{last time recap} 15/05 DRD

<https://www.sciencedirect.com/science/article/pii/S0168952503002695?via%3Dihub>

Crucial points:

Some biases are strictly related to the technology of competitive 2 colors microarray

Raw data cannot be used as it is, it has to be transformed in log scale (log base 2)

log2 is important since you can manage the data properly and the software can help you.

the best tool to see the data is MAplot, (M=minus A=average of the logarithm of the fluorescence, the red and the green fluorescence).

If cy5 /cy3 =1, it means that the logarithm is zero

If the signal of Cy5 doubles that of Cy3, the ratio is 2 and its log is 1 .

So from the graph you can see the fold change between treated and control signal

so is very useful since you can see immediately if your gene stays up or down the 0 zone , it means that you have some kind of difference in the signal between the two, so with 1 array you can look at 2 different biological sample and you can see if the signal of your gene is far from the median or not.

When point align along a sloped line, something wrong happened inside the array. There are different procedures (depending also on the level of the problem, so if it’s a matter of linear relationship, it needs to be scaled by the median)

it is a very important tool to see the distribution of your data.

this kind of skill will be applied also to other problems.

problem of normalization procedure: always present for each kind of arrays.

A powerful experiment needs good calibration of the arrays. The best procedure to normalize arrays is to center the data and then use the so-called quantile normalization {see yesterday}

best procedure: centre the data and then use the **quantile normalization**, this is the only possibility to compare among different arrays. we look for each array the lowest intensity and we replace the lowest intensity with the average of the lowest intensity of all the arrays; we will do the same with the second lowest signal..etc...we are changing the intensity of the signal.

Check this reference also foR A VERY CLEAR EXPLANATION, AT LAST, OF WHY IT FREAKING WORKS: <https://academic.oup.com/bioinformatics/article/19/2/185/372664>

What is our biological question in this point? Usually, to understand if expression changes in a population for a specific problem in comparison with a control (so we have a case/control design).

--------------------------------------------------

---------------------------------------------------

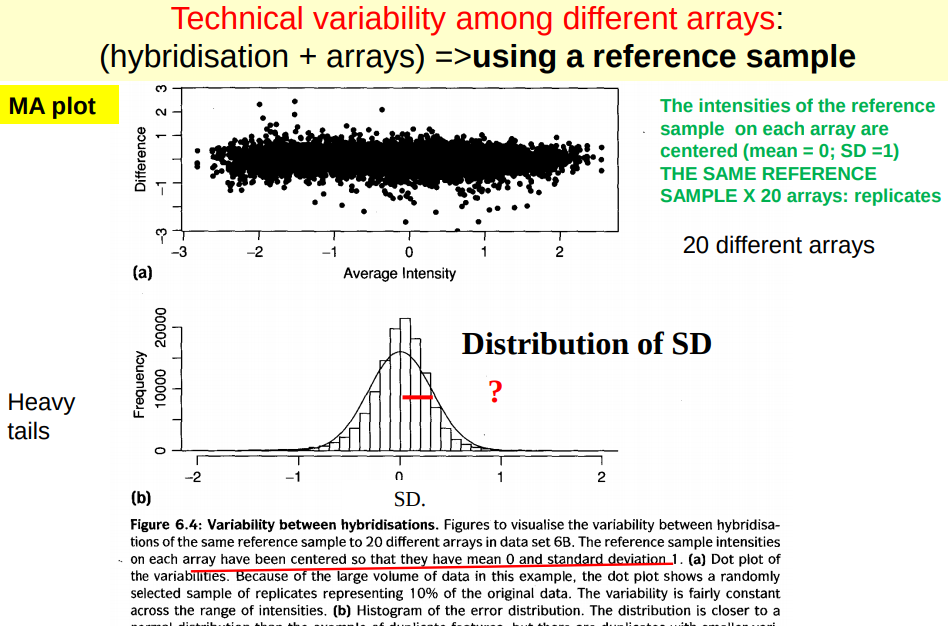
Check for technical variability :

in this case you need sample without biological variability.

here i need to perform replicates, that are the proper experimental design to look for technical variability. i’m using the same colors in different arrays: my task is to identify the coefficient of variation that is associated to my experiment.

one idea is to use this reference sample : we are comparing the same thing and my expectation is to find no differences among the different arrays.

MA plot: on y axes the log ratio and, on the x axes, the average of cy3 and cy5 in logarithm and apparently it appears a very good distribution



but there is a sort of variability

What is the coefficient of variation associated to my experiment?

I use the same reference sample among different replicates.

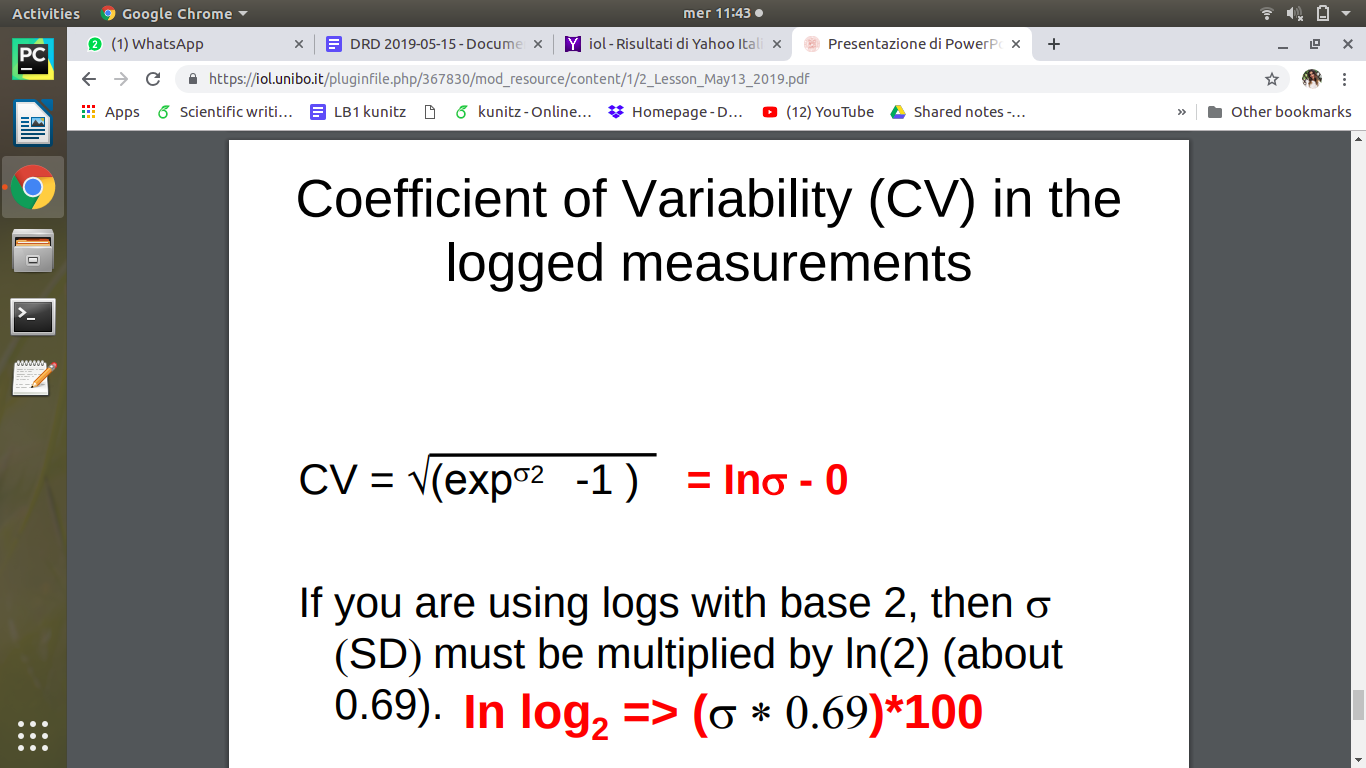
MA plot -> Y axis log (Cy5/Cy3); X axis log(average)

If we center the data (without quantile normalization, only mean = 0 and std = 1) we get a centered MA plot.

CV = (Std/mean)%

We can’t apply the formula in this case because we have log-normal distribution. So the proper formula is:

CV = sqrt(e^(σ2) - 1) = lnσ - 0

you have to multiply sigma to 0.69 to convert ln to log base 2

looking at the distribution of data you can estimate the standard deviation

coefficient of variability es =22%, is the quantity coming from technical variability.

what does it mean for the biological point of view? culo

we try to decrease the technical variability using other techniques like non competitive one color !

**Not competitive one-color arrays**

it does not exist competition since samples are separated in each array.

We have just one sample per array => no competition. We only have one fluorophore, so a lesser bias will come from this technique. This technology started in 1991, the creator was dr Stephen P.A. Fodor

The technique was setup in Affymax institute in California and the spin-off was named Affymetrix.

<https://www.thermofisher.com/it/en/home/life-science/microarray-analysis/affymetrix.html>

Affymetrix produced GeneChip (™)

genechip: is the trademark characterizing Affymetrix

almost all researchers interested in gene expression use affymetrix.

After the first burst and especially after NGS era, Affymetrix started to decline and unfortunately many employers had no work and the company was reduced. Now it has been acquired by ThermoFisher.

the genechip arrays are still used, they can be related on different type of organism or molecules :

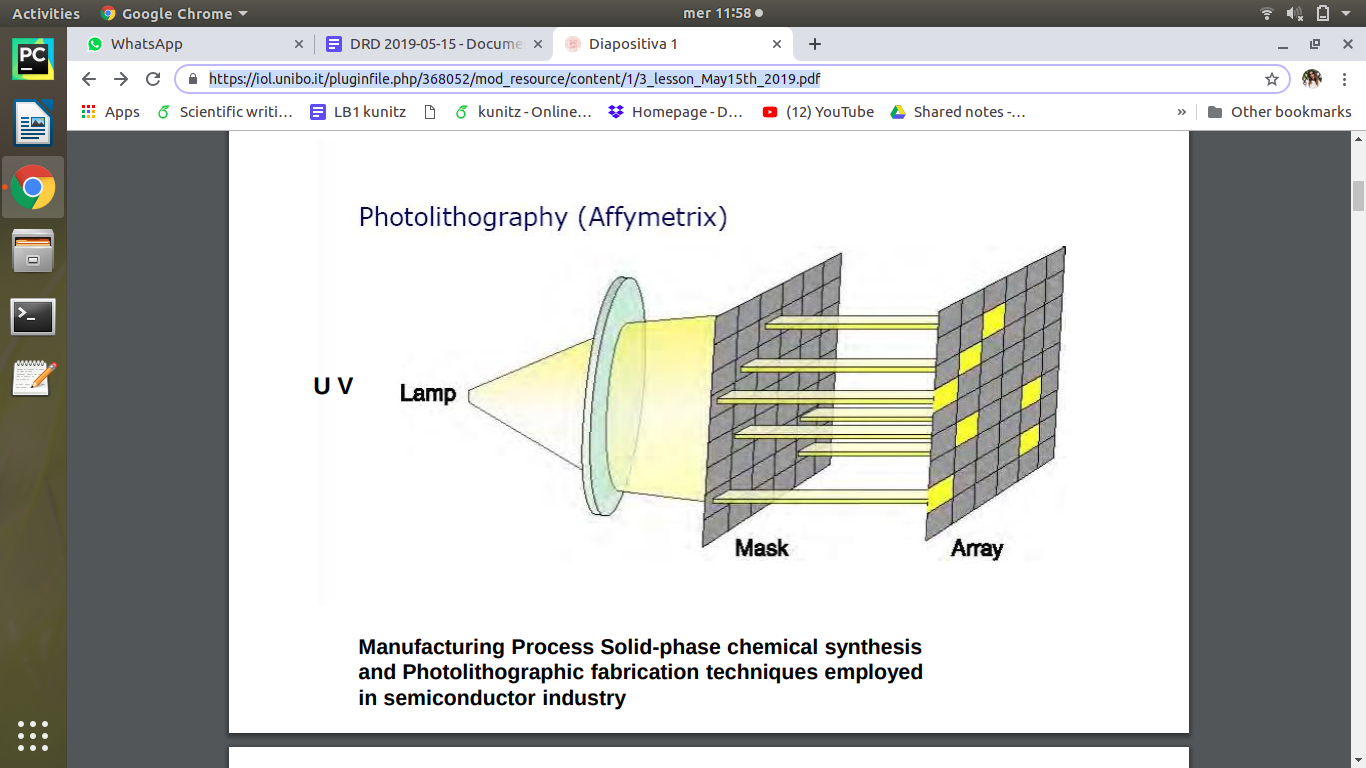
There is a good gene chip related to miRNA.

