Human genomics

Giacomo Fantoni
telegram: @GiacomoFantoni
Github: https://github.com/giacThePhantom/human-genomics

June 4, 2022

Contents

Part I

Notes

Chapter 1

Introduction

Genetics is the study related to specific genes or variants that do have an effect on ourselves. Genomics instead handles the function and structure of the human genome, evolution and anything relating to the whole genome e.g. coding regions, non-coding regions, linear/non-linear structure, anything that relates to the cell, generation of diseases during lifetime.

1.1 Genetics vs Genomics

Genetics Genetics is the study of heredity, or how the characteristics of living organisms are transmitted from one generation to the next via DNA. It dates back to Augustinian friar and scientist Gregor Mendel. It involves the study of a specific and limited number of genes or their part that have a known function.

Genomics Genomics is the study of the entirety of an organism's genes, the genome. Using high-performance computing and math techniques known as bioinformatics, genomics researchers analyse enormous amounts of DNA-sequence data to find variations that affect health, disease or drug response. In human that means searching through about 3 billion units of DNA across 23000 genes.

1.1.1 Differences

The main difference between genomics and genetics is that genetics scrutinizes the functioning and composition of the single gene, where genomics addressees all genes and their relationships in order to identify their combined influence on the growth and development of the organism.

The role of computational biology Computational Biology encompasses a wide range of numerical methods to analyze and integrate (large scale) data towards the understanding of molecular, cellular and structural biology. Examples: semi-quantitative simulations of metabolic pathways, 3D protein-protein interaction, characterization of 3D chromatic structure, discovery and characterization of disease related variants. The main subjects involved are biology; genetics; statistics; calculus; computer science / informatics; bioinformatics; technology. We will focus on human genomics, on how to mine raw (mainly sequencing) data, how to exploit data for Quality Control (QC), and how to interpret the results, in the context of human diseases, especially cancer.

1.2 Human Genomics - the Basis

1.2.1 Genetic Make-Up

The individual's genetic make-up is different in all of us and it is responsible for human diversity. SNPs (single nucleotide polymorphisms) and CNVs (copy number variants) contribute to make us all different. The majority of external phenotypes are from genetic variance that we inherit (but they can also be aquired).

1.2.1.1 Single nucleotide polymorphisms

Single nucleotide polymorphisms or SNPs are changes of one nucleotide in the sequence of a gene. They constitute 1% of the difference between two unrelated individuals' genomes and we can use them as quality control assets.

1.2.1.2 Copy number variants

Copy number variants or CNVs are the difference of the number of allele for a gene present in one individual. They contribute much more than SNPs in the difference between unrelated individuals, but they're less known as inherited variants, due to historical reasons (harder to quantify and detect). There are three different scenarios in the deletion CNV: 2 alleles, 1 allele, or 0 alleles (If both parents are monozygous in one gene the child could have zero copy of the gene). CNV span »1% difference between two unrelated individual genomes.

1.3 Inherited variants' relevance

Inherited variants can be characterized by penetrance and allele frequency.

Penetrance Penetrance is the proportion of individuals carrying an allele (or genotype) that also expresses the trait (or phenotype) associated with it.

Allele frequency Allele frequency is the ratio between the number of times the allele of interest is observed in a population over the total number of copies of all the alleles at that particular genetic locus in the population.

Recent studies have shown that genetic variance contributes to predisposition to certain diseases. What is also emerging is that if we are dealing with a very rare variant, if this variant is pathogenic, it also has high penetrance. Meaning, if the variant is pathogenic and very rare, it's very probable that all patients affected by the disease carry this mutation. This is shown in the top part of the diagram shown in figure 1.1. On the other hand, common variants could be associated to predisposition or susceptibility to the disease, but the penetrance is very low. In the middle on the diagram we find very well-known variants correlated to cancer. The majority of these have a moderate size effect (not everyone who has the variation develops the disease).

1.3.1 Differences in Genetic Make-Up, an example

One example of how the genetic make-up plays a role in deases is the ADME genes. ADME stands for *Absorption, distribution, metabolism and elimination.* It is a set of genetic variants that is

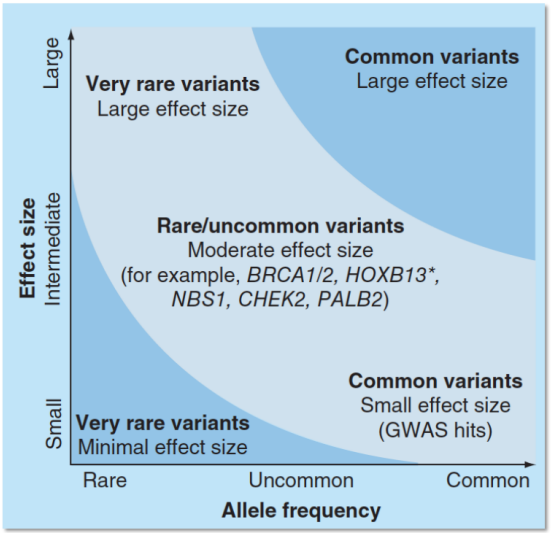


Figure 1.1: R. Eeles, Future Sci. OA (2016) 2(1), FSO87. Review on prostate cancer.
Common SNPs: $\frac{1}{4}$ has a homozygous for A1, $\frac{1}{4}$ has a homozygous for A2, $\frac{1}{2}$ heterozygous $\rightarrow$ minor allele frequency around 30-50%, low penetrance.
Rare SNPs: very large size effect i.e. if they are related to very specific traits they will have high penetrance e.g. deleterious variant.

able to change ability of the organism to react to certain compounds (pharmacokinetic variability),influencing the patient's treatment response. Both common and rare variants are involved.

Table 1. Comparison between pharmacogenomics approaches.

PGx Approach	GWAS	SNPs Panel	Candidate SNP
Sample size	Tailored for large populations	Tailored for small populations	Tailored for small populations
Number of investigated markers	Larger numbers	1–2 thousand	Smaller number
Hypothesis	Hypothesis-free and hypothesis generating	Hypothesis-free and hypothesis generating/PK and PD coverage	Selected on a priori knowledge
Study Design	Exploratory	Confirmatory/Exploratory	Confirmatory
Limitations	False Negative/control for multiple testing	Coverage of limited genes	False positive/non-replication of results/low genetic coverage

PGx: pharmacogenomics; GWAS: genome-wide association study; SNP: single nucleotide polymorphism.

Figure 1.2: From review: *Pharmacogenomic Profiling of ADME Gene Variants: Current Challenges and Validation Perspectives*

A therapeutic approach that considers these variations could be very useful in precision medicine.

1.3.2 Somatic Variants

Somatic variants are **not**inherited from parents and are not transmitted to offspring (most are harmless, but not all). They are also not present in every cell.
Somatic variants can be classified in:

- **Single Nucleotide Variants (SNV)**. SNVs and SNPs are basically the same thing (a point mutation), but SNVs are restricted to a certain population of cells, while SNPs are genetically encoded in all cells of the organism.;

- Indels (few nucleotides)
- Rearrangements, like gene translocation, chromosome breakage, chromothripsis (which falls in the subcategory of chromosomal rearrangements).
- Somatic Copy Number Aberrations (SCNA)

1.3.2.1 Types of acquired DNA aberrations

Translocation Translocation happens when a sequence is moved from one genetic locus to another. It can be a balanced translocation, meaning that the overall quantity of DNA is maintained (two sequences exchange locus), or unbalanced, where only one sequence move (insertion)

Inversion Inversion happens when a sequence inverts its orientation. It involves only one chromosome. Importantly, in the sequence of the inversion nothing changes, the change will be detected only at the head and tail of the inversion. Copy number changes (DNA quantity): duplication and dletion. COuld involve one or more chromosomes

Copy number changes It refers to a change in the quantity of DNA. In duplication a sequence doubles its copy number. In deletion a sequence is lost.

Chromoplexy From the Greek *pleko*, meaning to weave, or to braid. A class of complex somatic DNA rearrangements whereby abundant DNA deletions and intra- and inter-chromosomal translocations that have originated in an interdependent way occur within a single cell cycle.

Chromothripsis (From the Greek *thripsis*, meaning shattering into pieces). A clustered chromosomal rearrangement in confined genomic regions that results from a single catastrophic event, usually limited to one chromosome.

Kataegis (From the Greek kataigis, meaning thunder). A phenomenon that is characterized by large clusters of mutations (hypermutation) in the genome of cancer cells. An APOBEC family enzyme might be responsible for the kataegis process.[1]

1.4 Experimental techniques to detect variants/aberrations

1.4.1 Cariotyping

Basically all the aberrations described in 1.3.2.1 were discovered in the last 10-15 years because there's the need of NGS. Karyotyping indeed is not enough! Sequence specific variants, breakpoints, etc. could not be detected until NGS.

1.5 Sequence capture for cancer genomics

In the paper summarized in **??** it is described a typical sequence capture for cancer genomics.
One of the main realizations in DNA sequencing is that the test reference genome is the normal (non-cancer) DNA. Most of the times when studying somatic changes a reference is needed, and the best

[1] See Khurana E et al, NATURE REVIEWS | GENETICS, 2016

reference is the individual's own genome, usually retrieved from white blood samples. Intuitively, we can align both cancer and normal DNA so that we can detect if an aberration is cancer specific or it is present also in the normal DNA and discover somatic aberration. The **match normal** is used to distinguish SNV from rare SNPs, somatic and germline indels, but also to make sure that copy number variations are somatic. Baits are nowadays used in the sequencing step, in order to sequence only the exome (usually, money issue).

