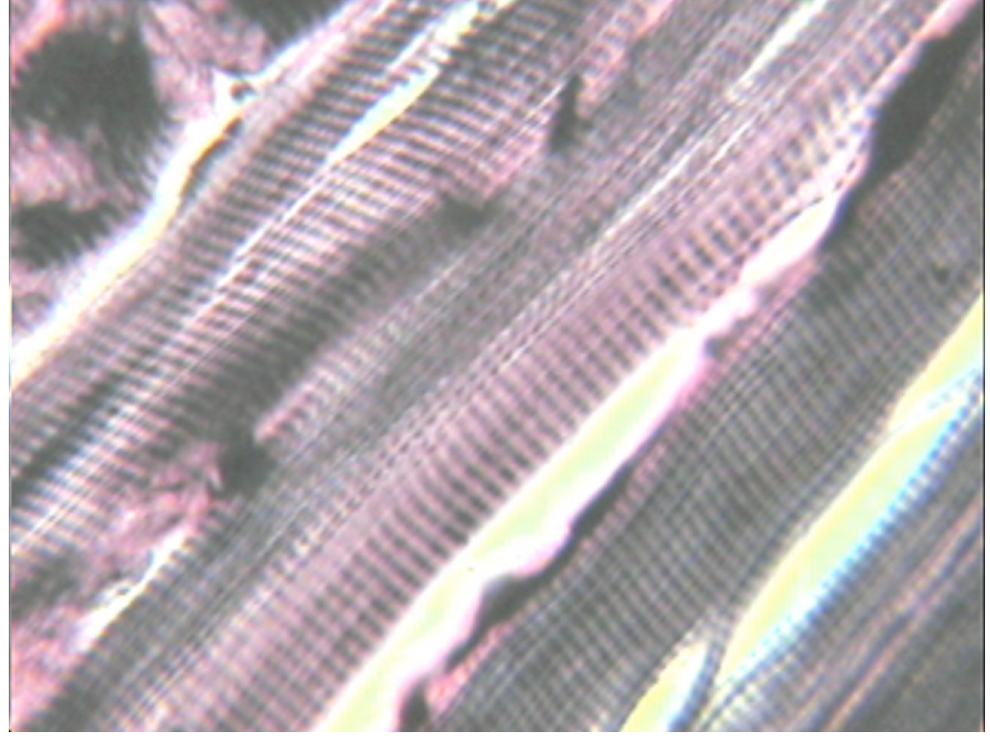
Statistical requirements:

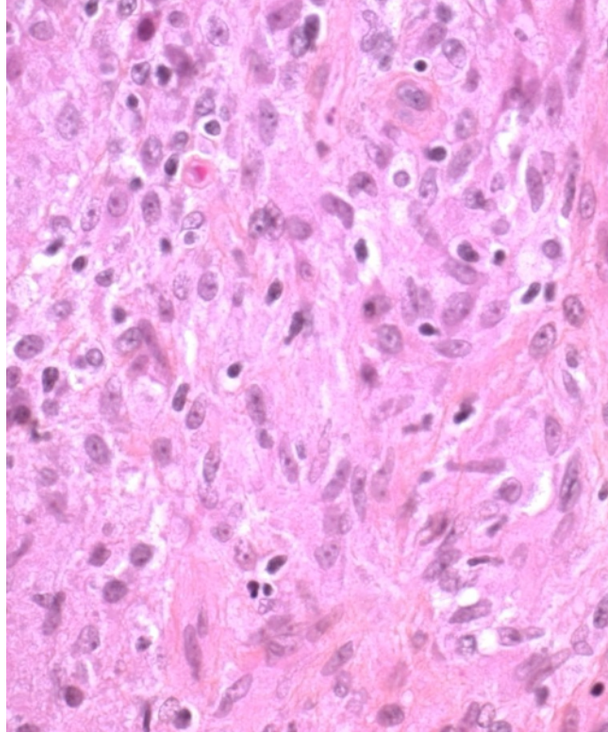
- Crucial step for commercial success: control error rates of test
- Essential: Data quality assessment for custom made measurement instruments

Afirma (Veracyte) in practice:

- Traditional diagnosis in thyroid cancer 30% inconclusive and lead to surgery (plus life long treatment), but 80% turn out to be benign tutors
- Afirma avoid half of these surgeries (plus morbidity)
- Potential economics impact for US: $122 million savings

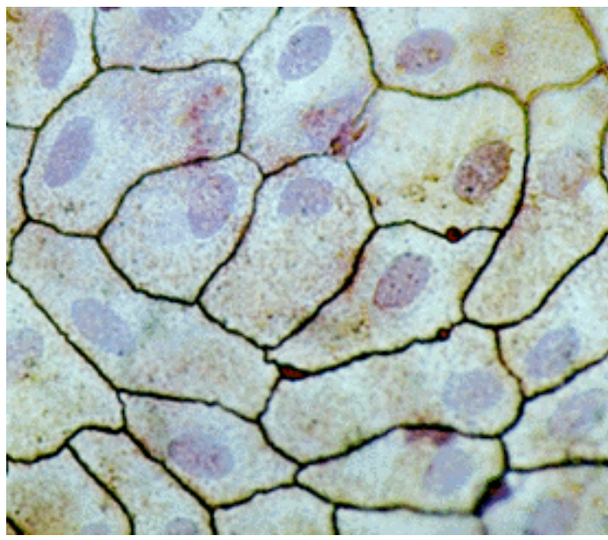


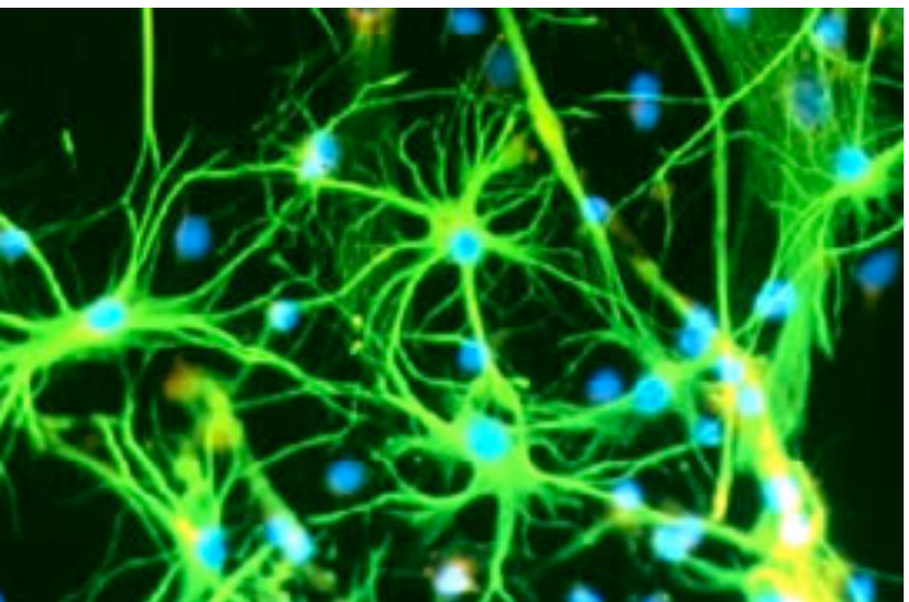


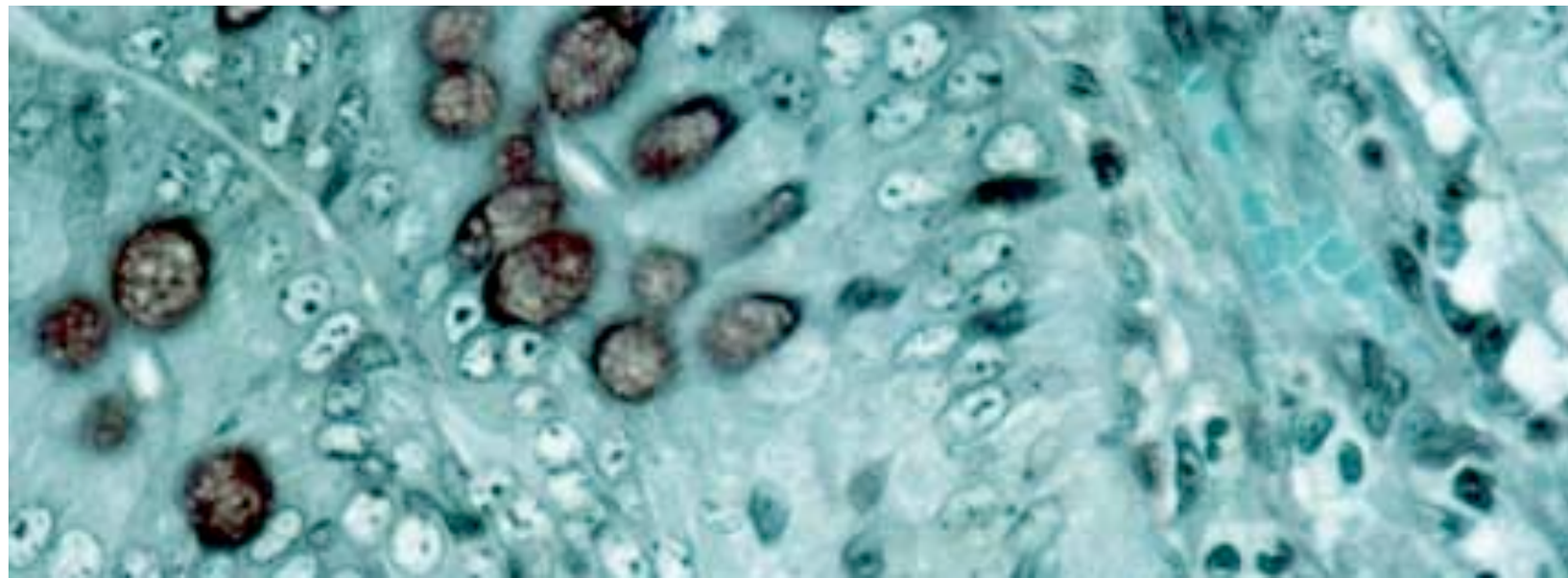


Genomics 101:

