17/05

Could use some revising as it was solo work & some parts might be missing

Some keywords

Affimetrix→ background collection, data transformation with log, quantile normalization,

looking at the same fluorescence but at different arrays.

affymetrix probeset name can have differents suffixes:

\_at: this probeset hybridize to unique transcript for this chip

\_s\_at: it means that this probeset can be shared and cross hybridize to set of sequences

\_a\_at : means alternative transcript

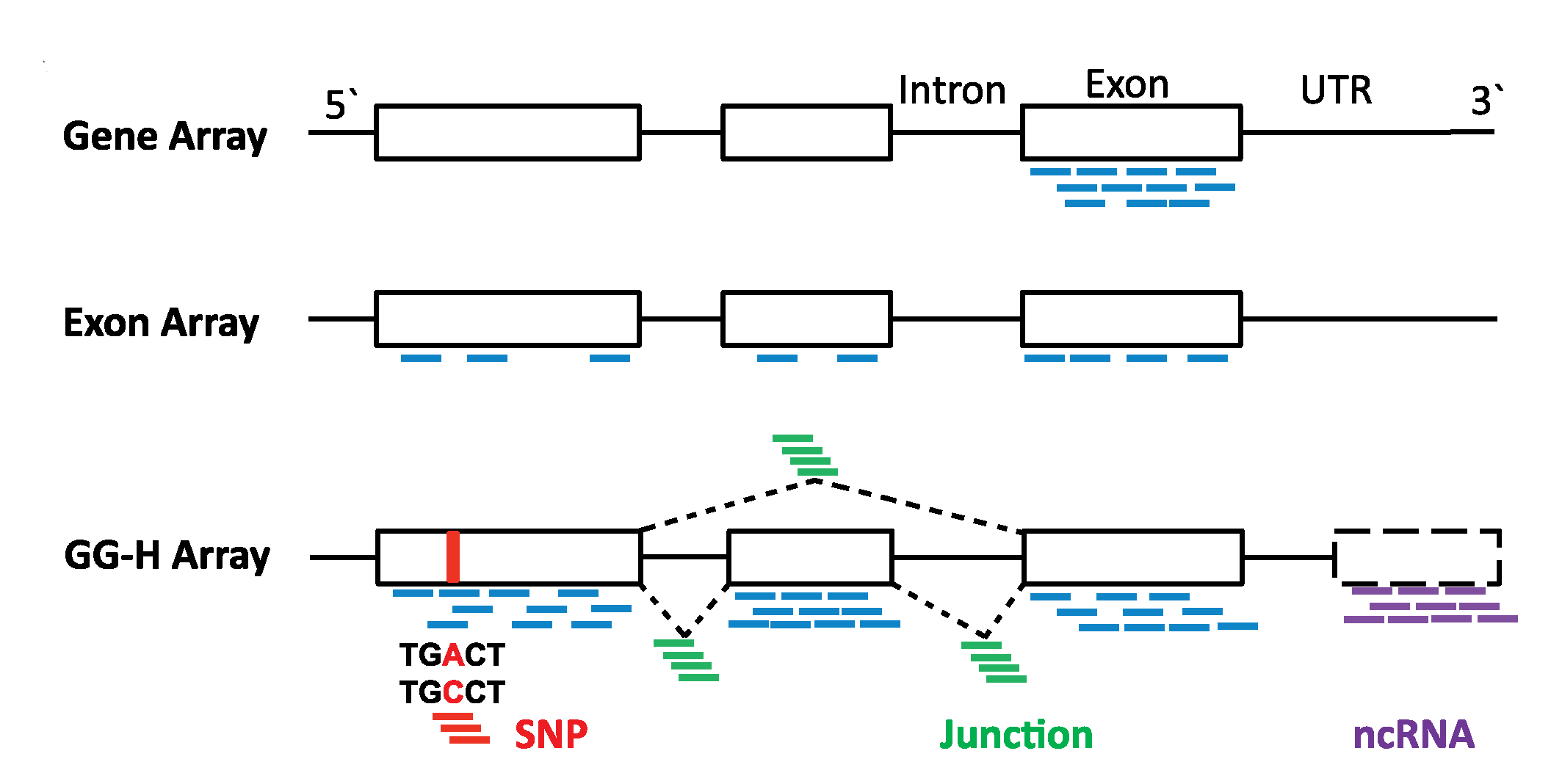
\_x\_at: some probes hybridized with other target sequences.

there is a specific dictionary and just to look if a specific probe set is really informative for one or more genes.

as evolution of the affymetrix gene chip, another gene chip is the HUMAN TRANSCRIPTOME ARRAY , it is a next generation transcriptome array.

affymetrix try to increase the production of this gene chip.

from the gene array to the gg-H array (next generation array)



in gene array array probe set are related just to the first exon of a specific gene

in exon array the probes are directed to all the exons of a gene

and in GG-h array the probes are related also to snp sequence, introns and non coding RNA. so the information is very huge!

Glue Grant Human Transcriptome Array (GG-H) is a collaboration result between

Stanford Genome Technology Center,Wing Wong's lab at Stanford, Affymetrix Inc and

the Inflammation and Host Response to Injury program ("Glue Grant"). The array has

been comprehensively designed to interrogate various aspects of the transcriptome,

including gene expression, alternative splicing, detection of coding SNPs and non-

coding transcription. With tailored protocol to work efficiently with small amount of

total RNA, the array provides a high-throughput but low-cost platform for clinical

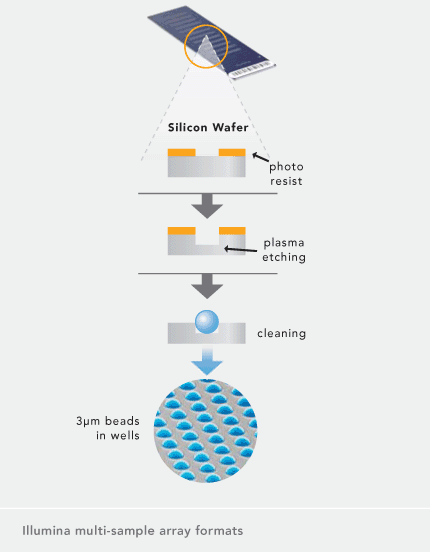
genomic studies.

MIAME → attempt at standardization for microarray data , with guidelines → what should be described in a paper

**ILLUMINA:**

they produce this kind of array named BEADCHIP since it contains thousands of bead containing thousand of probes of the same type (oligonucleotides).

we have the beads inside the well of the array.



**BeadArray® technology** is deployed in multi-sample array formats for DNA or RNA-analysis applications.

Multi-samples Uniform pits are etched into the surface of each substrate to a depth

of approximately 3 microns prior to assembly. Beads are then

assembled and held in these microwells by Van der Waals forces and hydrostatic interactions with the walls of the well.

Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences.

i can prepare beads in different tubes, i have specific bead for each bead, and then i can put them inside the slide. for each bead i have an optical fiber (able to emit light signal).

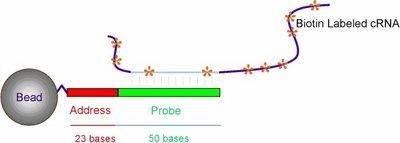
if we are interested in transcriptome :

for each bead we have the probe that should be hybridized with our target (of mRNA).

The probe is linked to a sequence (address) so for each bead there is a barcode technology that allows us to recognize the type of probe inside the bead.

they are used to test the level of the expression.

for each bead chip you use just one sample and one color (not competitive, just one color to estimate the level of expression).



Schematic view of a bead coupled with an oligonucleotide, consisting of the

address code and a 50 base gene-specific sequence (BARCODING TECHNOLOGY)

Each bead carries >1x10 5 identical ~ 79 mer oligonucleotides. For a 47k

gene expression array, 47,000 bead types are prepared, and equal

aliquots of each type are pooled.