Another concept is the need to sequence *deeply*, to find subclonal events that might give the cancer some fitness (like escaping immune system). Deep sequencing is necessary because not all cells present all the mutations characterizing the subclonal event, but also because if the sample comes from the tissue (tumor tisue), there are both cancer and some healthy cells and we need to able to distinguish them.

1.5.1 Single End (SE) and Paired End (PE) reads

Discalimer: the difference between PE and SE is not described
Paired end is useful especially for detecting structural variance. PE gives important information of relative position of a molecule wrt the reference genome, resulting in a necessary choice in designing the assay when we need to check for structural aberrations.
Interestingly, PE reads can be considered as two single reads from one molecule (PE protocol); but there's a loss of structural information. PE protocol is, however, twice as expensive!

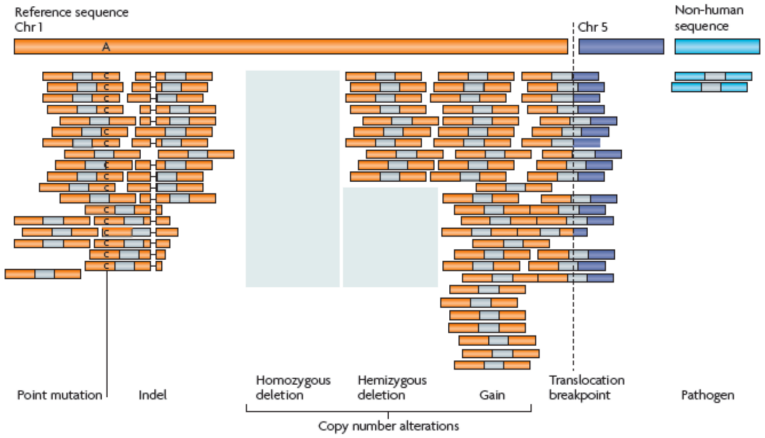


Figure 1.3: *Advances in understanding cancer genomes through second-generation sequencing*, Meyerson et al., Nature Reviews Genetics 2010

Figure 1.3 gives a nice graphical overview of genomic aberrations detectable by NGS, especially using PE sequencing.

Important to notice is that when performing PE sequencing is like having double the coverage. Both for homozygous and hemizygous deletions and insertions the single most important thing we need to care about is to have enough coverage on the whole experiment to perform significant downstream analysis. Notice in 1.3 the translocation breakpoint: without PE we would not be able to detect the translocation event.

Chapter 2

Coverage

2.1 Local Coverage and Allelic Fraction

Two key concepts needed when perfoming genomics analysis are the **local coverage** and **allelic fraction**.

Local coverage (cov) The local coverage (cov) at position (base) i is the number of reads that span p_i.

Allelic fraction (AF) The allelic fraction (AF) at position i is the proportion of reads that supports the reference base in p_i, and viceversa.

The Lander-Waterman equation to compute NGS coverage is:

$$C = \frac{L * N}{G} \tag{2.1}$$

Where G is the coverage, G is the haploid human genome length, L the read length and N the number of mapped reads.

2.1.1 Mapping in NGS

The number of mapped reads is always lower than expected. Errors, or major translocation will impair a good mapping of the reads.
There's a difference between physical and sequence coverage. Physical coverage is always higher, and it changes based on which protocol (PE or SE) is chosen for the assay. To put it simply, when calculating the sequence coverage we only take into account the actual ends, while when calculating the physical coverage we also account for the consequence part of the PE protocol. In figure 2.1 a shcematic representation of this problem is displayed.

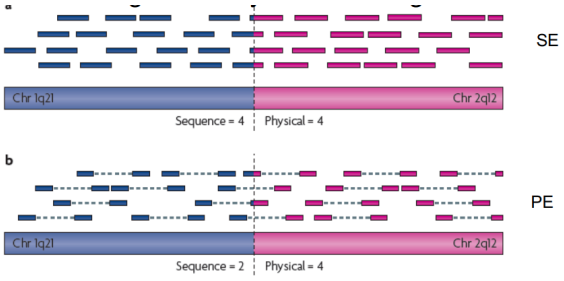


Figure 2.1: Schematic difference between sequence (on the left) and physical coverage (on the right). From Meyerson et al., Nature Reviews Genetics 2010

Formal definitions of sequence and physical coverage are:

Sequence coverage: Sequence coverage is the amount of oversampling (how many times a base is sequenced); to detect nucleotide alterations with high sensitivity, the 3 billion bases of the human genome have typically been 'covered' with at least 30-fold (30X) on average, meaning 90 billion bases of sequence data per sample.

Physical coverage: the expected distance between the paired reads is used to uniquely place the reads on the genome; unexpected read pairs are used to detect structural anomalies.

2.2 Tuning the intended coverage of a NGS assay

In some experiments setups there's the need to carefully control the amount of intended coverage. If we are looking for SNPs only, which by definition are present in all of the cells, we only need enough redundancy (local coverage) to detect them and distinguish the reference base and the alternative base (in case of an heterozygous SNP we will ideally find half of the reads supporting the reference and half supporting the alternatice). Indeed for SNPs we do not need more than 10-15 X coverage. However, if the sample comes from a tumor or hematopoietic events, we need to look for subclonal events. Subclonal events are not harbored by all of the cells but only by a fraction of them. If we expect 1/4 of the cells harboring the mutation, we need to increase the coverage.
The same reasoning goes for any monozygous mutation and any low abundant events, and for transcripts expressed at very low level (RNA-seq) and weak binding in ChIP-Seq.

2.2.1 Example on the importance of coverage

Here are represented 10 genes relevant for cancer. Each bar represents the average local coverage at 30 bp.

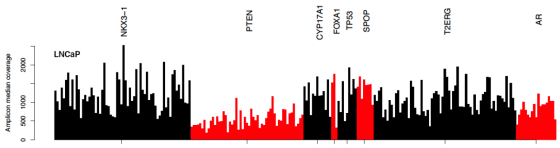


Figure 2.2: Example of local coverage. On the x axis genomic location (on top e ery gene) and the bars is the number of amplicons.

In figure 2.2 a barplot can be observed, representing the local coverage (y axis) in the gene locations (x axis). The *cov* is on average about 600x and it's "wavy", not evenly distributed (very typical). However, if one would do an average of the coverage for each gene, they would discover that one gene is abundantly underrepresented: PTEN.
What's happening on PTEN base on plot 2.2? Probably, the most accurate guess is a deletion. But of which kind? For sure not a homozygous deletion, since we would not be able to see any signal in the plot. From this data however we cannot say whether this is a monoallelic or biallelic deletion.
Note that this (and the following 2.4) plot is the actual way sequencing data is shown, while the figure presented in 1.5.1 was a schematic representation.

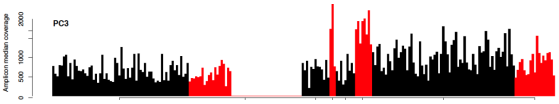


Figure 2.3: Example 2 of local coverage. The cell line is different (PC3 instead of LNCaP).

However, in the plot 2.3, from a different cell line, we can see a clear monoallelic deletion and a partial biallelic deletion of PTEN.

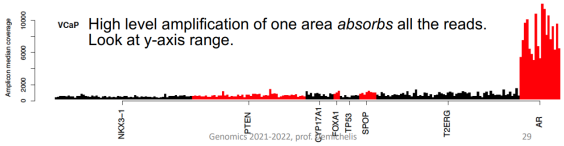


Figure 2.4: Example 3 of local coverage. The cell line is different (VCaP). The massive amplification of the AR region is typicall in advanced prostate cancer.

In figure 2.4 we can see a massive amplification of the antigen receptor. Probably it was a mistake in the assay: amplification on the AR does not allow the analysis/discovery of other copy number variations because all of the reads will go to the AR site.
These were amplicon-based approaches. With NGS instead, It is easy to increase the experimental coverage (i.e. the sequence depth) at later point in time. Provided our original sample/library is still available, we can perform another run of sequencing and then combine the output from different runs. Note that this isn't possible with array-based technologies.
However, there are some limiting factors of NGS DNA-seq experiments. Problems with repeated regions, but also not knowing the linearity of the genome. If the sequencing is done with longer reads we could tackle the problem by having longer molecules to work with, but there are more errors in the reads.

2.2.2 Databases

Two very known databases for NGS analysis are:

- **Genome Reference Consortium**: assemble a reference genome reflecting the most common sequences in population at each position while tracking information on polymorphisms.
- **USCS Genome Browser**: select a reference genome and query all known features.

2.3 Interpreting Pair Orientations

We will now look at some aberrations' discoveries performed using IGV.
In IGV, each vertical bar corresponds to a read. If there is a colored sign, there is a polymorphism or difference with respect to a reference. The bowser also gives info about the quality of the read and bases (and others from the BAM file).
While using a paired-end protocol, we can study inversions, duplications and translocations. A useful legend to navigate the subsequent example is reported in the list below.

- **LR** ($\rightarrow$ $\leftarrow$): Illumina (convention), the reads are left and right of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome;
- **LL**, **RR**: implies inversion in sequenced DNA with respect to reference ;
- **RL**: implies duplication or translocation with respect to reference;

2.3.1 Inversion

To detect major aberration, like inversions, we need reads that span the breakpoint (either long reads, or, better, PE reads).
What's happening exactly at the breakpoint in figure 2.5? What's the coverage when looking at the data? When we map them on the reference, we see that the direction is LL and RR, meaning that there is an inversion. We can also spot a drop in the local coverage. In the target molecule, either it does not exist or exist only in one allele and not the other one.
When interpreting structural variance, we need not to care only about end-orientation, but also coverage at the exact break point, which is due to the fact that the inverted breakpoint sequence does not exist in the reference genome.