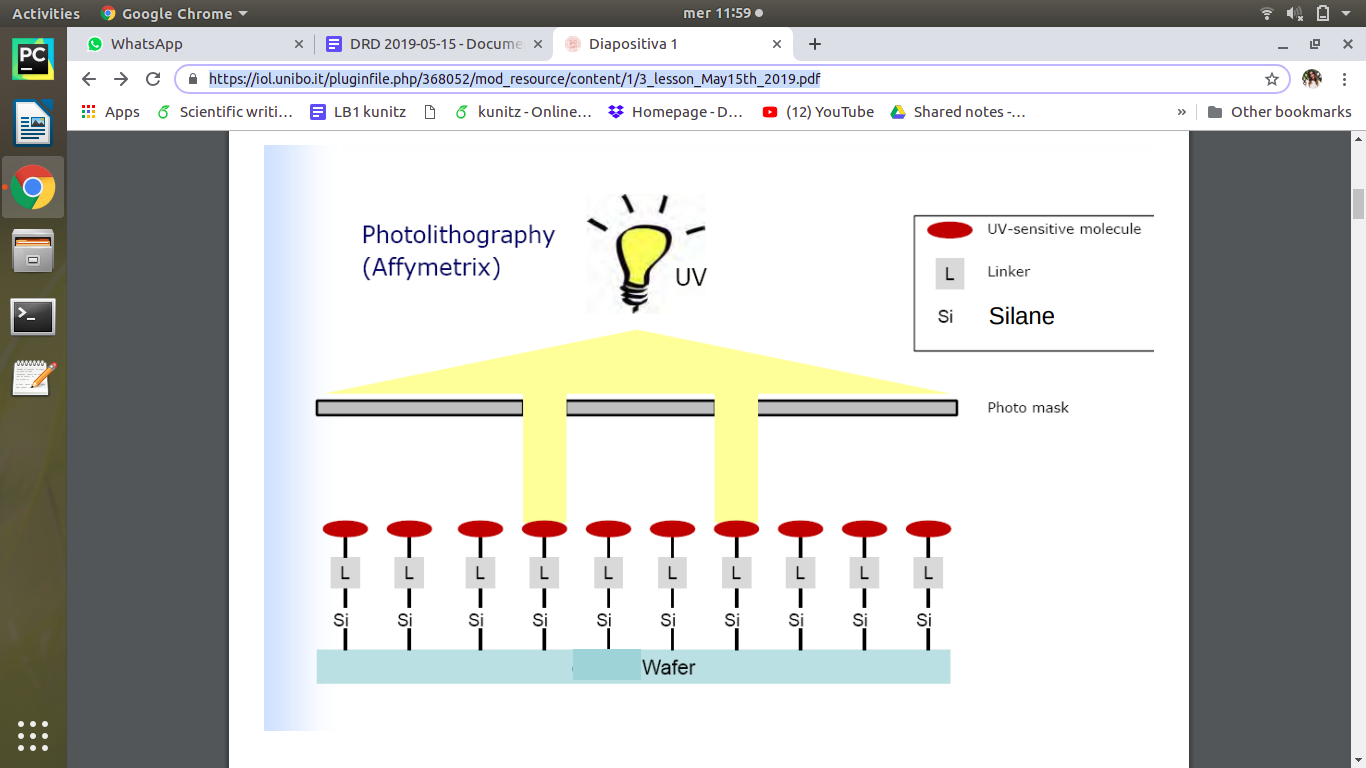
NGS Is the most used technique to obtain high quality data.

Non competitive array are characterized by photolithographic synthesis: **in situ** synthesis by photolithography

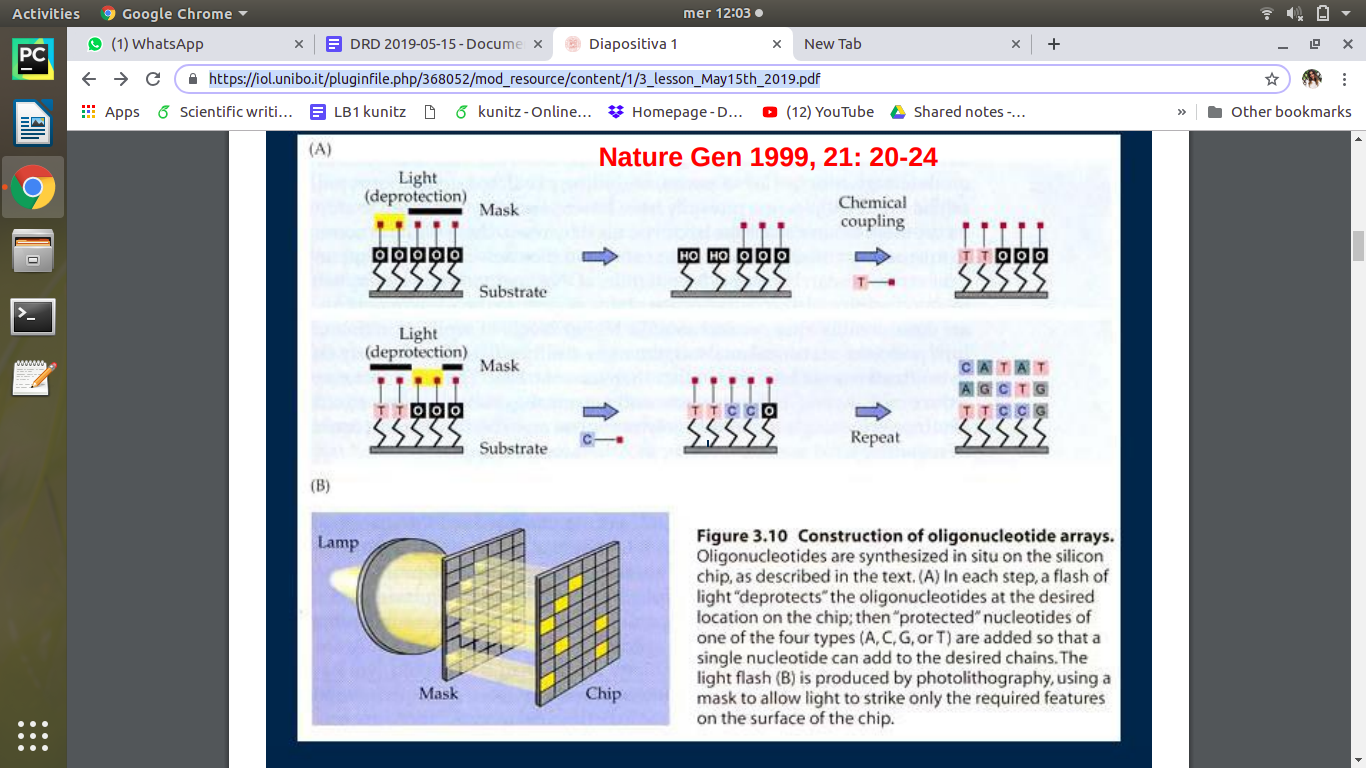
the type of fabrication is really different: photolithographic synthesis (in situ synthesis by a photolithography)

Photolithography means that you have a source of UV light, there is a kind of mask with holes that light can go through, on the array there are UV-sensitive molecules linked to the array by a layer of silane and a linker group which is usually very reactive. Light can erase the UV-sensitive molecule, so when you add nucleotides they can bind the linker.

the nucleotide growth one each other directly on the array using this mask that can be shifted producing



So in situ photolithography system means that nucleotide strands grow directly on the array by shifting the mask in order to enlight different positions to add the new nucleotides. (this is the way the array is built i.e. how the oligonucleotide probes are synthesized)



This kind of technique is completely different from the previous one which was based on spotting of the array.

REFERENCE: <https://www.nature.com/articles/ng0199supp_20>

the nucleotide probes are synthesized in a chain directly in the array.

Exam question -> culo differences between competitive and non-competitive.

*one* of the difference is related to the fabrication and the method to fabricate the different types of arrays.

