Chapter 1

Tumor evolution studies via NGS data

1.1 Why studying tumor evolution?

Understanding tumor evolution is useful for:

- **Academical purpose**; mainly for research
- **Clinical purpose**; the order of somatic events during tumor evolution can be relevant when considering the management of a patient, e.g. it can affect the treatment decided by the tumor board. A **tumor board** is an organism present mostly in research oriented ospitals (but also non research oriented ones and its popularity is increasing), and it is composed by different specialists (oncologist, genetist, radiologist, comutational biologist, pathologist and others) which manage the patients jointly. Aside from clinical purposes, this organism is also useful for training new experts.

1.2 Tumor heterogeneity

A tumor can arise due to:

- A single (or few) strong driver mutation (in oncogenes or oncosoppressor genes)
- Several mutations that gradually change the cell phenotype without leading to cell death

In both cases the mutations are somatic events due to stochastic processes, mainly due to carcinogenic substances that damage DNA therefore causing mutations (but also physcal phenomena such as radiations and others). These mutations are mostly associated to cell growth; this is why tumor cells often undergo clonal expansion and create a mass. The speed at which the mass grows and mutates is dictated by the mutations that occurred.

A tumor mass can be either homogeneous or heterogeneous; in general, the more aggressive and old the tumor is, the higher the degree of heterogeneity. Higher heterogeneity usually correlates to drug resistance (since some of the clonal populations mught be able to better resist the drug compared to others).

New Generation Sequencing allows to study all the somatic mutations that occurred in a cell, both the cancer related ones and the benign ones. By sequencing with an appropriate depth, you can infer in which fraction of the cell population a certain mutation is present; this allows to reconstruct the clonality of the tumor and the mutation history. Notice that since very deep sequencing can give errors, you usually need to check different loci in order to consolidate your result.

Tumor heterogeneity can be subdivided into:

- **Inter-individual heterogeneity**: tumor from patient A is different from that of patient B
- **Intra-individual heterogeneity**: tumors from the same patient might differ
  - **Spatial heterogeneity**: synchronous tumor masses in the same patient might differ
  - **Temporal heterogeneity**: a tumor might changes overtime, due to spontaneous or drug induced selection
  - **Intra-lesion heterogeneity**: an individual tumor mass might present different lesions (which display different clones with different mutations, therefore different phenotypes, treatment resistances and so on)

Almost always, many of these heterogeneities are present simultaneously.

Notice that genetic heterogeneity does not necessarely reflect morphological heterogeneity (e.g. different prostate lesions might look the same when stained but then display different markers using in situ immunochemistry). Moreover tumor mass size does not necessarely correlate with aggressiveness (hence imaging is not enough to study tumors). Heterogeneity might cause problem in the interpretation of the spectrum obtained via sanger sequencing, since the sample might contain different sequences for the same locus, hence leading to an overlap in the peaks. In case of different lesions, we can define tumor burden and features of each of them via individual sequencing.

Tumor evolution can happen in two ways:

- **Linear evolution**: genetic instability leads to new tumor clones and if those display some advantage compared to the previous ones, the older ones get replaced by the new ones (otherwies the new clone dies down). In this case you generally have low heterogeneity.
- **Branched evolution**: genetic instability leads to the formation, from an ancestral clone, of different clonal populations which can coexist in the same or different tumor masses. In this case you generally have high heterogeneity.

A metastasis can either have:

- **Monoclonal origin**: meaning that it originates from tumor cells coming from a single lesion. In this case you have similar features as the starting mass and overall low heterogeneity within the metastasis.
- **Polyclonal origin**: meaning that it originates from multiple tumor cells coming from different lesions. This phenomenon is called **multiclonal seeding** and it leads to high lesion heterogeneity. Moreover, the fact that it displays some of the features from each of the parental lesions makes the analysis more complex.

As previously mentioned, tumor heterogeneity plays a big role in defining treatment resistance. We talk about two types of drug resistance:

- **Primary resistance**: the pre-treatment tumor mass already contains cells that are resistant to the treatment; the treatment kills the non-resistant cells, hence the resistant clone expands.

- **Acquired resistance**: the pre-treatment tumor mass does not already contains cells completely resistant to the treatment; the clones that can survive the treatment the best could then mutate in order to acquire a treatment immunity mechanism.

In case of primary resistance, the tumor might already display some biomarkers pointing to some treatment resistance; this is useful for the tumor boards in order to avoid needless harmful treatments. However, no biomarkers for each treatment are known, plus the tumor can always evolve unpredictably and acquire a new resistance.