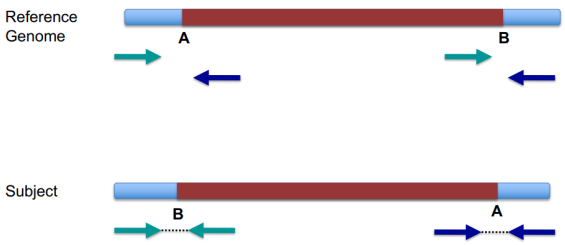


Figure 2.5: Inversion disocvering exploiting PE reads.

2.3.2 Tandem duplication

Notice how in the tandem duplication (figure 2.6) all the reads that do not cover the junction point align perfectly to the reference. The coverage will be 3/2 of the expected value, proportional to the extra copy. B junction and A junction will not have modifications. If we had a read mapping BA, we would observe a partial alignment at B on the reference. We observe no drop of coverage because the sequences also exist in the reference.

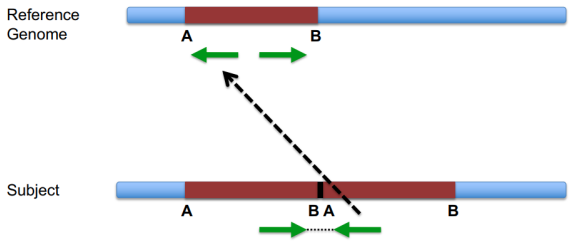


Figure 2.6: Tandem duplication discovering exploiting PE reads.

2.3.3 Inverted duplication

For the inverted duplication, in figure 2.7, we expect double coverage in the duplicated site in the reference genome.
Both A and B on the first segment on the subject are LR oriented, and the same occurs in the reference genome. When we add the second fragment the same holds, but direction will be LL and RR and the insert size will be significantly longer. In particular, we can notice that we have an overlapping of left and right reads on the reference. Furthermore, the coverage depth will highly increase due to the presence of multiple reads on the reference genome.

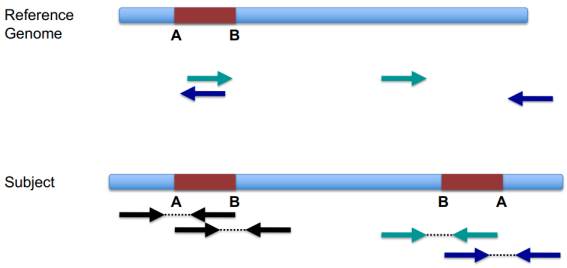


Figure 2.7: Inverted duplication discovering exploiting PE reads.

What if we have a deletion? How can you guess the size of it? We can look at the coverage or observed distance between the reads, which gives clean indication of the size of the deletion. For tiny deletions, smaller than the length of the read, we use the sequence within the reads and in this way discovering indels.

2.4 Summary

Summarizing, the elements to consider are:

- Pair ends relative orientation;
- Insert size length;
- Coverage within the aberrant region;
- Coverage outside of the aberrant region (flanking genomic segments);
- Coverage at the breakpoints.

Chapter 3

Genetic Fingerprinting

Genetic fingerprinting is a tecnique used to identify some characteristics of a genome (a pattern of variable elements), like SNPs or minisatellites, in order to uniquely characterize a genome.
Genetic fingerprinting can be used to compare a genome with a reference sample or to compare different genomes between each other, in order to determine their diversity or analogy.

DNA fingerprinting is applied in different fields:

- In Forensic, for identification purposes;
- In lineage related tests, for cells or humans. Eg. pternity test, hereditary tests.
- For the certification of the origin of cells used in the laboratory, to make sure that the cells are the right ones and that there are no major geentic drifts. Needed when using certain cell lines, for publishing purposes.

3.0.1 Variants used for genetic testing

There are different variants that can be used for genetic fingerprinting, such as Single Nucleotide Polimorfisms (SNPs) or inherited Copy Number Variations (CNVs). SNPs are substitutions of a single nucleotide at a specific position in the genome, whereas copy number variation is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals. Basically everything that is inherited and that is a polimorphism can be used in genetic testing, however some variants are more amanable than others. SNPs are the most amenable ones since they are simple, abundant in the genome and easy to detect in sequencing data at any coverage depth. For these reasons, in this lesson we will focus on the development of SNP-based genetic tests.

3.1 SNPs features

3.1.1 Hardy-Weinberg equilibrium and Minor Allele frequency

One property of SNPs which has to be taken into account when using SNPs for genetic testing is the **Hardy-Weinberg equilibrium**. In population genetics, the Hardy-Weinberg equilibrium states that allele and genotype frequencies in a population will remain constant from generation to generation under neutral selection, so in the absence of other evolutionary influences, like genetic drift, mate choice, sexual selection, mutation and so on.

In the simplest case of a single locus with two alleles denoted A and a with frequencies f (A) = p and f (a) = q, respectively, the expected genotype frequencies under random mating are f (AA) = p^2 for the AA homozygotes, f (aa) = q^2 for the aa homozygotes, and f (Aa) = 2pq for the heterozygotes. In the absence of selection, allele frequencies p and q are constant between generations, so equilibrium is reached. SNPs that respect this equilibrium are also the most studied, thus more informative.

3.1.2 Minor Allele Frequency

Also, when performing genetic fingerprinting, the aim is to maximize the probability to have different genotypes in unrelated individuals. For this reason, the more advantageous SNPs will be the ones in which the allelic frequency of the variants is the higher possible. Highest variability in the population allows to distinguish better more individuals.

Number-wise, a frequency of $\frac{1}{3}$ for each SNP would maximize the variability, but those SNPs wouldn't be in HW equilibrium and we might have missed calls. Therefore, the optimal SNPs to detect individuals' differences and similarities are those with genotype frequencies: $P_{AA} = 0.25$, $P_{BB} = 0.25$, $P_{AB} = 0.5$. 50% of individuals for that SNP will have a heterozygus genotype, 25% a homozygus genotype for the reference allele, 25% for the alternative allel.

This is equivalent to say that best SNPs will be the ones with **MAF** = 0.5. Minor allele frequency (MAF) is the frequency at which the second most common allele occurs in a given population.

Some useful projects:

- **dbSNPs**: is a database of small scale nucleotide variants. The database includes both common and rare singlebase nucleotide variation (SNV), short (=< 50bp) deletion/insertion polymorphisms, and other classes of small genetic variations. `https://www.ncbi.nlm.nih.gov/snp/`
- **HapMap3**: is teh third phase of the HapMap project whose aim is to develop a haplotype map of the human genome to describe the common patterns of human genetic variation in order to allows researchers to find genes and genetic variations that affect health, disease and individual responses to medications and environmental factors. The HapMap is a catalog of common genetic variants that occur in human beings. It describes what these variants are, where they occur in our DNA, and how they are distributed among people within populations and among populations in different parts of the world. `https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html`

3.1.3 Haplotype Blocks

Another important feature to consider for SNPs selection are **Haplotype blocks**. Haplotype blocks are blocks along the genome that tend to be inherited as segments (no recombination inside a haplotype block). In these sizable regions there is little evidence for historical recombination and only a few common haplotypes are observed.

So for example, if there are 10 SNPs in a block of 1 MB, the genotype of one specific SNP in that block gives an indication the genotype of the other SNPs in the same block, since they are inherited together. Hence, if there is a haplotype block, there is no point in sequencing all SNPs in that block, it is sufficient to select some specific SNPs. Also, when running a fingerprint assay, there is no point in using all SNPs in a haplotype block since they won't bring additional information independently.

SNPs in the same HB are said to be in **Linkage Disequilibrium** (LD). Linkage disequilibrium measures the non-random associations between alleles or polymorphisms at different loci. A higher LD indicates a SNPs with a stronger tendency to co-segragate. Haplotype Blocks are therefore commonly represented with *linkage disequilibrium plots* 3.1. In these plots, SNPs are represented in a way that does not respect the genomic distance, but the order along the genome (position of each SNP relative the others).

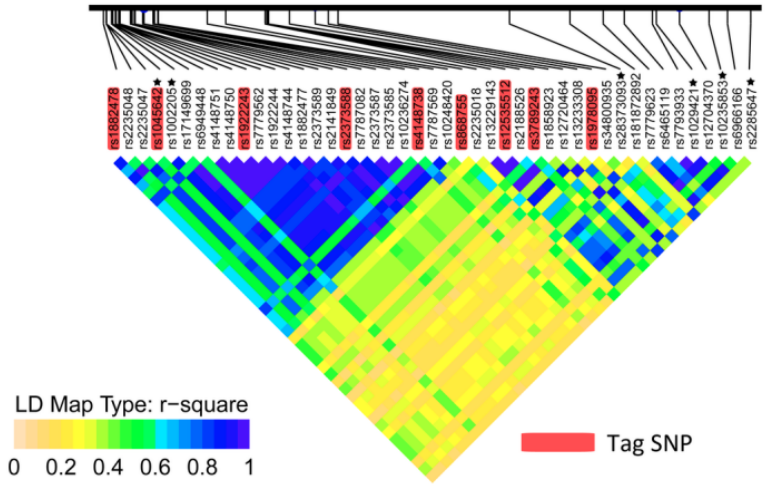


Figure 3.1: LD plot of SNPs with top-ranked bayes factors in CHB (Han Chinese in Beijing) of 1000 Genome Phase I. The colors indicate the strength of pairwise LD according to r2 metrics. The SNPs marked with asterisks represent independent strong associations. Tag SNPs are here shadowed in pink.

