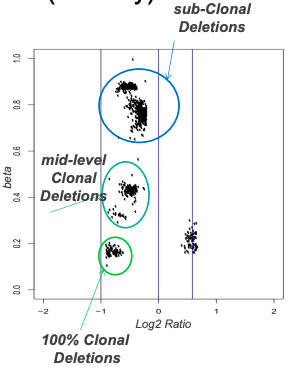
**TUMOR EVOLUTION STUDIES VIA NGS (CONTINUED)**

***Recalls from the previous lecture:***

At the basis of tumor evolution is the concept of how to use informative SNPs: SNPs for which a specific individual has heterozygous calls so that set of SNPs is unique for every individual.

This property is connected to the fact that when we have the loss of an allele, the allelic fraction of the informative SNPs within that lesion will be informative of the lesion and its depth (clonality = what’s the fraction of tumor cells that very likely harbor that lesion).

We can also have different population of cells, when a set of lesions is present in every population it is said to be clonal whereas when a specific set of lesion is harbored only by a subpopulation it is defined as subclonal.

*Estimate of DNA Admixture*

*Log2 Ratio* is the log2 of the ratio of the tumor over the normal that applies to array data signals (intensity of the signals) but also to the local coverage of a tumor BAM file over a normal BAM file.

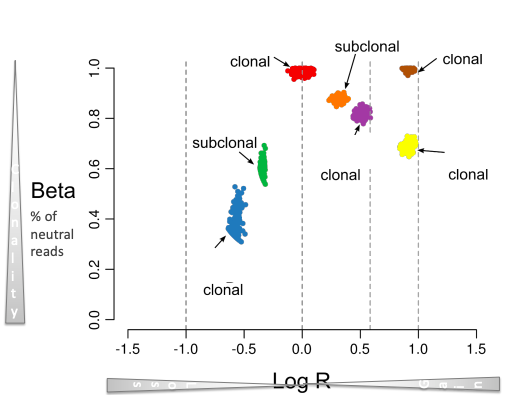
In the figure each dot is a genomic segment or a gene that clusterize in the space and when dots are in a same cluster it means that they very likely share the same copy number status and also the same level of clonality.

*Beta* is a variable that goes from 0 to 1 and provides information of the number of reads that equally represent the two alleles; when beta is equal to 1 the concept of admixture (1-purity) is equal to 1 meaning that purity is equal to 0 if we are at the top of the y scale it means that there’s no signal related to tumor content, while the lower we go, so the closer we get to 0, the higher the tumor content and the level of clonality is.

If we use this equation we can assess the level of clonality of a cluster.

So the graph in the figure puts in relation the copy number status (log2 ratio) and the purity/clonality of the sample (Beta); the more we go towards the left the fewer number of copies, the lower on the y axis the higher the clonality.

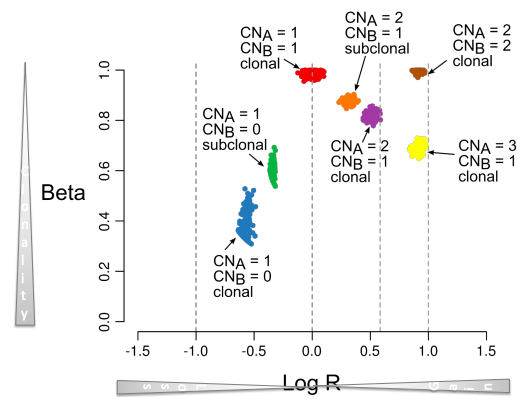
The best proxy of the quantity of tumor content present in a sample is done using the lowest cluster.

We have losses and gain of DNA copies, moving on the x axis.

The beta is related to the clonality so the lower we go the more clonal the signal is.

The only difference from the previous figure is the presence of extra clusters:

* The blue cluster with deletions is the most clonal one
* Both blue and green clusters had deletions, since they have a negative log2 ratio, but the green ones are less clonal than the blue ones
* In log2 R = 0 and ß = 1, where there’s the red cluster, we have a status of no copy number changes (wild-type status in terms of copy numbers). This basically represents a total number of alleles which is the same in both the tumor and normal sample.
* All the other clusters with a positive log2 ratio had a gain of DNA



In this figure the number of copies that correspond to all the clusters in the space is also reported.