But the beadchip needed to study the methylation of the DNA is different from the one used for the transcriptome.

Barcode technology?

you can see the combination of 3 different colors, which repeated different times gives the barcode for the single gene (red , green and white).

<https://miteshshrestha.files.wordpress.com/2018/03/microarrays-_-microarray-analysis-techniques-and-products.pdf>

for bead you will have an optical fiber.different

The most recently ILLUMINA gene chips are quite good tech for gene expression estimation. We want to see the gene chip for DNA methylation analysis.

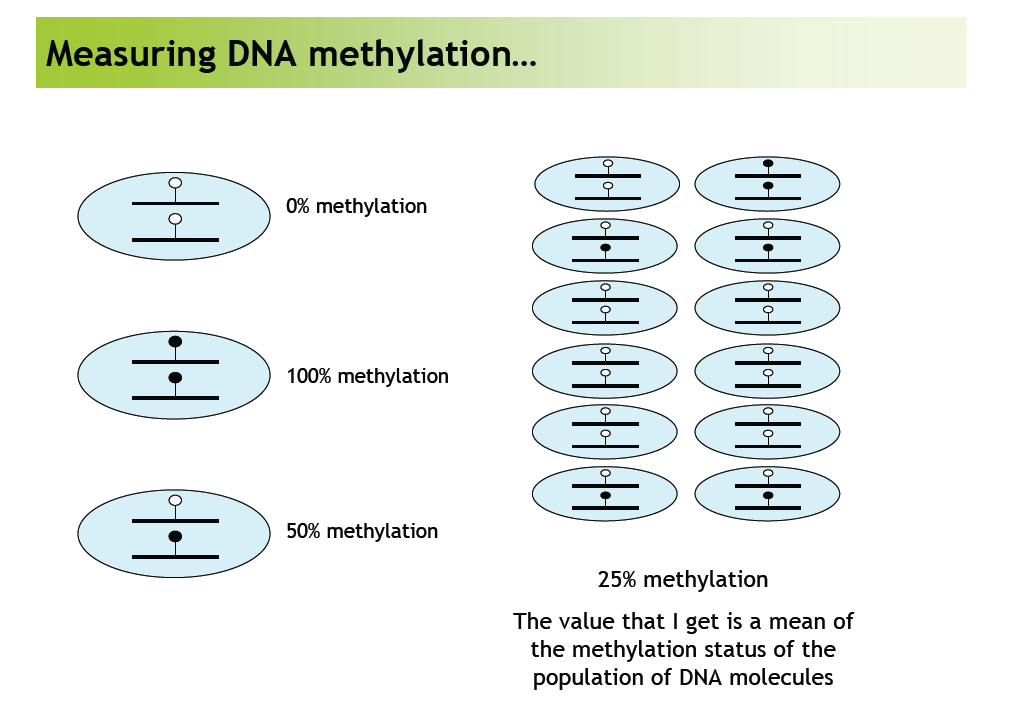
**DNA METHYLATION**

when you are looking for dna methylation we are looking for CpG methylation because the C near a G can be methylated or unmethylated.

in one cell you will have 2 alleles that can be non methylated (white circle), or methylated both (black ones) and we can have 1 methylated and one unmethylated (50%).

the most of the times i can have very different condition in different cells , so the final methylation status for each single CpG comes from the total amount of dna i’m looking for, and if you have to consider the level of methylation in a site specific chromosome location, i can in principle have different percentage of methylation:

es if i have 12 cells it means 24 alleles. if you consider the number of blacks over all, you can derive the percentage of methylation over all the cells. (in this case 25 %).

If we are looking in one cell we can have both allele with C methylated, unmethylated or half methylated (one meth and the second no). But the point is that i have no one cell, i’m looking for DNA for many cell. So i can have very condition. 

Studying DNA methylation

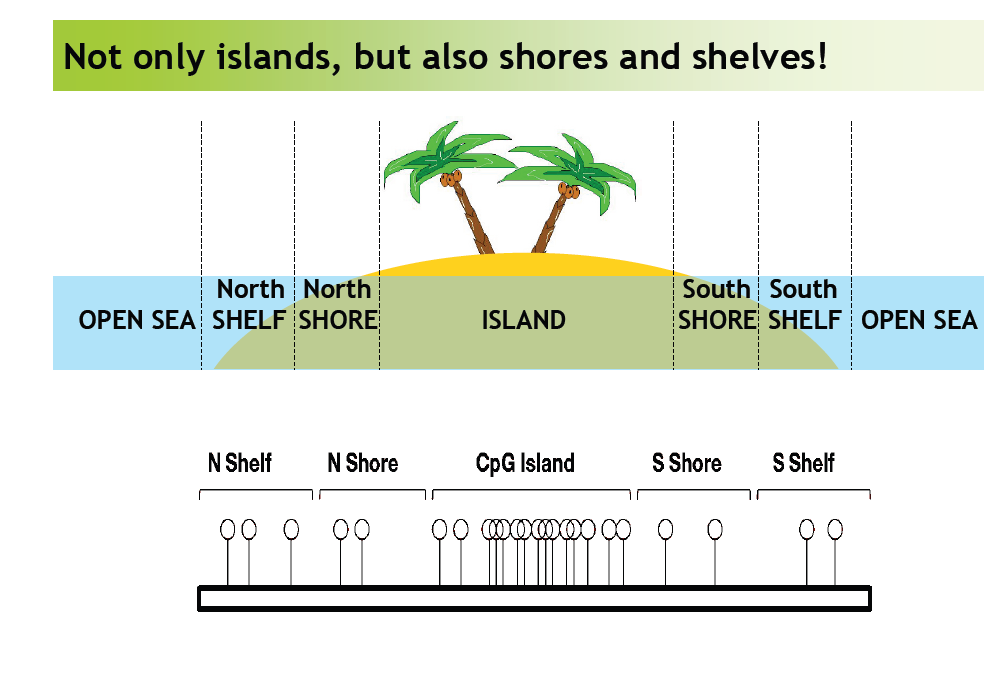
If you are looking for 1 cell and have 2 alleles you can have CpG in both alleles, and the alleles can be both non methylated, one methylated or both methylated. Usually i am not looking for a single cell analysis but rather for DNA from many cells → i can have many different conditions

From the total amount of DNA im looking for i can w

this mix of cells I’m analyzing (often from the same tissue, from a biopsy ) i should derive the level of methylation maintaining that you have different cells differentially methylated!

The estimation of the level of methylation in different tissue: the result can be the same even if the pattern of methylation is different (we have to consider some constrain, we have a low resolution).

If you consider the alleles here the **methylation pattern is different** but the **percentage of methylation can be different**.

Where are cpg? they are spread among the genome but they are also contained in the so called CPG island, since there are specifics area with CPG 

for each island we have also specific zones related to the north and the south of the island with specific names like SHELF and SHORE .