The colors indicate the strenght of pairwise linkage disequilibrium (LD) according to r2 metrics. In fact, not all of the SNPs are informative to distinguish between individuals. In 3.1 **Tag SNPs** are shadowed in pink. A Tag SNP is representative of a region with high linkage disequilibrium and represents a group of SNPs (called haplotype).

3.1.4 Other SNPs features

- Choose SNPs that are in areas that are not likely to undergo somatic aberrations. So exclude chromosomal locations which undergo frequent somatic aberrations. Eg. areas commonly deleted in tumor will produce LOH but probably also no calls, since there is no DNA.
- Choose SNPs equally represented/spread all around the genome (not in specific chromosome regions).
- Select autosomal only SNPs (not on chromosome X).
- Select SNPs in exons. If we were to run a targeted assay, this would cover more exons instead of introns. It will also be more probable to have signal from a non-DNA assay, for example if calling a genotype from RNA sequencing data (even though it is not always done).
- Exclude/include disease or drug response associated loci.

- Include/exclude loci with significantly different MAF in different ethnicity. If we include them we can also have a lineage type of tests in the same assay.

3.1.5 Number of SNPs to select

If we want to build a test to run genetic fingerprinting using SNPs, **how many polymorphic loci (SNPs) should be tested?** We want to make sure that the measure of the test will be able to differentiate unrelated individuals. But we must also remember that many variables must be taken into account, possible mismatches in particular. Those can be due to the sequencing process itself (experimental mismatches) but also to changes due to somatic events (biological mismatches). All these events can be used in the test with a different weight, based on how likely they are.

3.1.5.1 Experimental mismatches : Genotype call error rate

During sequencing, each machine will produce some errors, resulting in some loci for which no data will be available. If those loci include some SNPs of interest, then no call will be associated to that SNP. Experimental mismatches are related to the error rate of the technology used, they are platform dependent.

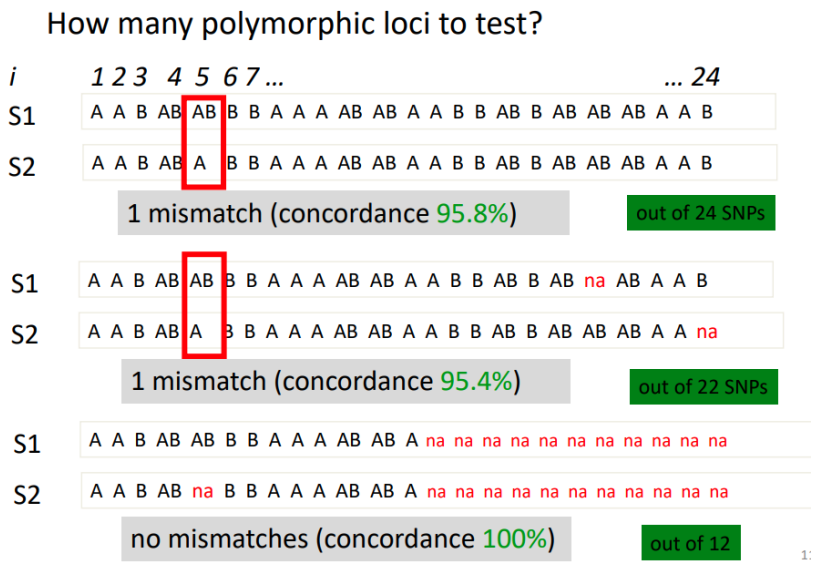


Figure 3.2

Some examples: In each example in figure 3.2 there are two samples with the same number of potential SNPs: 24. To detemine the difference/similarity of the two samples we can look at the genotype for each position and count mismatches.

Legend: 'A' stands for 'AA' (e.g. homozygus genotype for the reference allele); often refered to as Aa. 'B' stands for 'BB' (e.g. homozygus genotype for the alternative allele); often refered to as Bb. 'AB' stads for heterozygus.

- First Example: over the 24 loci, there is only one mismatch. This translates to a level of concordance of 95.8%. Those 2 individuals are highly related or DNA comes from the same saples.
- Second example: there is only one mismatch but there are some 'na', indicating that for some positions we don't have a call (not available data). Therefore, in this case the concordance is measured out of 22 SNPs and is equal 95.4%.
- Third example: here a lot of nas are present, leading to have only 12 SNPs available. This brings to a concordance of 100%.

Different examples produced different levels of concordance. What do we trust the most?

The first set of SNPs is the one that we trust the most, because it has the higher number of available SNPs. Wider number of SNPs provides the most reliable information.

Biological mismatches

In the context of desease samples and tumors, many somatic events can happen, like deletions, gains of copies, homozygus deletions, ecc. Some common ones are:

- Loss Of Heterozygosity (LOH): event that results in loss one parental copy of a region which results in the genome having just one copy of that region. If that region contained a heterozygus locus (e.g. SNP), there will be loss of Heterozygosity.

  AB -> A.
- Gain Of Heterozygosity (GOH): due to a mutation in a site often polimorphic through inheritance. These are pretty rare. A -> AB.
- Double Mutation (DM): very rare.

Biological mismatches can be properly modeled in our assay. We can, in a data driven way, assess the error rate for the genotyping for some specific SNPs or run tests. We can also think in terms of SNP-specific or tissue-specific probabilities.

The main point is that all mismatches must be taken into consideration. For this, all implemented tests use *more than the minimal number of SNPs* that allow to identify individuals.

Genetic Distance

Having defined the number of SNPs to use, with maximum MAF and other amanable characteristics, the genetic test should provide a measure of some sort, which will be the output metric, associated with a probability of the measure to be correct.

As a simple measure, we can count the number of loci where two samples show different genotype and normalize on the total number of queried loci, defining a certain level of discordance (or concordance). The output value will be the 'genetic distance' between the two samples given the selected loci. The distance is proportional to the number of discordant calls.

In figure 3.3 we can see an example of a typical graph used to measure the genetic distance using SNP-based genetic testing. We have 4 samples with a set of 5 SNPs for each one. The distance is measured among all possible pairs, whose indexes are reported on the x-axis.

- s1 and s2 have 3 A in common, one locus has no call and another one produces a mismatch. 1 mismatch out of 4 produce a distance of 0.25.
- samples s1 and s2 have 5 mismatches out of 5, so a distance (or discordance) of 1.

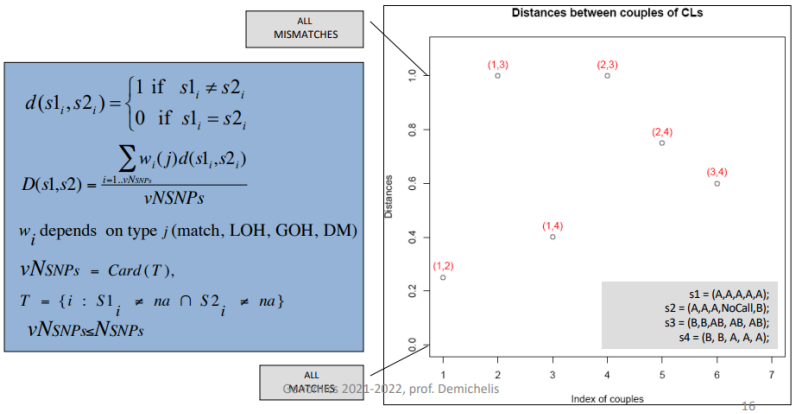


Figure 3.3: Genetic Distance graph with 4 samples

If we put that into an equation will have that: for each position i (SNP) bwteeen sample 1 and 2 we can have 1 if the genotype is different, 0 if they are identical. Then we determine the distance D by summing up the different scores obtained for each SNP. We can associate different weights w_i to different mismatches or we can put all equal to one. Then we devide by the total number of SNPs for which we have available calls, vNSNPs, which will be lower or equal to the total number of SNPs, NSNPs.

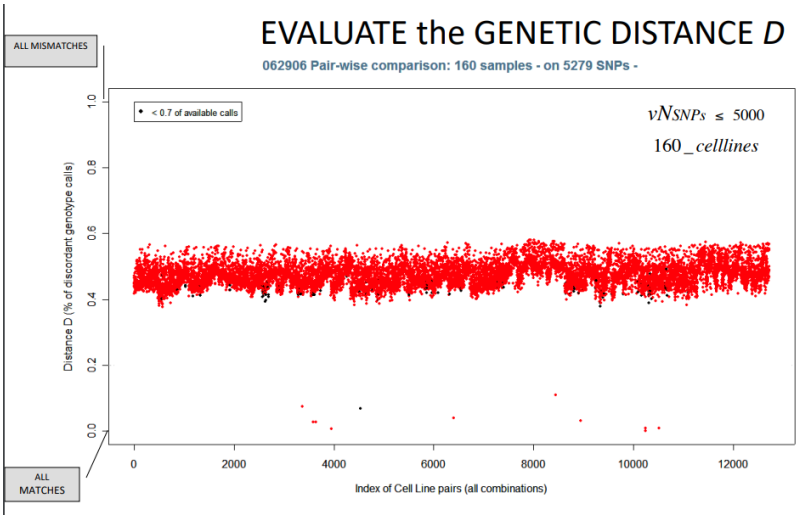


Figure 3.4: Genetic Distance graph with 160 samples

This other example at figure 3.4 shows the distance, measured by genetic fingerprinting, of a

collection of 160 samples of cell lines.