From DNA to cells





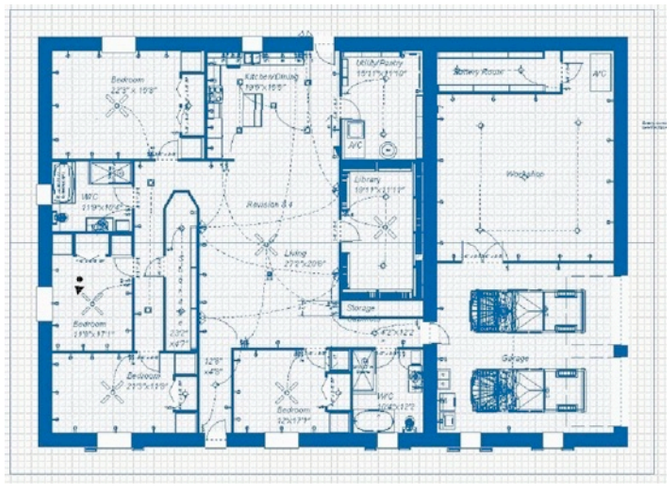


Metaphor: Architecture

Textual description

…, stone walls, roof, divided into room, glass windows, wooden frames, hardwood doors,…

Plan



Product

Casa Loma, Toronto

Design team

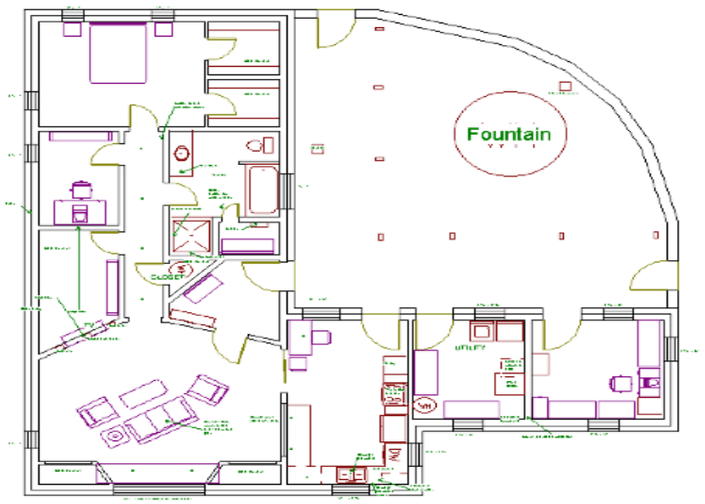
Construction

Process and product are not deterministic

Textual description (same)

…, stone walls, roof, divided into room, glass windows, wooden frames, hardwood doors,…

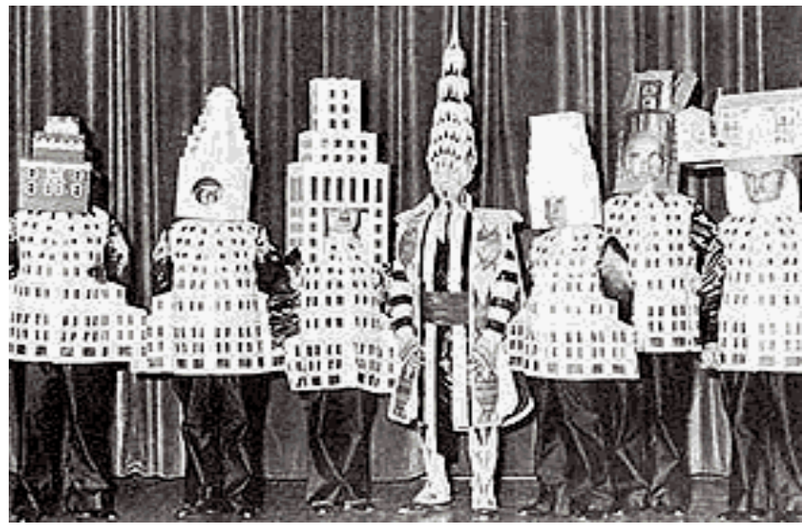
Plan (different)



Product (different)

Hackesche Höfe Berlin (built1906/07)

