

Introduction to Microarray Analysis

Methods Course: Gene Expression Data Analysis

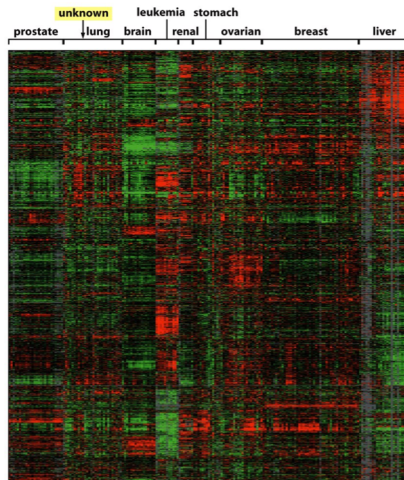
-Day One –

Rainer Spang

In this course you will learn ...

How to analyze and interpret high dimensional molecular profiles

How to use R and Bioconductor to document your computational analysis



There are profiles of ...

RNA Transcripts (mRNA, miRNA, lncRNA, ...)

Proteins (total expression, phosphorylation, ubiquitination ...)

Metabolites (intra cellular, secreted, ...)

Epigenetics (DNA methylation, histone methylation, histone acetylation)

Transcription factor binding (ChIP)

DNA copy number variation

Immune cell infiltration

Microbiomes (16S rRNA, Metagenomes, ...)

... and many more

Profiles are lists of quantified molecular features

Feature 1	9.342
Feature 2	6.766
Feature 3	0.001
...	
Feature 21451	3.881

RNA Transcripts (MYC, ACT, BCL6, ...)

Proteins (Myc, pAct, ...)

Metabolites (glucose, pyruvate, ...)

Epigenetics (CPG-island, genomic region)

Transcription factor binding (binding site)

DNA copy number variation (genomic region)

Immune cell infiltration (M1 macrophages, Th2 cells, ...)

Microbiomes (Clostridium sporogenes, Methanosphaera stadtmanae, ...)

Profiles can be generated by different technologies

RNA Transcripts (microarray, nanoString, RNAseq)

Proteins (MassSpec, protein array)

Metabolites (NMR, MassSpec,...)

Epigenetics (ChIP-seq, bisulfate sequencing, ATAC-seq)

Transcription factor binding (ChIP)

DNA copy number variation (aCGH, NGS)

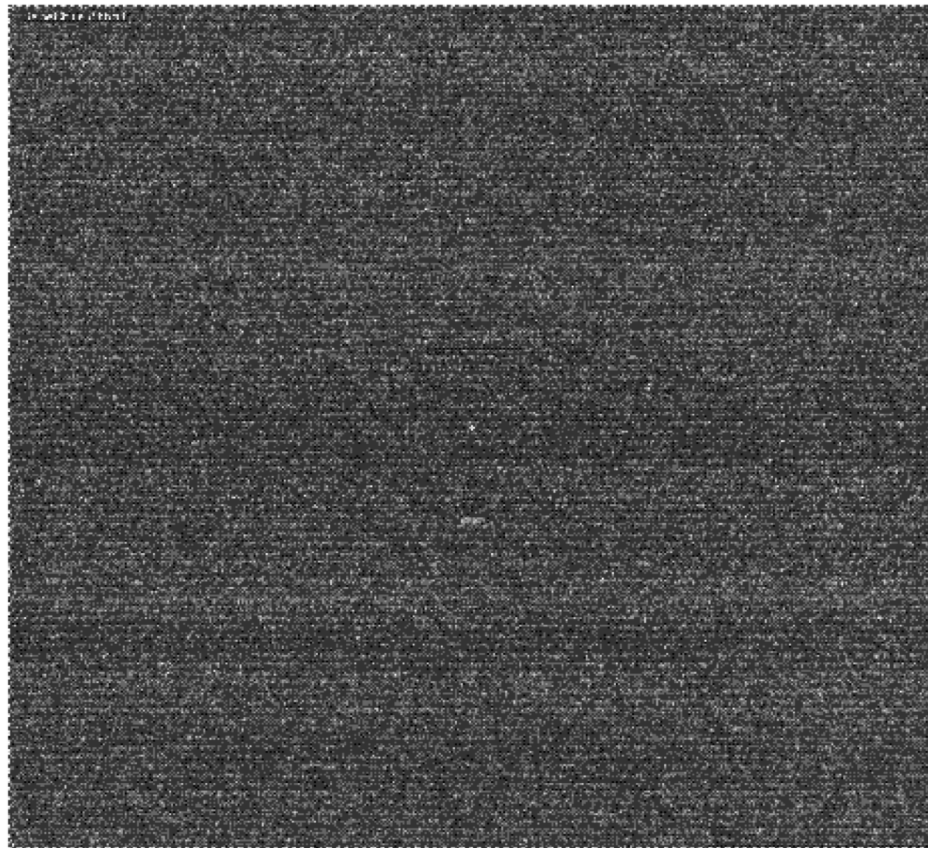
Immune cell infiltration (FACS, imaging, proteomics)

Microbiomes (arrays, 16S rRNA-seq)

... and many more

How can all this be covered in one week?