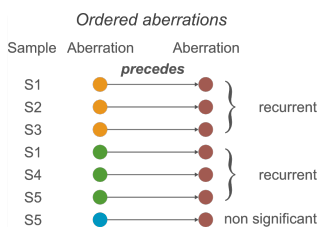
* Blue one: one copy of DNA, so we have a deletion
* Green one: also one copy of DNA but with subclonality

This is how we can map in the space the status of clonality and the number of copies for a specific segment in the genome.

So again, the lower we go the more clonal the clusters are, the more left the deeper they are in terms of loss of DNA.

We can use these information to build *evolution maps*.

The first thing to do is to look, within each individual, at concomitant deletion where one is subclonal to the other one.



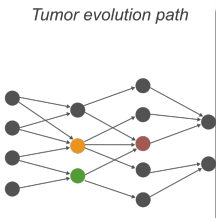
In the figure:

* In sample 1 the brown lesion is subclonal to the orange one, and that same lesion is also subclonal to the green one.
* In sample 2 we have again the support of the relation between the brown and orange lesion with the same level of subclonality (brown subclonal to orange).
* In sample 3 is the same as in sample 1 and 2.
* Samples 4 and 5 have the same concomitant green and brown lesions again with the same level of subclonality.
* In sample 5 only we also have another concomitant lesion (blue subclonal to brown).

So we perform this analysis for all the concomitant lesions in our sample and we start drawing the arrows to keep track of what is subclonal to what. We compile this list across all individuals and look for how many times we see support for the same relationship in the same direction.

In our case we can say that the relationship going from orange to brown is supported by 3 out of 5 individuals; the same can be said for the green going to brown. The blue one is instead not significant since it’s supported by only one individual.

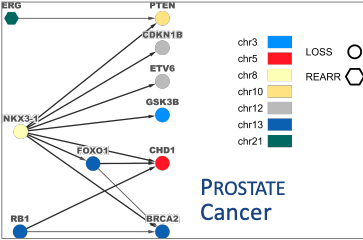
So having multiple observation supporting that aberration x precedes aberration y (i.e. aberration y is subclonal to aberration x) we can build an evolution chart.



The orange and the green which have no relationship between them, are at the same level on the x axis in the path and they both go into brown.

So one can assume that the more clonal a lesion is the more likely it is that it occurred earlier during the evolution (time is on the x axis of the path), and we can look for recurrent relationships among lesions.

In principle we can say that the grey ones at the beginning happened at the same time point and then at a second time point, the tumors in our set of samples, underwent loss of orange and green genes and then later they both underwent loss of the brown gene.

If we do that in large datasets (lung cancer melanoma, prostate cancer …) we can come up with all the dependencies that were observed and that were supported by more than one individual (e.g. in prostate cancer we can say that a loss in NKX3-1 precedes the deletion of PTEN).

Even if we have hundreds of BAM files on whole exon sequencing data from large collections all that we can build are evolution maps with at most three layers (pretty disappointing).

This has multiple reasons, one of them is that:

* To build a relationship which is statistically significant between two genes we need to have multiple instances of that relationship (in many samples) which means that we need to have co-occurrence of the two lesions and subclonality of the second lesion with respect to the first in a significant number of individuals compared to the total number of individuals that have co-occurrence. So if co-occurrence occurs in N individuals and subclonality of the second lesion to the first one occurs in a fraction of those, only if this fraction is significant with a proportion test out of the total number, then we can build the path.

Therefore we are tremendously limited by co-occurrence of lesions.

To boost the reconstruction of these paths gene families or pathways have been exploited.

E.g. if we are dealing with PTEN which is a tumor-suppressive gene relevant in a specific pathway (PF3K), then it doesn’t matter if we have deletion or inactivation of the same genes in the same pathway, what matters for the tumor evolution is that that specific pathway is altered and so what we can do is start aggregating signals from genes that belong to the same pathway.

So if individual 1 has a relationship between gene A and some gene in a specific pathway (PF3K) and individual 2 has a relationship between gene A and a second gene in that same pathway, then we can assume that maybe they have the same effect and so we can aggregate the information on the landing gene.

So instead of going from gene 1 to gene 2 we go from pathway 1 to pathway 2, and in terms of numbers what we gain is that the co-occurrences are counted including all the gene lesions with the same function in pathway 1 and all the gene lesions with the same function in pathway 2 (if we consider the inactivation of the gene then we have to consider all the lesions that inactivate the gene and not others).