Design team

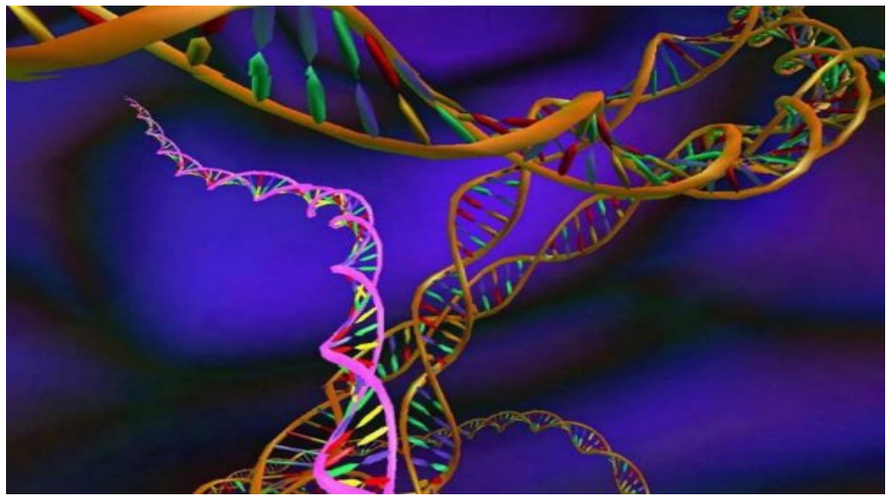


Construction

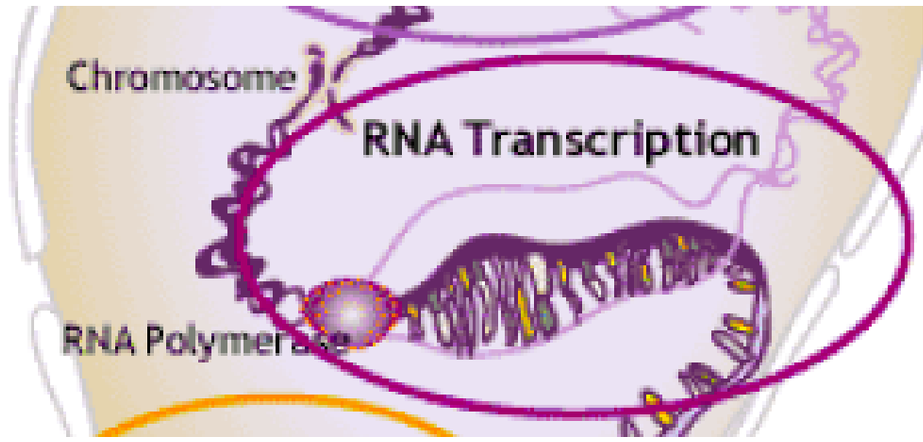
Central Dogma of Molecular Biology

Textual description

DNA

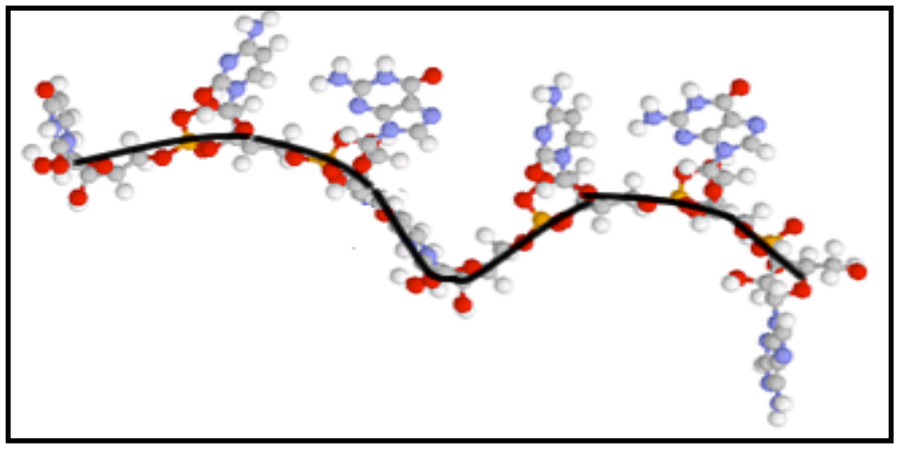


Design

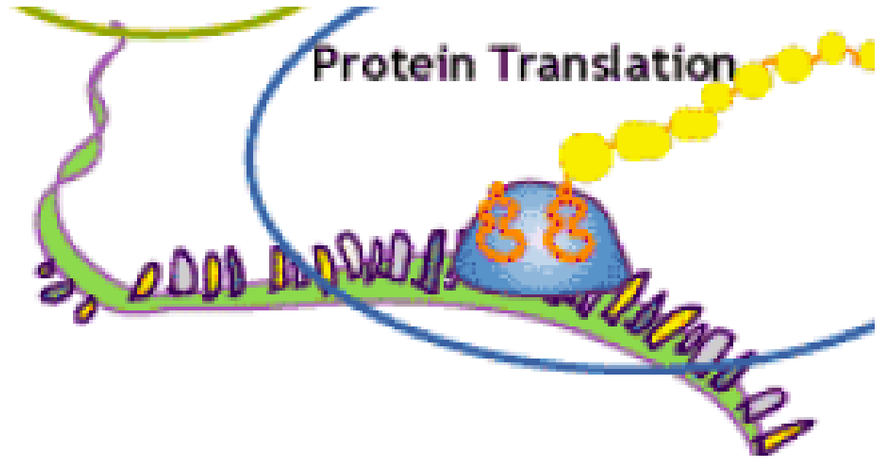


Plan

RNA

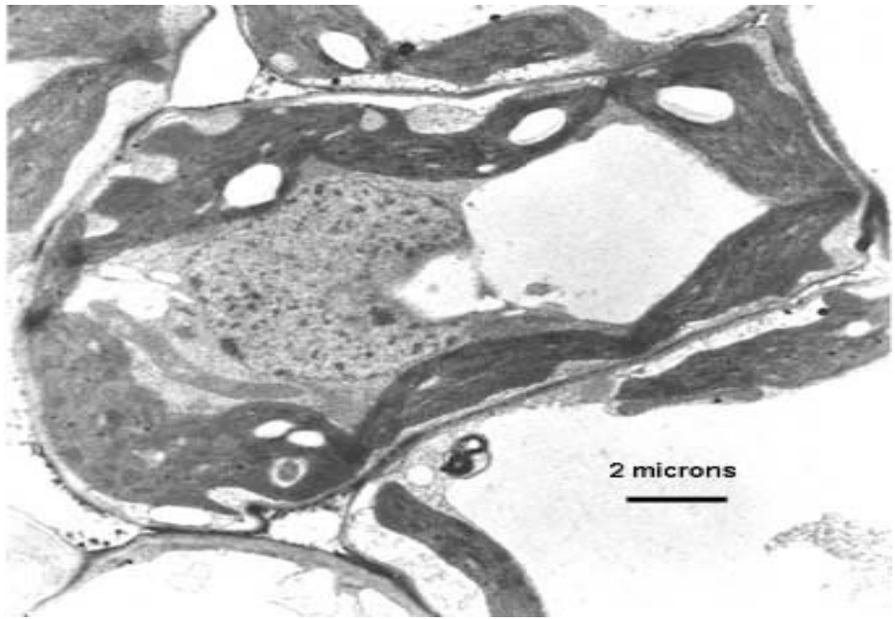


Construction



Product

Cell



Gene expression

Gene expression =

the gene's degree of biochemical activity

(here: amount of RNA produced by the gene)

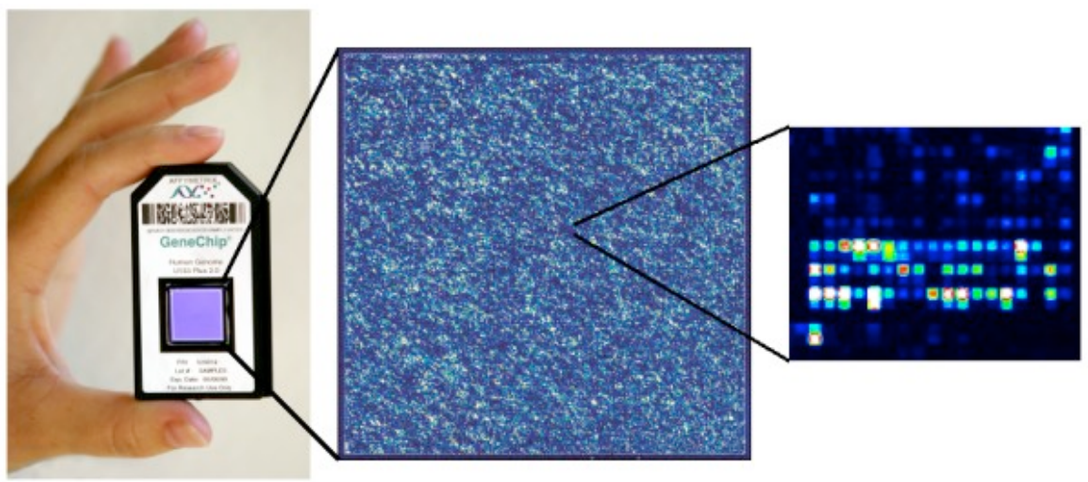
Depends on **factors** such as:

- Type of the cell
- State of cell
- Developmental stage

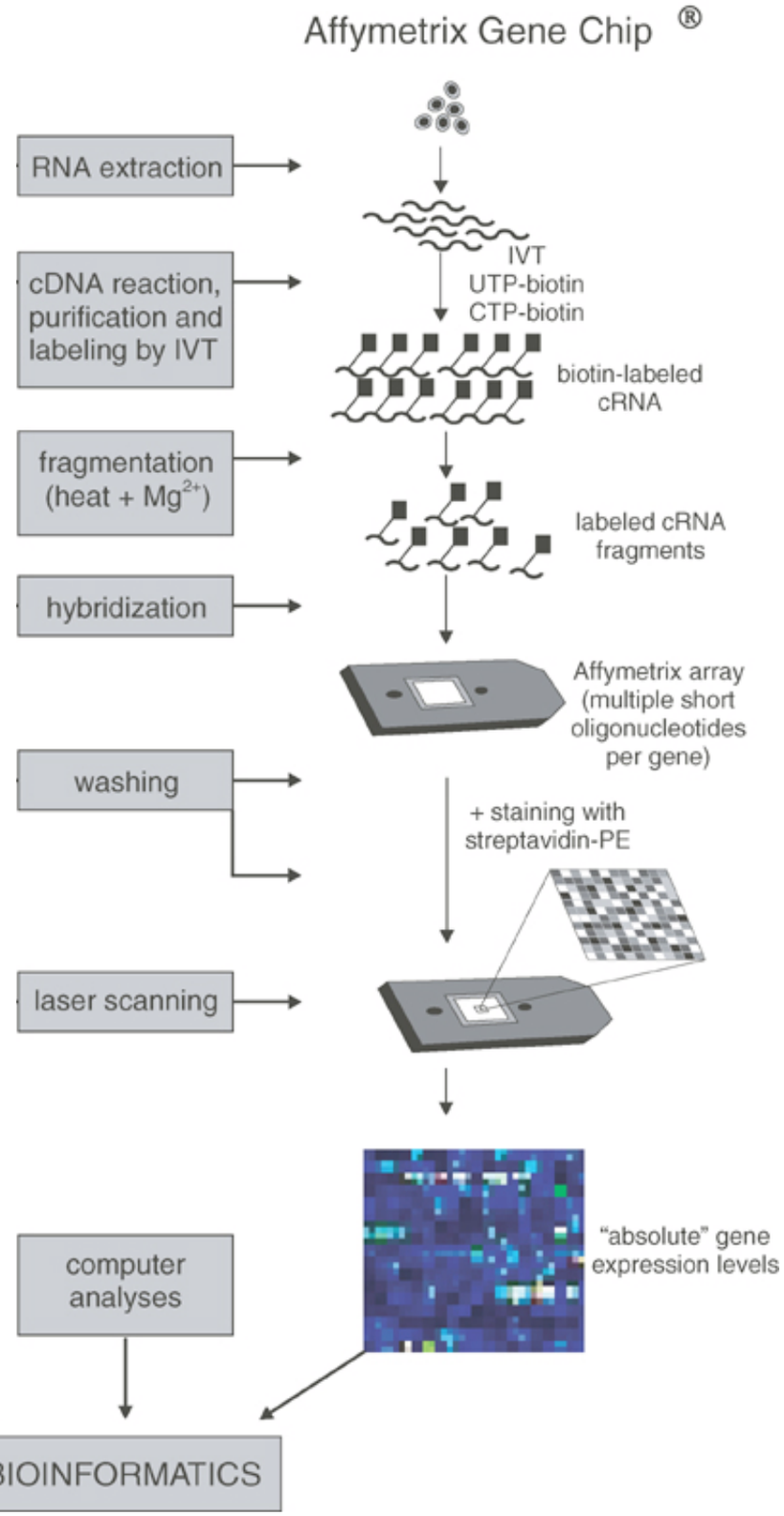
Use to gene expression to **detect genes** involved in cellular processes, diseases, development etc.