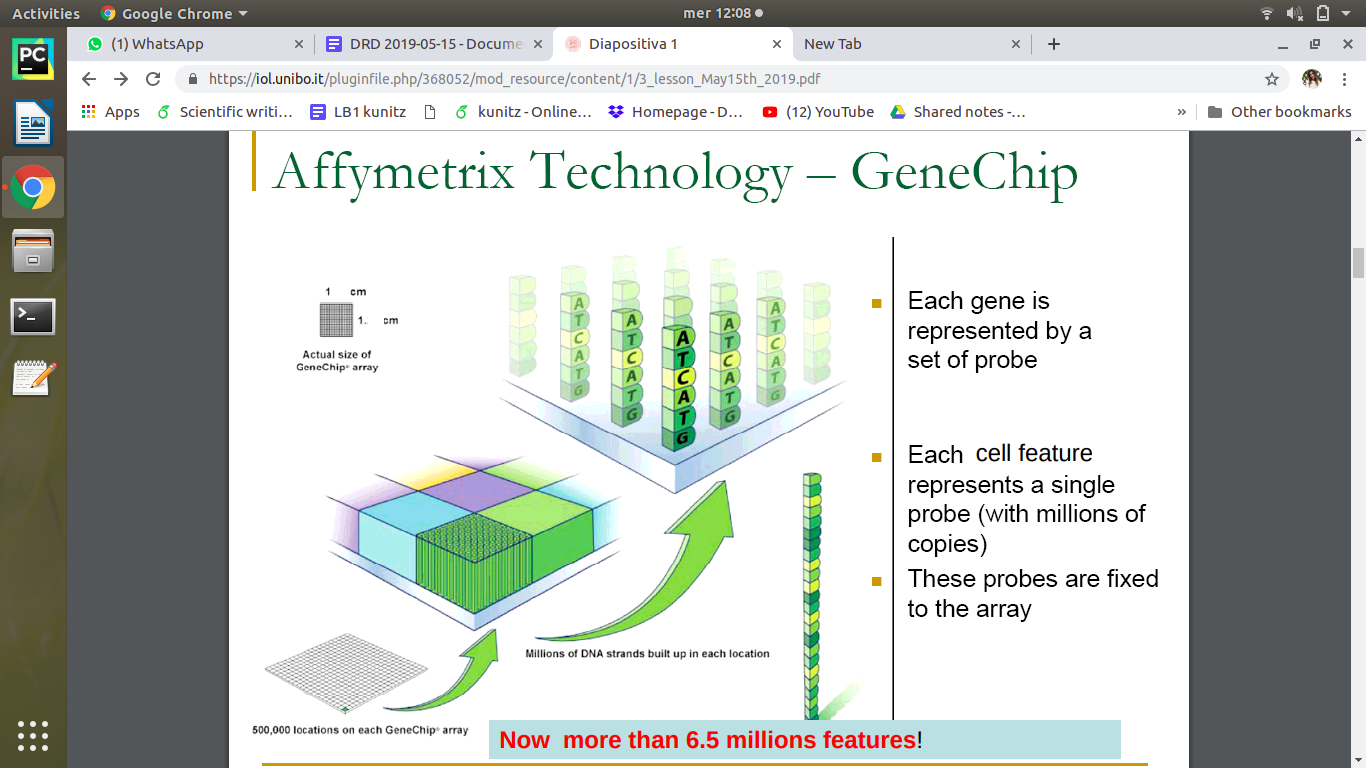
(spotting vs photolithography)

and now we see the other differences

The size of the chip is usually 1 cm^2, it contains millions of *features* or *cells*



Each feature is a single probe in many many copies.

each chip contain this kind of very small square (each one is named feature or cell ) and each of it contains nucleotide in situ synthesized. so the chain grows up for each single feature. and the length of each oligonucleotide depends on the genechip but usually is 20-25 bp.

this oligonucleotide are the probes in situ produced by photolithography.

GeneTitan microarray system -> integrate system for hybridization, washing and imaging.

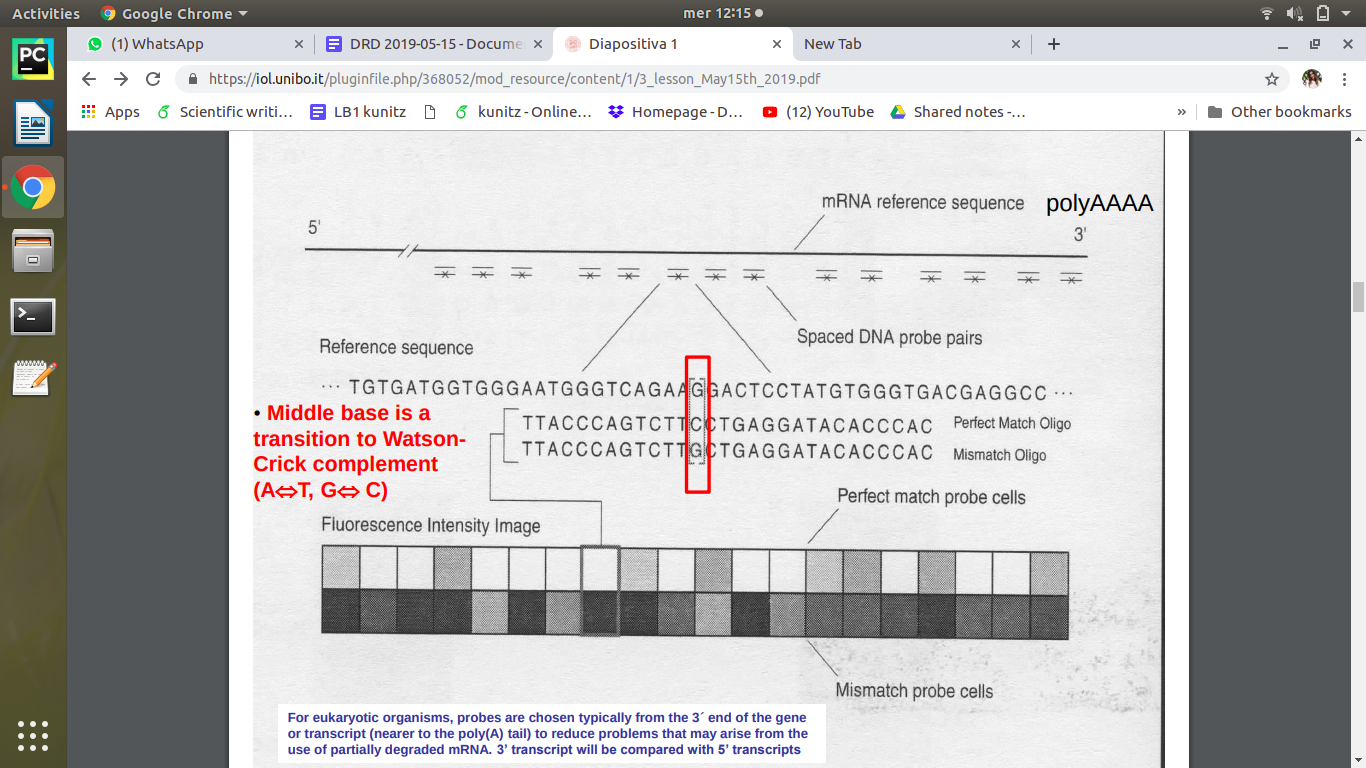
HumanGenome HG-U133

expression arrays for 1300000 DNA oligos (probing 47400 transcripts)

for this type of genechip the word pair means that you have 2 different probe (match or mismatch probe)

and another characteristic is that to identify one transcript you should use a probe set (not just a probe)

first strategy of affymetrix:

use relatively small probes but identify one transcript with a probe set. In the scheme if you look at the original sequence of the mRNA the strategy was to use more than 11 different probes at different positions for the same transcript. the probe set should give information on the transcription.

the probe named mismatch means that if you look at the reference sequence of mRNA , one probe has a perfect match with the sequence while in the mismatch we have the same sequence except in the middle where we have a transition of nucleotide (c->G as mismatch).

the basic strategy:

* identify a probe set for a transcript , not only the perfect match probes but also the mismatch probes where we have a different nucleotide that usually have a transition in the middle. why to use the mismatch? since i can catch the background signal , the aspecific signal that can be related to each perfect match probe.
* the idea is: i can look at the match and the mismatch and i can say that the mismatch can be the background signal , BUT THIS IDEA WAS WRONG since these mismatch often give high positive signal. the background signal had an higher signal than the signal itself.
* so affymetrix decide to produce an algorithm to change the result in order to have a NOT negative result every time (when the mismatch is higher than the perfect match it means that your signal is not significant

in each feature, identify a transcript with a probe *set*. The strategy was to produce oligos probing different regions of the same transcripts.

The probe named match is exactly complementary to the reference sequence, the mismatch probe instead has one nucleotide transition at some point of the probe.

The strategy is to identify a probe set for each transcript (a certain number of different probes for a transcript) and for each probe in the probe set, make also a mismatch probe.

(reminder that transitions are A <--> T and C <--> G). The idea is that I can catch the background, aspecific signal that can be related to each perfect match probe.

After the hybridization, when e obtain the final image of the array, I can look at the perfect match and the mismatch, and I assume that the mismatch represents the background signal to subtract from the perfect match.

Unfortunately, this idea was W R O N G. (eh sì, è proprio wronga)

The mismatch very often overestimated the background (very often the problems raised because what should be the background signal was higher than the actual signal). Everything started with this wrong idea. Immediately after, Affymetrix decided to construct an algo to change the results in order to obtain everytime non-negative results (identifying a threshold, the perfect match was subtracted by this threshold when the mismatch was greater than the perfect match). What happened was that you always had positive results aka usable data, but they were not significant.

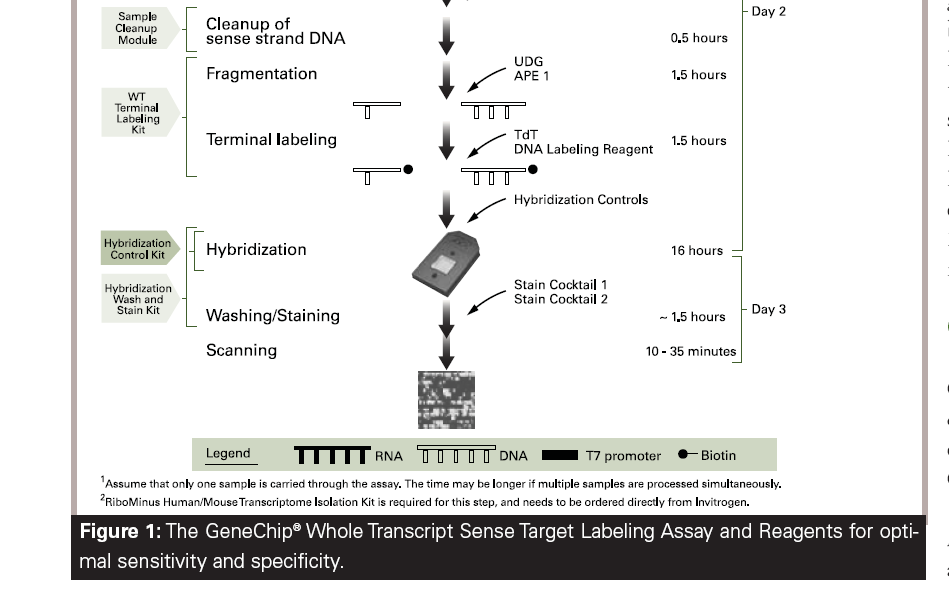
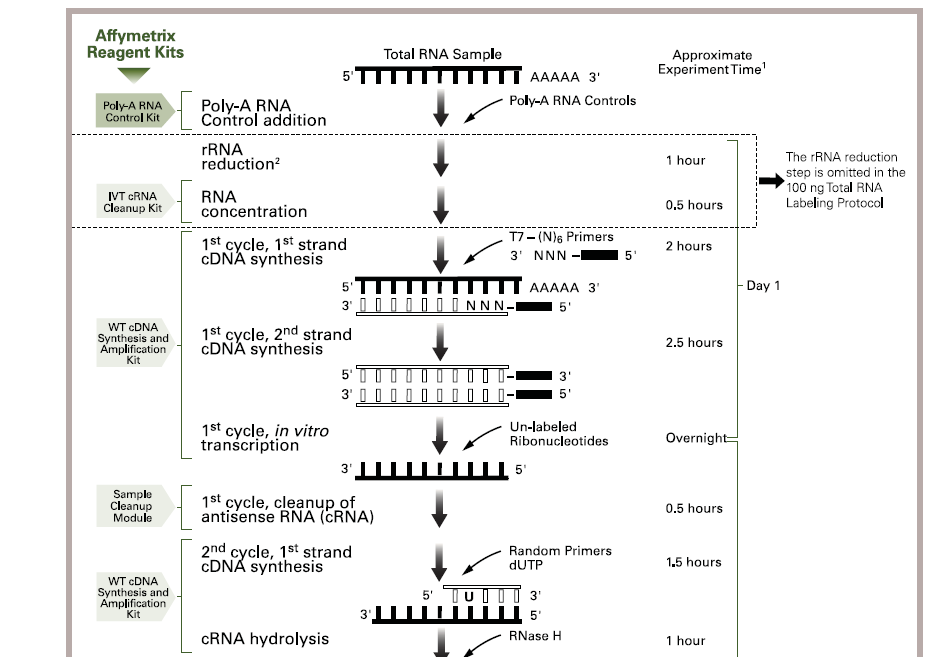
Finally, affymetrix produced chips without the mismatch probe because they realized it was complete bullshit. When you download data you can see if there are match and mismatch probes (for example if you take data produced using an old chip). Nowadays mismatch probes are not used anymore.