We can then run a simple test to build our path.

With this method we start having some more data to look for major changes during the evolution of the tumor pathway.

E.g. in prostate cancer we’d identify a set of pathways that are more or less at some level altered in earlier staged disease and that then trigger or are precedent to our pathways. Doing so we can learn more in terms of the biology of the disease evolution.

We can also decide to go for a mix model or a mix approach, where for certain genes we go at the pathway level while for other we treat them separately.

There are also more complicated ways to make inference of tumor evolution. Some try to avoid the hypothesis that the more clonal a lesion is the more likely it is to happen early, because we know it’s not always the case; it might be in untreated samples but not in treated samples. In a treatment regiment, because of drug pressure selection, specific resistant clones harboring a specific lesion can take over due to their higher rate of proliferation, so in this case if we see a lesion that appears to be more clonal it doesn’t really mean that it happened earlier, it may be that it had a higher proliferation and so it’s taking over (and we see it as apparently clonal but it’s in fact a late event) -> important concept in precision medicine.

So simplicistic approaches like the one discussed are proper for untreated (in terms of drugs) primary diseases.

Evolution charts can also be boosted via the combination of multiple molecular layers.

**THE EFFECT OF PLOIDY AND PURITY CORRECTION ON LOG2(T/N) DATA**

*How can we use measure of the tumor purity and the effect of the tumor ploidy?*

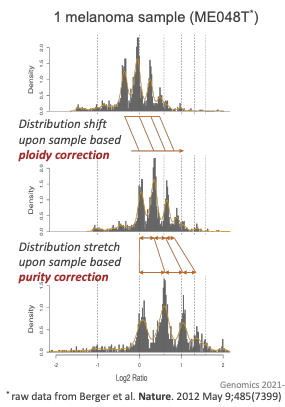
*How can we compare two different samples for which we quantify completely different levels of tumor content?*

E.g.: we have a sample a 100% pure and with 50% of clonality (a lesion present in 50% of the cells) and a second sample with a tumor purity of 10% and a clonality of 100% (a lesion present in 100% of the cells), we need a way that allows us to compare numbers without having to convert everytime for every lesion the depth of the lesion based on the tumor content, so we need an equation that we can apply to every individual data that puts everything on the same level

(same concept as gene expression normalization).

The coverage makes data coming from different samples comparable because we normalize everything to the total coverage, but when we deal with diseased cells we can have contamination from the admixture, so we need an extra step.

The step, once we know how to assess the tumor purity and ploidy, is quite simple: we need to adjust the data for tumor purity and ploidy.

*Schematically*

In the figure we are looking at one tumor sample: a whole genome sequencing of one melanoma sample.

We see multiple peaks which correspond to different copy number states.

Let’s suppose we have a genome with a backbone of three copies but we sequence a bulk and we don’t have 100% purity but 80% (so 20% is contamination).

***Ploidy correction***

Computationally we assess the ploidy through the copy number space and then correct the data.

From the tumor and the normal we obtain something like the first graph, and we could wrongly assume that the main peak is always in 0 (wild-type state of the genome), but it shouldn’t.

In fact, if we assess the ploidy and overall we see a backbone state of three copies for our genome, then the main peak should be shifted toward three.

So, the *ploidy correction shifts the distribution* towards the right (second graph).

***Purity correction***

We correct our data and the *purity correction causes a stretch between the peaks*, since tumor admixture dilutes the signal. So, the effect of purity correction is a wider spread between the peaks (third graph).

+ add the example graph

* If we have one extra copy in our tumor, the log2 ratio will be around 0.58 and so we would expect that the signal will peak around that value; for two extra copies we’d expect a peak around 1 and so on.
* We’ll have the peak of the normal state around 0 and then if we have an underrepresented allele in our tumor we’d get another peak around -1 for the hemizygous deletion and then the homozygous deletion.
* If our signal is not 100% pure tumor (so diluted by normal cells), the peak at -1 and 0.5 would be closer to the 0 peak for uncorrected data.

*When we correct for tumor purity we stretch the distribution to go to the correct positions.*

E.g.: 25 whole genome sequencing of melanoma samples

Immagine che contiene testo, antenna

Descrizione generata automaticamente

* 1st graph: The distribution of the log2 data of uncorrected signal, every melanoma sample is highly aberrant with a ploidy that is different between different individuals and a purity that is also different between different individuals. But we do have the tumor ploidy and purity so we can correct the data.
* 2nd graph: we correct for ploidy
* 3rd graph: we correct for purity too

If we don’t correct our data we’ll see much noise (as in the first graph).