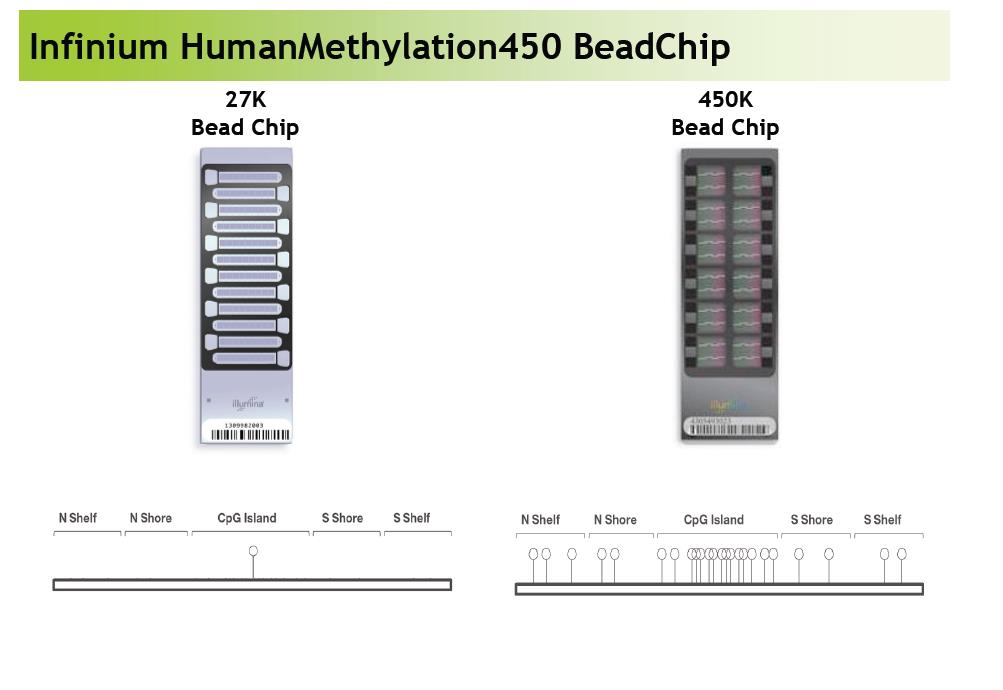
this condition produce a sort of functional genome and usually these cpg highly correlates with gene expression. in particular shores are highly correlated with gene expression!.

DNA methylation:

sometimes there is correlation: DNA unmethylated -> more gene expression.

it depends if the methylation is on promoter.

data seems to suggest that **shore are highly correlated with gene expression** (usually *un-methylation correlates with more gene expression*)

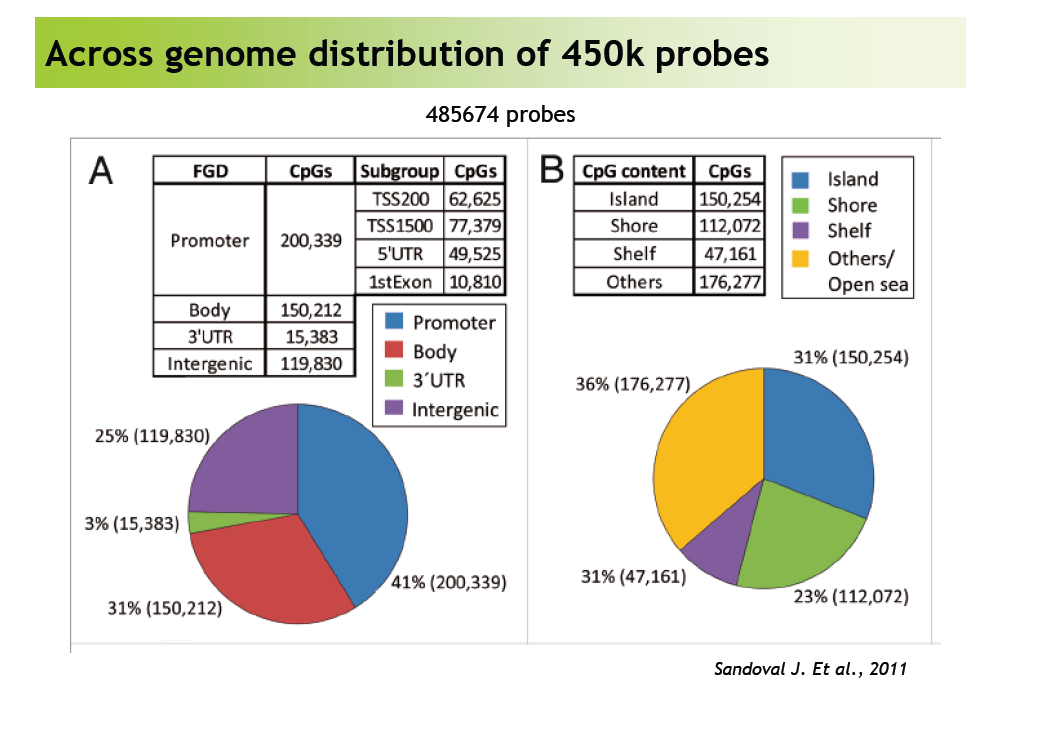


Each bead is for a specific CpG

Different beadchips -> one of the 1st was the 27k

27000 cpg -> the first produced from illumina (the cpg we are looking for are 27K)

the other is 450 k cpg and if you compare the cpg island you can see immediately the difference of coverage.

Looking at the bead chip array:

in panel A you have a scheme of the number of CPG and you can see the distribution of the cpg in the promoter of the gene, the body of the gene, UTR 3’ gene and intergenic.

You know exactly where the cpg is located!

These are the number you can menage using this kind of data.

Illumina produced an evolution of this bead chip: **the infinium of methylation EPIC**

Comprehensive Genome-Wide Coverage,Including the following content categories requested by methylation experts:

•CpG sites outside of CpG islands

•Non-CpG methylated sites identified in human stem cells (CHH sites)

•Differentially methylated sites identified in tumor versus normal

•FANTOM5 enhancers (functional annotation of mammalian genome project)