1.3 Algorithms to study tumor evolution

You can study tumor evolution using information from:

- Samples from the **same subject**, from different time points or lesions; this way you can reconstruct mutation order and metastatic processes within the individual (base on shared or not mutations).
- Samples from **different subjects** affected by the same pathology (e.g. prostate cancer); you use recurring patterns across individuals, this way you can reconstruct more generic features of the pathogenesis, for instance:
  - Very common mutations in the pathology (those shared across many individuals)
  - Mutations that tend to happen in a specific order (take for instance two mutations A and B; if in the majority of tumors which present both lesions, B is almost always subclonal to A, then probably A tends to happen prior to B).

For more in depth reading (clickable links):

- *The evolutionary history of lethal metastatic prostate cancer, Gundem et al, Nature 2015*
- *Punctuated evolution of prostate cancer genomes, Baca et al, Cell 2013*

NOTE: I did not add some pictures even though they were commented in class because I think that the relevant part was understanding the points listed above.

In general, when you have some tumor data, you try to see which of your models best fits the progression.

There are several aspects that must be taken into account during this type of analysis; most of them are useful in comprehending the pathology and its mechanisms, but at the same time they make analyzing the NGS data more difficult. Some of these aspects are:

- **Heterogeneity** (inter-patient, intra-patient, intra-tumor)
- **Time dependence** (tumor changes overtime)
- **Treatment status** (was the tumor treated, if yes how?)
- **Admixture DNA** (presence of non-tumoral DNA, *explained more in depth below*)

In a tumor biopsy you could have (and this is generally the case) other cells that are not tumoral (healthy tissue cells, stromal cells, leukocytes...). It is then defined the concept of **admixture**, which is *the fraction of non-tumoral DNA within the sample.* Admixture is then used to define **tumor purity**, which is

$$\text{tumor purity } = 1 - \text{ admixture}$$

To sum up, a fully tumoral sample would have admixture equal to zero and purity equal to one. The opposite holds for healthy tissue (purity equal to zero, admixture equal to one). Deconvoluting the sequences derived from admixed DNA complicates NGS data analysis, but also provides useful information:

- Aggressiveness of a lesion; in general, the lower the purity, the better the outcome
- Defining whether a mutation is actually part of a subclonal tumor population or it is just admixed DNA

The most useful feature from NGS for characterizing tumor evolution (clonality, purity and so on) are:

- Copy number mutations
- Point mutations
- Single cell sequencing
- Polymorphic information (which SNPs does the tumor have)

The algorithms used to study tumor evolution use **informative SNPs**, meaning:

- SNPs for which the individual is heterozigous (hence they vary from individual to individual)
- SNPs for which the allelic fraction is easily measurable

Making parsimonious assumptions (mainly that all clones have the same growth rate), these algorithms allow to study any form of genetic aberration.

REMINDER: Do not confuse the following concepts:

- **Minor allele frequency**: frequency of the alternative allele for a locus in the **entire population**
- **Allelic fraction**: frequency of the alternative allele for a locus in a **single individual**. It is a local property of the individual. In terms of NGS this becomes:

$$\text{Allelic fraction} = \frac{\text{locus reads with minor allele}}{\text{total locus reads}}$$

Other important concepts to consider are:

- **Neutral reads**: reads equally representing parental chromosomes
- **Beta fraction**: percentage of neutral reads. Beta goes from 0 to 1; the closer the value to 1, the closer the reads are to a 50/50 split among parental sequences, the closer the value to 0, the closer the reads are to a 100/0 split in favour of either parental sequence.

NOTE: The way to compute beta values is in the slides but skipped during the lecture
The **allelic fraction** for an informative SNP can be:

- 0 if the alternative allele was deleted
- 1 if the reference allele was deleted (the non-alternative one)
- Around 1/2 if both alleles are present in equal proportion

- Some other value in the range (0,1), that could be due to duplication, heterogeneity (admixture and/or subclonality), errors and so on. In this case some further information might be required (for instance the coverage)

The **beta fraction** can be:

- 0 if either allele was deleted (hence you have only one)
- 1 if both alleles are equally-represented (normal condition)
- Any other value in the range (0,1), and this is also due to heterogeneity and other factors.