From the corrected data we learn that:

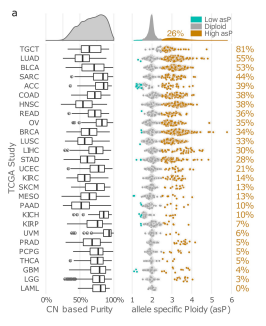
* A lot of tumors have a backbone ploidy of two
* There are some hemyzigous deletion not perfectly centered in one but closer to one in the 3rd graph if compared to the 1st
* Some signal is compatible with homozygous deletion
* Whave a reasonable amount of signal for three copies which could come from a threeploid status of some tumors.

These corrections are part of standard preprocessing.

***Tumor Ploidy and Purity adjustment, corrected TCGA data***

*How commonly does suboptimal tumor purity affect proper copy number data analysis?*

*How common it is that purity is not equal to 100% and ploidy is not equal to 2 in any primary disease*



In the figure we can see a list of tumor types, where every draw is a tumor type (lung carcinoma, bladder cancer, colon cancer, ovarian ecc.). On the x axis we have tumor purity (1-admixture) going from 0 to 100% and for each type we can see the distribution of the tumor purity analysis of all the samples from the TCGA dataset.

Every tumor type has a different number of sample profile

Looking at the GBM (glioblastoma multiforme), the middle vertical line is the median signal of the distribution, there are outliers shown and the black horizontal line represents the interquartile range.

Altogether across 27 tumor types they were able to assess the tumor cellularity, clonality and all in about five thousand of those, meaning that a great fraction of those had some optimal data (very strict criteria)

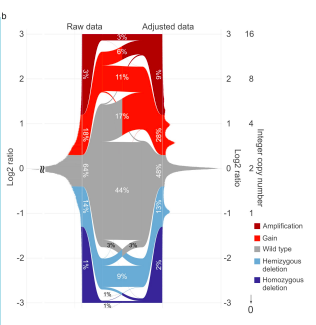
* The majority of the median distributions are above 50 %.
* The overall tumor cellularity was almost 70%.

*If we look at ploidy: what is the fraction within each tumor type with a ploidy significantly above two*?

In the graph they are sorted by decreasing percentage of tumors with a ploidy higher than two; for example, for the first and second tumor type, more than 50 % of the primary tumors have a ploidy status above two so either they underwent whole genome duplication (4 or more copies) or at least we have three.

Then we have some tumors with very low ploidy (blue dots) where at least one copy of the entire genome is completely lost -> low allele specific ploidy assessment.

The figure shows what happens to data when we correct for ploidy and purity

* On the y axis we have the log2 ratio
* On the left side we have the raw data
* On the right side the adjusted data

We can see where correction for ploidy and purity takes the signal.

Focusing just on the first half we can see that

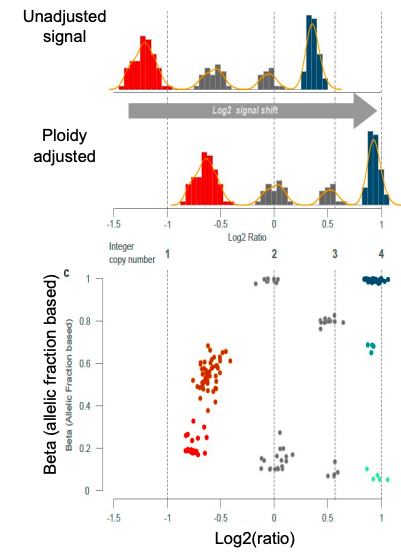
we have the same noise we’ve seen for the melanoma uncorrected data.

*The correction of the data results in the reclassification of 30% of the totality of the segments* (if we don’t correct we have a wrong copy number classification in 30% of the cases)

Then there are certain copy numbers which are more or less affected by these corrections.

What’s interesting is that the correction led to the doubling of the homozygous deletions that we were able to observe (these are very important because it means that the proteic product won’t be there at all).