•ENCODE open chromatin and enhancers

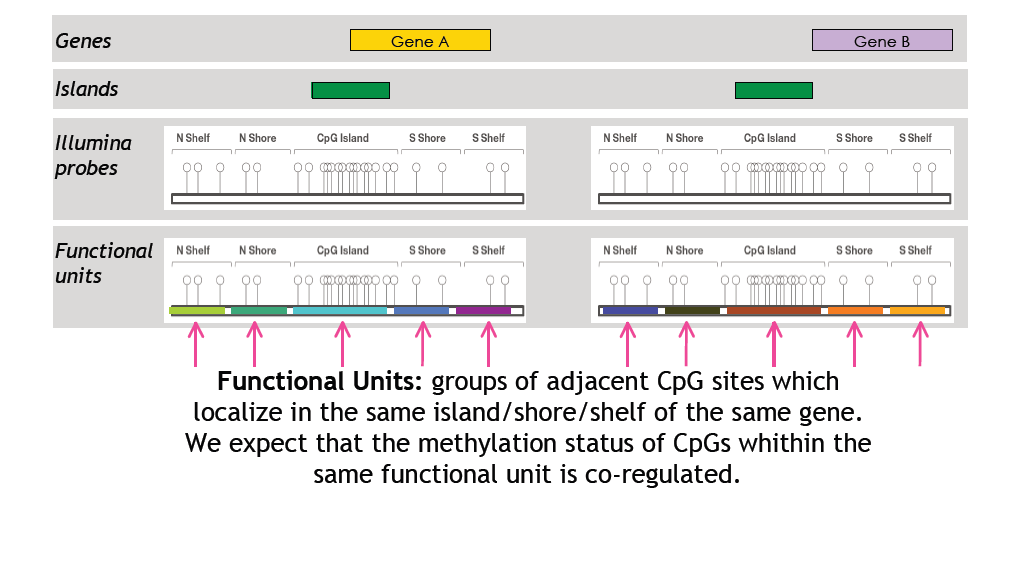
•DNase hypersensitive sites

•miRNA promoter regions

•> 90% of content contained on the Illumina HumanMethylation450K BeadChip

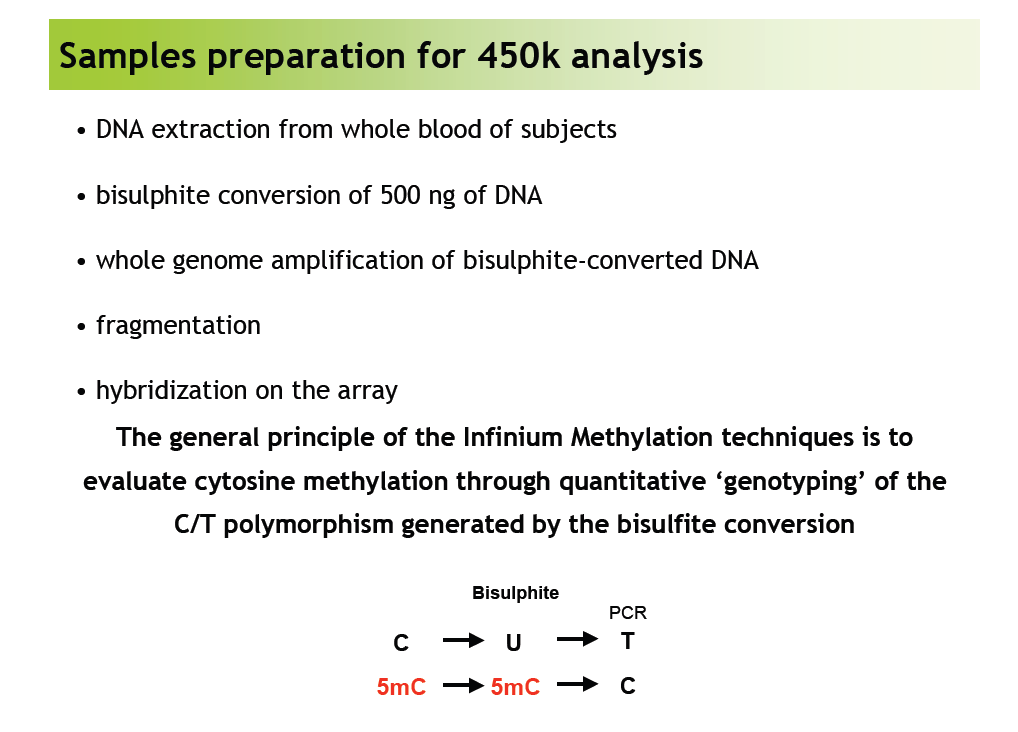
Each beadchip costs around 300 euro

This technology to analyse the methylation is really low (in comparison to the deep methylation sequence techniques).



We can assume that cpg are highly correlated to each other.

From the functional point of view: in each island cpg coregulated in terms of methylation level.

for methylation analysis we are using DNA:

**Bisulphite step**

The bisulphite step is important since we have the modification of some nucleotide: each single cytosine is transformed in uracil and after PCR we obtain instead of uracil, timidine!

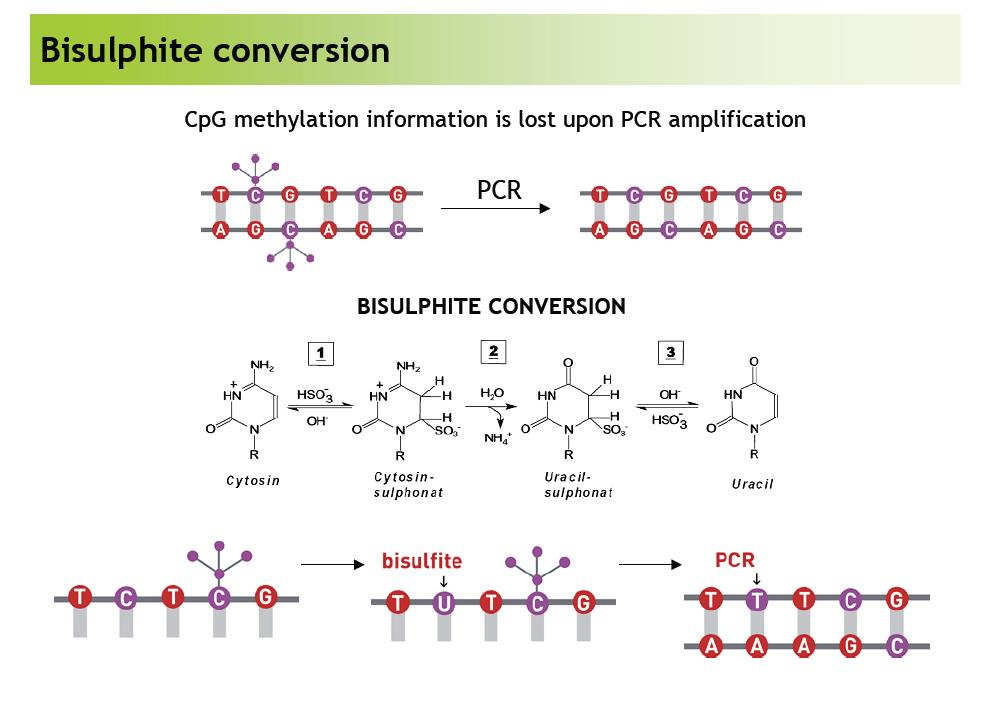
What happen if the cytosine is methylated? Bisulphate deletes methylation and at the and we have a single cytosine unmethylated.

However methylated cytosine become cytosine!

(so we lost the methylation profile using PCR).

When you amplify dna with methylated cytosine you lose the information of methylated sites.

**using the bisulphite conversion you can maintain the methylation profile.**

So any time you see a C it means that originally there was a metC, whereas when you see a T it could have been either a T or a C.

The bisulphide converted dna fragments are hybridized on the beadChip array.

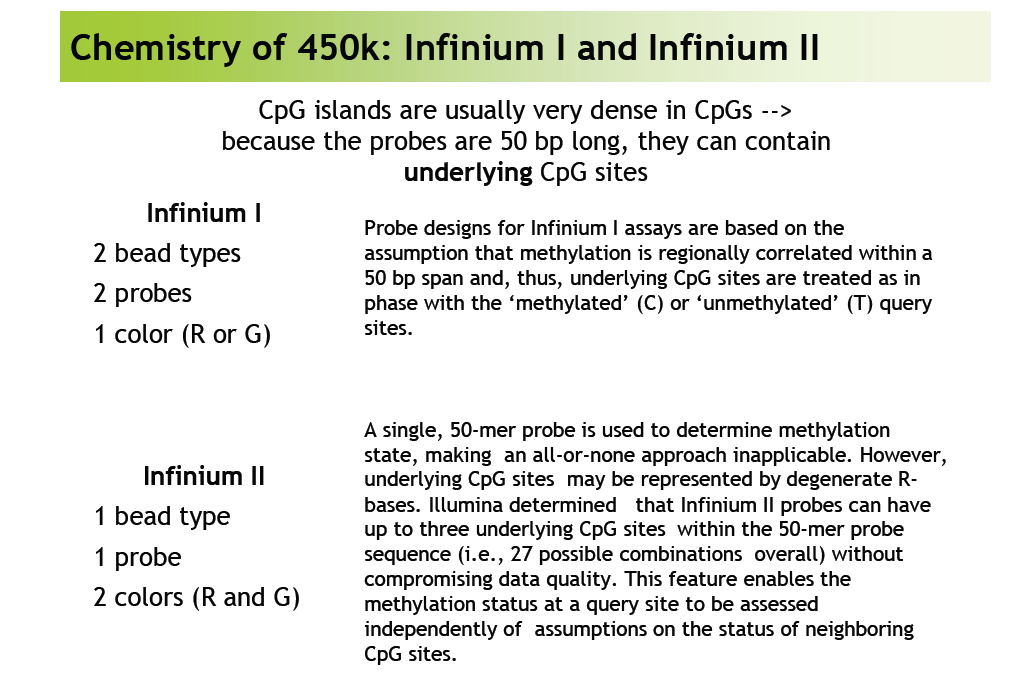
Looking at the type of bead chip we have:

2 kind of strategy and 2 kind of chemistry

* infinium 1 → it contains 2 different bead types with 2 different kind of probes
* infinium 2 → just one bead type with one probe and will use 2 colors

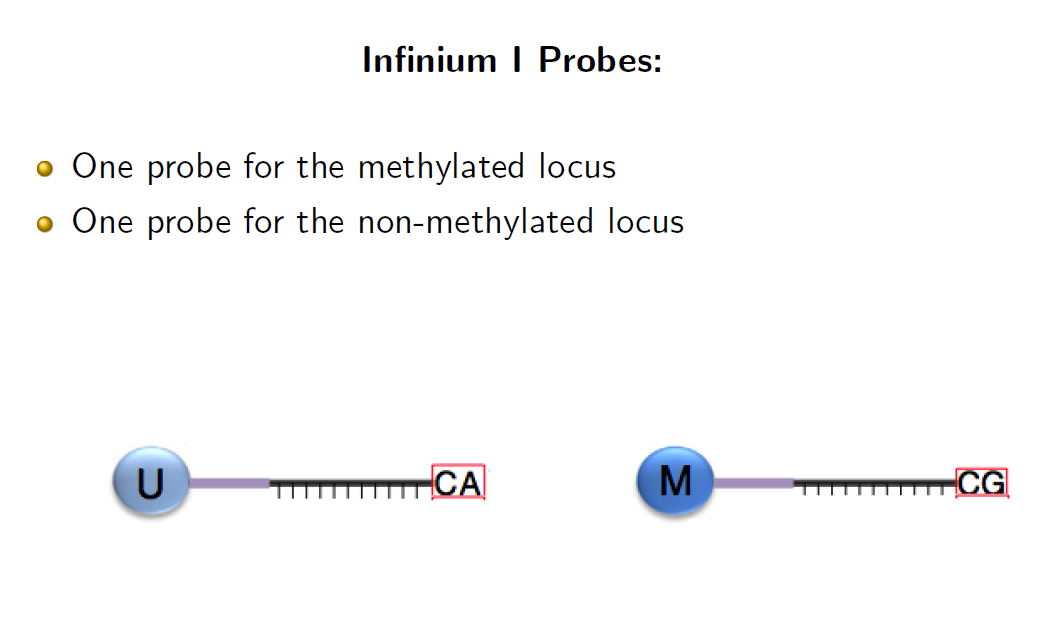
Now we are looking for a single cytosine methylated or not, it is like looking at the single nucleotide.

But each cytosine is near to a guanine and all the probe you have are really specific for each CpG (well known cpg among the genome).

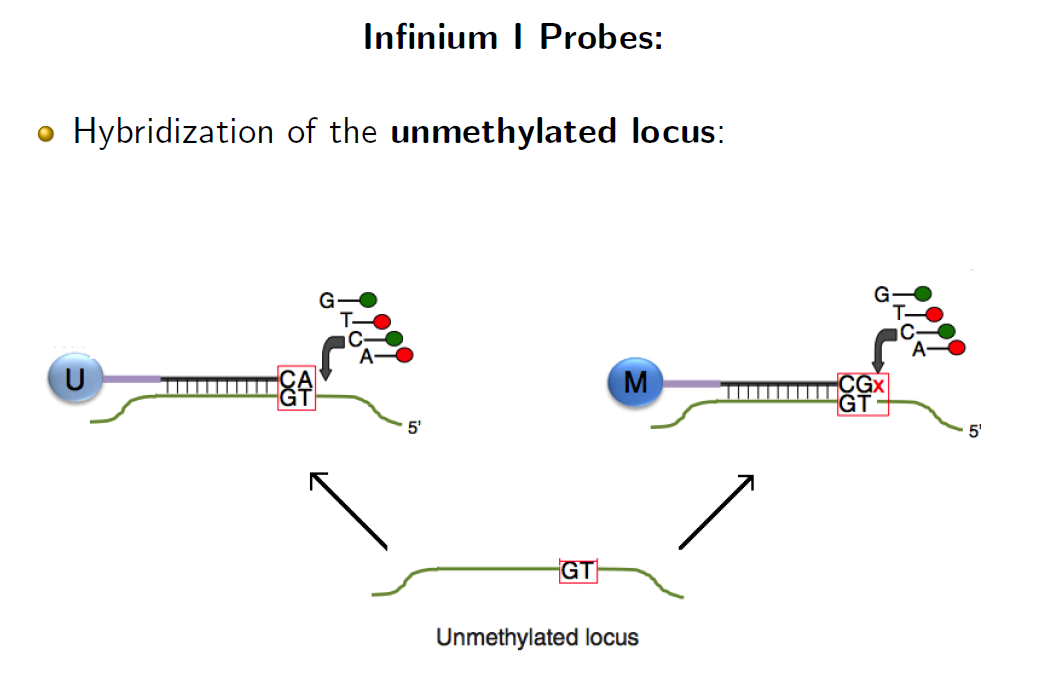
**Infinium 1**

Looking for the 1st type of strategy : it uses 2 different beads (1 methylated and the other unmethylated) and the probe have the same sequences except for the last nucleotide.

This last nucleotide allows to discriminate methylated from non methylated.



2 different beads supporting the same sequence except for the last nucleotide.

because this should interrogate the end of the fragment in order to see if the cytosine is methylated or not!

here i have the unmethylated cpg locus.

remind that the fragment of dna was converted by bisulphate, so if the C was unmethylated it will be converted into a T, so the probe complementary seq will be CA.

Instead if the CpG was a mCpG the amplified seq will be CG so the complementary probe will be GC

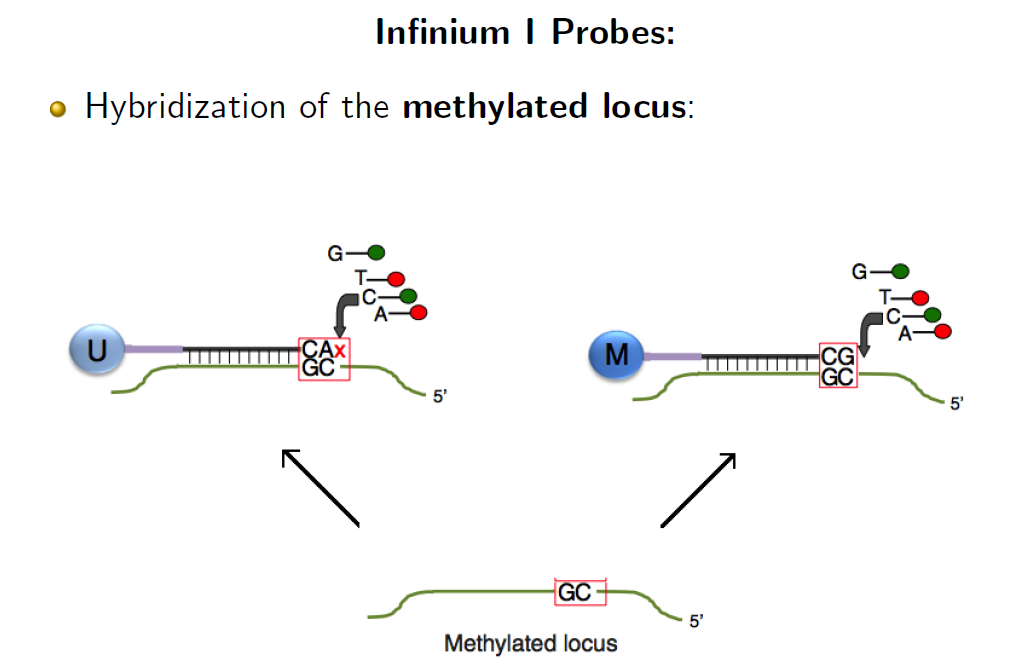
a thymidine originated by a unmethylated cytosine.

so at the unmethylated you have the right pairing, in the methylated one (you have not the right pairing) .

a methylated cytosine should be recognized by the nucleotide on the probe, and the last nucleotide should identify the original cytosine in the original DNA .

so only in the matching pair the PCR reaction can continue, otherwise if the original cytosine was methylated the mismatch will stop the reaction.