We will use gene expression data from microarrays as an example



Microarrays measure mRNA abundances in a tissue or a cell culture

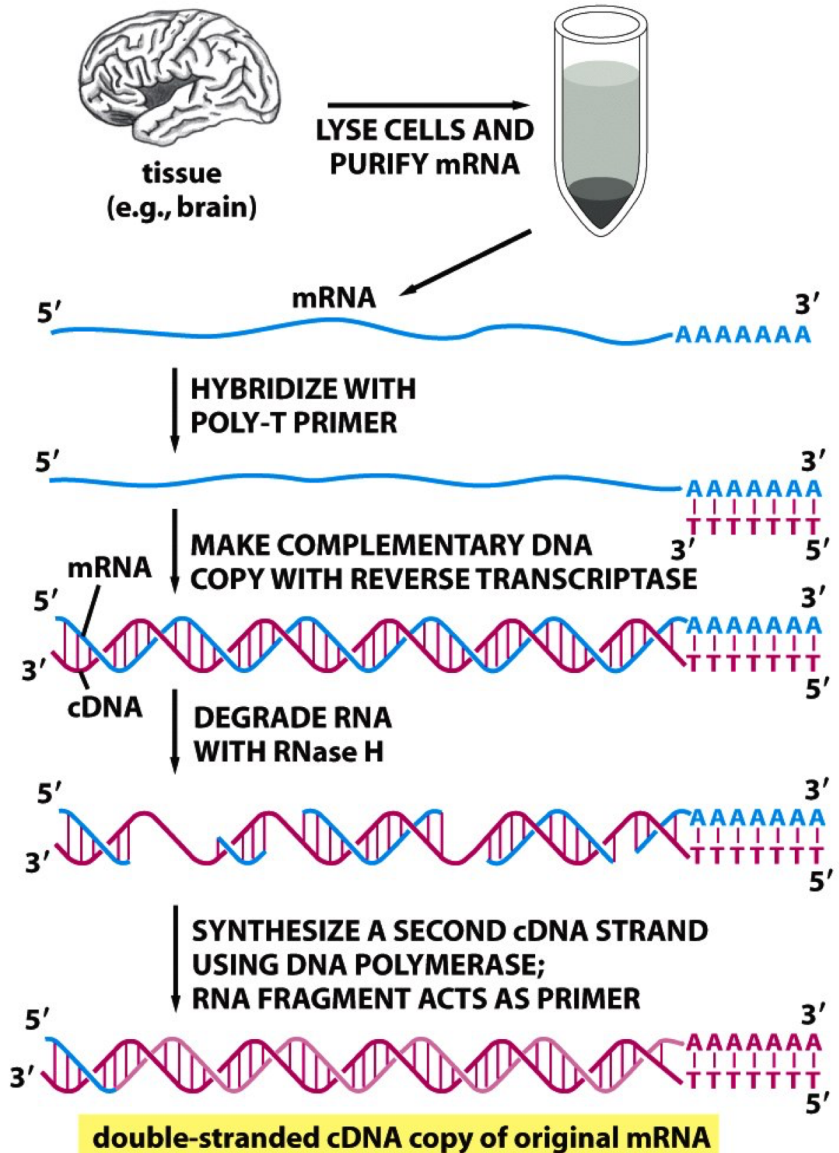
DNA



RNA Transcripts

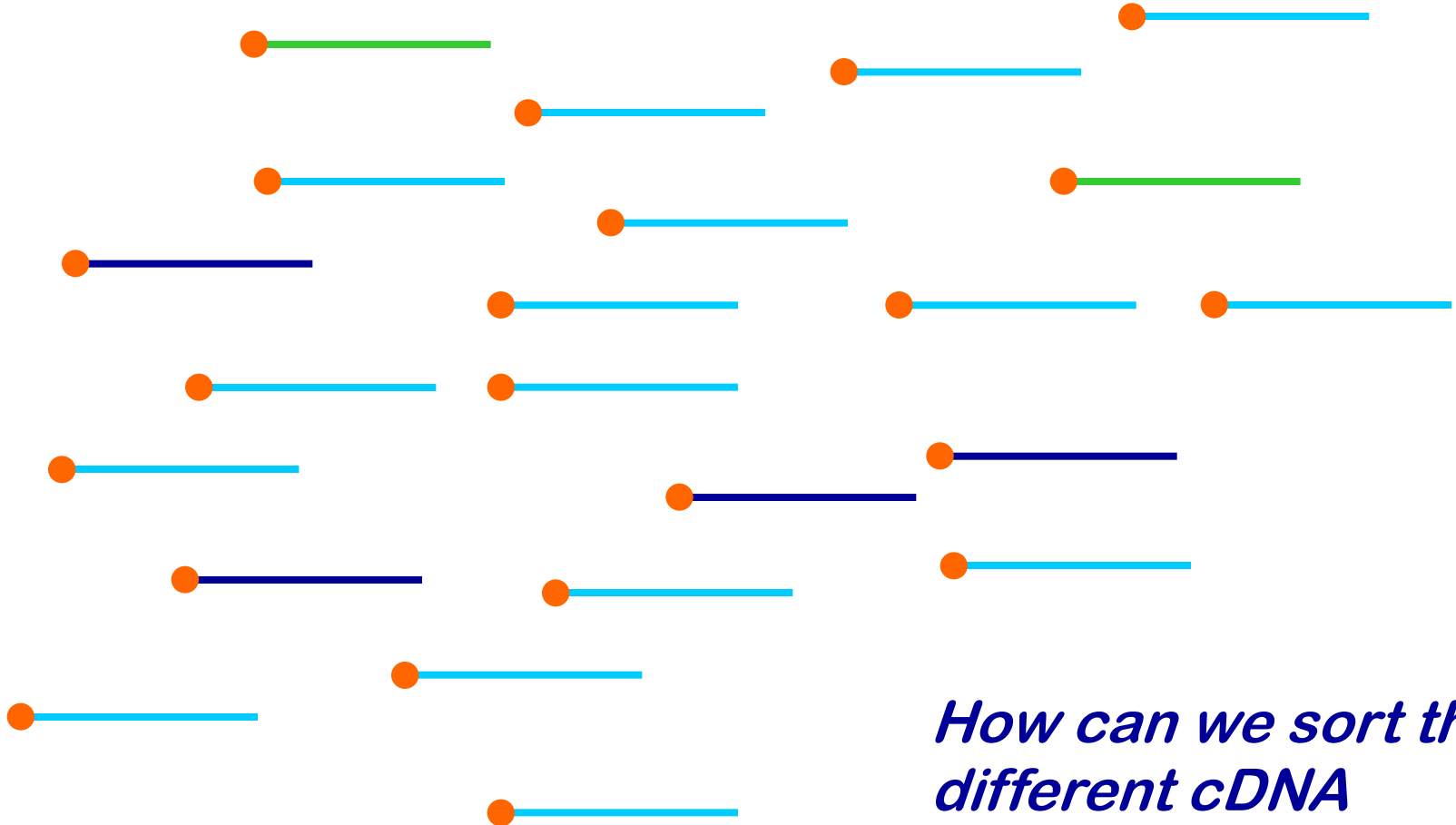
Purify mRNA and transform it to cDNA clones

You always start with same amount of RNA



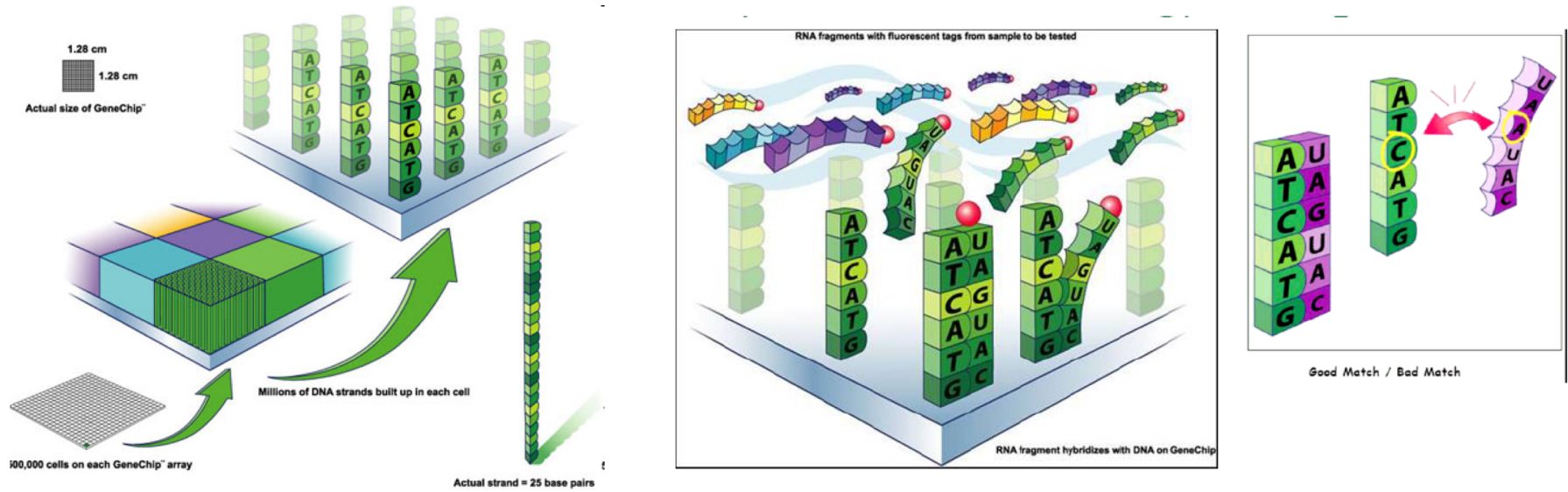
FLOURECENTLY LABEL THE DNA

*This gives us a complex probe of
fluorescently labeled cDNA*



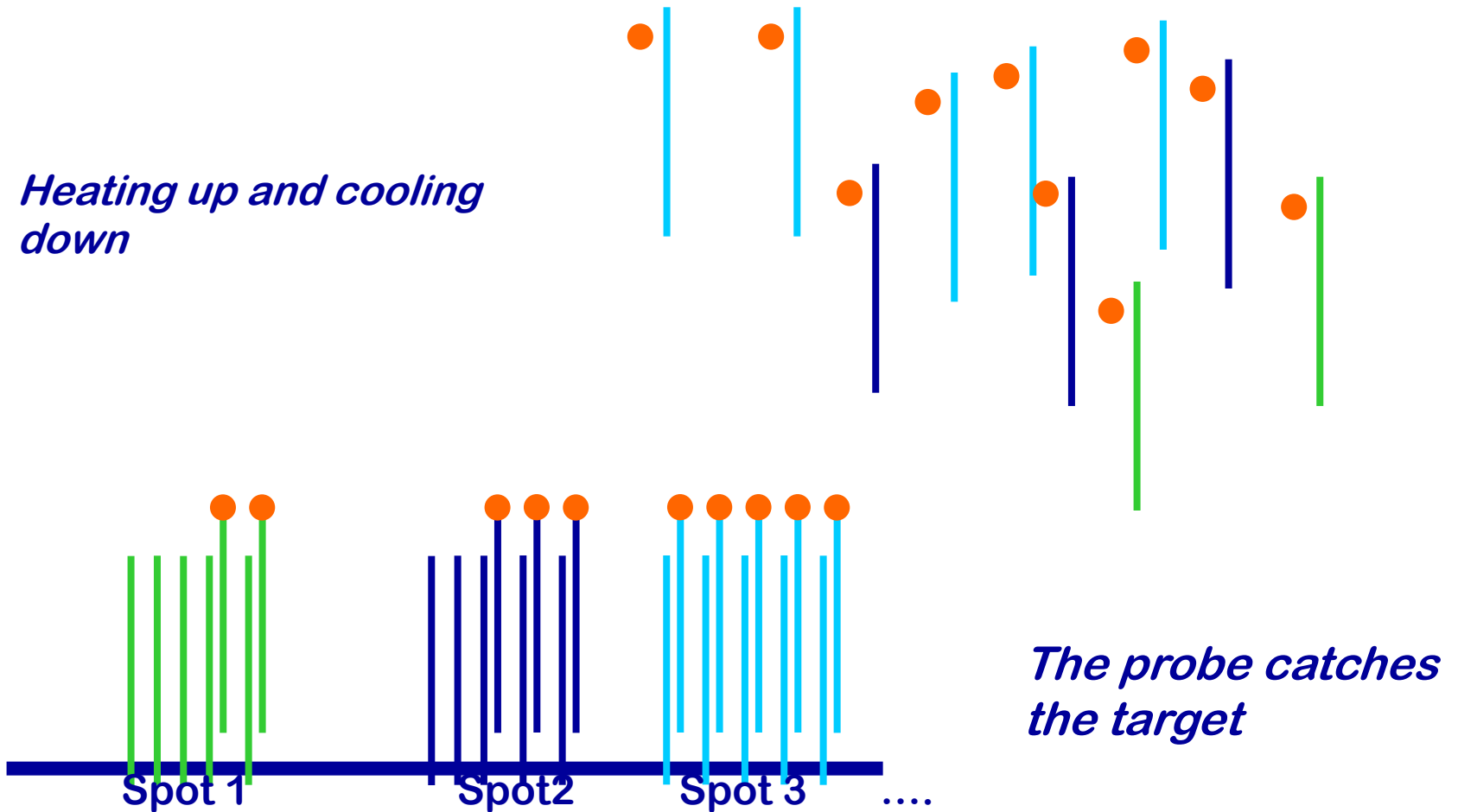
*How can we sort the
different cDNA
molecules?*