High throughput gene expression measurement with microarrays

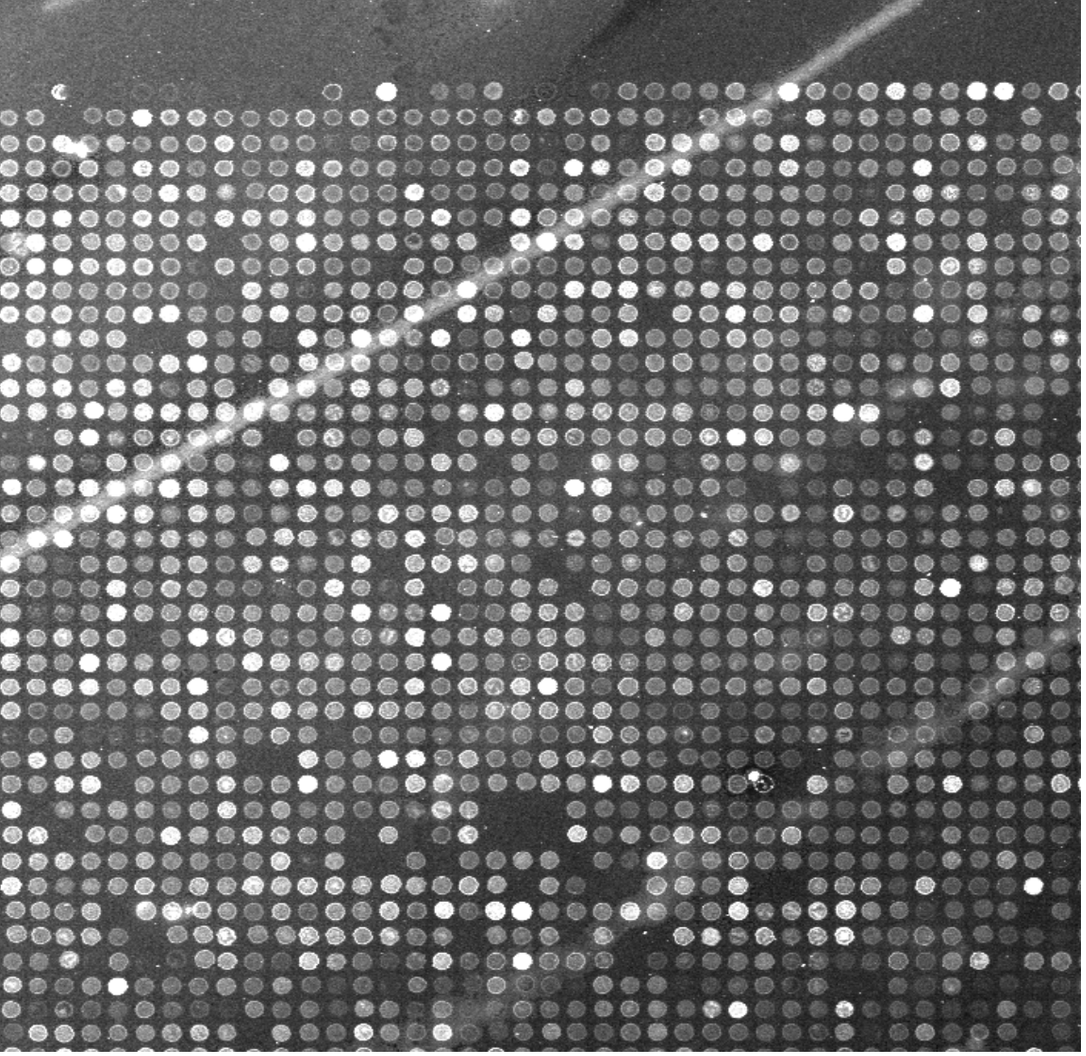
- Assesses expression levels of tens of thousands of genes
- Simultaneously in one experiment