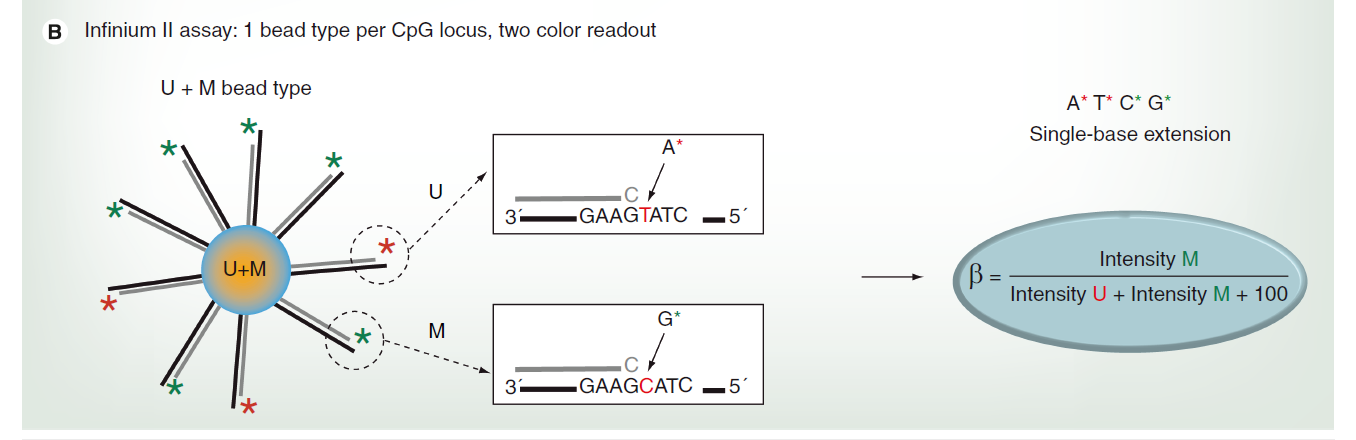
the probe is actually 1nt longer then the site being interrogated

yep, since the last locus is the one that changes between the 2 kind of probes 

if you have a cytosine it should be recognized by the right brobe. so in the unmethylated probe the pairing is not correct so the elongation reaction is not possible.

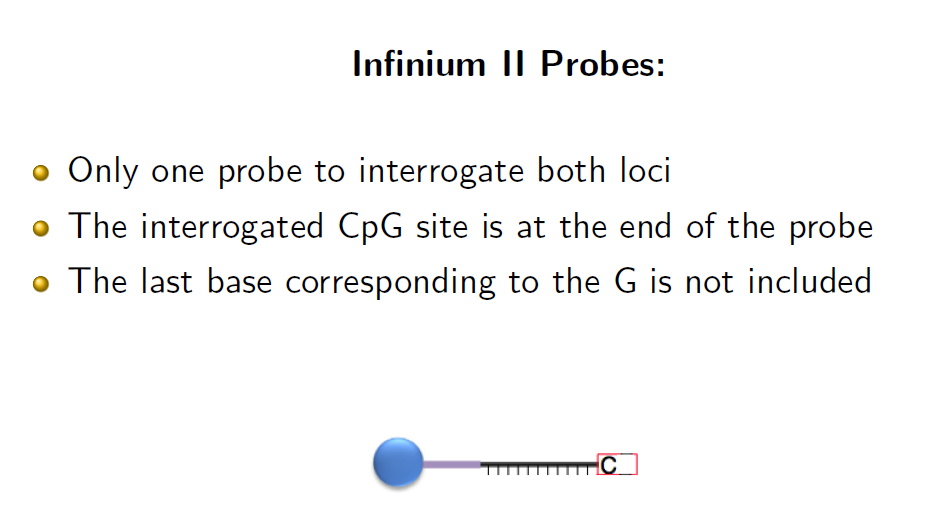
otherwise the correct probe, the one for methylated fragments will pair correctly with a perfect match.

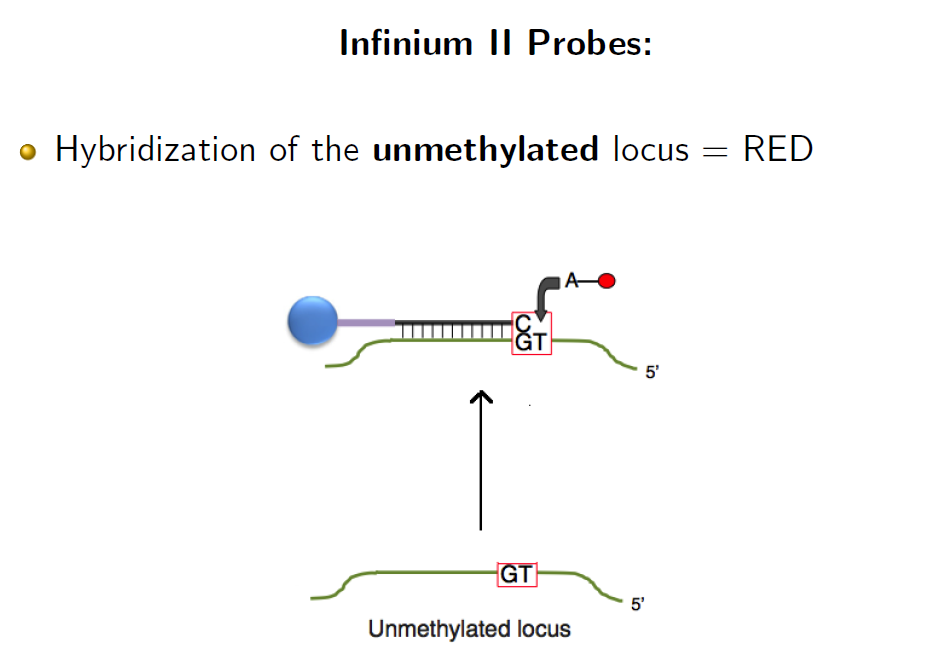
so 2 beads for the same sequence probe except for the last nucleotide that interrogate exactly the same nucleotide.

**Infinium 2**

Then we have the infinium 2 assay, with just one probe to interrogate both the loci.

you have just one bead, with a probe that ends with a cytosine. and you want to interrogate what happens after this cytosine.