The idea was that the type of gene chips produced are now without mismatch probes (e due). Why increased fluorescence for the mismatch probes? Nobody ever answered to the problem of why the fluorescence would go up for the mismatches.

In nowadays chips, there are control probes for the calculation of background signal.

The aim of Affymetrix originally was to measure **absolute** gene expression with just one chip (hence the match/mismatch probes)

Steps of the protocol: start from total RNA; erase rRNA; retrotranscription => first complementary DNA; obtaining ds cDNA; obtaining the cRNA (obtained using unlabelled ribonucleotides, so it will give the possibility to obtain exclusively the *sense* complementary DNA. The *sense* cDNA is the target for the gene chip and we get rid of the *antisense* .

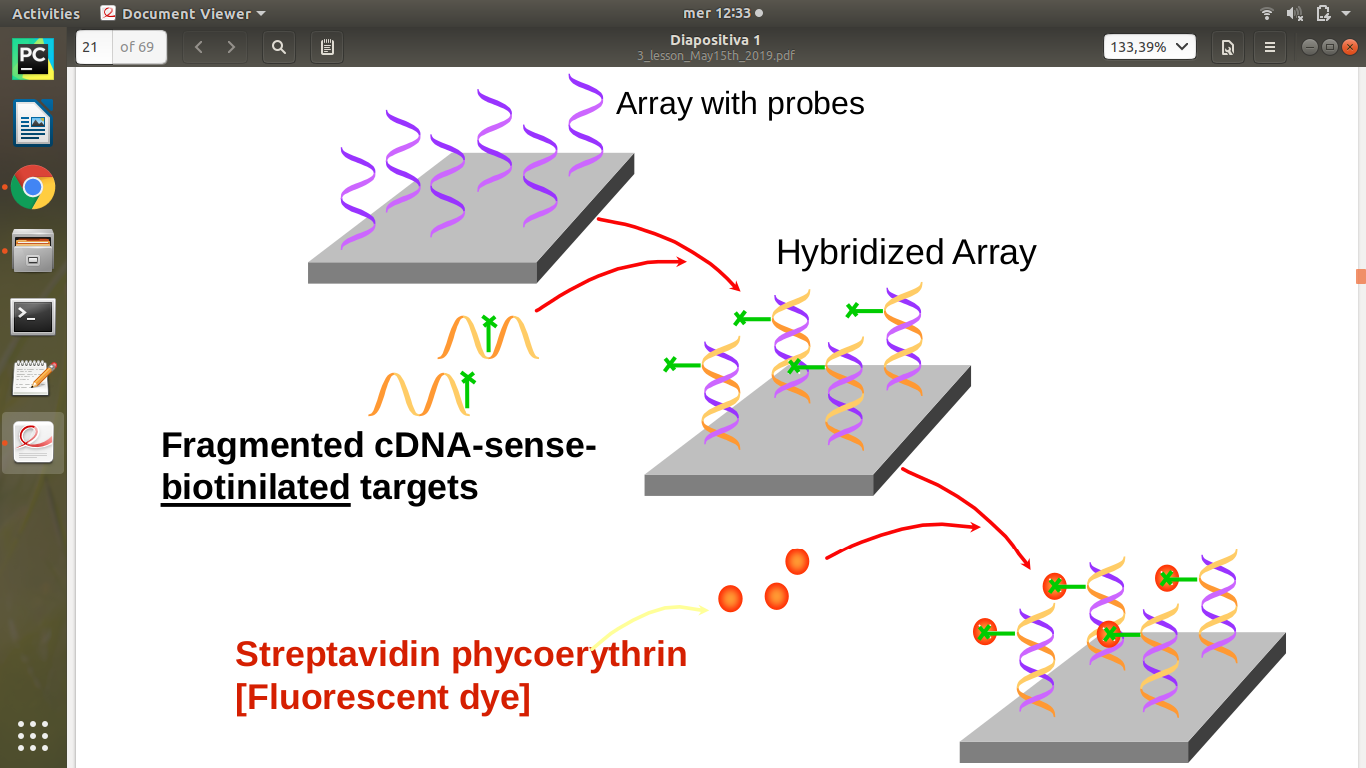


Fragmentation of the cDNA -> performed using a specific enzyme that cuts near uracil residues. The terminal nucleotides are modified with biotin.

wow never heard of this mysterious biotin molecule

Terminal deoxynucleotidyl transferase(Tdt)

The sample is added to the gene chip for hybridization (step that lasts 16h), then the chip is washed and stained (last step). The staining works because the fluorochrome is conjugated with streptavidin.

This protocol means three days of work. Phycoerythrin is the fluorophore, it emits light in the red.

each single array has one sample and one colour.

QA:

Visual inspection of image

• Control positive probe sets: checkerboard,edge and array name cells are all o.k.

• Quality of features: discrete squares with pixels of slightly varying intensity

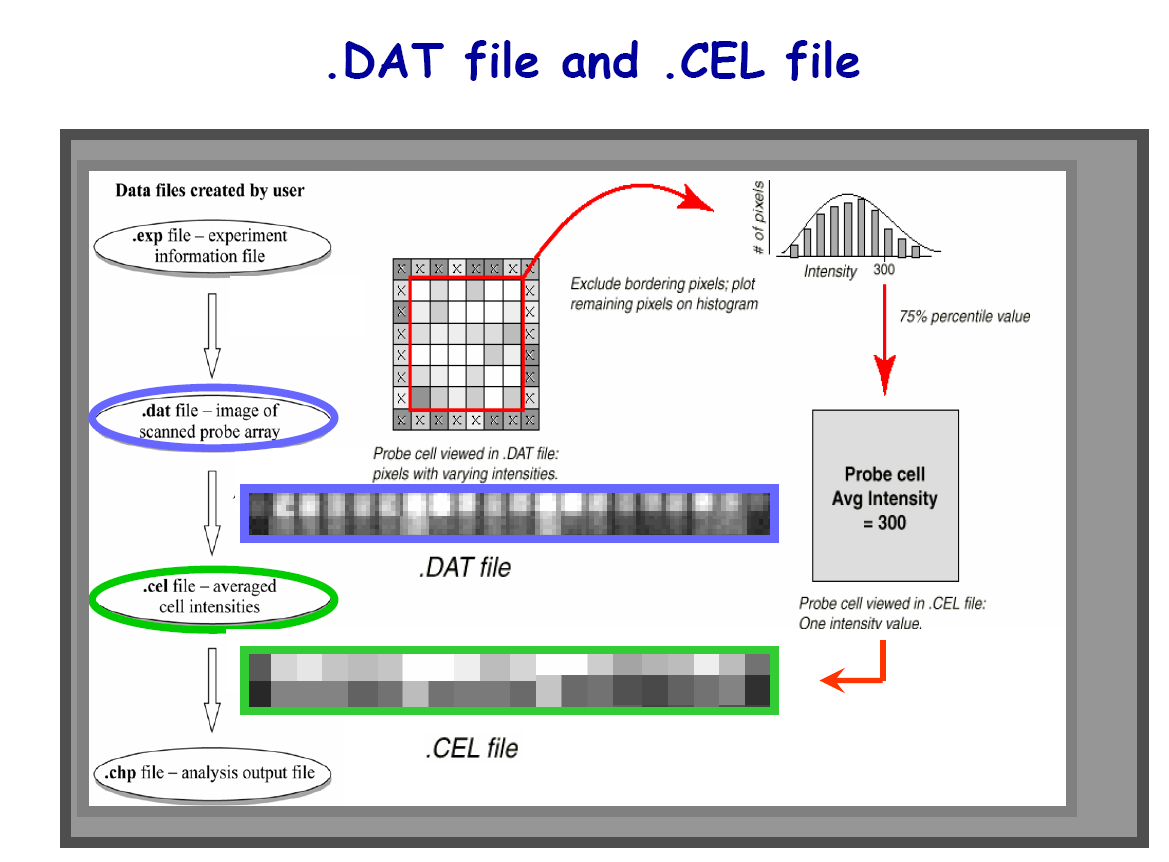
• Grid alignment (the grid is superposed to the image with a software and it identifies each feature; only the central 8x8 pixels are used for signal)

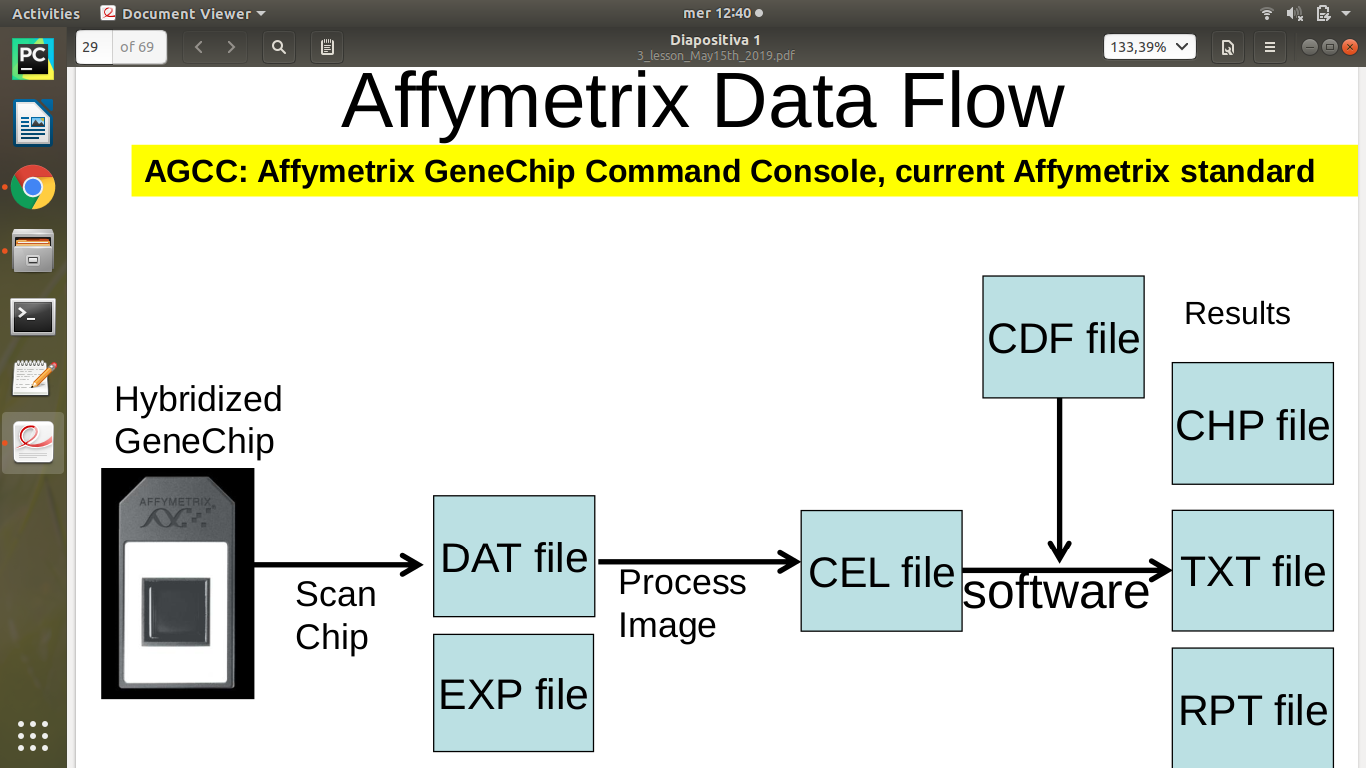
• General inspection: scratches, bright SAPE (streptavidin phycoerythrin) residue (masked

out)

we can not only use a reduced number of pixels inside each feature but also use a lower percentage of intensity signal??? using pixels whose intensity is above the 75° percentile