**ALLELE SPECIFIC ANALYSIS (CNA, CNB SPACE)**



Thinking in terms of allele specific data:

1. We have unadjusted signal
2. We adjust
3. Then we can go to the beta-log2 ratio space where we can see that the data underneath the peaks are belonging to specific clusters

This suggests that by only looking at the log2 ratio we are unable to distinguish the presence of clusters with different clonalities.

The most interesting information is the lower cluster (on the x=0 axis):

* Even when the T/N = 1 (tumor/normal ratio) what we can have is a status of one copy and one copy or something that equally gives a log2 ratio equal to 0 but which still represents copy neutral loss of heterozygosity (CN-LOH), so two copies on one allele and zero copies on the other.

+ example figures (will be added soon, I have to draw them)

1st figure:

We have the loss of an allele on A so we’ll have 2-1-2 copies

2nd figure:

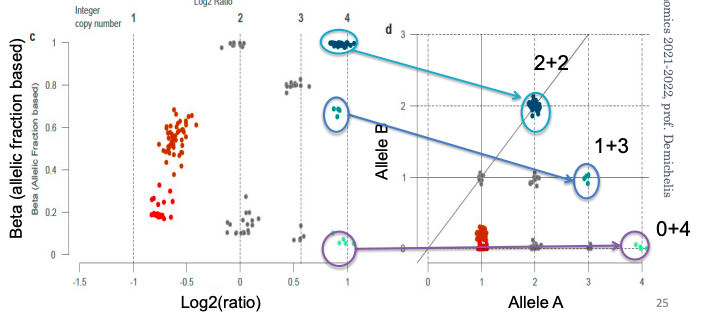
We have the same situation on allele A but allele B is doubled so we’ll have 3-2-3 copies

So, in this situation, the gene x will have two copies but both of them coming from the same allele (B).

Computing the log2 ratio in this situation we’ll have the log2(2/2) which will lead to the collocation on the 0 axis but on the lower part (due to the clonality).

The log2-beta statuses allows us to distinguish the copy-neutral LOH.

Also for the gain is the same (three copies from the same allele and zero from the other)

There are equations that allows us to go from here to a space where our coordinates are the number of copies of allele A and number of copies of allele B.For four copies we can have different combinations:

* 2 copies of A + 2 copies of B,
* 3 copies of A + 1 copy of B
* 4 copies of A + 0 copies of B

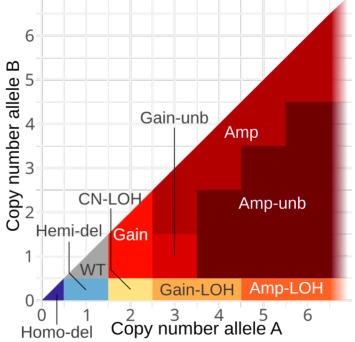
The equations are not important, what’s important is that once we have corrected the data then we can shift our analysis up to the level of number of copies of each allele for each gene.

*Why is this important?*

E.g.: Let’s imagine that for gene X we have one copy lost on allele A and a point mutation on the allele B which leads to unfunctional product so full loss of the protein.

If we instead are in the second case and the point mutation happened after the duplication then we’ll still have an allele functioning, whereas if it happened before the duplication, we’d have again full loss of functional protein.

If we are able to distinguish the alleles we are able to also distinguish in which situation we are (which means we can distinguish between what’s functional and what’s not).

* Extra graph with the same space allele a/ allele B where we can divide the space in terms of total number of copies and also what happens on both.

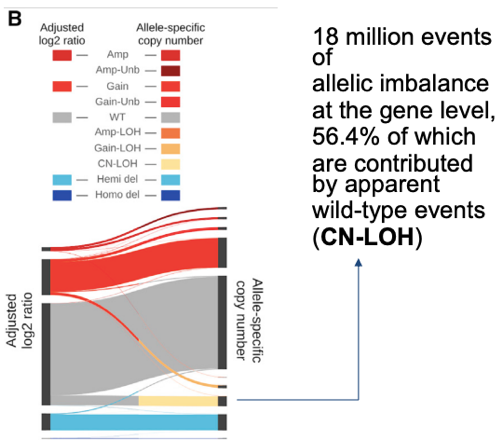
So, this whole computation allows us:

* To reclassificate copy number status in the space by shifting and stretching
* To also assign a copy number A and B to every segment of the genome, which means to every gene