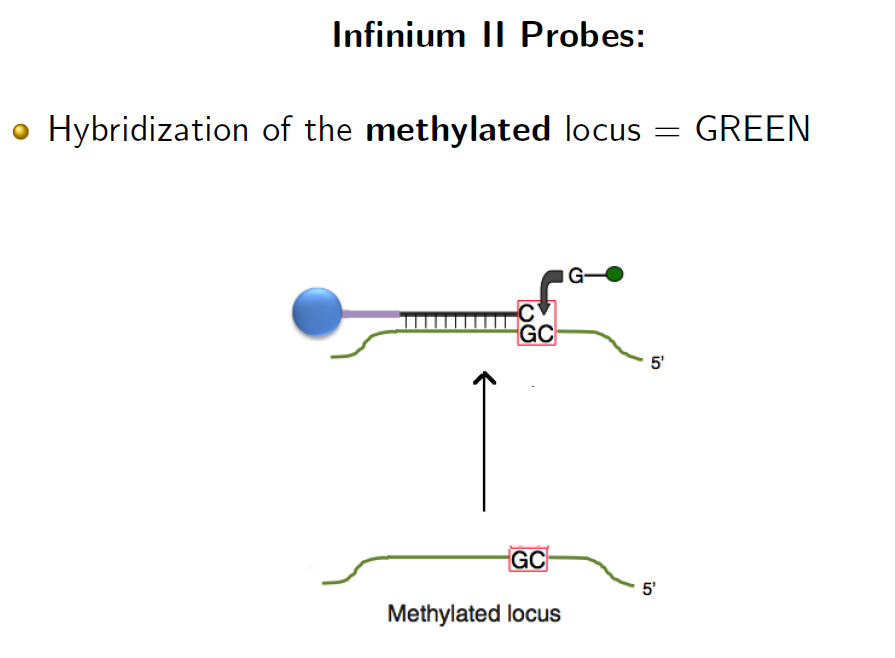
what kind of nucleotide is after the cytosine?

if you had an unmethylated cytosine originally, here you add the adenine nucleotide (with red fluorochrome) it will pair.

so red color->unmethylated.

in the same bead we can have methylated cpg , that will produce a cytosine that will paired with a guanine nucleotide (with green fluorochrome).

so green color-> methylated cytosine

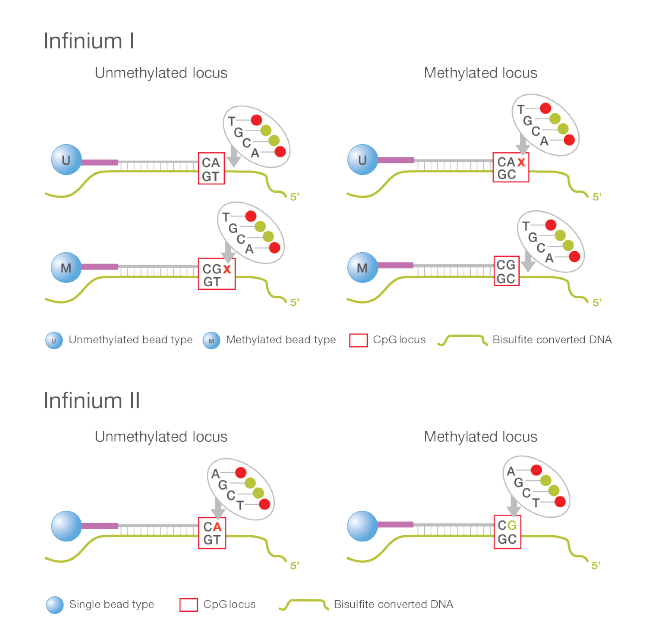


if you have a methylated C now you have a single base extension with a G marked with the green fluorochrome.

Inf1 gives the possibility to interrogate both cases at once: for the same CpG we have two signals

Same goes for the inf 2 actually

so summing up:



the signal derives from the evaluation of the 2 different cases of the fluorescence.

Signal intensity:

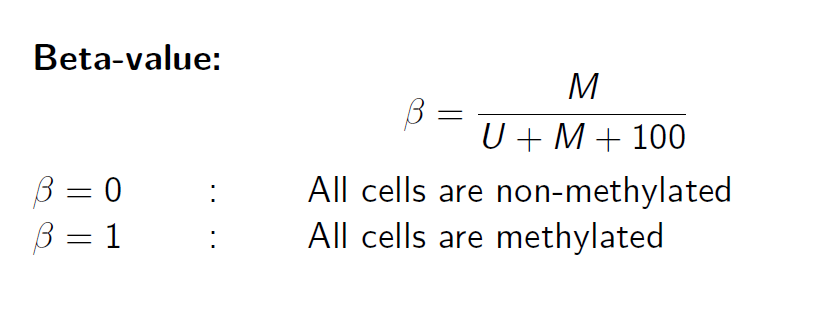
- Infinium I = for each methylated or unmethylated CpG probe a different

level of red/green intensity can be found due to the presence of DNA from different cells

- Infinium II = Methylated is Green and Unmethylated is Red for each CpG

probe could have a different level of red/green intensity due to the

presence of DNA from different cells

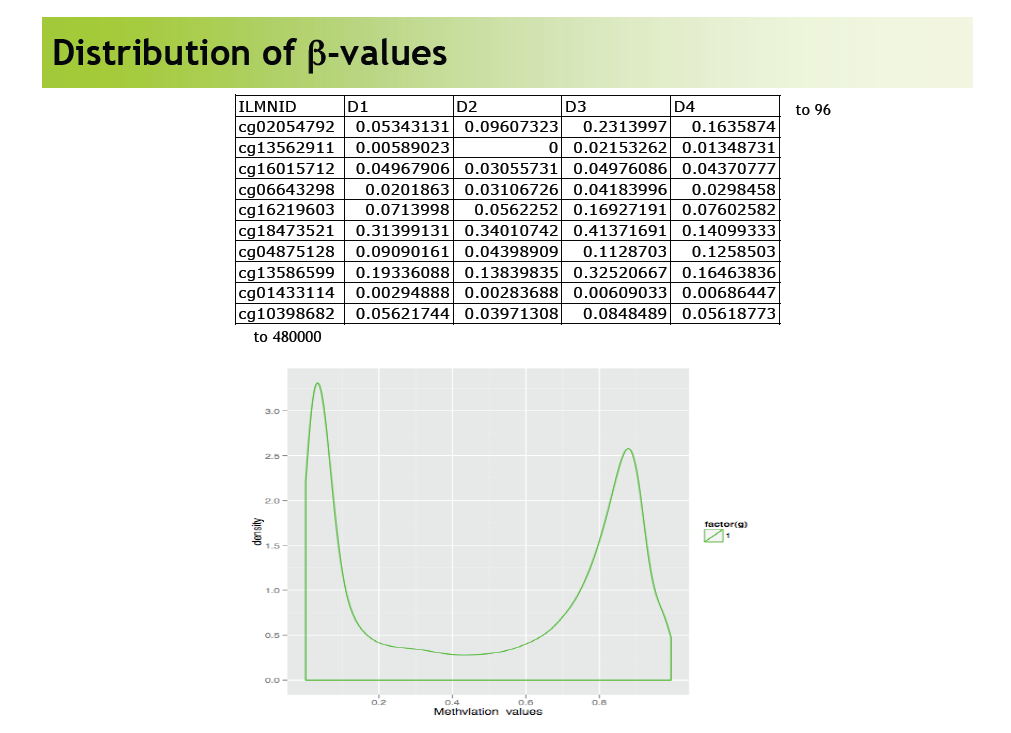


beta value: goes from 0 up to 1. zero means completely unmethylated and 1 means completely methylated.