Transcriptome wide oligonucleotide libraries glued to a chip catch the cDNA in a hybridization reaction

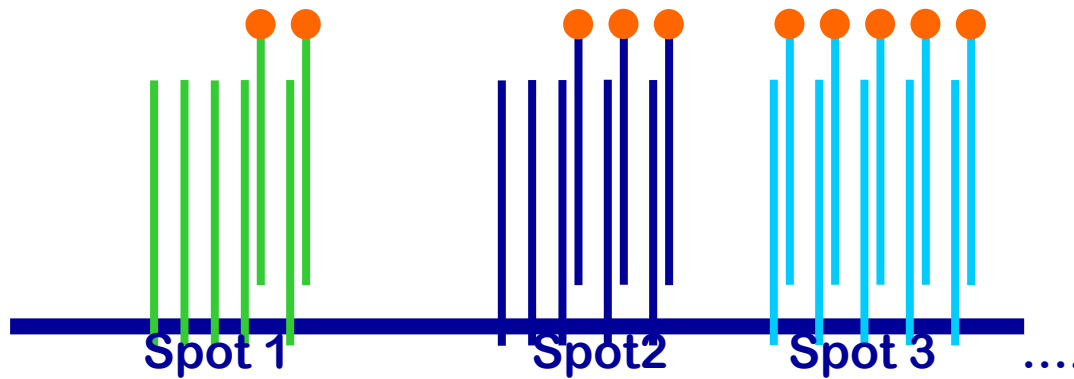


The microarray is a cDNA sorting device

Heating up and cooling down



High expression high fluorescent intensity of the spot ... low expression low fluorescent intensity of the spot



Gene expression is read out as fluorescent intensities

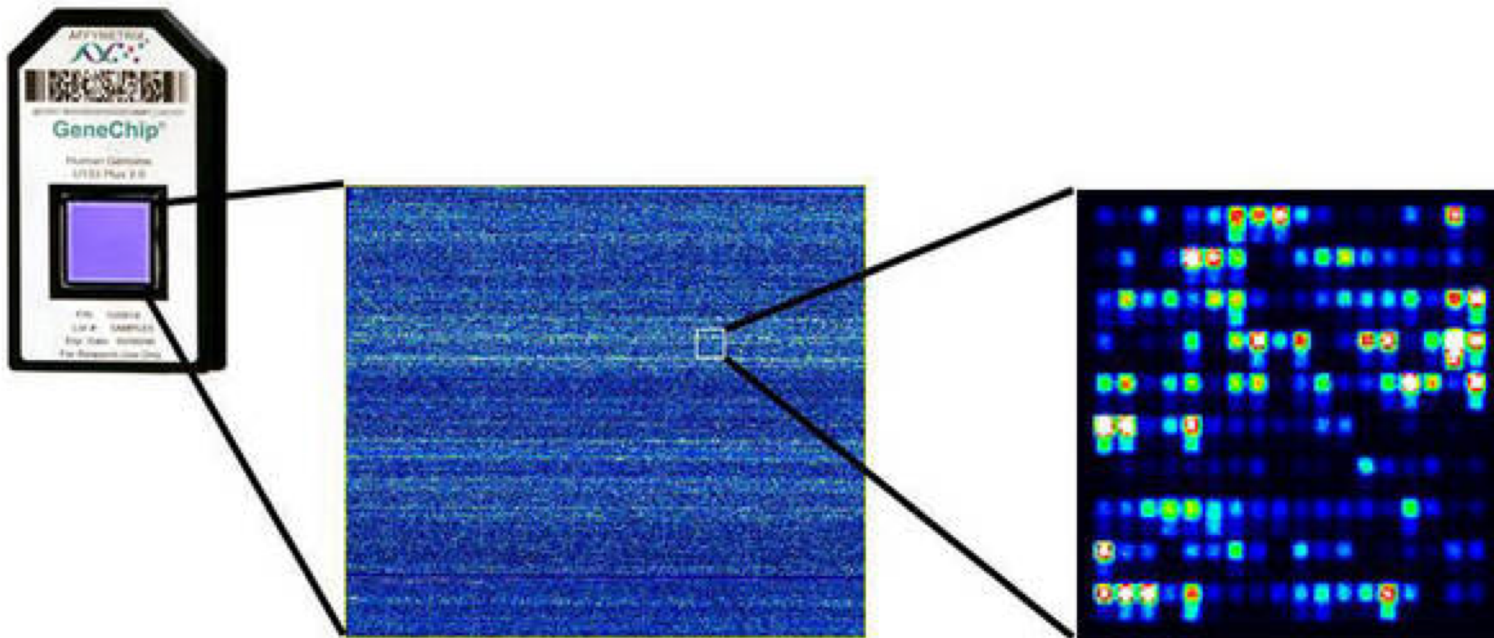
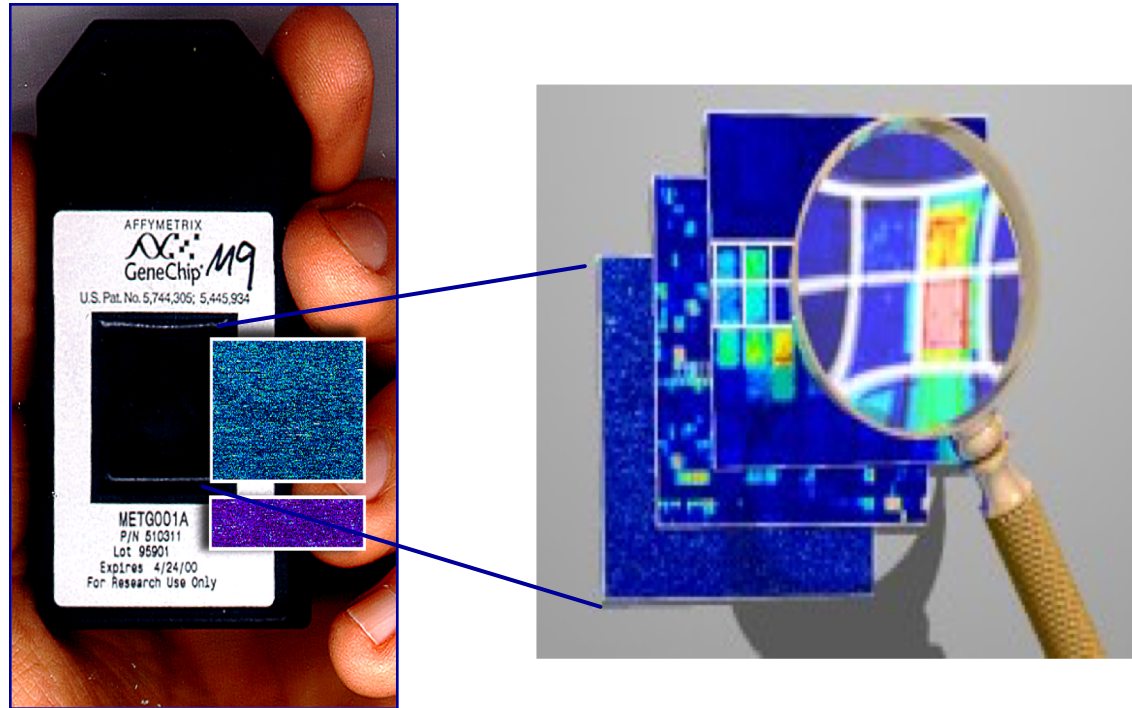


Image Analysis



Let the Affymetrix software do it

The .cel file is loaded by R

X	Y	Mean	STDV	NPixel
0	0	166	30.8	16
1	0	13135	1216.2	16
2	0	165.3	25.5	16
3	0	13706	1305.2	16
4	0	95	24.9	16
5	0	155.8	21.8	16
6	0	11675.8	1296.9	16
7	0	184	24.3	16
8	0	11465.5	1533.1	16
..