The data files in output are the .dat file (all pixels) and the .cel file (final processing of the image given by reduced number of pixels for each feature and averaging the pixel intensities for each feature).





CDF file contain the exact sequence of the probe for each feature.

highly informative on the probe you are managing.

So CEL file and CDF file are crucial

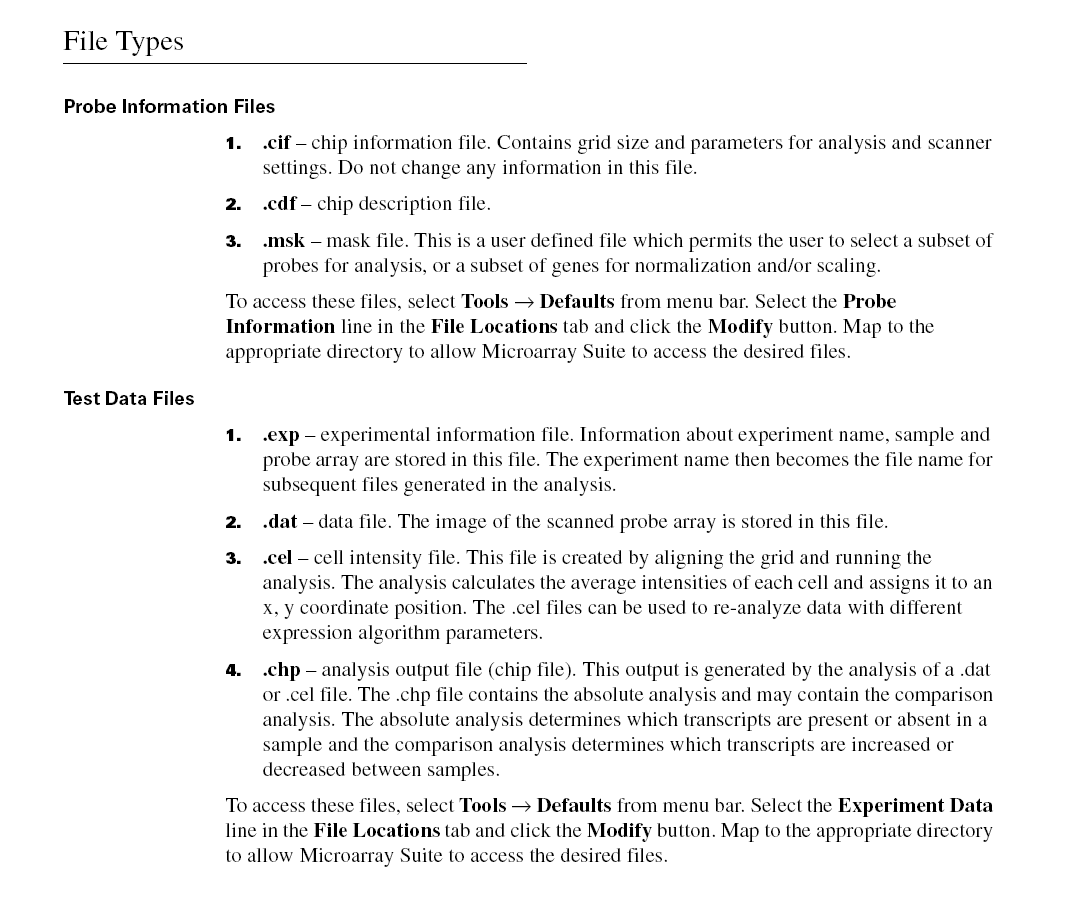
The CDF file is provided when you buy the gene chip. When you process the cel and cdf files together you get other results in specific formats.

when you want to analyse this kind of data you have to receive a CEL file.

Anyway in repositories there are **CEL files**.

Possible question → what is a cel file? and we are expected to know what it means

CEL file contains the information, the signal intensity for each probe after processing of the image, after the scansion.

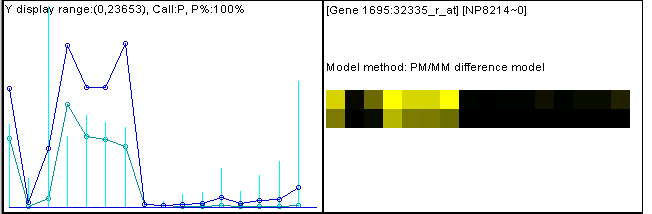
Affymetrix claims that you can get the results with just one chip but the rest of the world doesn’t agree.

Question → difference between CEL file and DAT file

The problem of affymetrix genechip is that of different probes for the same transcript! Many probes for the same transcript arises some questions. If I look at the same probe set (we take for granted that we speak only about perfect-match probes), some probes have an intensity and other probes have a very different one, although all probes should refer to the same transcript.

if i look at the same probe set (perfect match) , the perfect match (blue line in the graph) should give information about the transcript. it means that if you look at the probe set, some probes have high level of intensity and others have not.

the sequence is not the same! but this different probes in the probe set should give the same information, but how is it possible if the level of intensity is so different?



The problem is that the probes have different sequences, in particular a different %GC which means a different ease in hybridization. A very interesting experiment showed that different probes had different affinities to the same transcript.

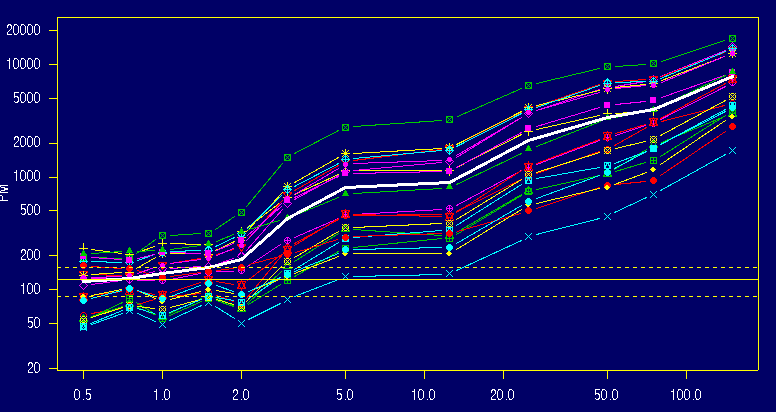
X axis -> log(concentration) of a target ([spike-in](https://en.wikipedia.org/wiki/RNA_spike-in)) you are testing one messenger with different concentration.

Y axis -> log(intensity)

for each probe set you will have different probes.

one probe set should give you the information about the signal intensity. look at the different level of signal intensity for each probe:

increasing the concentration , the probe green is at the same intensity but if you compare it with the other probe inside the SAME ARRAY, this difference is bigger than comparing for different concentration (that means different arrays)

inside each probe set we have different level of signal intensity, this variability is much bigger than the variability looking across the arrays.

give result considering just one array is not possible but if you improve your experimental design using more than just one array maybe you can catch the expression level

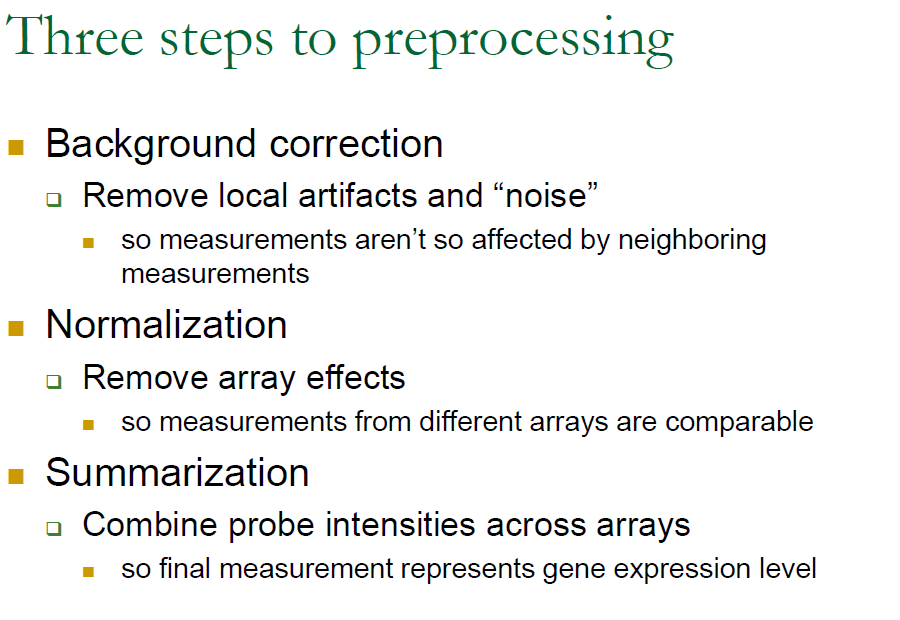
The key issue here is that variation of a specific probe across multiple arrays could be considerably smaller than the variance across probes within a probe set.

The difference of signal inside each array is greater than the difference of signal across arrays. (in other words, if the variation of intensity within the array at a concentration x overrides the variation of intensity observed for two arrays with a slightly different concentration, it means that in an array where I don’t know the concentration, I have a very big uncertainty about the quantity of RNA that I am observing.)

For this reason a procedure of normalization is needed , using more arrays! the bigger is the number of arrays the best is the result.

8 different arrays for 8 different concentrations.