The number of possible pairs corresponds to: 160 * 159 / 2 (number found in the x-axis).

By applying this measure to a larger collection of samples like this one, with many SNPs, we expect to find an **average distance** among all possible pairs that very unlikely will be close to 1. The MAF of the SNPs is 0.5 but it will never happen that, with a high number of SNPs, the discordance will be 1. We will have an average distance that in this case around 0.5, since by chance we all share some genotypes on a large number of SNPs.

Here they found certain pairs with a very low distance, sometimes almost equal to zero (dots at the bottom). This was a surprising result because it shows that those pairs, which were suppose to be different cell lines, were actually not different cell lines (only less than 70% of SNPs have available calls).

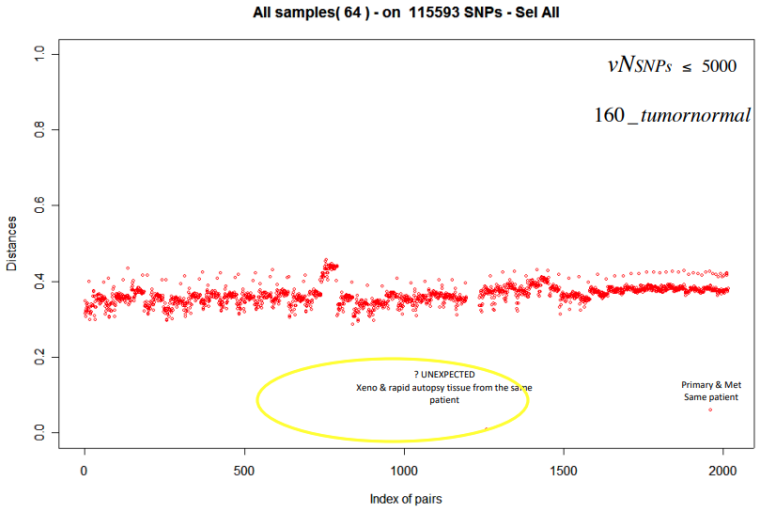


Figure 3.5: Genetic Distance graph with tumor samples

In this last example at figure 3.5 genetic fingerprint was performed on a collection of 160 tumor samples, with a larger SNP array (more that 100.000 SNPs).

From the analysis, two samples with very low distance were observed. One of the two samples came from a Rapid Autopsy Progam and the other one from a xenograft model.

RAP are programs for which patients at the end of their life agree to donate their tumor tissues which can be used for research. In these very complex but highly valuable programs, the material must be taken within two hours after death. Those sample are usually higly caractherized but after a while the track of the patient's identity is lost. Here, what happened is that one man who donated tissue by this program was sequenced and for some of those metastasis models were generated and implanted into a mouse and a xenograft model was derived. Thanks to fingerprinting it was possible to determine the same origin between xenograft and patient.

The power of this tecnique is very high, it allows also to identify and remove things that we don't want in our study. Eg. if running a study (like a GWAS study) on a certain interesting geopraphic area, we will want to remove the members of the same family because that would skew the results. Genetic fingerprint can be used for this purpose.

3.1.6 Some questions

Q1: Would the average of unrelated samples distance increase or decrease after selection of ideal SNPs?

If we maximize the likelihood that SNPs have a different genotype among individuals and we use these to determine the measure, then the average distance of unrelated individuals will increase.

Q2: Is it likely to obtain a genotype distance D = 1?

We get distance 1 only if we are looking at too few viable SNPs. Whereas with a well selected pool of SNPs, and a high enough number of SNPs, it is very unlikely that the distance is equal to 1.

3.1.7 Further considerations

How does the genetic distance among different samples change when varying the number of selected SNPs used to perform the test?

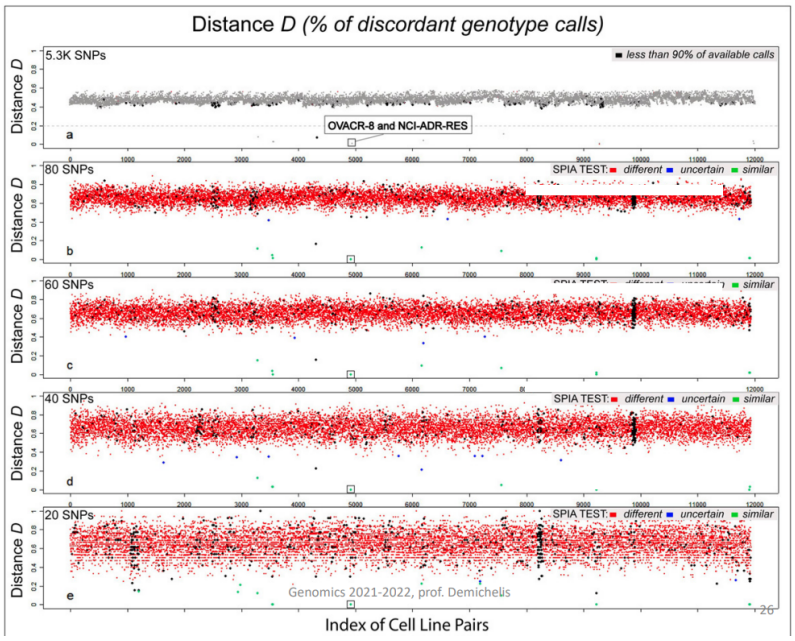


Figure 3.6: Genetic Distance graph at deacreasing number of selected SNPs

The genetic distance among many samples, with an array of 5.3K SNPs, was measured, using a decreasing number of SNPs (from the initial total number of SNPs to decreasing numbers of highly selected SNPs) 3.6.

It is noticeable that, in the second plot where 80 SNPs matching the required characteristics were selected, the average Distance across all pairs is higher than in the previous example, in which all available SNPs were used (0.45 vs. 0.65). Also, the standard deviation of greater. Decreasing the number of SNPs to 60, then to 40 and 20 leads to have the same average distance between pairs, which settles around 0.66, but higher standard deviation.

In reality we always need enough SNPs, enough information, in order to prevent unexpected issues and to be sure that for any pairs of sample we have enough information to trust our measure.

3.2 Building a SNP-based genetic test

Building an identity test base on SNPs is a MULTI-STEP process, consisting in:

1. Definition of a genotype/genetic distance to compare samples;
2. Definition of SNPs requirments, based on the intention of the assay.
3. Selection of SNPs:
    - This can be done in a data-driven manner, through an iterative procedure of training and test on known sample set;
    - Or, performing the selection based upon MAF and Hardy-Weinberg equilibrium. For example, using HapMap data.
4. Implementation of a probabilistic test (different, uncertain or similar)
5. In silico validation on independent/multiple dataset.
6. Validation on cell lines genotyped on independent platform.

We have already seen some of the steps needed (1, 2, 3), we now pass to the following ones.

3.2.1 Implementation of a probabilistic test

Other important questions which we have to answer to when designing a genetic test are:

- What is the threshold on the genotype distance to call two samples 'identical' ('similar') or'different'?
- How confident would the call be?
- What is the minimum number of loci needed for a robust test?

It could be useful to have a probabilistic test to determine if the measure of the test is correct at with which level of confidence. We can use a probabilistic approach to compare observations with expectations (gold standard).

Under the assumption that SNP calls at different loci are independent, we can think in terms of Binomial distribution. Each SNP can be considered as a trial, n = number of SNPs in the assay, k = number of matches, p is the probability of match and (1-p) of mismatch. Then the probability of having k matches (successes) out of N SNPs (trials) follows the binomial distribution.

With n, np and np(1-p) large enough, we can use the Gaussian approximation of the Binomial distribution with K_{mean} = np and sd = np(1-p).

With somtehing that simple we can add a probabilistic test in our assay, defining an area of conficence given by $K_{mean} \pm m * sd$ where m is the number of standard deviations used to define the thresholds which will lead to have a smaller or wider confidence area.

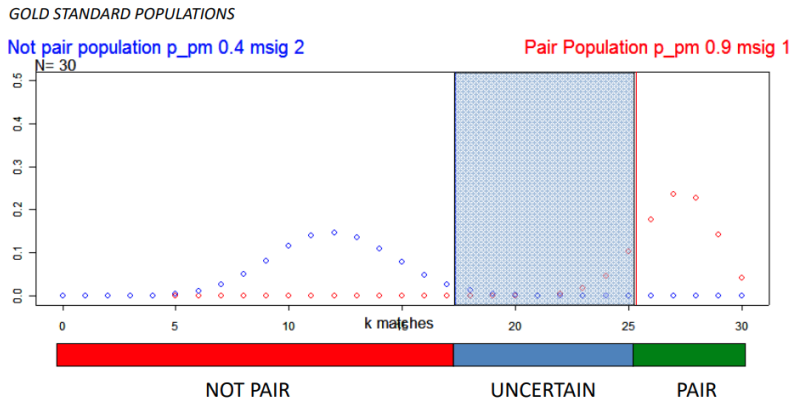


Figure 3.7

So for example: given two unrelated samples, we reason in terms of 'what is the probability of having a certain number k of matches over a total number of n SNPs, therefore a certain value of D?'.

The probability mass function for unrelated individuals is shown in figure 3.7 with a blue dotted line and indicates that there is a low probability of having both a very low and a very high number of matches.

We can also think in the opposite term: given two related samples, what is the probability of having matches? As represented by the red dotted line, in this case there will be a high probability of having many matches.

Using these probabilities we can set two thresholds which will define 3 regions:

- A **'not pair'** region for which the two samples will be considered as 'different'
- a **'pair'** region for which the two samples will be considered as 'similar'
- and an **'uncertain'** region, a grey zone, for which no certain result can be produced.