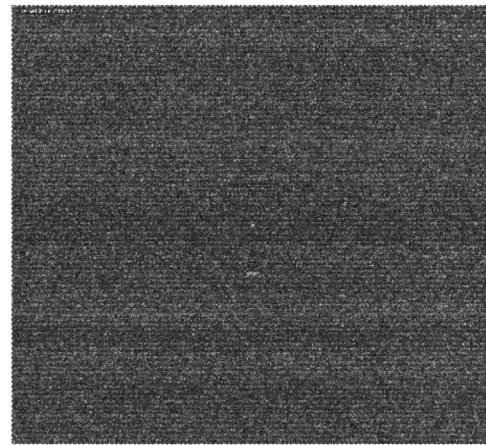
You always use the same amount of RNA

Array A



1 μ g RNA

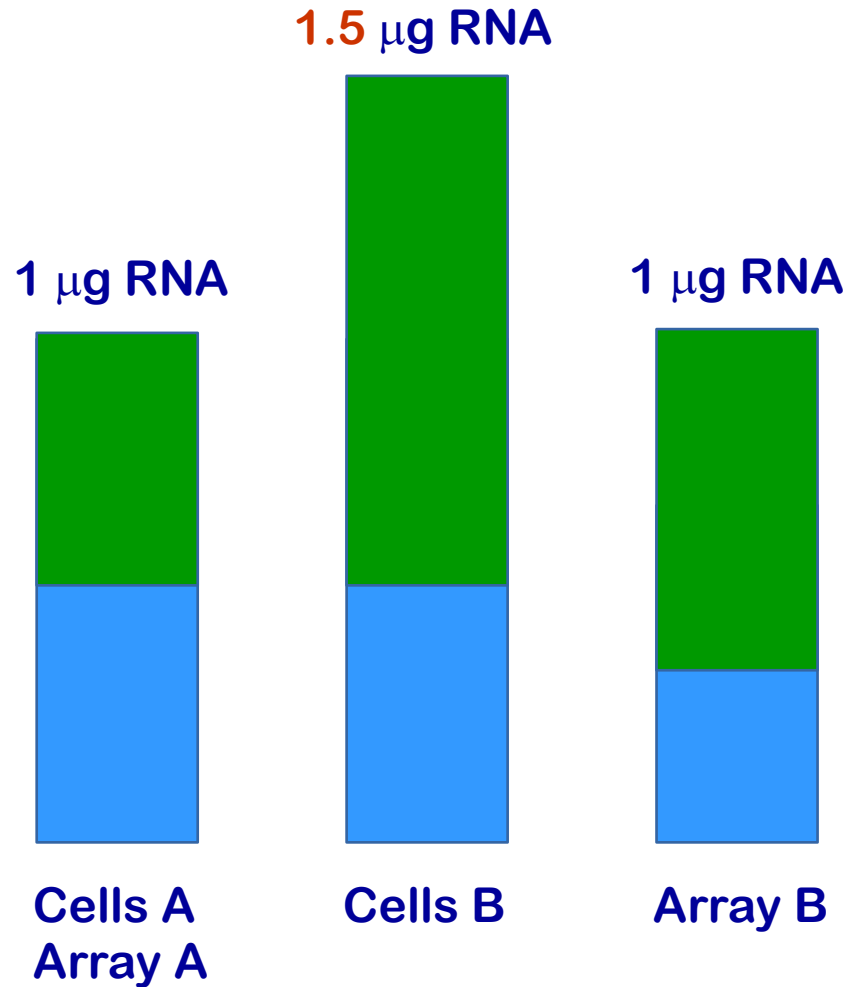
Array B



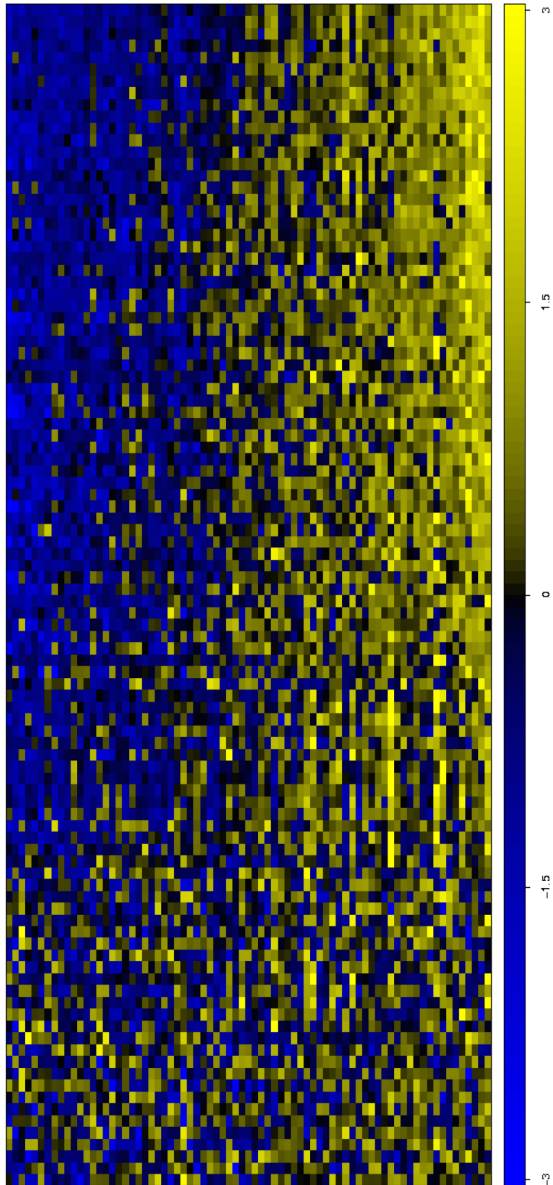
1 μ g RNA

When comparing profile A vs B, what do you see if 50% of the genes go up 2 fold and all others stay the same?