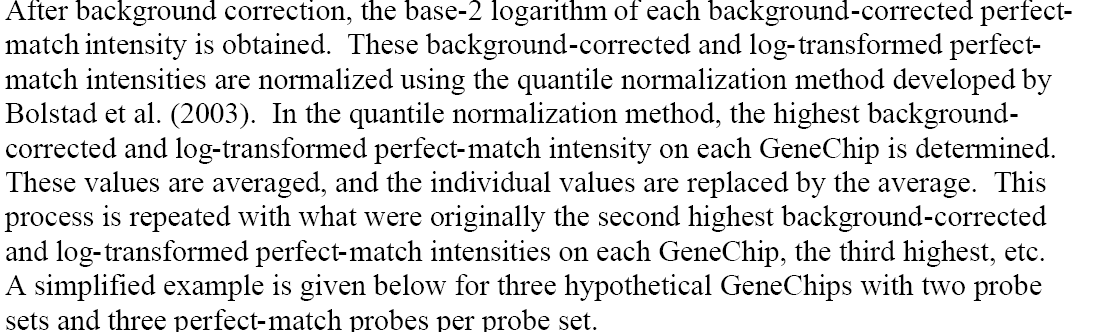
guideline: you have to background correct the array, normalize the signal intensity and identify a method that summarize the signal intensity.



the idea was : in order to obtain the best result we have to use a multiple array experimental design (RMA: robust multichip average analysis) by Bolstat.

REFERENCE: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC150247/>

suggestion: use the software for background correction, use data related to log2, adopt the quantile normalization.



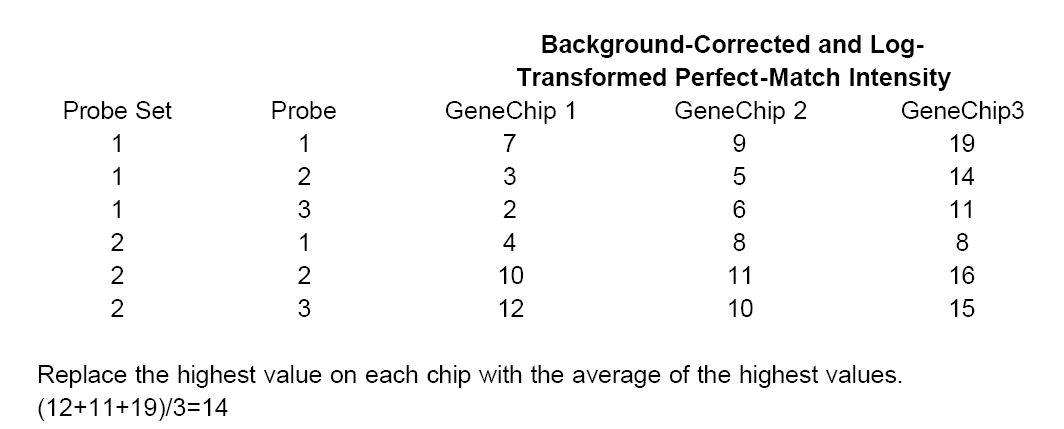
* **background correction**
* **normalization of signal intensity**
* **summary of probe set signal intensities** (the different intensities given by probes in a probe set need to be summarized in one number)

In order to obtain the best result, a multiple array experimental design is necessary (RMA - Robust Multichip Average Analysis. Bolstad, 2003; Irizarry, 2003)

The idea is to adopt the quantile normalization

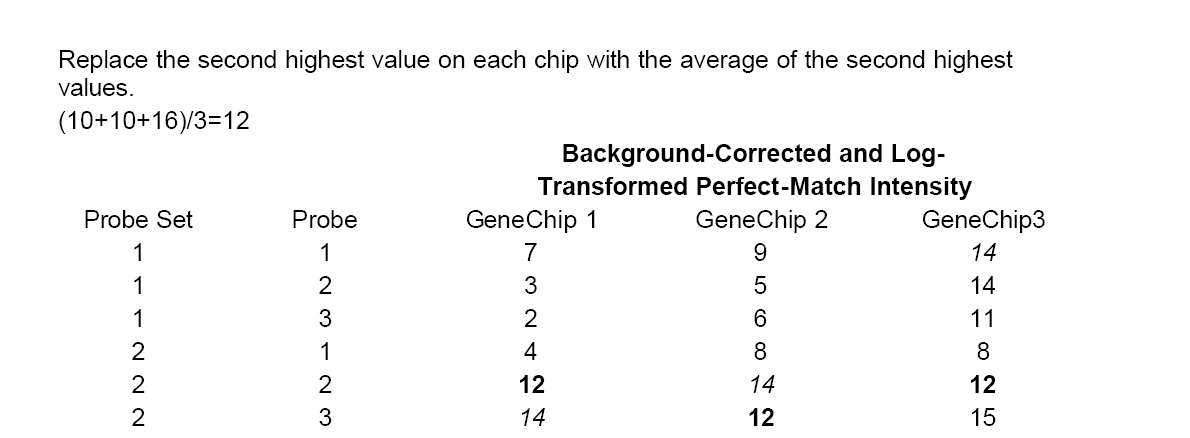
probe set 1 and 2 give information about 2 different transcript, for each probe set here you have 3 different probes, and we have 3 different genechip.

the idea is to start from the highest value (instead of the lowest in this case) and you can manage the quantile normalization

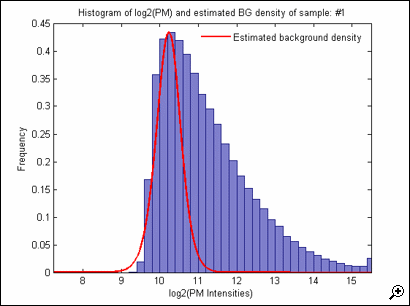


the 3 highest values are replaced by 14 (the average highest value form the 3 different gene chip).

then you have to select the second highest value.



and you go on until you complete the all values you have from the highest to the lowest…

at the end the procedure of quantile normalization is completed. now if you look at the distribution of your probe signal intensity. you obtain this distribution of probe signal intensity after quantile normalization.

you have many probes with low intensity and not so many probes with high intensity.

Now we have to find the final value for the summarization of the probe inside each probeset.

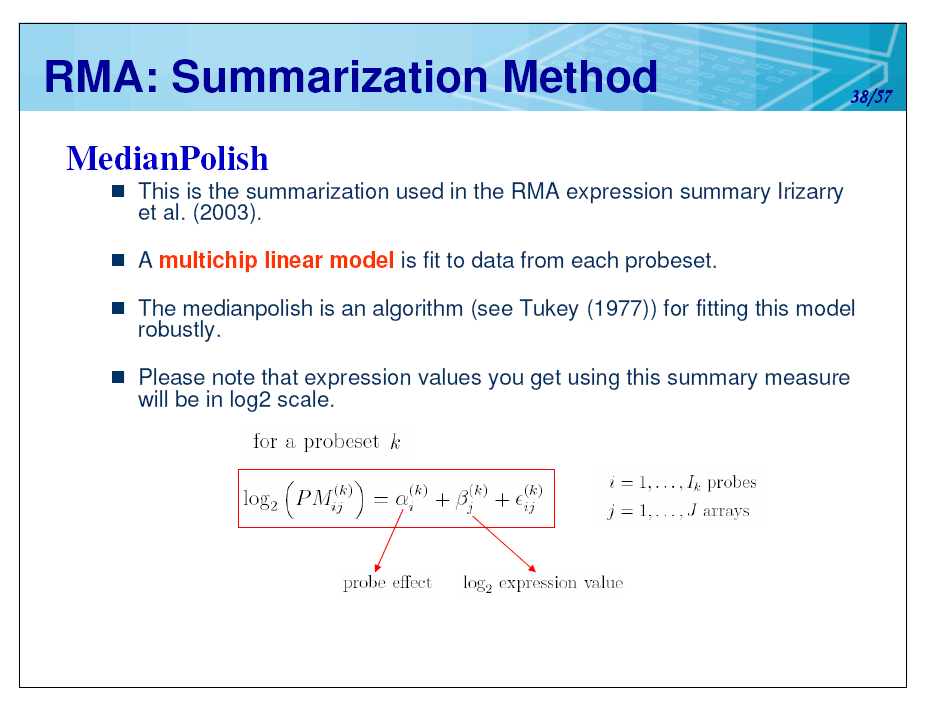
and from this probeset we should derive a number (the real signal intensity )that should give you a quantification of the intensity of the signal.

We have three gene chips to normalize. In principle we should be able to say now what happens. We look at chip values from the highest to the lowest; the three highest values will be replaced by the average of the highest values.

Then we do the same with the second highest value, the third highest value etc etc etc.

The distribution obtained after quantile normalization is similar to (it looks like a gamma function to me).

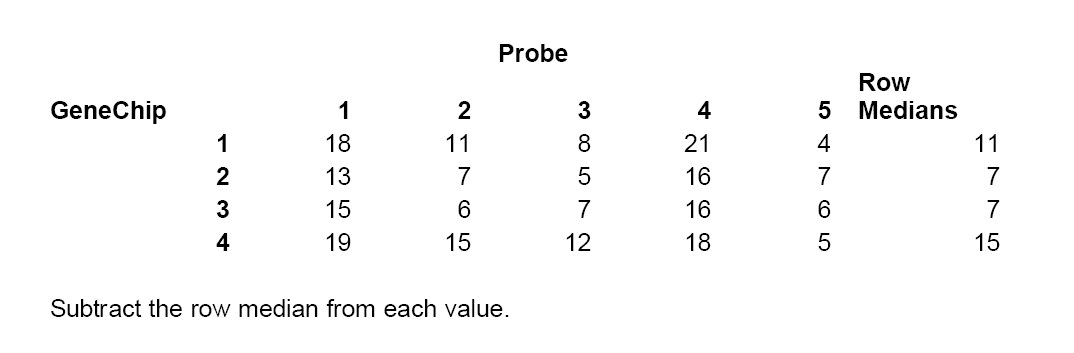
In the RMA, the final step is the summarization method which is based on a specific modeling, assuming that each probe signal intensity has an additive quantity of signal from the affinity of the probe to the target. we should take into account a possible variability that origin from the sequence.

additive model:

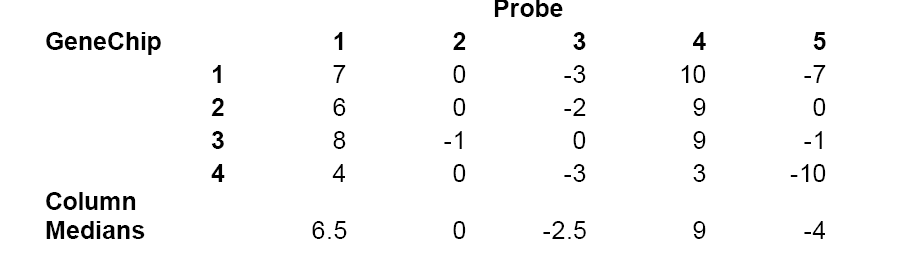
each probe has an intensity that come from the 2 intensity plus an additive quantity derived from the affinity of the probe with its target plus an error quantity.

the [median polish](https://en.wikipedia.org/wiki/Median_polish) is an algorithm that should identify the affinity in order to obtain the true signal intensity.

This algorithm should determine the affinity term to be added to the intensity value; in this way the probe set can be averaged to obtain the true signal intensity.

Applying the **Tukey median polish algorithm**, the median of each row is subtracted from each value in the row.