Then we can move the grey area based on what we want to be certain of and on how many SNPs we have.

By decreasing the number of SNPs, the grey zone will become more tiny, making the result more difficult to interpret. For example, a difference of only 2 matches could lead to opposite conclusions.

By contrast, with more SNPs the area will be wider and easier to interpret. Hence using a number of SNPs greater than the minimum number is better, otherwise there will be many uncertain calls.

Further considerations and examples

In the past, before sequencing area and SNPs array area, short tandem repeats were commonly used for genetic fingerprint. They were used on gels to distinguish related and unrelated individuals, eg, for the initial paternity test.

Inherited copy number variants can be used too for a fingerprinting test, but not all of them. The more amenable for this test are the loss type of CNV. In the population there will either a copy number of 2 or 1 or 0. If both parents have heterozygus pair of CNVs it will be possible that I have

a homozygus deletion. If both parents have 2 copies at a site that is polimorphic in the population, we will have a genotype equal to 2 copies.

If we think about gain of CNV then it becomes messy, because when combining multiple copies and have a add up we cannot distinguish what comes from what pair, so we cannot use them to identify an individual.

3.2.2 Example 1: Cell line passages

A mass use of these genetic tests is done to assess genetic changes in in-vitro cultivation (also, in studies in tumor evolution, lineage plasticity, etherogenity across metastasis across individuals or a single tumor). Cell lines go through multiple passages in which they are used and stored. Genetic fingerprinting can be used to assess if among different passages the cells have remained the same, if they were mislabeled or if major genetic drifts happened.

In this example, two types of prostate cell lines which underwent multiple passagges were used: N15C6 (passages from 48 to 63) and N33B2 (passages from 21 to 39). The cell lines were profiled with a SNPs array and the assay was run. All passages of each cell lines were compared with all other passages. We expect all passages to have the same genetic fingerprinting in the same cell line.

However the results obtained using the full array of SNPs (50k), showed that some pairs which should be exactly identical (distance equal to zero) are actually a bit different (points at the bottom-left). By contrast, by using a set on only 54 SNPs, this diversity is not detectable, indicating that using the perfect number of SNPs could make us loose some information.

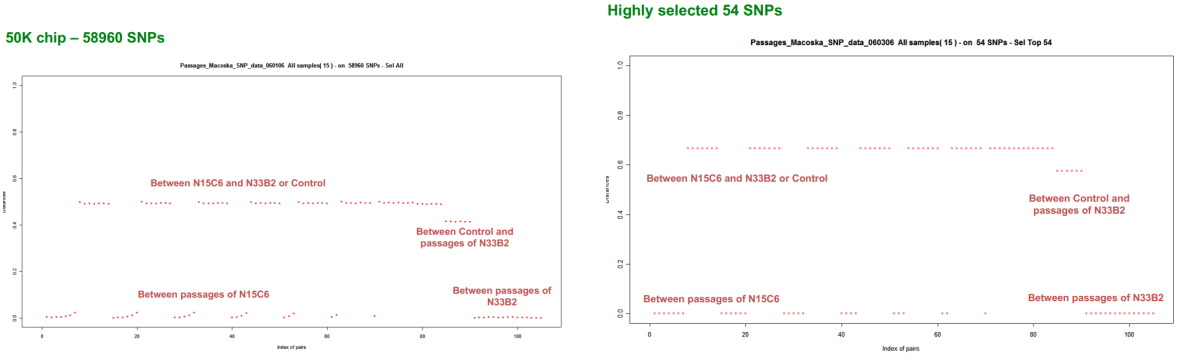


Figure 3.8

In order to understand this increase of distance, they looked at each chromosome to see if there were problems that justified increase the increased distance expected to be equal to zero in that cell line. All chromosome were tried. If we focus only on the SNPs spread across Chromosome 11, we observe that there is a major difference for certain passages with respect to the initial ones, only for one cell line (N15C6). This was due to the way the cells were immortalized (insertion in chromosome 11).

3.2.3 Individual's Relatedness (genotype-distance)

The HapMap consortium sequenced hundreds of individuals for different ethnicities and also used trios. Trio sequencing is a technique which involves the sequencing of the genome of mother, father

and son/daughter. Trios provide major information for haplotype blocks, for identifying regions related to inheritance, ecc.

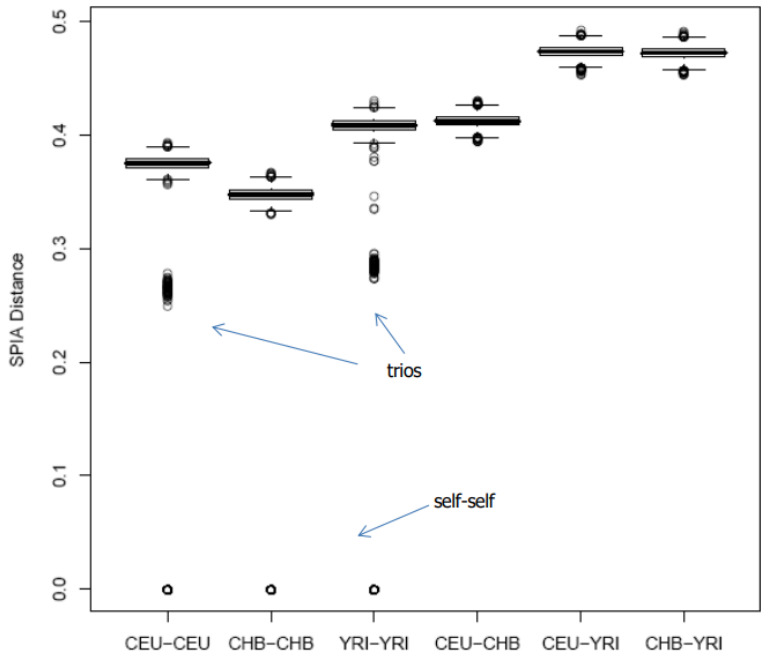


Figure 3.9

By looking at the data based on SPIA Assay (a genotype base assay which measures distance) at figure 3.9 we see that self-self pairs have distance zero, as expected; samples within each ethnicity have a certain average distance, which is lower that the distance observed among different ethnicities. Differences in distance among mixed samples are due to the fact that the SNPs used had on average higher MAF in some populations than in others. We also notice that in trios the distance is not 0 and is not equal to the median distance of unrelated individuals. This can be used for paternity tests or even in forensic science.

3.2.4 Example 3: Cancer supscettibility test

The data showed refers to a study were they were looking for polimorphisms that increase the likelihood of prostate cancer. In these studies, if relatives are present in the cohort, only one of them is taken to avoid skewing the results. When looking for signs of cancer susceptibility by performing genetic fingerprinting, the division based on the degree of relativeness was determined 'for free' and could be used to remove unwanted samples from the cohort.

3.2.5 Genetic structure of the human population

One relevant aspect of the human genome is that it contains everything needed to learn about the genetic structure of the human population.

Undestanding the geentic structure of human populations is of fundamental interest to medial, forensic and anthropological sceinces.

Some of the reasons as to why knowing the genetic substructure of data is important:

- The goal of association studies is to identify DNA variants that affect disease risk or other traits of interest. However, association studies can be confounded by differences in ancestry.
- Misleading results could arise if individuals selected as sidease cases have different ancestry, on average, than healthy controls. If in a study all controls are of the same ethnicity and the test is done on an individual of a different ethnicity than the test is biased.
- If we run a GWAS study using two ethnicities and we want to uset the same markers of sucettibility worldwide, it won't work.

Especially in medicine and in the study of human evolution it is important to track the genetic background of individuals that are involved in studies in order to understand if the individuals are form a homogeneous population or from genetically distant ones. More and more, clinical studies must have declarations of the checks and interpretation of the data of the genetic background of the individuals present in the study. It is very important to come to results for which we know exactly what is the applicability. To avoid spurious results, association studies often restrict their focus to a single continental group.

Advances in high-throughput genotyping technology have improved the understanding of global patterns of human genetic variation and suggest the potential to use large sample sets to uncover variation among closely spaced populations. One important piece of information to consider when developing methods to understand the genetic structure of a population, is to think in term of variance, which is also relevant for human diseases. Many SNPs have different MAFs in different populations. If we use those, and are able to have all of them in a simple computational way, we could be able to infer what is one individual's genetic background in terms of origins (e.g. chinese origins).

The easiest mathematical approach to assess how well SNPs can distinguish ethnicity is by using **Principal Component Analysis (PCA)**. By running a very simple PCA on a set of SNPs including SNPs with different MAF in different populations we can, in a space, distinguish different ethnical groups. And we could also start thinking at individuals' origins.

How accuratley can one predict an individuals geographic-ethnical background based upon his/her geentic barcode?

3.2.5.1 Example paper: 'Genes mirror geography within Europe'

In the study seen during lectures they used a 500.000 (500k) single nucleotide polymorphism array. Information about the country of origin of grandparents, parents and other relatives was used to determine the geographical location that best represents each individual ancestry. They run a combined study where they used a supervised search to find the best SNPs to make inference and then they tested it on another set of individuals. By using high confidence data (individuals with high confidence origin data) and by using the genotypes of highly informative SNPs for specific region-related inheritance, they were able to rebuild the map of some of the countries in Europe 3.10.