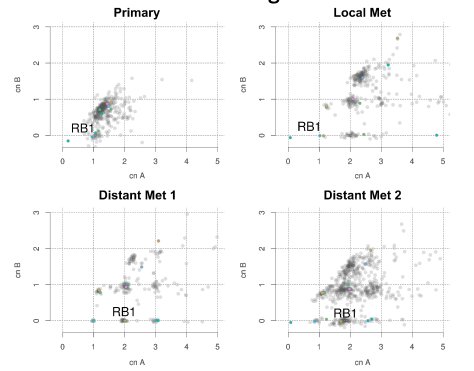
If we do that we can see that many of the segments that have a total number of copies equal to two are in fact 2+0 and not 1+1. This means that there is a significant fraction of the genome which is apparently wild-type but which actually underwent loss from one an allele and a gain on the other. This event is called copy-neutral loss of heterozygosity (CN-LOH).

Copy-neutral because the number of copies doesn’t change but there’s been loss of heterozygosity.

From the TCGA data, they observed a relevant fraction of high copy number levels (4-5 copies) which all came from the same allele (one allele was lost and the other underwent multiple cycles of duplication).

So, looking at the copy number only we’d say there’s a gain (which is true) but we wouldn’t have all the complete information (we also have to perform the allele analysis).

These information are relevant in precision medicine because there are ways to target genes exploiting loss of heterozygosity and up until now it was only used for deletions but now that’s known, even if we have an apparent CN-LOH or we have a copy number gain LOH we can still consider to use the same approach.

**Case study – CNA, CNB real data example with multi-sample data from the same patient**

We have one patient and we’re looking at a primary sample, for which we plot the whole sequencing data in the copy number allele space and what we see (from the first plot) is that:

* There’s a cloud of dots (every dot is a gene) which has a total number of copies around two
* There’s a cluster that underwent hemyzygous deletion so we only have one copy of all the genes in there
* There’s one gene with a homozygous deletion (0,0).

Then we have three other metastatic sites for which they had biopsies so that they could run whole genome sequencing and perform the analysis of the data in the same space.

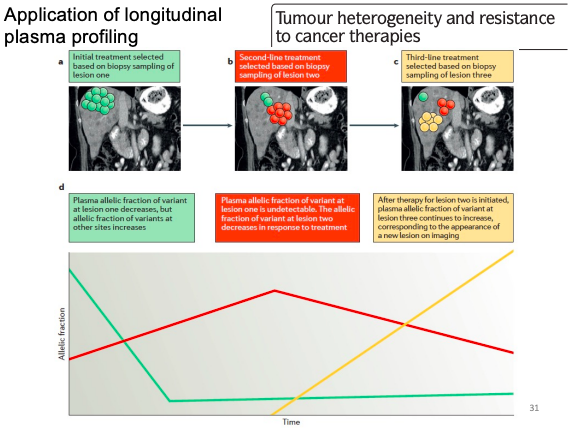
We have a local metastasis and two distant mets.

What we see:

* In distant met 1 there’s no homozygous deletion\*
* In both the distant mets the gene RB1 gained an extra copy on allele A
* In all the mets there are extra gains of copies of all the genes (maybe there’s been a whole genome duplication of some sort)
* In distant met 1 the data are as clean as to allow us to state that the data point in yellow/grey over the 1 is subclonal (if we have genes with 1+1 copy is equivalent to say it’s a subclonal hemizygous loss, it means that all the cells have at least one copy and then some cells also have a second copy)
* In terms of evolution, very likely extra copies of the whole genome also in the local met after the loss of the second copy of the gene
* CN-LOH of many genes, including RB1
* Level of subclonality overall not high

\**How’s possible that there’s a homozygous deletion in the primary tumor which is then absent in the distant mets?* No DNA can be regained, it’s impossible that the gene is reacquired, so probably the seeding of the distant mets happened before the loss of the gene.

Another way to track evolution is to have *serial time points.*



If we deal with biopsies over time we can track the evolution using the allelic fraction of a lesion.

E.g.: reasoning in terms of point mutations, let’s say we have a point mutation at time point 0 in certain allelic fractions, which correspond to different subsets, we track the fractions over time.

Doing this we can make inference of which subsets appear during the treatment and are taking over (red one in the example figure).

Allelic fraction at any time point needs to be corrected for tumor content, otherwise we would not be able to compare multiple time points from the same patient.