At the end you will produce an excel file:

with the first column with all the probes, 2nd column has the results of beta-val for the array. Each column is the beta-val of one array right?

for each CpG i know the exact location in the genome!

the code for each cpg contain information also on the location of the cpg island in the 

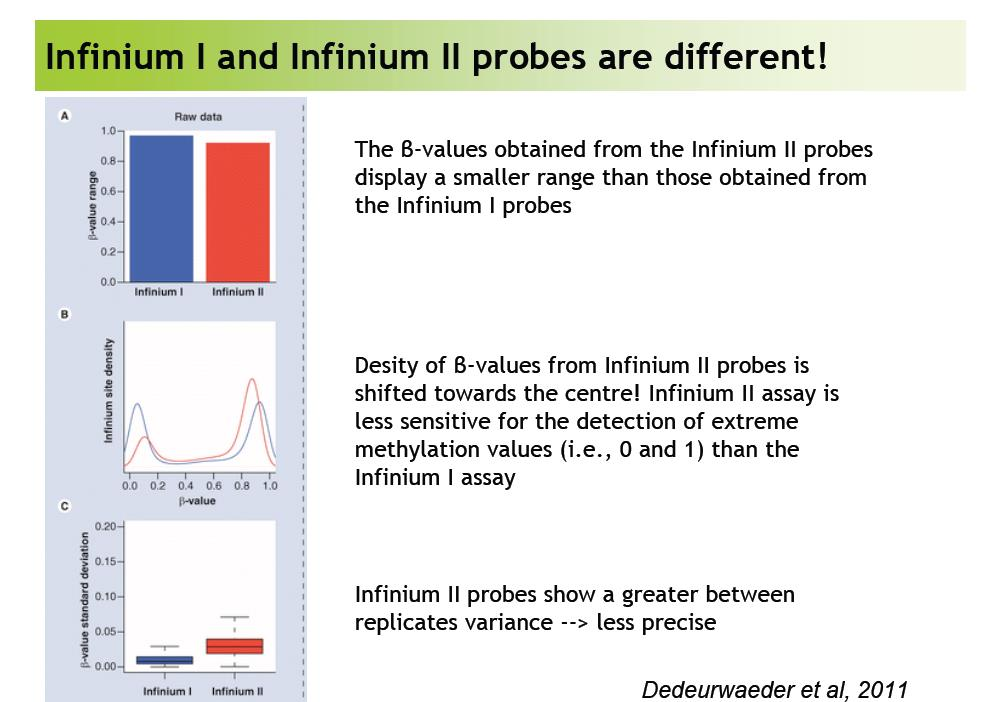
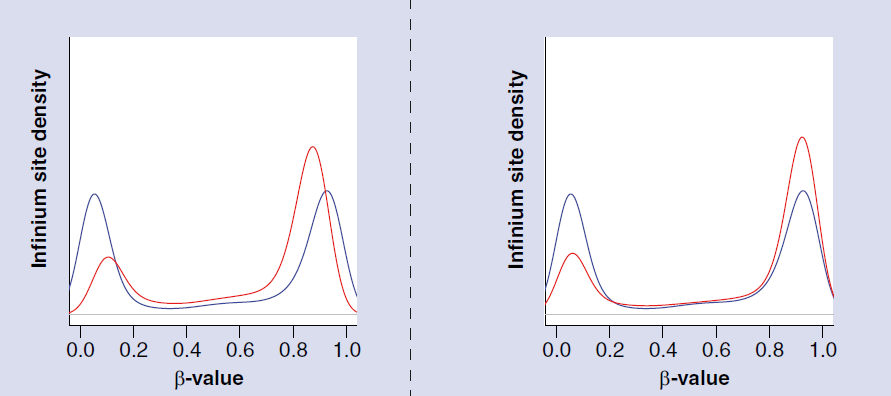
chromosome.

if you look at the distribution you can see 2 peaks, on the y axis is the frequency and on the x axis the Beta-val.

so from the distribution you can observe that the b-values clusters near zero and near 1 (meaning that inside the same tissue, the mixture of cells will have a similar methylation profile) ??

compare inf1(blue)/2(red): they are different! the two peaks are Not well aligned nor of the same heig. the range of infinium 2 seems to be smaller.

this is a big problem for this technologies.



so they tried to solve this problem: inf 2 has a smaller beta-val as clearly seen in the box plot.

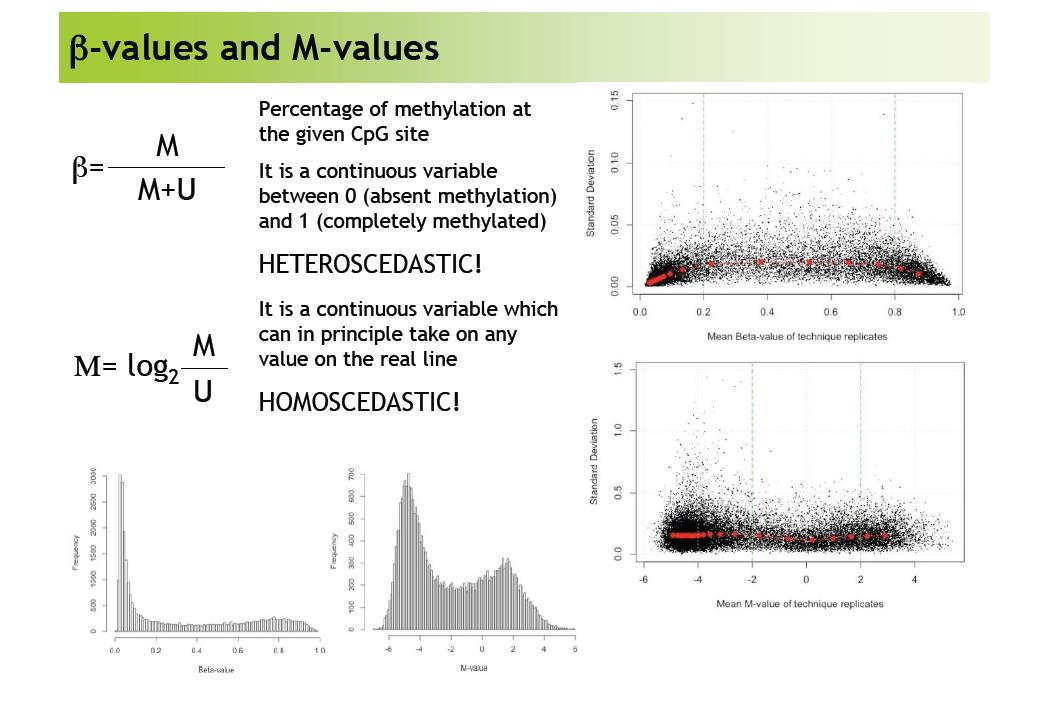
Also the peaks are shifted towards the center, same explanation as above.

So we have to do intra array normalisation to align the peaks.

x → mean of the values

y → std dev

most of the dots are near 0 and 1 → heteroscedastic variability

M = log2(M/U) → this improves the ??

now to normalize data we start from the row beta value, they will be converted in M values and these latter will be normalized in order to align the 2 peaks.

we will do it in the practicals!

Why this difference though? Why is it worth it to use inf2 anyway if it is less reliable?

for both the inf1 and 2 the biological q is the same.

why do we have to “fix” the inf2? She says that in this way we can compare the values inside the same array

INFINIUM 1 AND 2 ARE IN THE SAME FUCKING ARRAYY because she is not in the brain of illumina people but they want to interrogate the same position with 2 different techniques.

It magnifies the probe capability to detect the methylation and the two methods validate each other. But why do we assume that the inf1 is more reliable than the inf2? **There’s a slide (the one with box plots) talking about Infinium II probes showing a greater between replicates variance: I guess it’s a matter of precision and since inf2 is less precise, we assume inf1 is more reliable (Davide’s responsabilities)**.

similar problem to the cy3/5 array???

we don’t consider the real signal but just the ratio of met/unmet ? or rather the beta-val

compare CpG

inf2 and 1 detect the same thing however 1 is more sensible and gives a wider range of values, whereas the 2 simply is less sensible so it gives back a smaller range. So the same numerical signal doesn’t correspond to the same biological situation: for this reason we have to “normalise” the inf2 to give the same bio meaning to the same numerical value.

This is to the best of my understanding, it could be wrong and sensibility could be the wrong word for it.

However as Alessandro is saying this argument doesn’t hold as nicely if the signal is actually a ratio because then it doesn’t matter the actual values measured with the 2 infinium tech, since in proportion the signal should be the same.

there is still no uniformly accepted practice