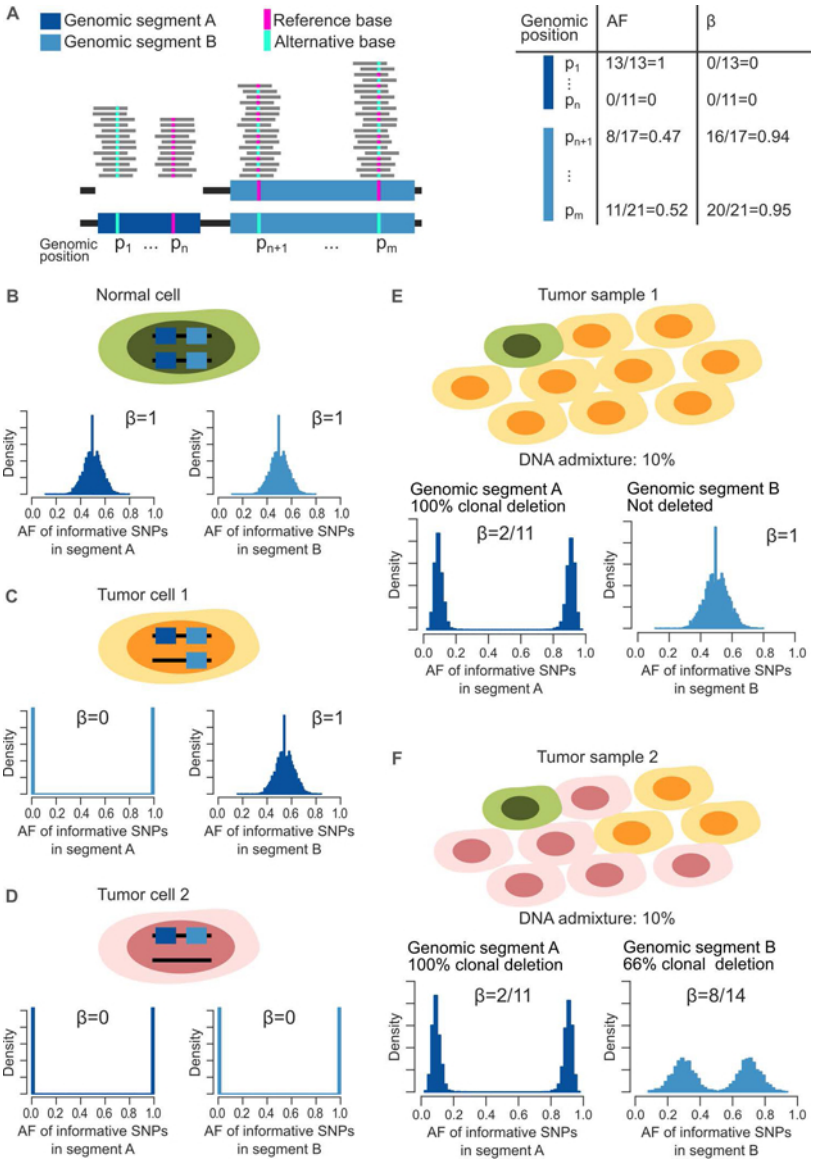
Notice that allelic fraction and beta fraction give similar information, but the beta fraction is not allele specific, hence it is better suited to study genetic abnormalities.

Using allele frequency and beta fraction, informative SNPs can be used to reconstruct the genealogy of the mutations. If there is a deletion of a region then you have a loss of heterozygosis for all SNPs in that region (since you chose informative SNPs, hence heterozygous ones); then based on the mutations present or absent in the different clones of the lesion (since you do not have a perfectly homogeneous mass) you can reconstruct their order.

When designing a test you need multiple informative SNPs for each genomic fragment of interest. Moreover you have to choose alleles that have high MAF (hence the minor allele frequency is still rather high), since those are more likely to give you information.

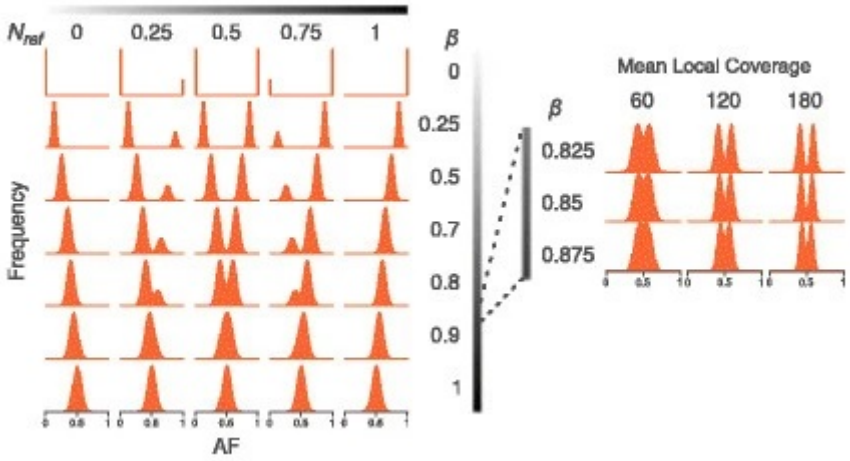
For more in depth reading (clickable links):