*We will observe both up and down
“regulation”*



*Analyzing gene expression
with the eye*



The Heat Map

Rows: Genes

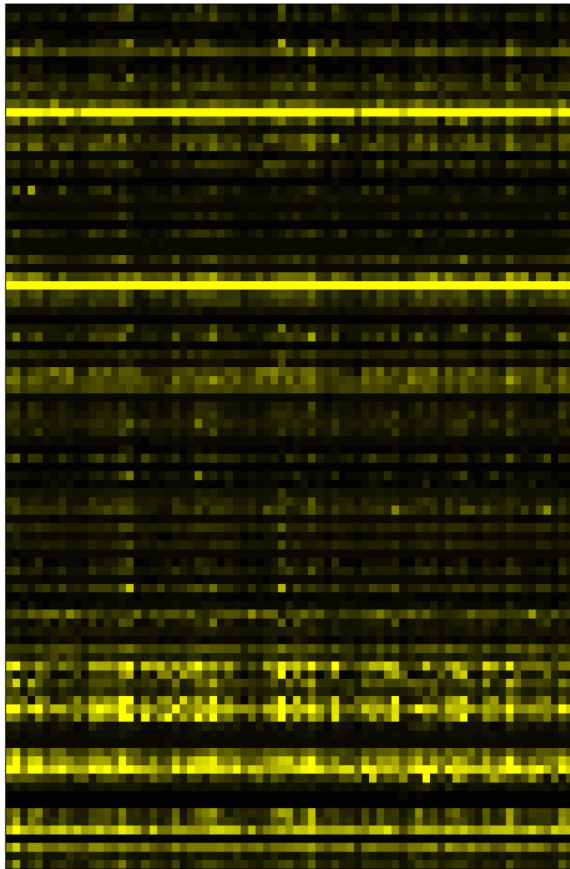
Columns: Samples

Color: Expression

High

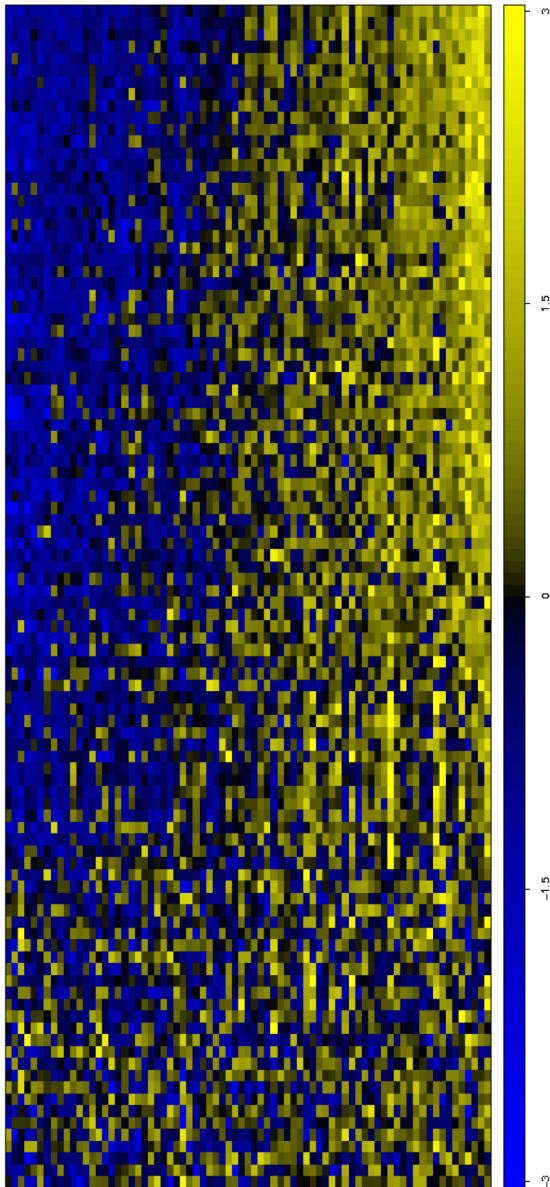
Low

*The color encodes expression levels,
but not globally ...*



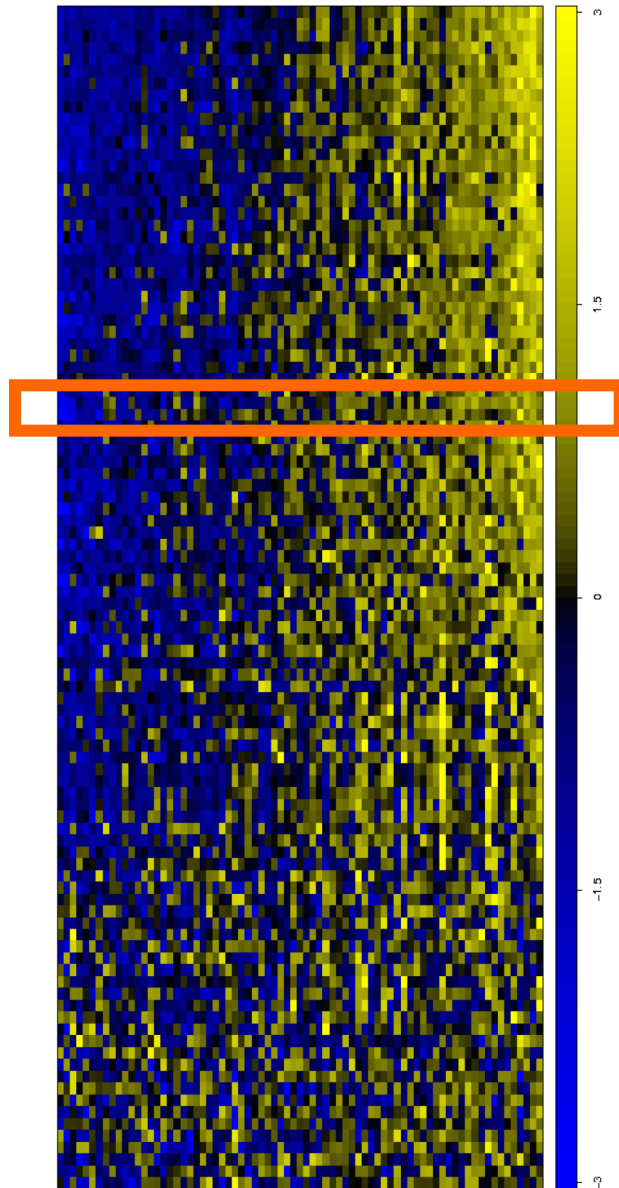
*... otherwise the heat
map looks like this!*

*The largest expression
differences are between
genes and not within a
gene across samples*



Using ranks gene by gene

The **highest value** of a gene across samples is bright yellow, **the lowest** is bright blue



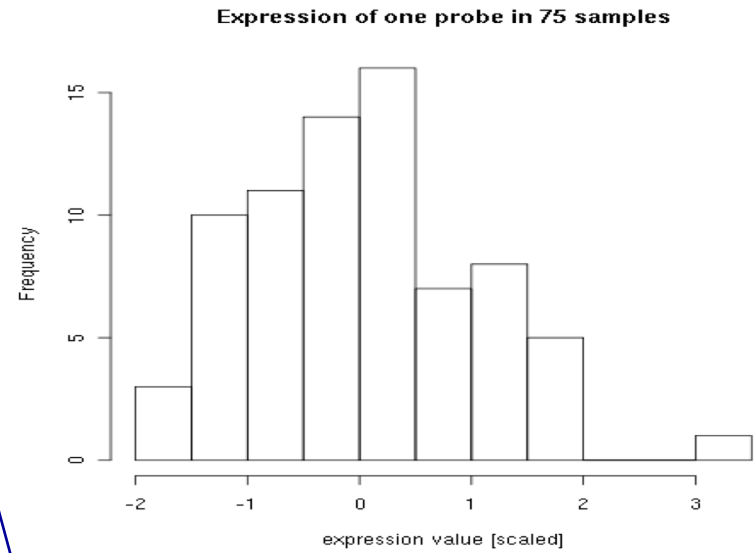
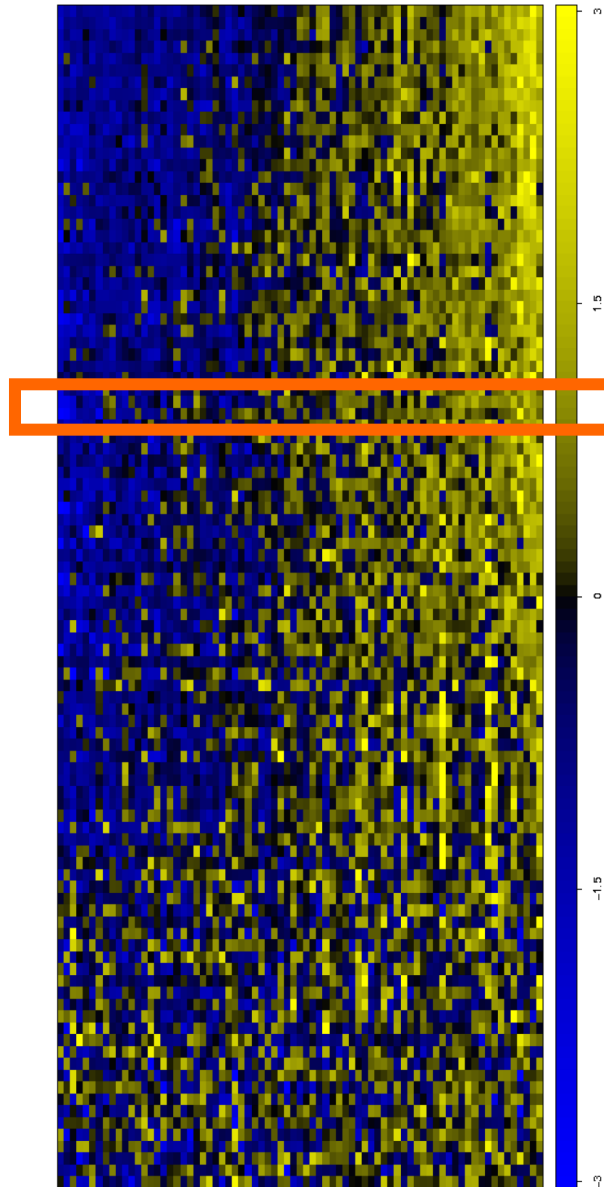
The colors suggest