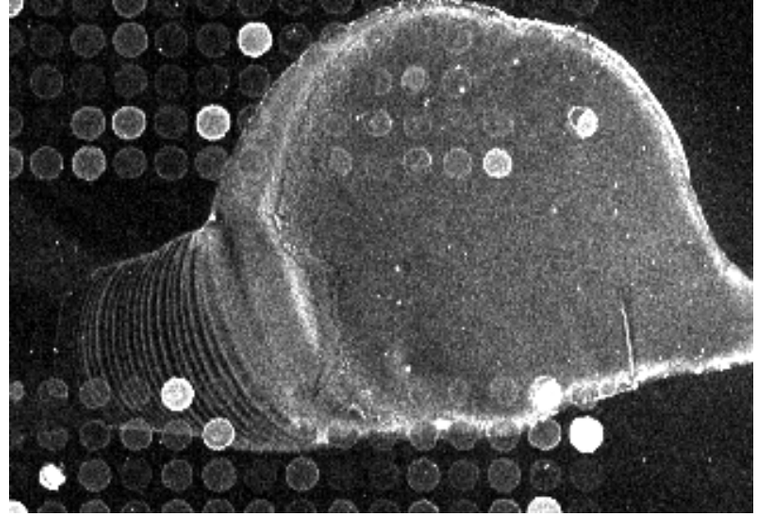


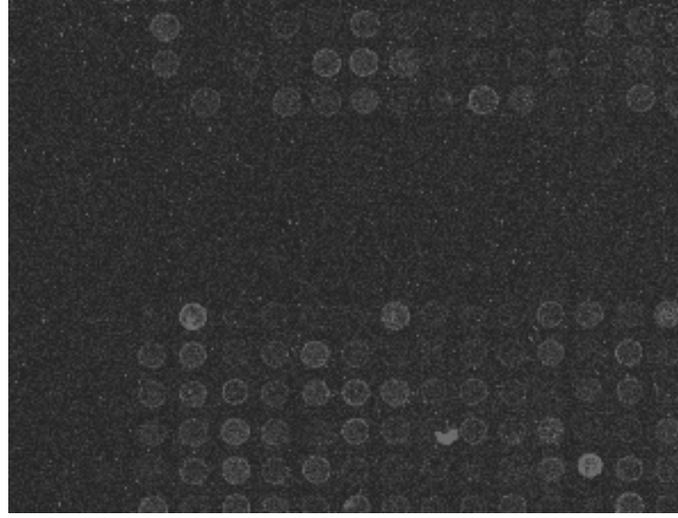
Workflow

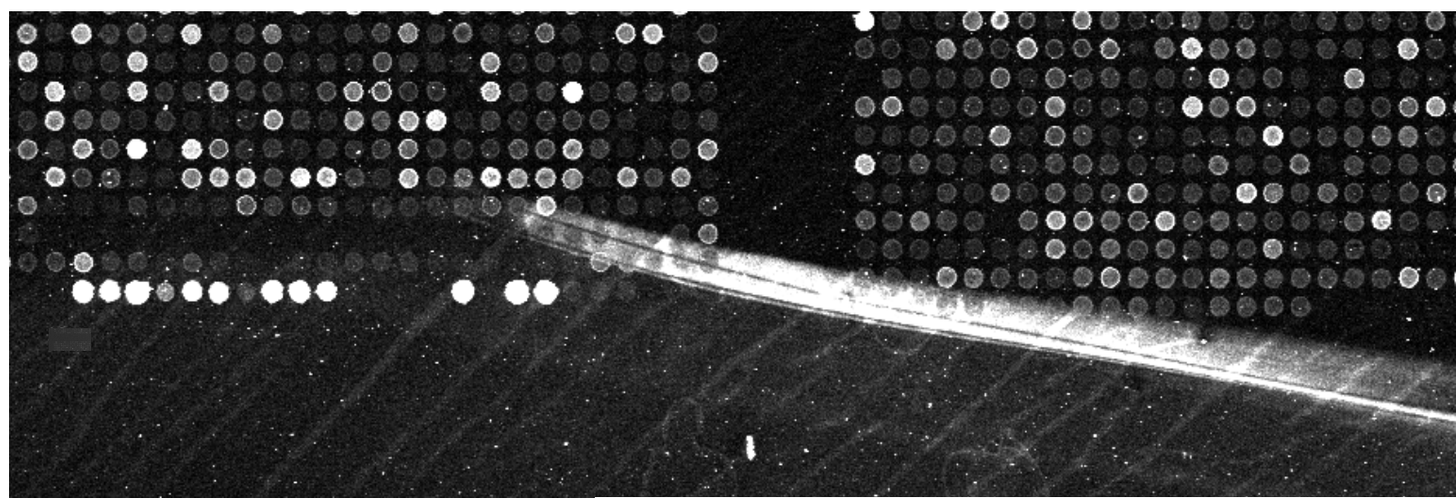


http://www.nature.com/leu/journal/v17/n7/images/2402974f1.jpg



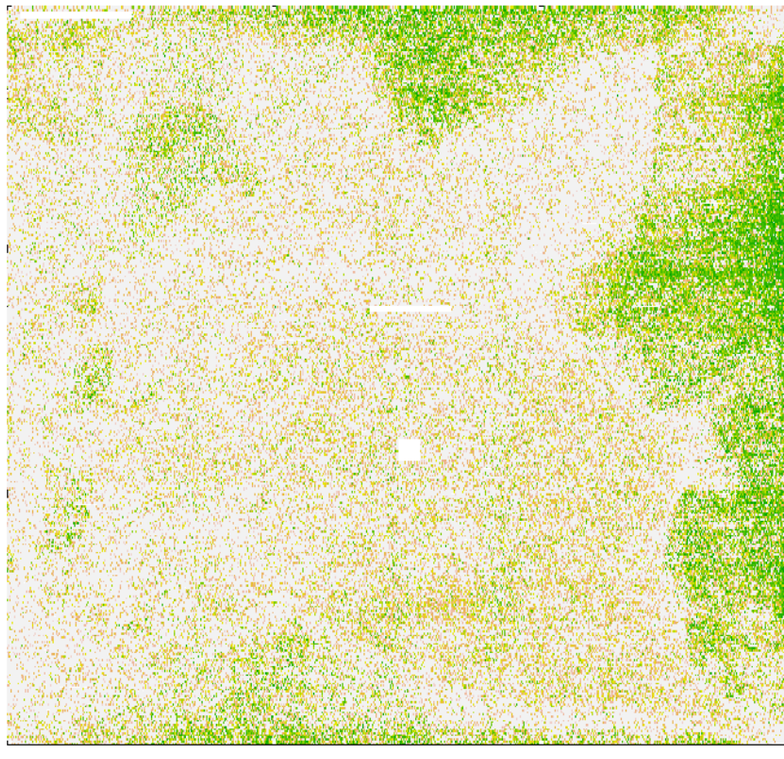






High throughput gene expression measurement quality assessment toolbox

16qgrega



High dim genomic data QA/QC challenges

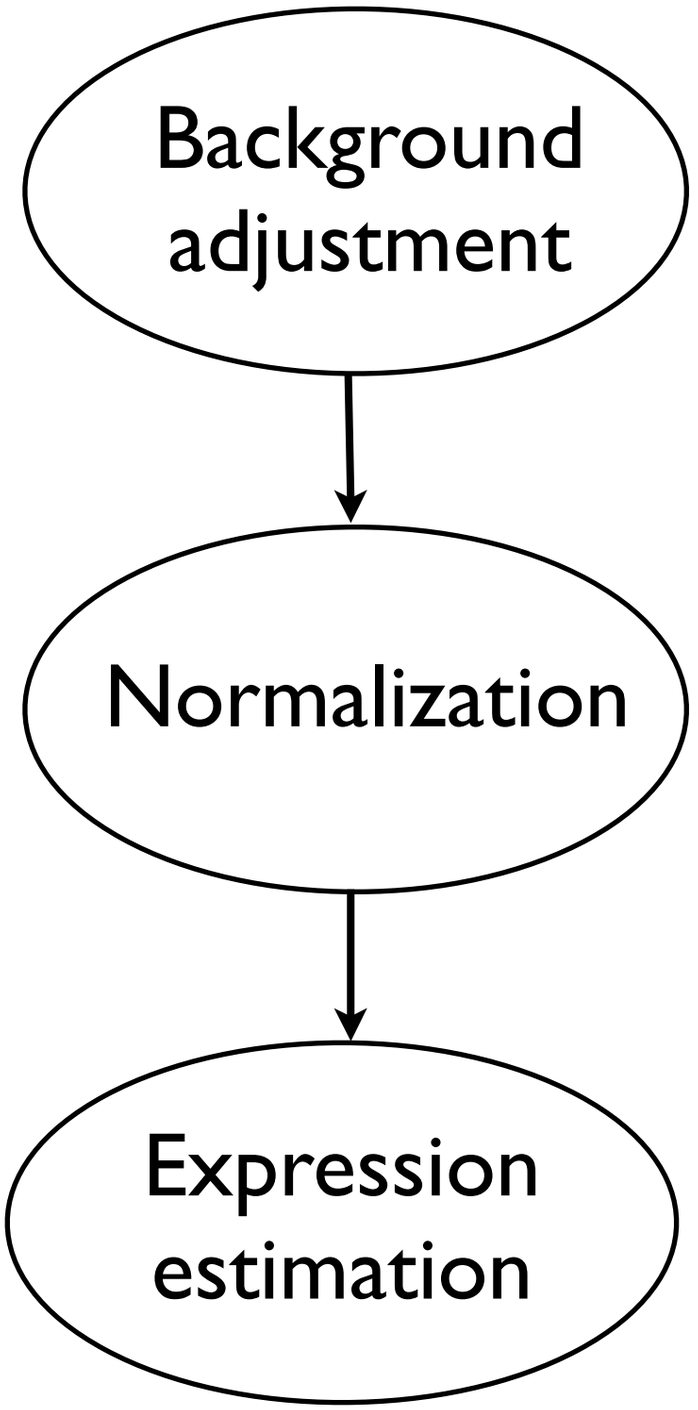
- Simultaneous measurements of huge numbers of genes
- Missing or partial 'gold-standards'
- Unknown correlation structure
- No agreement on models for microarray data
- Measurement taken in a multi-step procedure
- Divorcing technical variation and biological variation
- Systematic errors more relevant than random errors
- Platform specific
- Data collections (risk of being swamped with poor quality

	log probe intensities array 1	log probe intensities array 2	. . .
Gene 1	6.0097 7.8997 4.7292 6.0237 5.0233 5.5657 7.6687 7.3411 4.7232 5.9112 6.2232	5.2322 6.2234 5.3233 4.5443 2.8389 7.8223 8.2548 8.9967 7.6755 6.7445 6.7899	. . .
Gene 2	4.5557 7.8661 3.4554 7.6998 7.8556 9.3441 8.7552 6.8887 6.7233 5.6677 4.5446	7.8556 7.7675 5.6652 4.5565 4.5578 6.1823 6.4154 5.6231 4.5557 3.6569 9.1329	. . .
. . .	. . .	. . .	. . .

- Tens of thousands of genes
- 10-1000 arrays
- Various biological conditions (e.g. disease/ control, time points)
- With technical replicates
- Note: heterogeneity among probes within the same probe set

	log probe intensities array 1	log probe intensities array 2	. . .
Gene 1	6.0097 7.8997 4.7292 6.0237 5.0233 5.5657 7.6687 7.3411 4.7232 5.9112 6.2232	5.2322 6.2234 5.3233 4.5443 2.8389 7.8223 8.2548 8.9967 7.6755 6.7445 6.7899	. . .
Gene 2	4.5557 7.8661 3.4554 7.6998 7.8556 9.3441 8.7552 6.8887 6.7233 5.6677 4.5446	7.8556 7.7675 5.6652 4.5565 4.5578 6.1823 6.4154 5.6231 4.5557 3.6569 9.1329	. . .
. . .	. . .	. . .	. . .

Data analysis



RMA Model

("Robust Multi Array" (RMA) by Irizarry et al. 2002)

Fix gene (probe set).