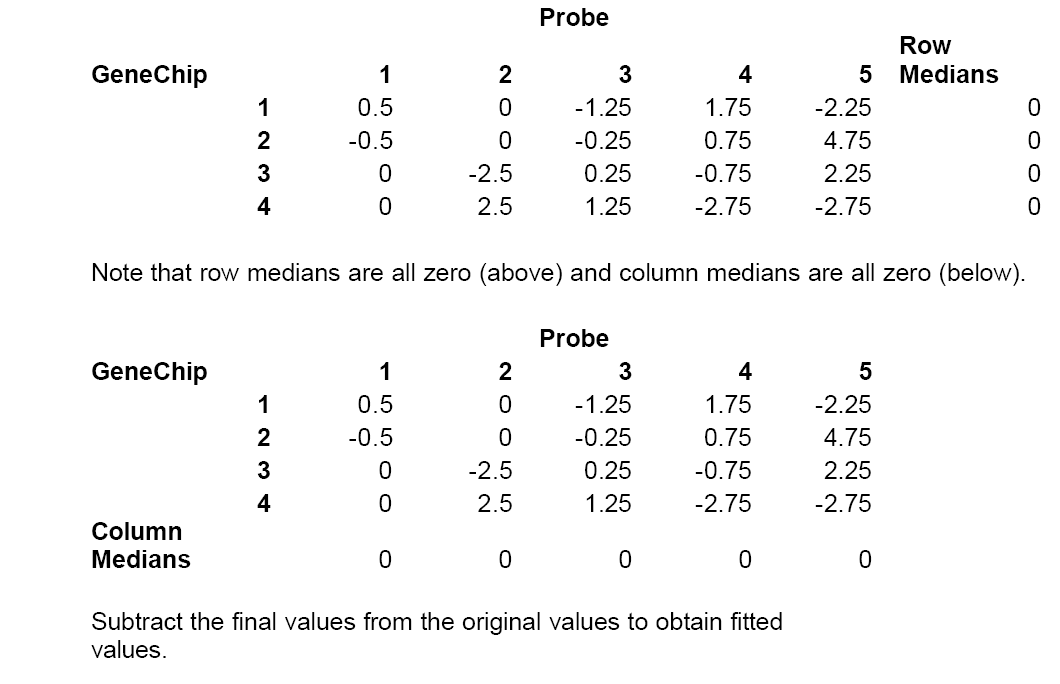
then identify the column median for each probe in different genechip, and you have to subtract the median from each column.



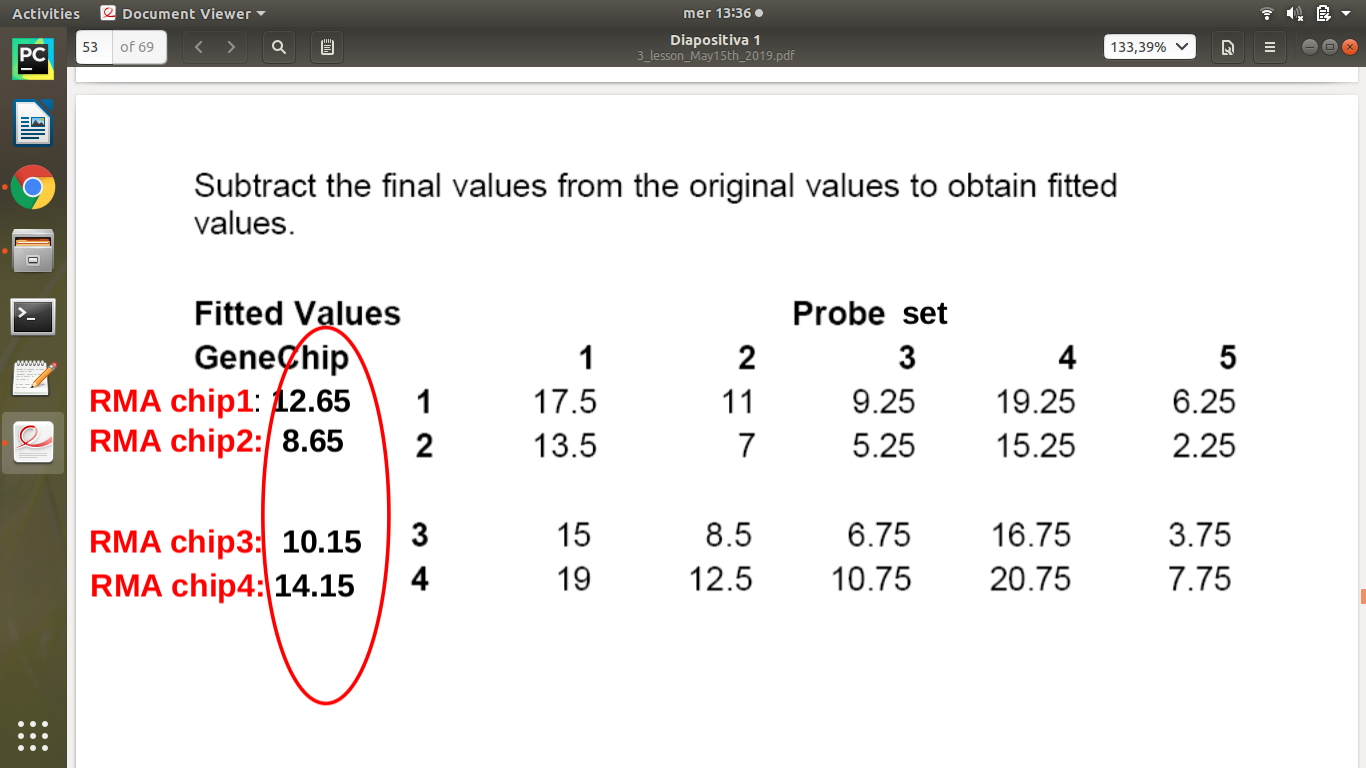
This algorithm is iterative and it goes on until you obtain median 0 in the rows and median 0 in the columns. (for all rows and columns)

until you have all the rows and columns median equal to zero.

and at the end you find that the probes have a final kind of ‘residuals’.



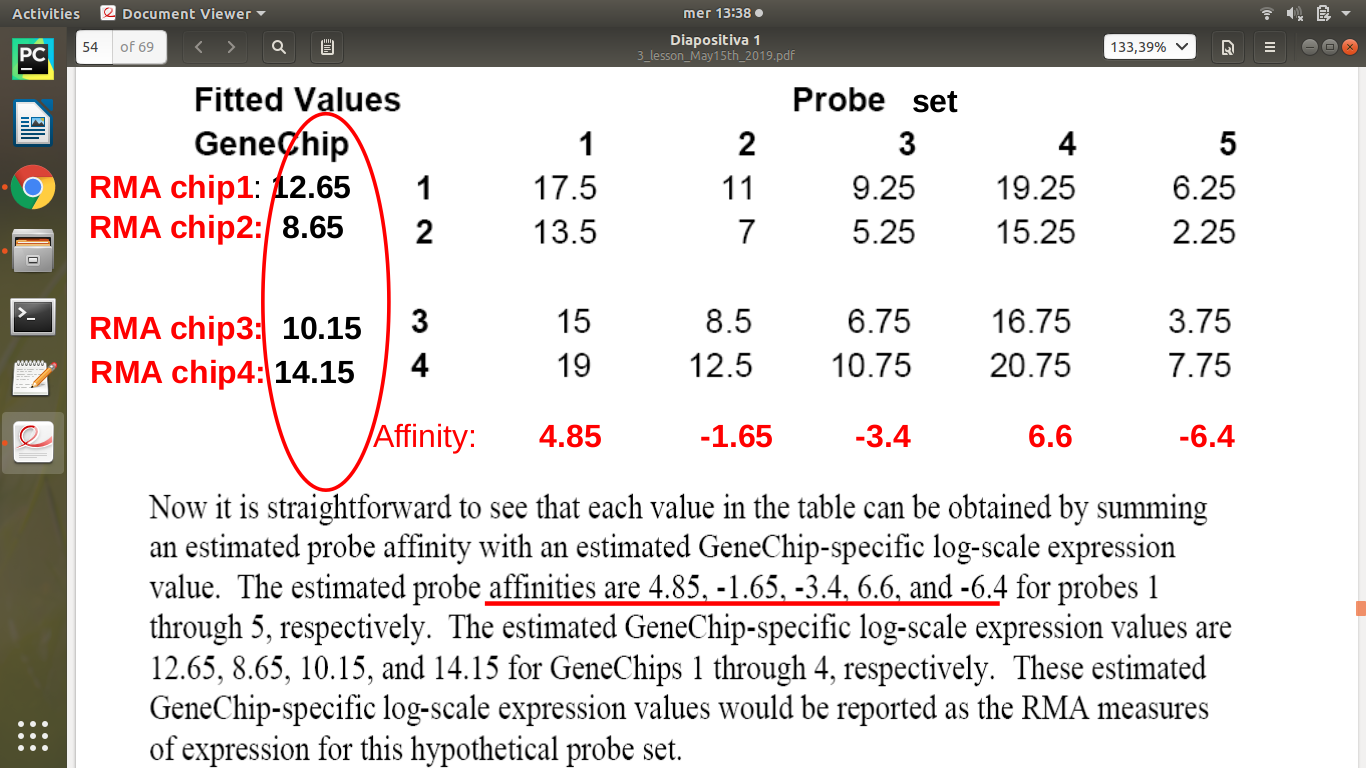
The residual values at the end of the algorithm are subtracted to the original data! So in the end the probe set is normalized and by averaging the values of the probe set, you get a trustworthy average of the probe set.



Average for each probe set => the signal value of same transcript

But what about the affinity? the model said that we have the signal plus an affinity value of the probe for the target.

the affinity is calculated , the subtraction of the values, gives us the affiniy of the probes inside the probeset:



The affinity is calculated on the basis of the model. It’s computed thinking about the subtraction (???)

The aim of the whole procedure is to identify the component of signal given by the affinity.