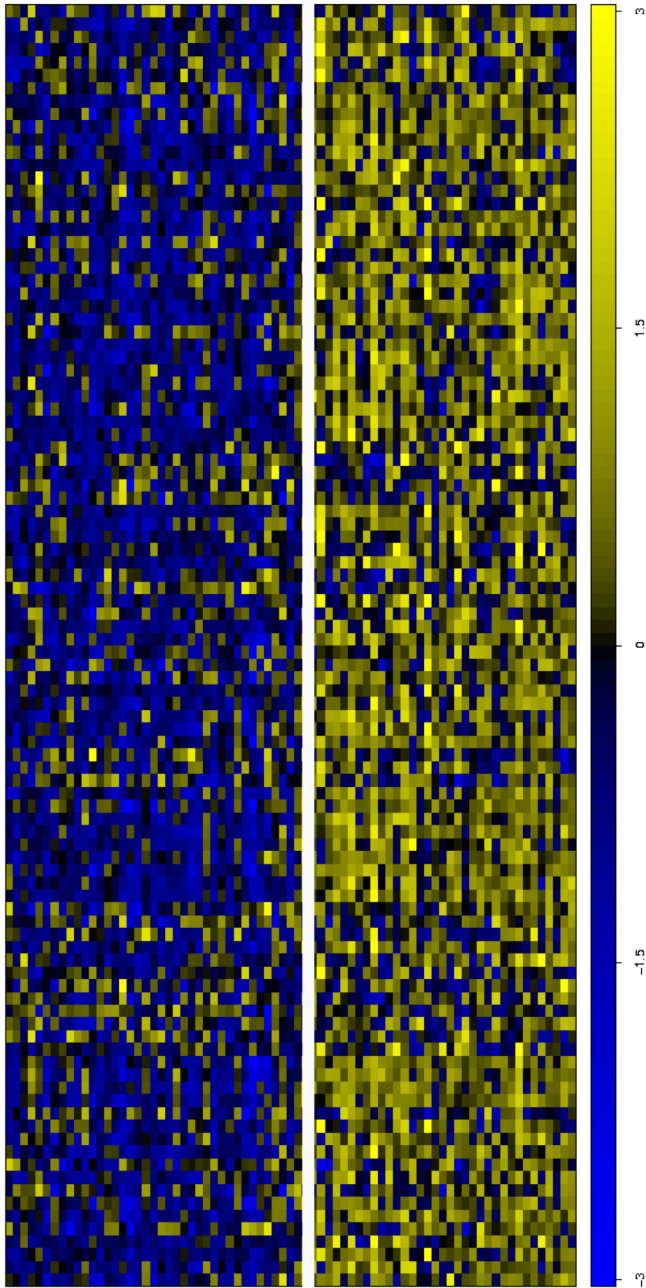
...

... that these genes have two well separated expression levels

It is low for the left half of patients (right) and high for the other half (left) ...