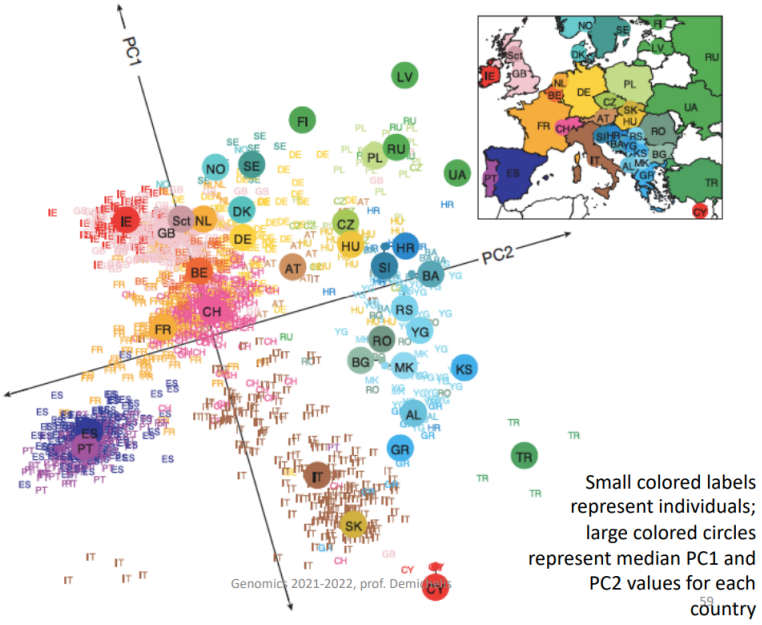


Figure 3.10

This result might be a little bit of a push, but it is true that by using properly selected variants it is possible to distinguish individuals coming from different countries. The way those SNPs are selected is very similar to the process saw for genetic fingerprint, but pushing for the selection of variants that are different in terms of MAF in different populations.

Clusters that are a bit more dense and distant from the others (like the Spain/Portugal cluster) could be due to the fact that many SNPs selected are typical of that area and are therefore able to maximize the difference with respect to that area (so it is a data-related 'issue').

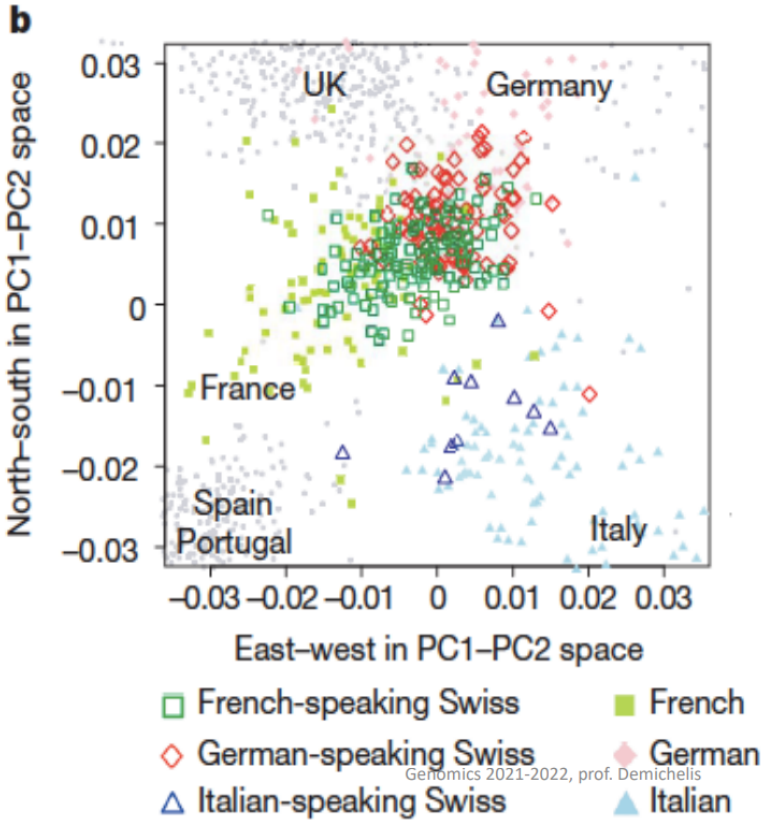


Figure 3.11

Focusing on Switzerland, they could even make inference on the linguistic canton 3.11. Again this is a bit of a push, but it is possibly true that in country where some regions have very different habits (e.g. marriage within the same area) might lead to have similar genetic fingerprint.

3.2.5.2 Summary and notes

Low-frequency alleles tend to be the result of a recent mutation and are expected to geographically cluster around the location at which the mutation first arose. Hence, they can be highly informative about the fine-scale population structure.

Despite low average levels of genetic differentiation among Europeans, close correspondence between genetic and geographic distances was found. When mapping the genetic basis of a disease phenotype, spurious associations can arise if genetic structure is not properly accounted for.

Chapter 4

IGV (Integrative Genomics Viewer)

Written by Maurizio Gilioli

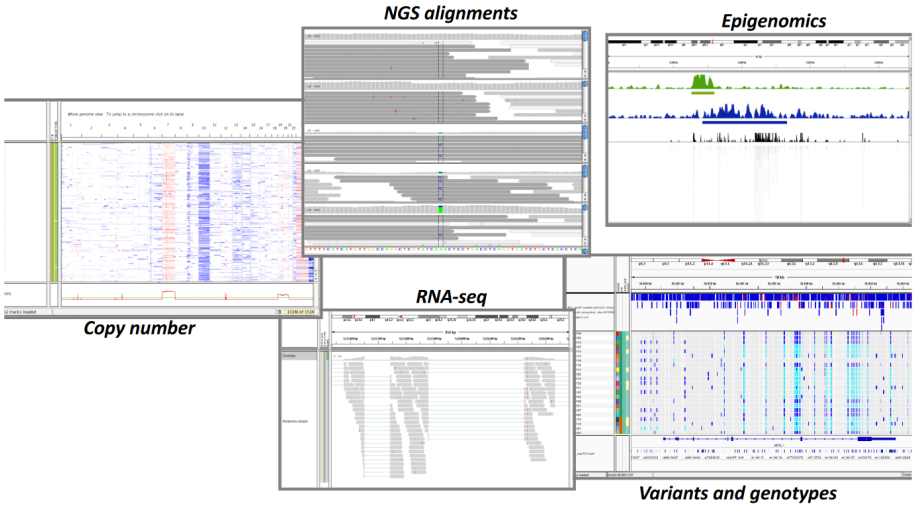
4.1 Main characteristics

The human genome nowadays is being explored extensively thanks to exons and whole-genome sequencing, epigenetic surveys, expression profiling of coding and noncoding RNAs, single nucleotide polymorphism (SNP) and copy number profiling, and functional assays. Those findings are essential to pave the way for the future **precision medicine**, which is an approach for desease treatment and prevention that takes into account individual variability in genes, living environment, and lifestyle for each person. The scope is to administer the right drug, at the right time and at the right dose for each individual.

Below, some of the main utilizations of IGV, also represented in figure 4.1.

- **NGS alignment**
- **Epigenomics studies**
- **Copy number evaluations**
- **RNA-sequencings**
- **Identification of variants and genotypes**

Figure 4.1: All the important usages of IGV



The IGV software is an **high-performance lightweight visualization tool** for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including next-generation sequence data, and genomic annotations. Data sets can be loaded from local or remote sources, including cloud-based resources.

It allows to move, zoom in and out quickly over different genomic scales (subfigure 4.2), and also to jump in precise positions of the sequence. It is possible to search for genomic coordinates or gene names. For each resolution scale ("zoom level"), the aggregated data is divided into tiles (subfigure 4.3) that correspond to a region viewable on a typical user display. Each tile is subdivided into bins, with the width of a bin chosen to correspond to the width represented by a pixel at that resolution scale. The corresponding data tiles for each zoom level are stored in the binary Tiled Data Format, or TDF, which has been optimized for fast tile retrieval.
*A tiled data file (**TDF**) file* (.tdf) is a binary file that contains data that has been preprocessed for faster display in IGV. TDF files are generated by using the *igvtools* package (*toTDF* command).