Example -> take column 1 and subtract the mean values for the probes:

17.5 - 12.65 = 4.85

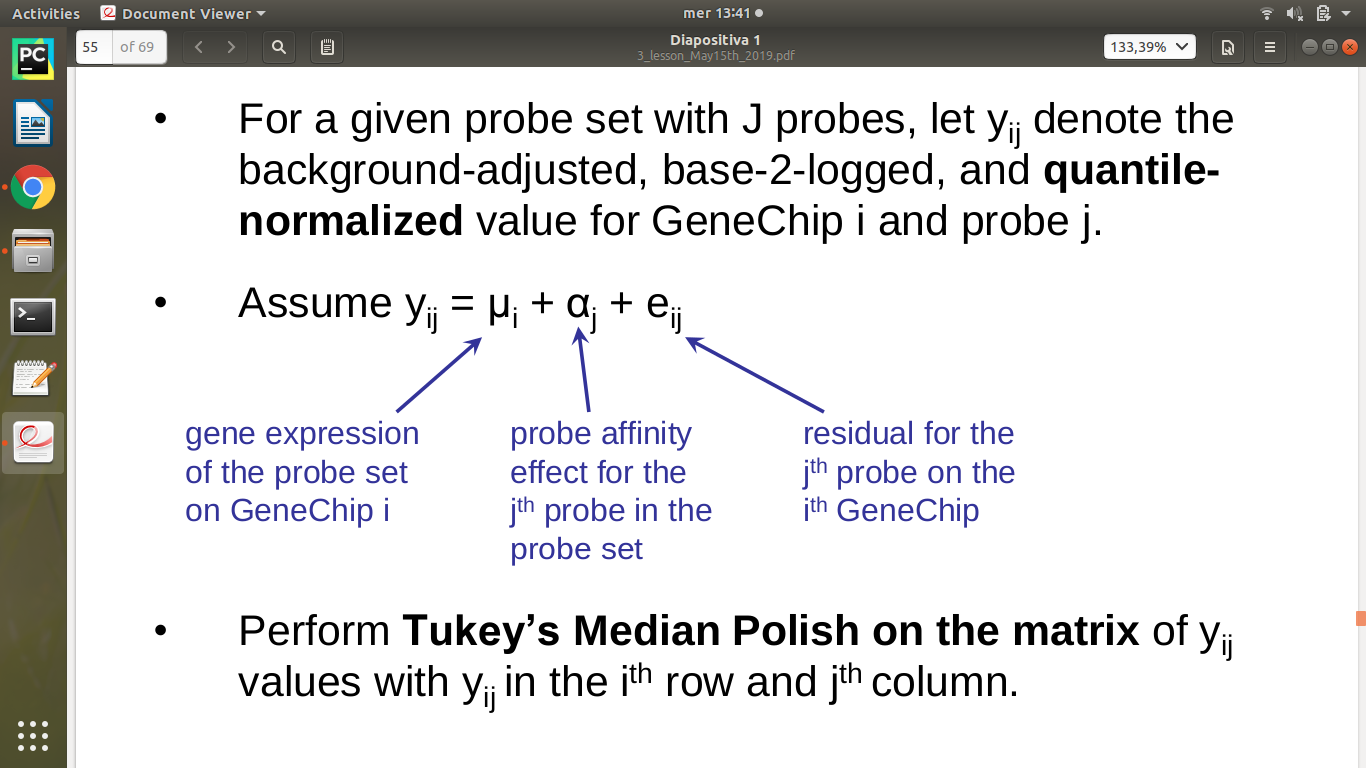
13.5 - 8.65 = 4.85

15 - 10.15 = 4.85

19 - 14.15 = 4.85

and 4.85 is exactly the affinity. This holds for all the other columns. Peccato che la prof non ce lo spieghi, come non ci spiega perché questa cosa funziona.

RMA give you the final result based on a model: for each probe you have a certain signal with an affinity. and we can compute the single probe affinity by this subtraction .



**Methods for computing expression**

• **Affymetrix MicroArray *Suite MAS*:** v.5

– robust average of probes on one chip; it does a reasonable job on probe sets that are bright

**• (multi-chip) methods**

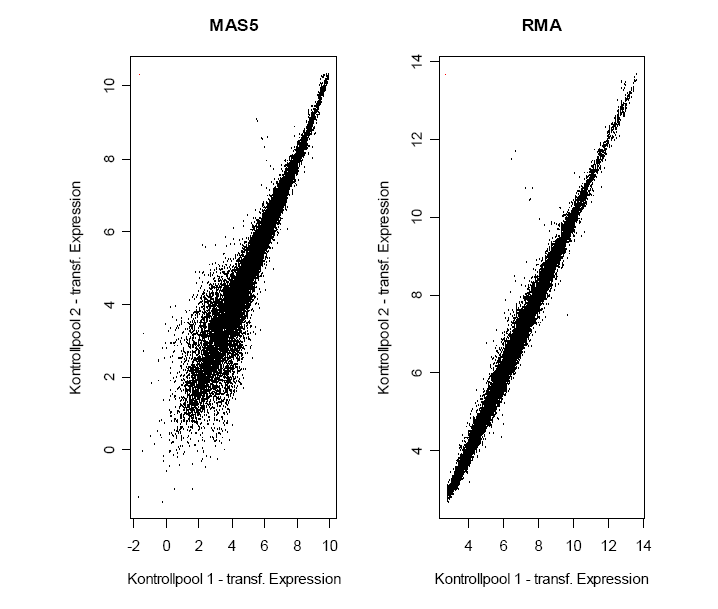
– RMA- Robust *Multichip Average*: use the original raw values as given in *CEL.files*

(Bioconductor affy package); it do better on genes that are less abundant

• Irizarray et al. (2003) Biostatistics 4, 249-264

• Many others published

if we compare these 2 methods we can see some differences:

we are comparing 2 RNA pool:

if everything works well we should have the same results for both.

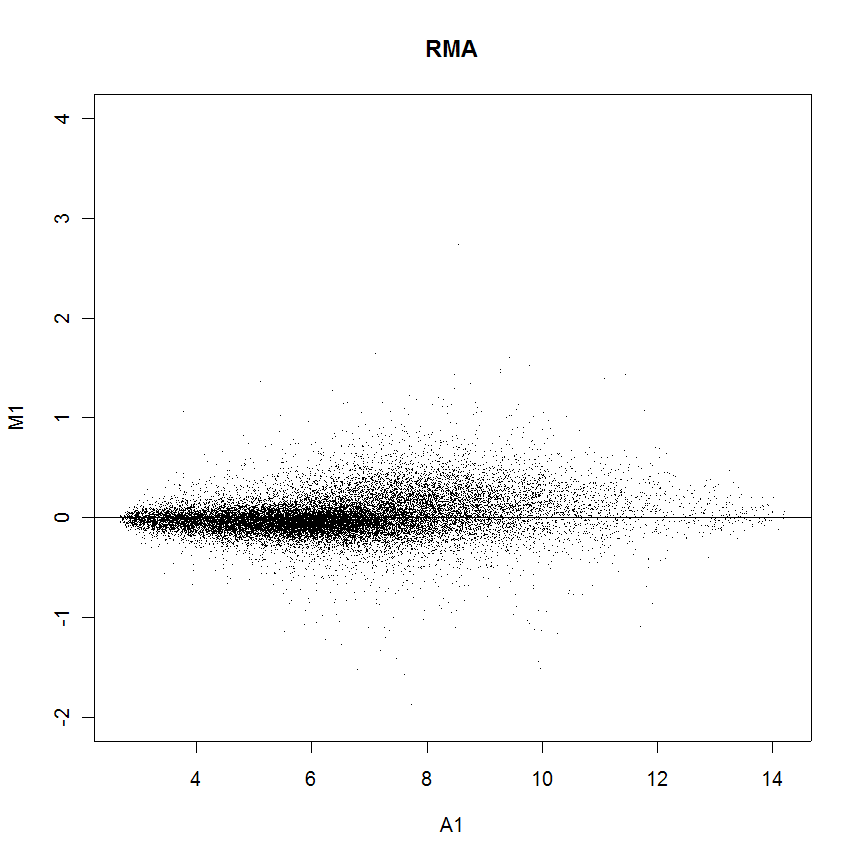
in RMA the result seems to be very good since dots are fitting , and the variability is apparently very small,

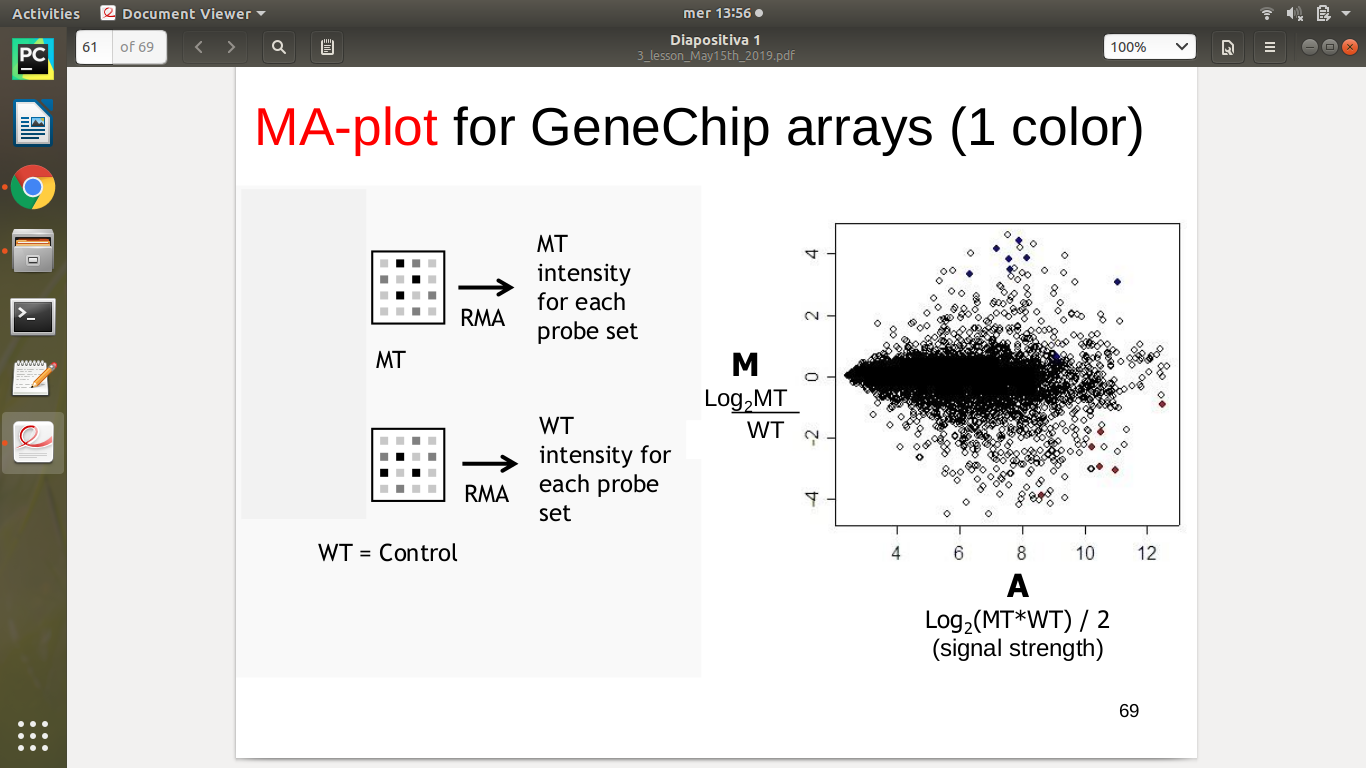
but in mas5 the variability is very large.

if i use an MA plot? could it be easier to see the the differences?

here we have just one fluorescence, so how can i manage this plot? i can have fluorescence from 2 different samples ! so i can merge data and produce this kind of graph.

the best result is from RMA compared with affimatrix software (mas5). at low intensity mas5 doesn’t work very well, but it works better at high intencity.





e comunque non ha capito che non vogliamo che ci ripeta quello che ci ha già detto ma vogliamo della statistica vera

queste lezioni sono il motivo per il quale siamo lo zimbello degli statistici

ne parliamo a martelli? e se sì cosa gli diciamo/come glielo diciamo?