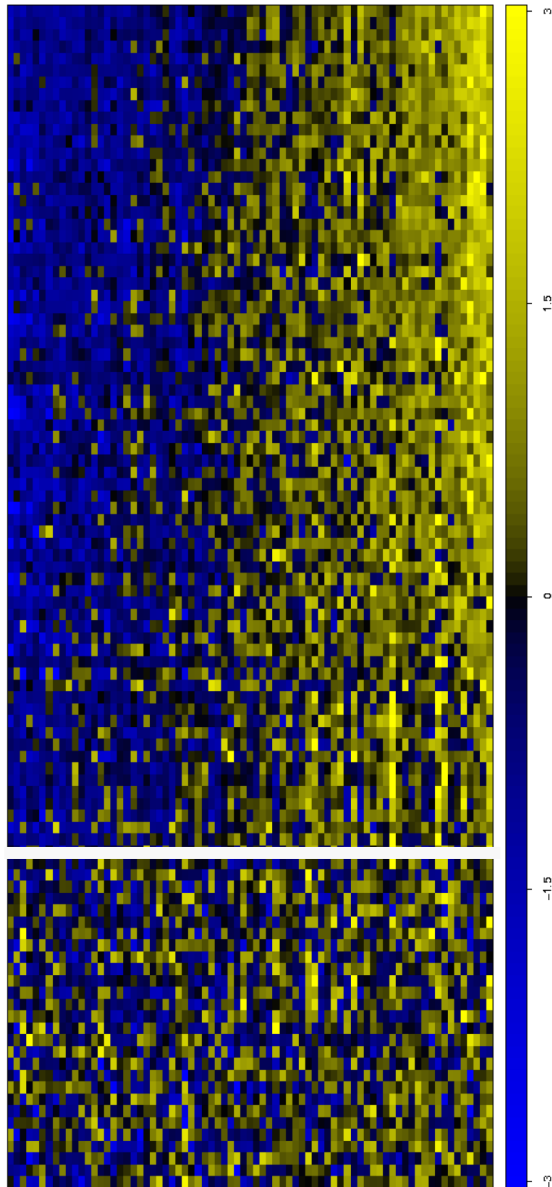
... well this is not the case





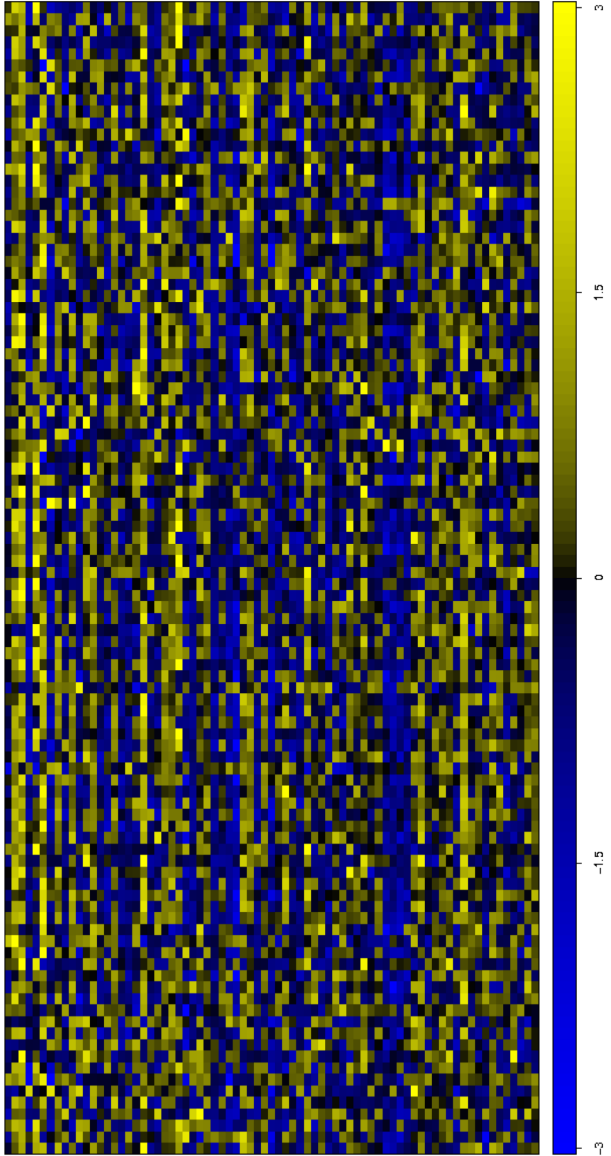
Two classes of samples

All genes differentially expressed



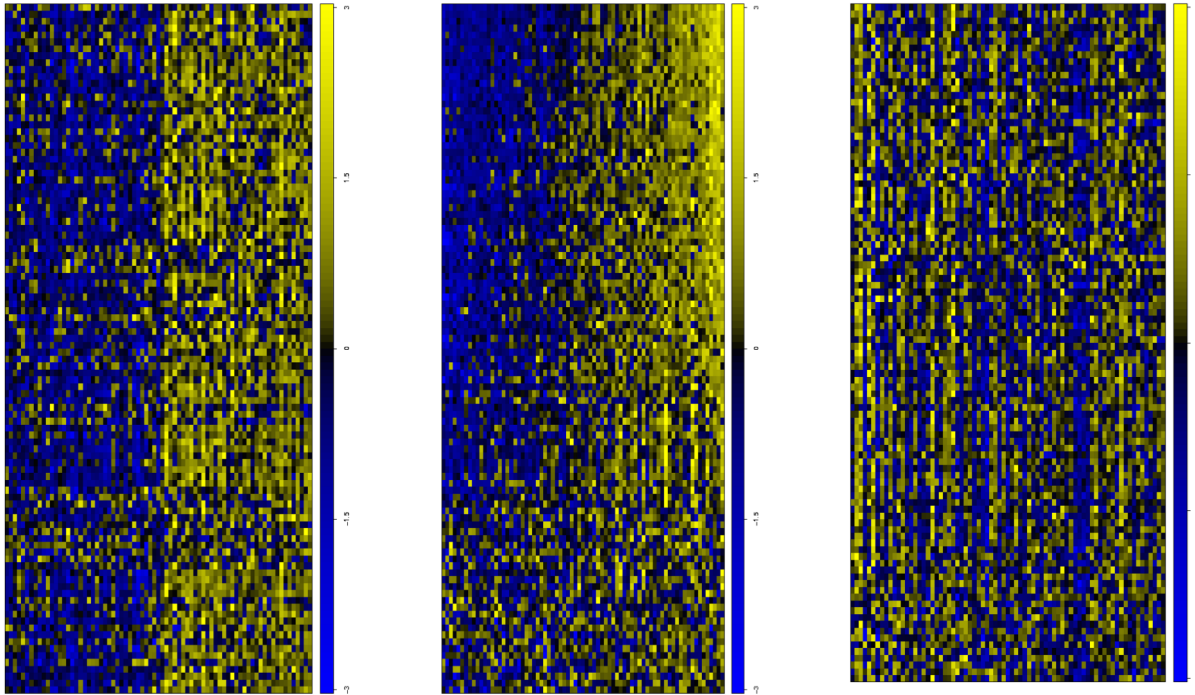
*A continuum of
samples*

*Two groups of
genes*



*Nothing but
noise*

Deceiving the eye

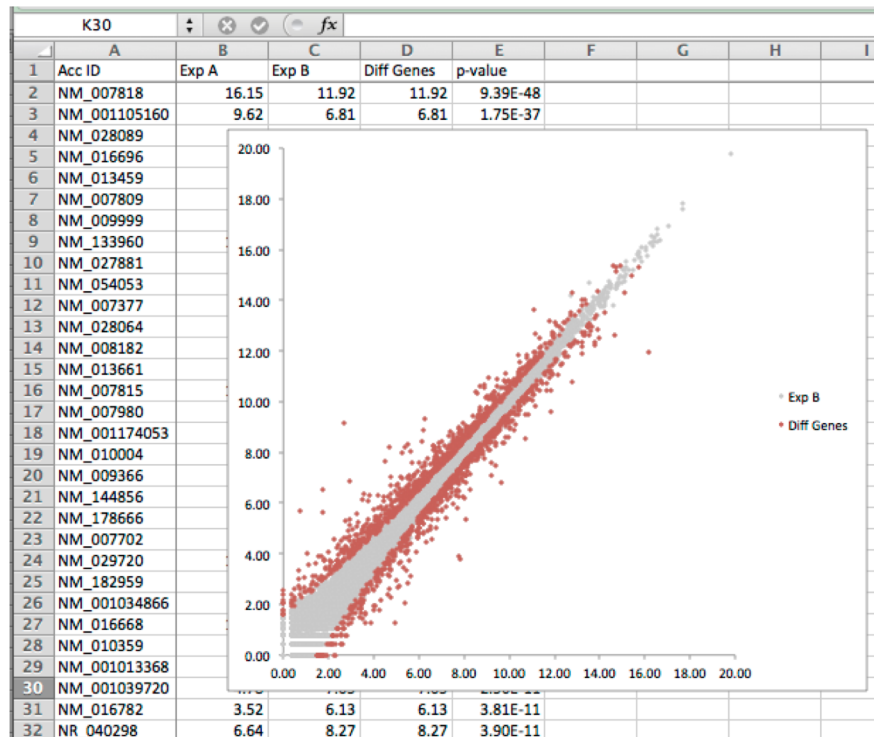


*It is all **the same data** ... just sorted differently*

Why using R?

```
> library(MCRestimate)
> NestedCV.svm <- MCRestimate(known.patients,
+   "mol.biol",
+   classification.fun = "SVM.wrap",
+   variableSel.fun = "varSel.highest.var",
+   poss.parameters = list(gamma = 2^(-2:2)),
+   cross.outer = 3,
+   cross.inner = 3,
+   cross.repeat = 3)
```

Isn't something like this easier?



Nightmare Part 1

You sit down to finish writing your manuscript.
You realize that you need to clarify one result by running an additional analysis.

You first re-run the primary analysis.

The reproduced primary results don't match with what you have in your paper.

<http://biostat.mc.vanderbilt.edu/wiki/pub/Main/TheresaScott/ReproducibleResearch.TAScott.handout.pdf>

For how long can you remember the exact sequence of clicks you made to get your results?

Nightmare Part 2

When you go to your project folder to run the additional analysis, you find multiple data files, multiple analysis files, & multiple results files. You can't remember which ones are pertinent.

<http://biostat.mc.vanderbilt.edu/wiki/pub/Main/TheresaScott/ReproducibleResearch.TAScott.handout.pdf>

For how long can you remember what exactly is in which file?

Nightmare Part 3

You've just spent the week running your analysis & creating a results report (including tables & graphs) to present to your collaborators. You then receive an email from your PI asking you to regenerate the report based on a subset of the original data set & including an additional set of analyses – she would like it by tomorrow's meeting.

<http://biostat.mc.vanderbilt.edu/wiki/pub/Main/TheresaScott/ReproducibleResearch.TAScott.handout.pdf>

Re-clicking is no fun!

Nightmare Part 4

Finally you have submitted your manuscript. The reviewers are positive but require that you compare your analysis to one that uses the method of Lique et al (Bioinformatics 2015)

Good luck with doing that in excel!

Gene expression

Group and sparse group partial least square approaches applied in genomics context

Benoît Liquet^{1,2,*}, Pierre Lafaye de Micheaux³, Boris P. Hejblum^{4,5,6,7}
and Rodolphe Thiébaud^{4,5,6,7}

¹School of Mathematics and Physics, The University of Queensland, Brisbane 4066, Australia, ²ARC Centre of Excellence for Mathematical and Statistical Frontiers, QUT, Brisbane, Australia, ³CREST, ENSAI, Campus de Ker-Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz cedex, France, ⁴Inria, SISTM, Talence and ⁵Inserm, U897, Bordeaux, ⁶Bordeaux University, Bordeaux and ⁷Vaccine Research Institute, Creteil, France

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on March 14, 2015; revised on June 26, 2015; accepted on September 3, 2015

Abstract

Motivation: The association between two blocks of ‘omics’ data brings challenging issues in computational biology due to their size and complexity. Here, we focus on a class of multivariate statistical methods called partial least square (PLS). Sparse version of PLS (sPLS) operates integration of two datasets while simultaneously selecting the contributing variables. However, these methods do not take into account the important structural or group effects due to the relationship between markers among biological pathways. Hence, considering the predefined groups of markers (e.g. genesets), this could improve the relevance and the efficacy of the PLS approach.

Results: We propose two PLS extensions called group PLS (gPLS) and sparse gPLS (sgPLS). Our algorithm enables to study the relationship between two different types of omics data (e.g. SNP and gene expression) or between an omics dataset and multivariate phenotypes (e.g. cytokine secretion). We demonstrate the good performance of gPLS and sgPLS compared with the sPLS in the context of grouped data. Then, these methods are compared through an HIV therapeutic vaccine trial. Our approaches provide parsimonious models to reveal the relationship between gene abundance and the immunological response to the vaccine.

Availability and implementation: The approach is implemented in a comprehensive R package called sgPLS available on the CRAN.

Contact: b.liquet@uq.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.