- *Ploidy- and Purity-Adjusted Allele-Specific DNA Analysis Using CLONETv2, Davide Prandi, Francesca Demichelis, 2019*



1.4 Estimating admixture and clonality

Notice that the beta fraction correlates with the shape of the distribution of the allelic fractions of the informatives SNPs in the read; with beta = 1, you have a normal distribution with mean 0.5, with beta = 0 you have two sharp peaks at 0 and 1, with any intermediate value you have to peaks which can be partially overlapping for values close to 1. Notice that increasing the coverage does increase the resolution of the peaks. For this reason increasing the coverage (with beta constant) does increase the ability to distinguish clonality, especially of populations that are only some degree of difference from each other.



To estimate the admixture/clonality of a cell population:

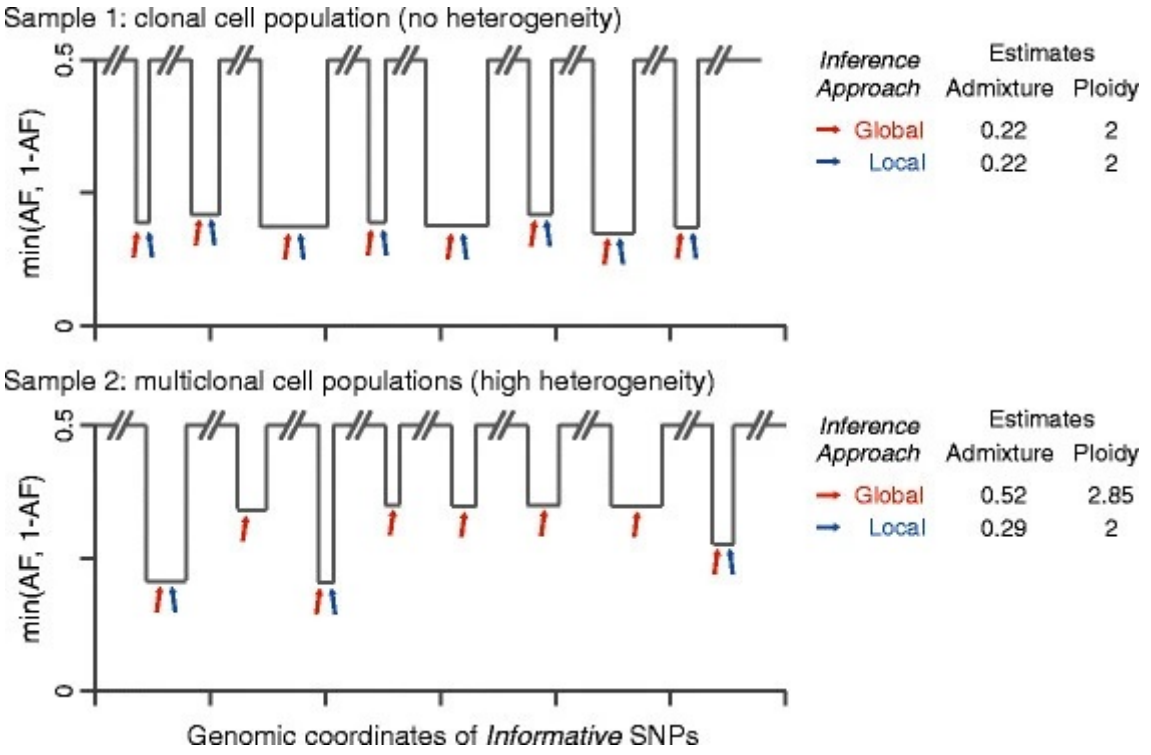
- Measure the allelic fraction and beta fraction of each informative SNP of a genomic region
- For each region try which of the models fits your data the best (basically map the distribution of the allelic fractions of the region against prefitted reference distributions)
- You can then compute the local and the global admixture:
  - **Local admixture** is a measure of the fraction of cells displaying a certain lesion with respect to another; for this reason local admixture is used as an estimate for **clonality**
  - **Global admixture** is a measure of how many cells, on average, have a lesion; this can be used to estimate the **DNA admixture** (purity) of the sample

Hence this technique allows you to distinguish purity and subclonality.