$Y_{jk} = log_2$ normalized background corrected PMs

Probe effect β_j and **Array** effect α_k,

and error

$$Y_{jk} = \beta_j + \alpha_k + \varepsilon_{jk}$$

(and sum zero constraint on probe effects)

Fit with iterative reweighed least square algorithm returning weights

1. Relative Log Expression (RLE):

Median Chip: median expression over all arrays (gene by gene)

RLE (gene A) in array k =

log ratio gene A's expression in array k and gene A's median expression

Idea: use RLE distribution for quality assessment (QA)

Interpretation on distribution level,
based on two biologic assumptions:

(A) majority of genes similar between different samples

(B) # upregulated genes = # downregulated genes

Then, good quality is indicated by:

Med(RLE)=0 small IQR(RLE)

2. Normalized unscaled standard error (NUSE):

$$\text{NUSE} = \frac{1 \Big/ \sqrt{W_k}}{\text{med}_{k'}\, 1 \Big/ \sqrt{W_{k'}}}$$

Note:
Normalization because of heterogeneity in # effective probes

Interpretation based on biologic assumptions

(A) majority of genes similar between different samples
(B) # upregulated genes = # downregulated genes

Then, good quality is indicated by:

Med(NUSE)=0 small IQR(NUSE)

3. Quality landscapes

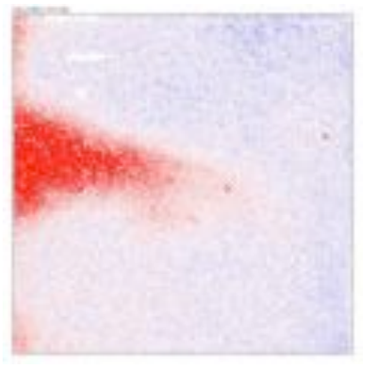
Weight images:

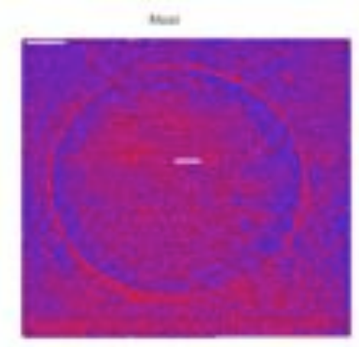
Colour a rectangle by probe weights
according to their spatial location on array.

dark green = low weights (poor quality)

Residual images:

Same, but with residuals.
red = positive residuals
blue = negative residuals





Very helpful for revealing causes for poor quality…

Residual images illustrating poor quality

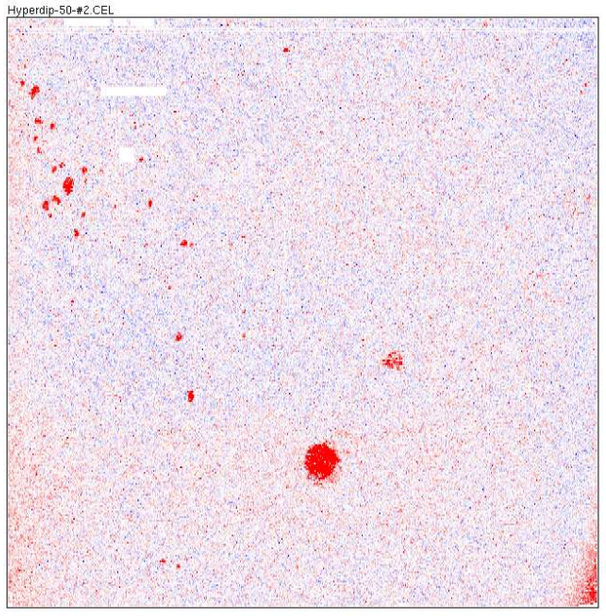


Fig. J1: "Bubbles"

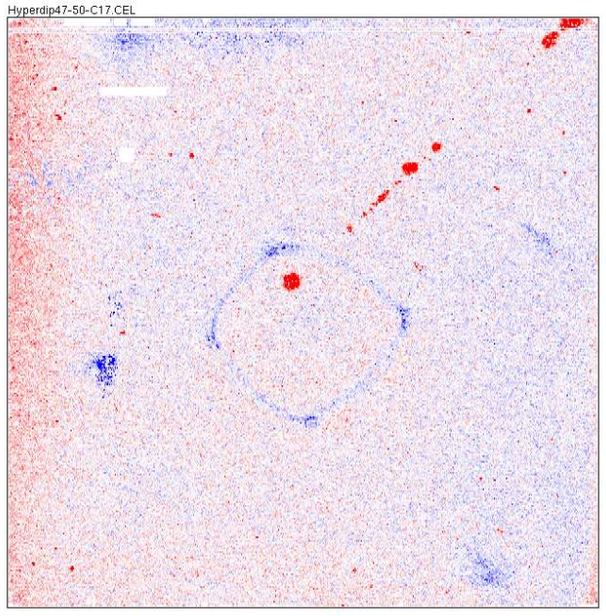


Fig. J2: "Circle and Stick"

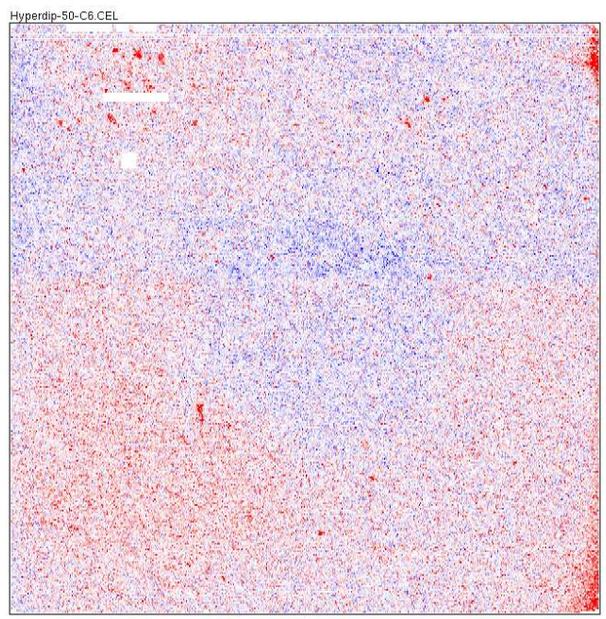


Fig. J3: "Sunset"

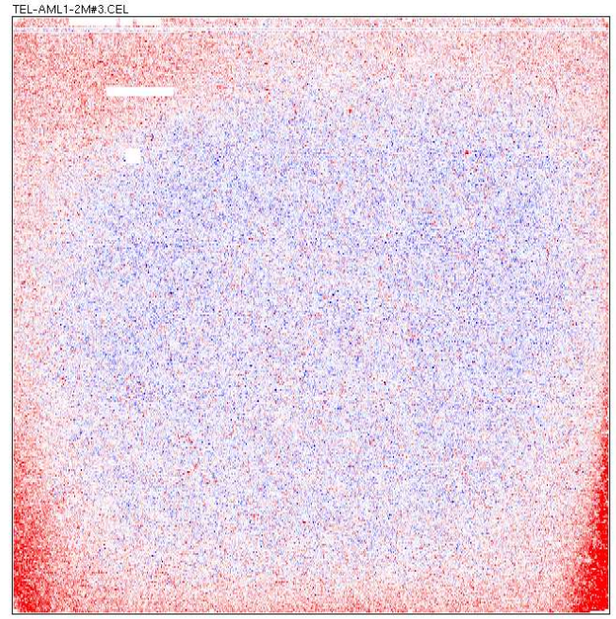


Fig. J4: "Pond"

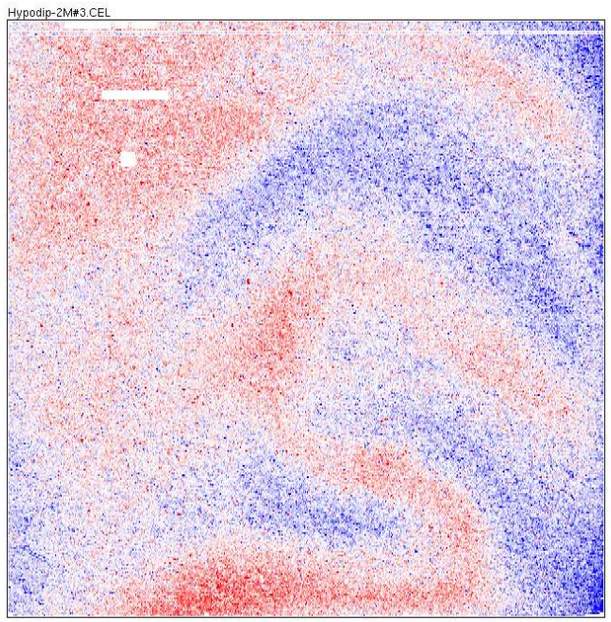


Fig. J5: "Letter S"