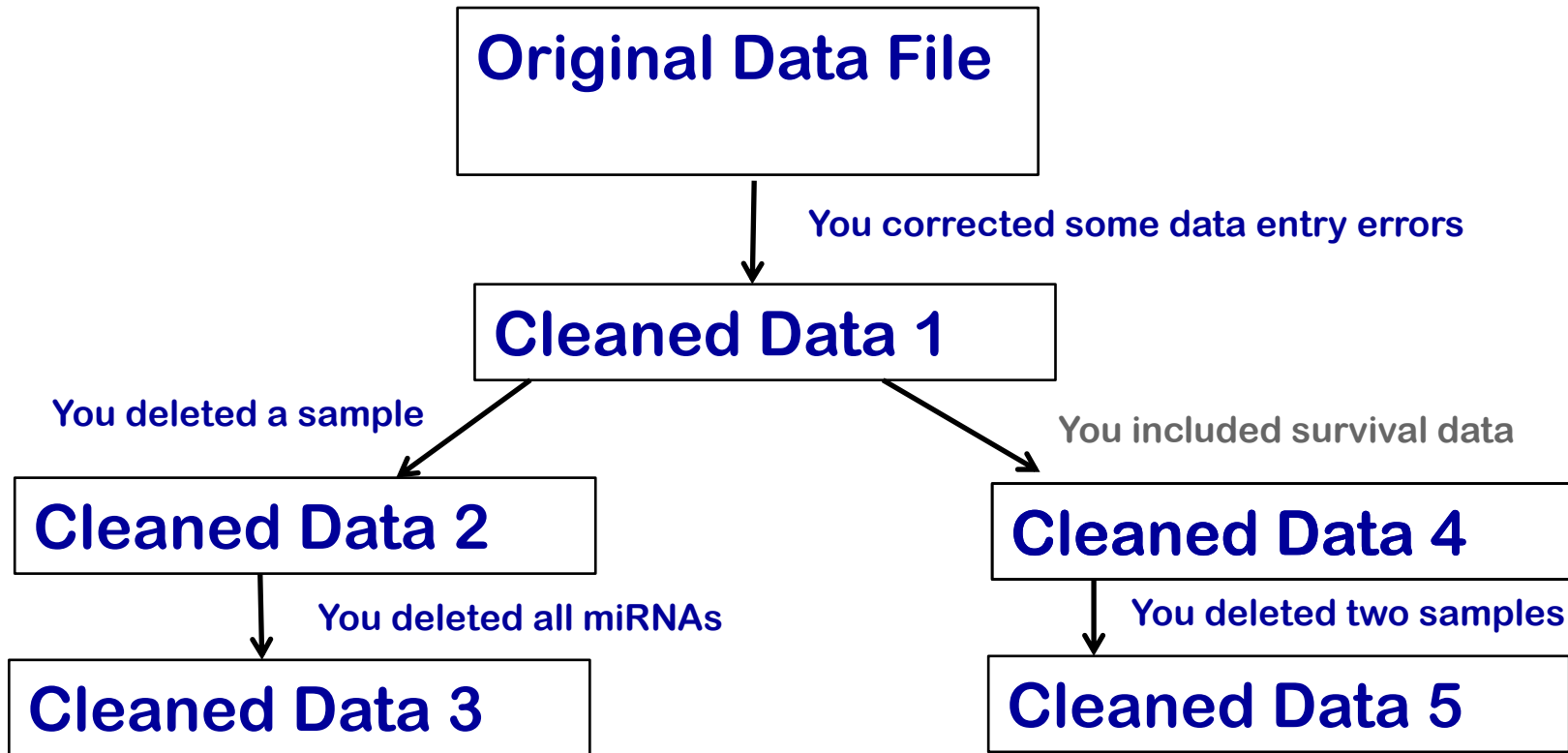
A Nightmare (only for decent people)

**Your paper is published and widely acknowledged.
You receive an email from a colleague who wants to
run your computational analysis on her own data.
She asks for instructions and help.**

What do you give her?

My project is small. I can easily keep track of the few analysis steps in my head. Essentially, I only computed one p-value per gene

You did more ...



You write every step into your lab log

**What about using an
R-file as your lab log**

You can rerun R-files



Repeatability of published microarray gene expression analyses

John P A Ioannidis¹⁻³, David B Allison⁴, Catherine A Ball⁵, Issa Coulibaly⁴, Xiangqin Cui⁴, Aedin C Culhane^{6,7}, Mario Falchi^{8,9}, Cesare Furlanello¹⁰, Laurence Game¹¹, Giuseppe Jurman¹⁰, Jon Mangion¹¹, Tapan Mehta⁴, Michael Nitzberg⁵, Grier P Page^{4,12}, Enrico Petretto^{11,13} & Vera van Noort¹⁴

© 2009 Nature America, Inc. All rights reserved.

Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public data availability. Several journals require public data deposition and several public databases exist. However, not all data are publicly available, and even when available, it is unknown whether the published results are reproducible by independent scientists. Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in *Nature Genetics* in 2005–2006. One table or figure from each article was independently evaluated by two teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis. Repeatability of published microarray studies is apparently limited. More strict publication rules enforcing public data availability and explicit description of data processing and analysis should be considered.

research, the Uniform Guidelines of the International Committee of Medical Journal Editors state that authors should “identify the methods, apparatus and procedures in sufficient detail to allow other workers to reproduce the results”¹². Making primary data publicly available has many challenges but also many benefits¹³. Public data availability allows other investigators to confirm the results of the original authors, exactly replicate these results in other studies and try alternative analyses to see whether results are robust and to learn new things. Journals such as *Nature Genetics* require public data deposition as a prerequisite for publication for microarray-based research. Yet, the extent to which data are indeed made fully and accurately publicly available and permit confirmation of originally reported findings in many areas, including gene expression microarray research, is unknown.

In this project, we aimed to evaluate the repeatability of published microarrays studies. We focused specifically on the ability to repeat the published analyses and get the same results. This is one important component in the wider family of replication and reproducibility issues. We evaluated 18 articles published in *Nature Genetics* in 2005 or 2006 that presented data from comparative analyses of microarrays

Data and materials availability All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*. All computer codes involved in the creation or analysis of data must also be available to any reader of *Science*. After publication, all reasonable requests for data and materials must be fulfilled. Any restrictions on the availability of data, codes, or materials, including original data obtained from other sources (Materials Transfer Agreements), must be disclosed at the time of submission. If there are any MTAs pertaining to data or materials produced in this research, or


Science

AAAS

 OPEN ACCESS

EDITORIAL

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve , Anton Nekrutenko, James Taylor, Eivind Hovig

Published: October 24, 2013 • DOI: 10.1371/journal.pcbi.1003285



Rule 1: For Every Result, Keep Track of How It Was Produced

Rule 2: Avoid Manual Data Manipulation Steps

Rule 3: Archive the Exact Versions of All External Programs Used

Rule 4: Version Control All Custom Scripts

Rule 5: Record All Intermediate Results, When Possible in Standardized Formats

Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds

Rule 7: Always Store Raw Data behind Plots

Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected

Rule 9: Connect Textual Statements to Underlying Results

Rule 10: Provide Public Access to Scripts, Runs, and Results

...SCIENTISTS AND THEIR SOFTWARE

A survey of nearly 2,000 researchers showed how coding has become an important part of the research toolkit, but it also revealed some potential problems.

> **45%** said scientists spend more time today developing software than five years ago."

> **38%** of scientists spend at least one fifth of their time developing software.

> Only **47%** of scientists have a good understanding of software testing.

> Only **34%** of scientists think that formal training in developing software is important.

Computation is not my scientific focus. I will not use it frequently. I am a wet lab person

Data and text mining

Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades

Jonathan D. Wren^{1,2,*}

¹Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104-5005, USA, ²Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on January 27, 2016; revised on April 1, 2016; accepted on April 21, 2016

Abstract

Motivation: To analyze the relative proportion of bioinformatics papers and their non-bioinformatics counterparts in the top 20 most cited papers annually for the past two decades.

Results: When defining bioinformatics papers as encompassing both those that provide software for data analysis or methods underlying data analysis software, we find that over the past two decades, more than a third (34%) of the most cited papers in science were bioinformatics papers, which is approximately a 31-fold enrichment relative to the total number of bioinformatics papers published. More than half of the most cited papers during this span were bioinformatics papers.

Yet, the average 5-year JIF of top 20 bioinformatics papers was 7.7, whereas the average JIF for top 20 non-bioinformatics papers was 25.8, significantly higher ($P < 4.5 \times 10^{-29}$). The 20-year trend in the average JIF between the two groups suggests the gap does not appear to be significantly narrowing. For a sampling of the journals producing top papers, bioinformatics journals tended to have higher Gini coefficients, suggesting that development of novel bioinformatics resources may be somewhat 'hit or miss'. That is, relative to other fields, bioinformatics produces some programs that are extremely widely adopted and cited, yet there are fewer of intermediate success.

Contact: jdwren@gmail.com

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

NEXT

First Steps in R

Questions