Graphically you obtain a plot with:

- On the x axis, the cromosomal coordinates indexed by informative SNPs. The longer the horizontal segment, the bigger the considered region.
- On the y axis, the MAF values for the informative SNPs. MAF values are mirrored on 0.5, since you do not care about distinguishing the alleles. Any drop below the 0.5 value means that the region does not have a 50/50 split. The deeper the drop the deeper the difference in the representation of the alleles.

In the example picture, the top subplot shows drops which have very similar depth, hence global and local admixture are similar and there is very low heterogeneity. In the bottom subplot you have differences in local and global admixture, hence we can infer the presence of different clonal populations.



For this type of analysis is always useful to have the **match normal DNA** (the non-tumor DNA of the subject): match normal DNA is usually obtained from leukocytes in the blood, otherwhise one could somehow deconvolute the signal of the admixed cells.

Another graphical representation in bidimensional space is the following plot:

- On the x axis the **log2 ratio**, meaning

$$\text{log2 ratio} = \log_2 \frac{\text{local tumor coverage}}{\text{local normal coverage}}$$

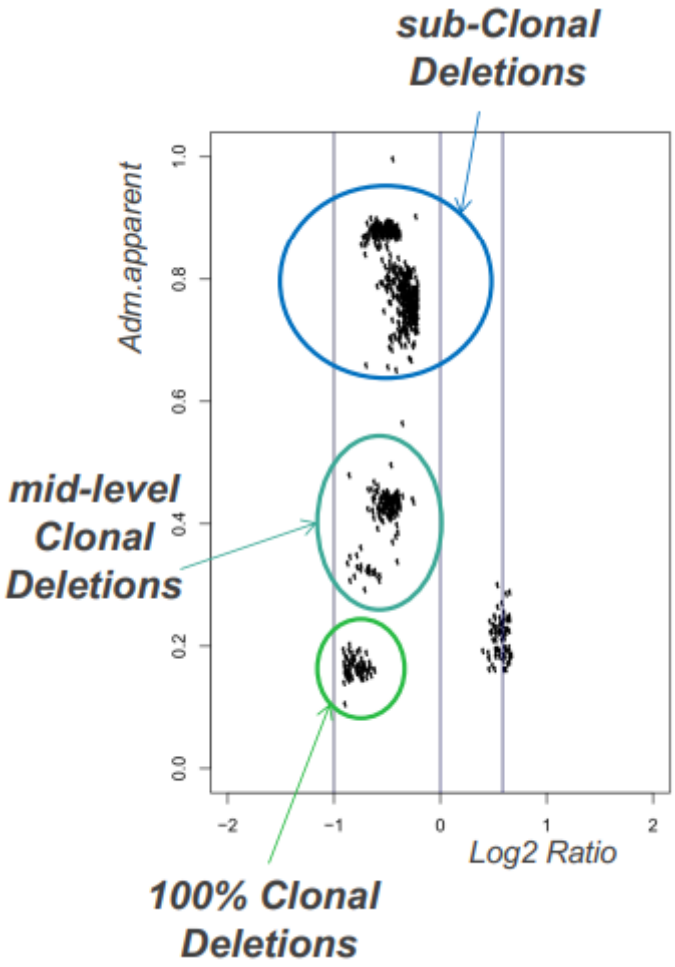
  This indicates how abundant cancer DNA is with respect to healty DNA (gain of DNA if above zero, loss of DNA if below zero).

- On the y axis the **apparent admixture**, which is defined as

$$\text{Adm. apparent} = \frac{\beta}{2-\beta}$$

  Notice that this measure refers to each individual deletion/abnormality.

- The dots which represent the individual genomic segments. The dots tend to create multiple clusters and the closer two points are, the more probable the events they represent are close to each other in time.



You can compute clonality using the formula:

$$\text{clonality} = \frac{1 - \text{Adm. apparent}}{1 - \text{Adm. global}}$$

An example of how heterogeneity can lead to difficult to interpret results can be found in the following paper: *Unraveling the clonal hierarchy of somatic genomic aberrations NOTE*: The case study was rushly explained during the lecture, but in my opinion it did not provide any further information; this is one of the papers uploaded on moodle.