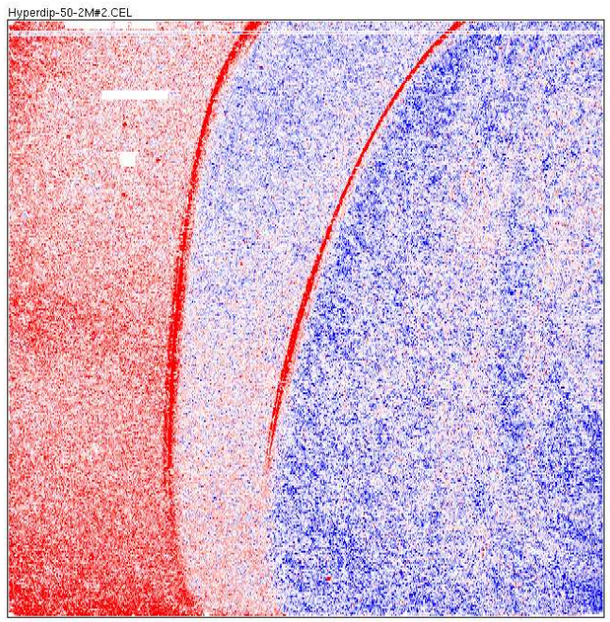


Fig. J6: "Compartments"

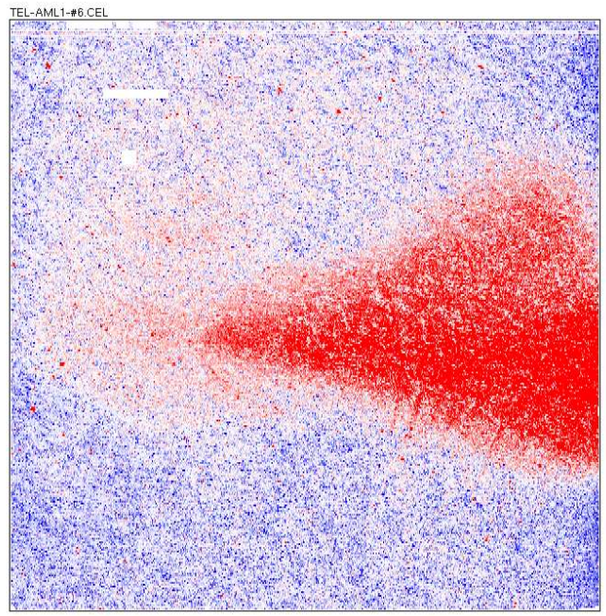


Fig. J7: "Triangle"

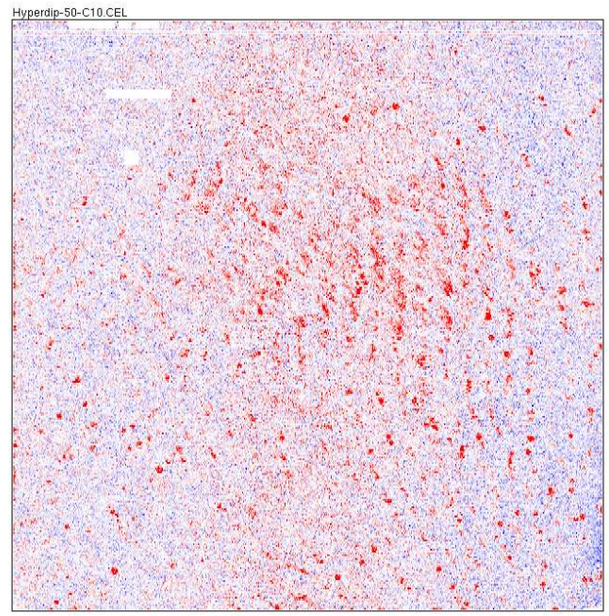
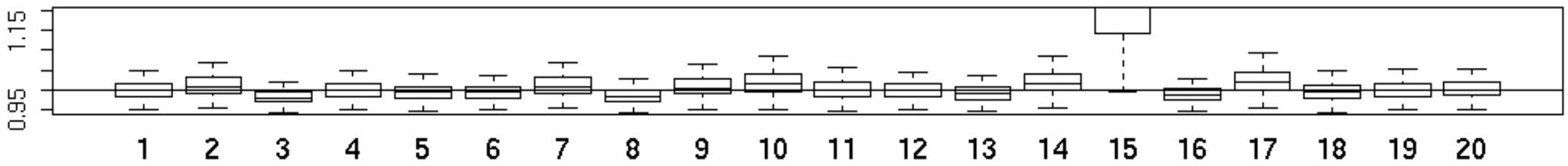


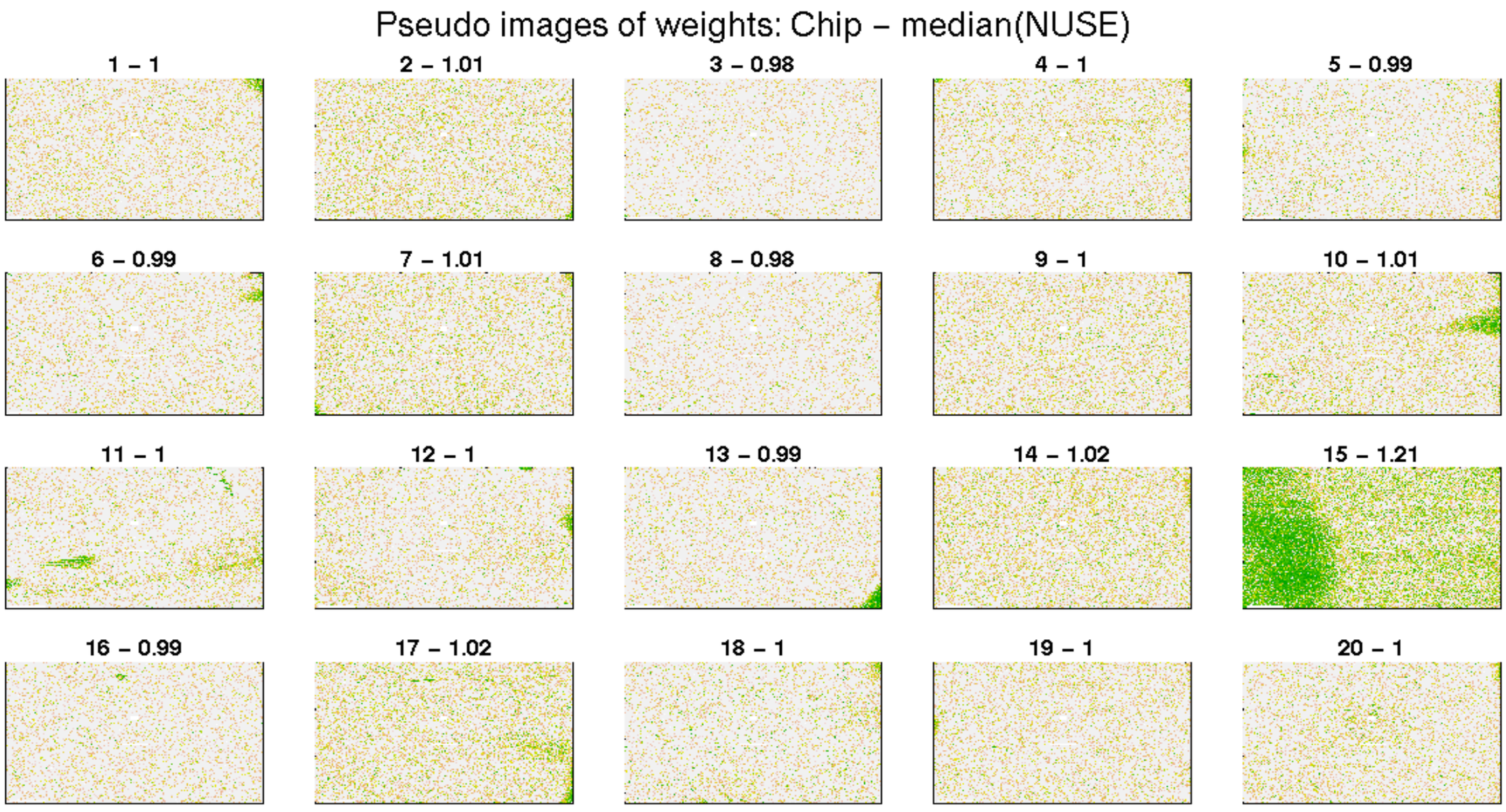
Fig. J8: "Fingerprint"

www.stat.berkeley.edu/~bolstad/PLMImageGallery/index.html

NUSE



Weights



Example for data quality variation between biological conditions

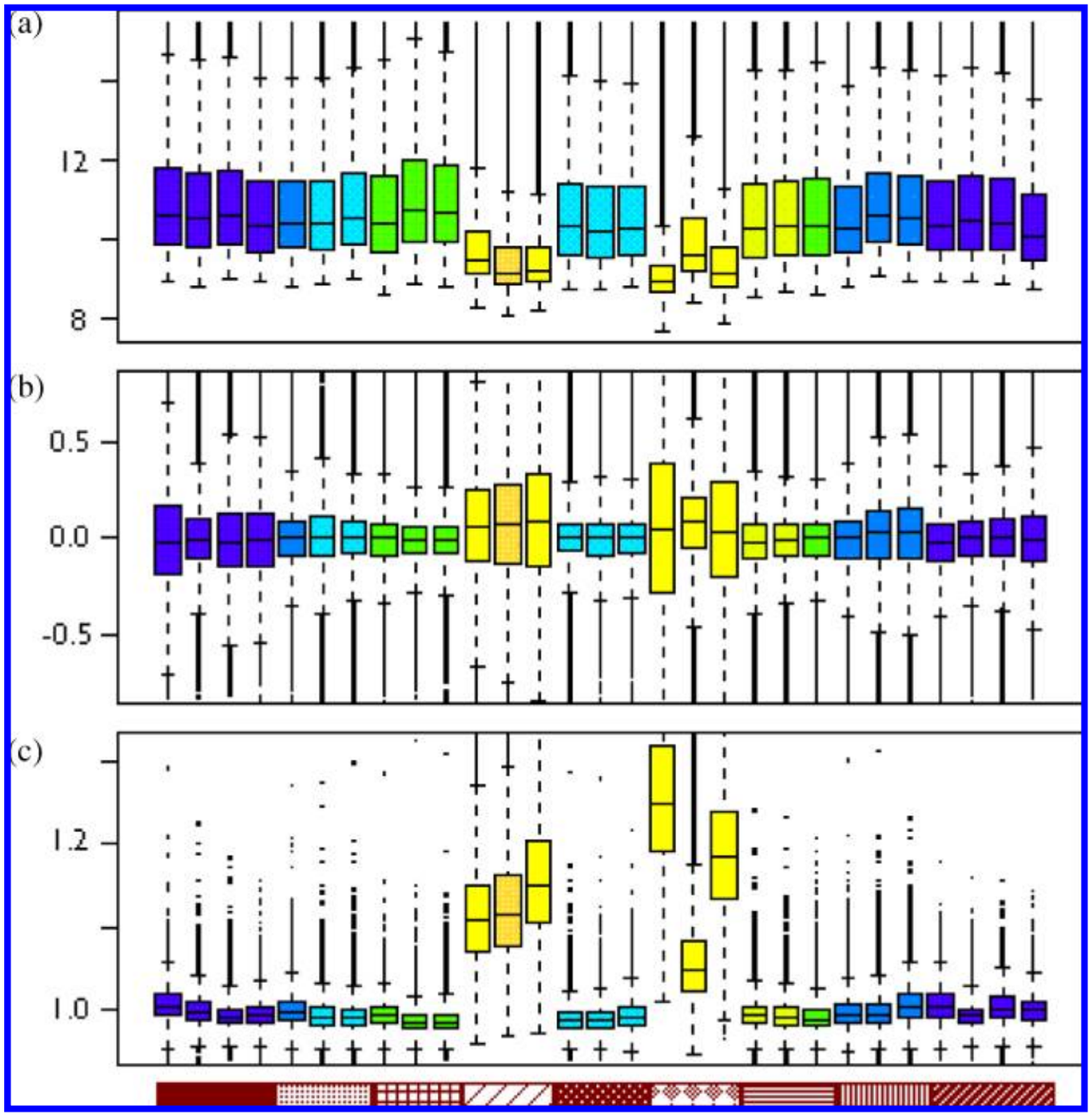


Figure F1. Series of boxplots of log-scaled PM intensities (a), RLE (b), and NUSE (c) for a comparison of nine fruit fly mutants with three to four technical replicates each. The patterns below the plot indicate mutants, and the gray levels of the boxes indicate hybridization dates. Med(RLE), IQR(RLE), Med(NUSE), and IQR(NUSE) all indicate substantially lower quality on the day colored white.

Example for a lab bias in data quality

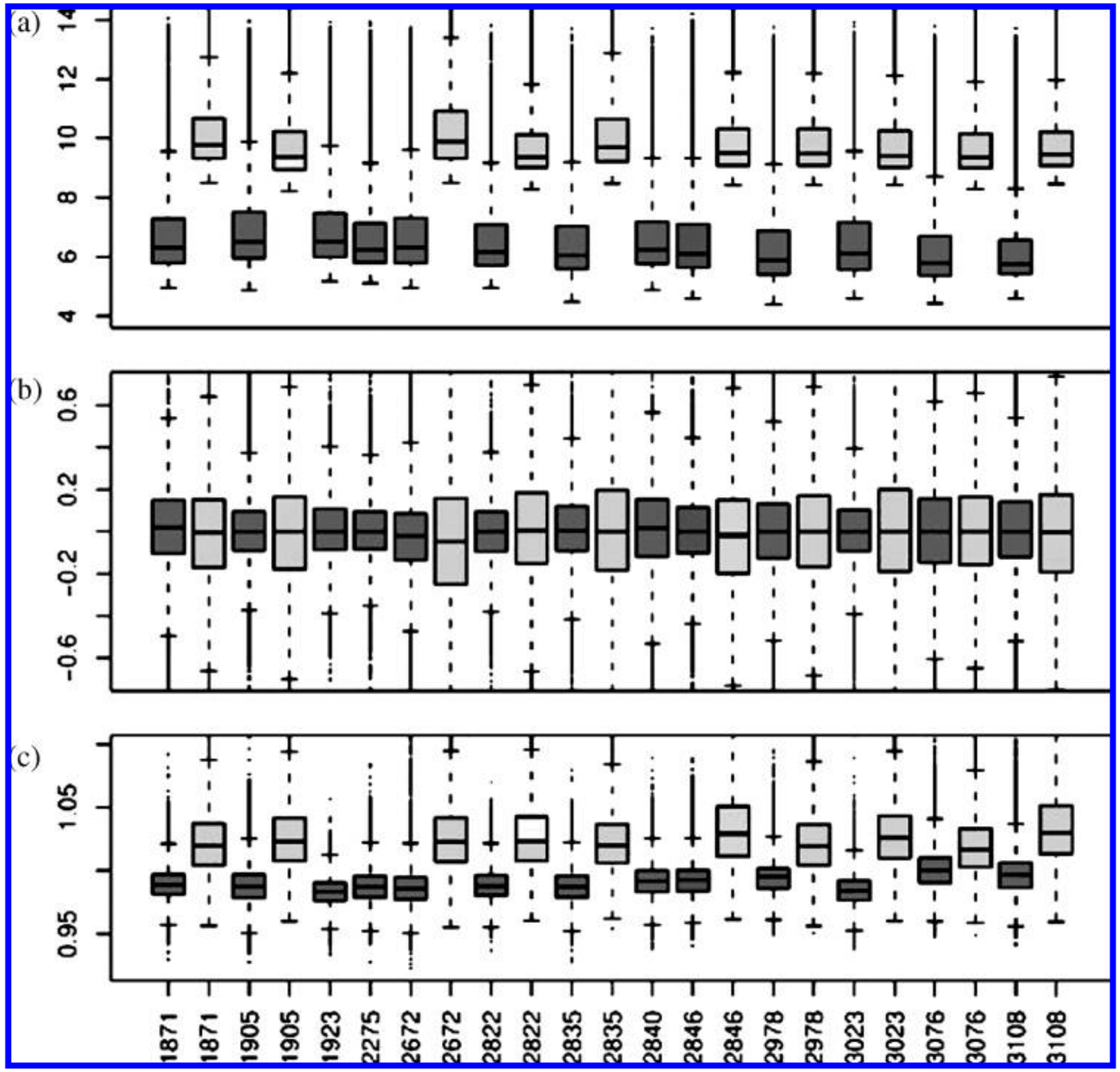


Figure H1. Series of boxplots of log-scaled PM intensities (a), RLE (b), and NUSE (c) for Pritzker gender study brain samples hybridized in two labs (some replicates missing). Gray level indicates lab site (dark for Lab M, light for Lab I). The log-scaled PM intensity distributions are all located around 6 for Lab M, and around 10 for Lab I. These systematic lab site differences are reflected by IQR(RLE), Med(NUSE), and IQR(NUSE), which consistently show substantially lower quality for Lab I hybridizations than for Lab M hybridizations.

Practical uses of our QA toolbox

- Small labs: concrete feedback on design and conduction of experiments
- Core facilities: biases, efficiency, process control
- Controlling error rates in genomic diagnostic tools (Afirma, OncotypeDX etc.)
- Quality benchmarking

Context dependency of quality

Ask: What would be the consequences of poor quality?

Shewhart (1927), Pioneer of industrial QC:

"The applied scientists knows that if he were to act upon the meagre evidence sometimes available to the pure scientist, he would make the same mistakes as the pure scientist makes in estimates of accuracy and precisions. He also knows that through his mistakes someone may lose a lot of money or suffer physical injury or both. [...]
***He does not consider his job simply that of doing the best he can with the available data; it is his job to get enough data before making this estimate*."**

Thanks to

Gene expression group at UC Berkeley

Biologists (for really bad microarray data)

R and Bioconductor communities (for packages)

Inside out group at University of Warwick

Perkin Elmer (for lots of dead pixels)