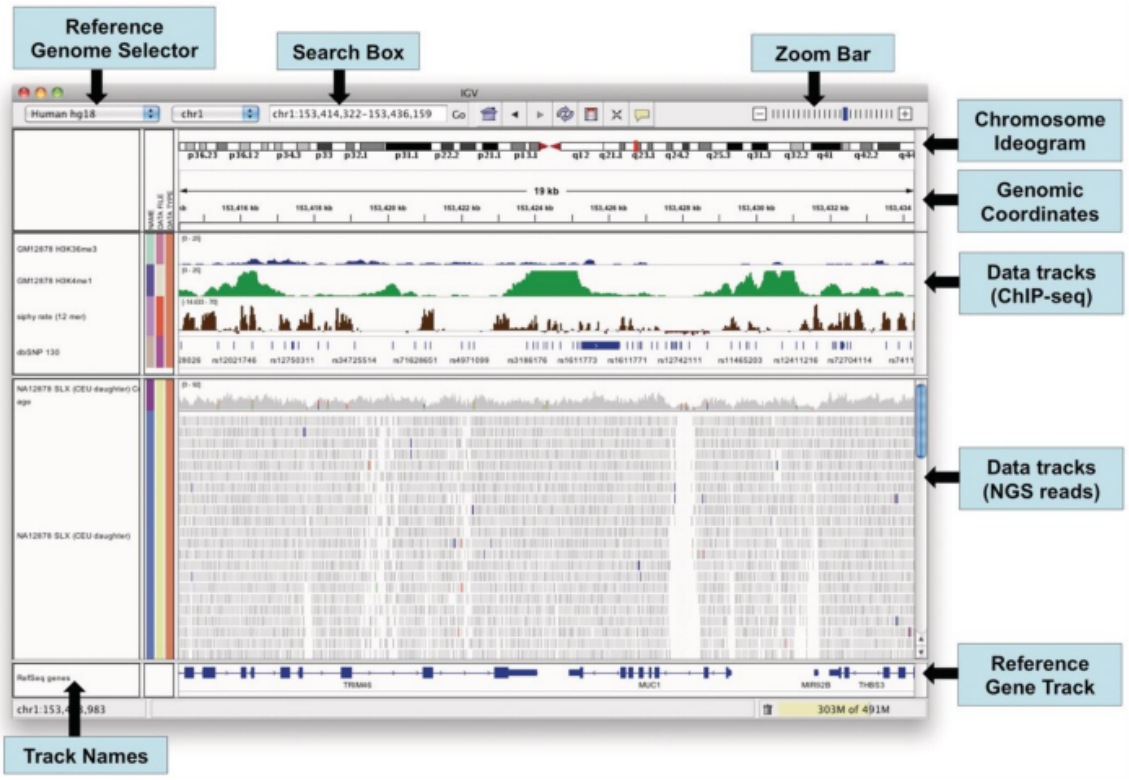


Figure 4.2: IGV interface main features

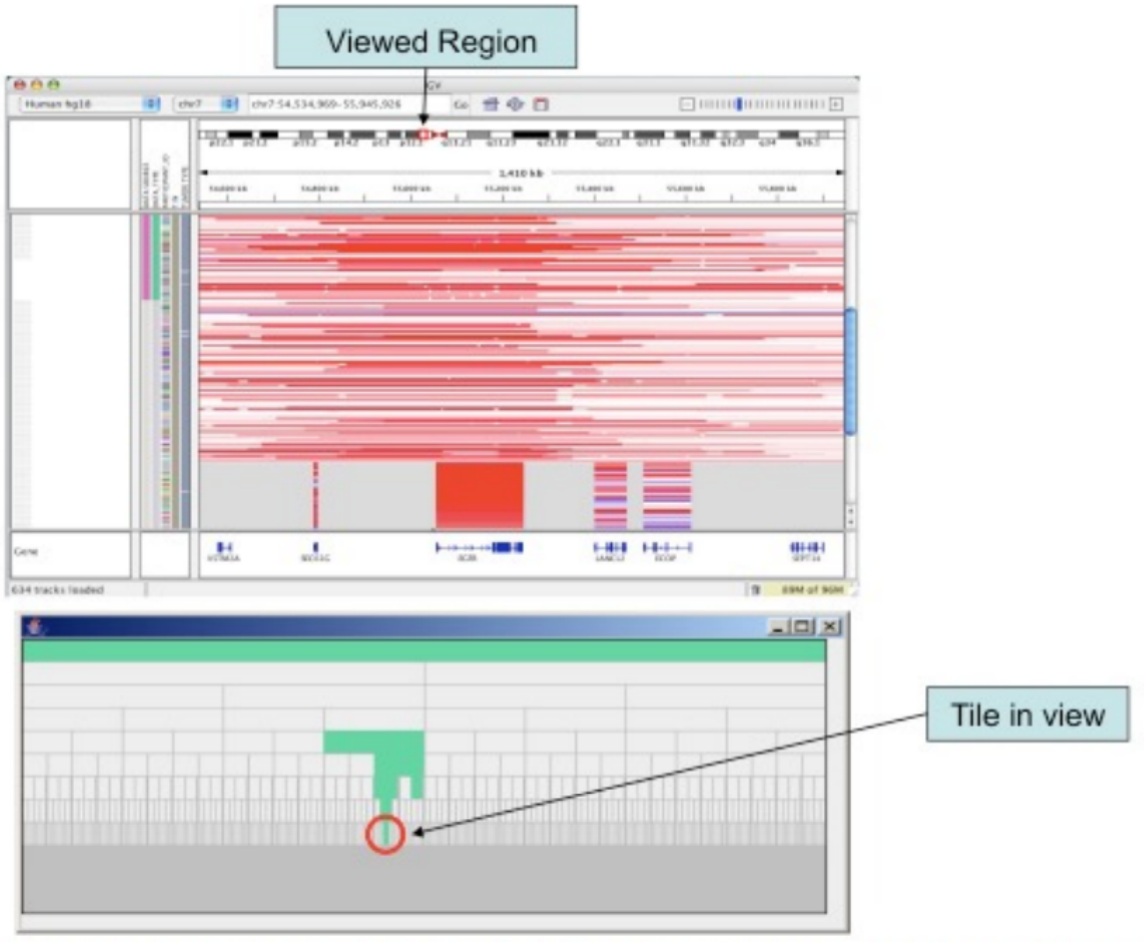


Figure 4.3: Tiles view in IGV

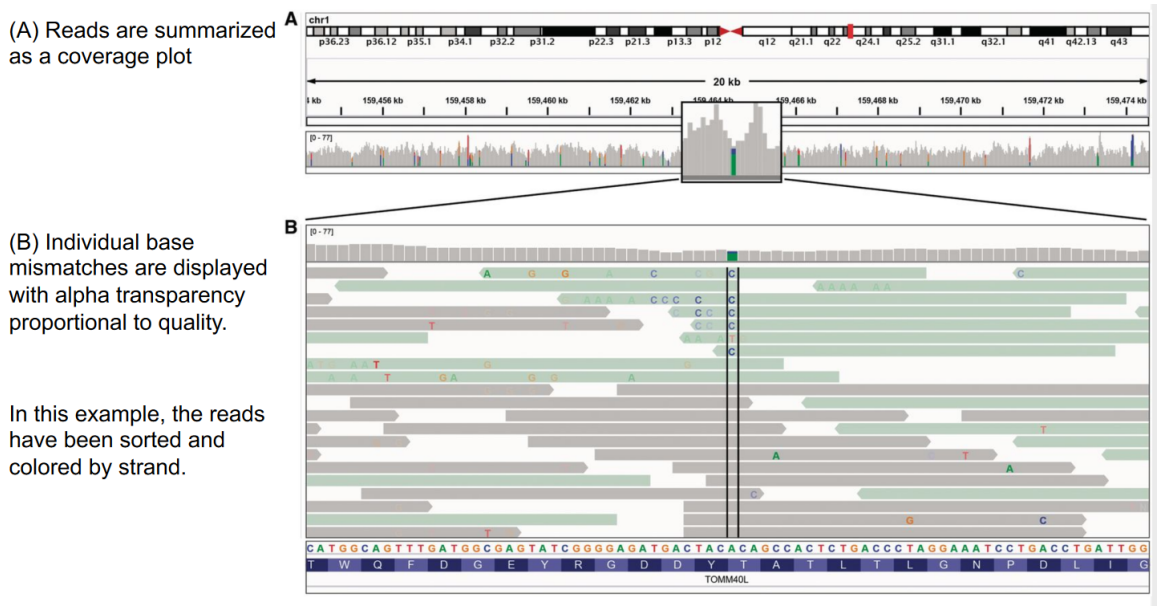
Importantly, **tile** sizes for each zoom level are constant and small, and also, a single tile at the lowest resolution (spanning the entire genome) has the same memory footprint as a tile at the very high zoom levels (might span only a few kilobases).
Tiles no longer in view are discarded as needed to free memory. Navigation through a data set is similar to that of *Google Maps*, allowing the user to zoom and pan seamlessly across the genome at any level of detail from whole genome to base pair.

Pixel resolution errors, occuring when data density exceeds the constraint given by the number of pixels available for display, could be solved through data aggregation. As the user zooms below the 50 kb range, individual aligned reads become visible. It is possible then to zoom further, and see the bases at each position.

Annotations for specific genomes could be found consulting the UCSC Table Browser (UCSC table).

Other information regarding IGV are present in the Supplementaty information - Integrative Genomics Viewer pdf file.

Figure 4.4



4.1.1 Igvtools

Igvtools comprises a set of utilities to prepeare large files for efficient display.

Figure 4.5: igvtools possible operations, the "count" function allows to generate coverage data, and it takes in input a BAM file. The obtained file could be then loaded with the "Load pre-computed coverage data" commandq

count	– Computes alignment coverage from BAM files – Produces TDF or WIG files
toTDF	– Converts sorted data file to binary tiled data (TDF) – Supported file formats: WIG, bedGraph
sort	– Sorts file by genomic start position. – Supported file formats: BED, GFF, GTF, PSL, SAM, BAM, VCF
index	– Creates index for large genomic annotation files and alignments. – Supported file formats: BED, GFF, GTF, PSL, SAM, BAM, VCF

4.1.2 Session Files

Sessions are an integral part of IGV, allowing users to share their data and views with other users simply and accurately. Session files describe the session in **XML**.

Figure 4.6: Structure of the XML file

Required - These elements are required in a session file. All session files must follow XML standards.

- <Global>: Contains information about the general state of IGV when the session was saved
  - genome= The genome id
  - locus= The genomic range selected when the session was saved
  - version= The session version (this must equal '3')
- <Resources>: An enclosing element for all Resource elements
- <Resource>: Contains the location and other important information for your data files; for instance, a Resource could be a DAS server, BED file, or sequence alignment
  - name= The name of the track for single track files
  - path= The path IGV uses to access the resource
  - url= The URL path to the resource / UCSC Track Line Url

Optional - These elements are optional in a session file and are added by IGV to help determine the placement of the data and visual style choices.

- <Panel>: Contains information about the placement of Tracks in the visual panels
  - name= The display name for the Panel
  - height= The default height for the Panel
  - width= The default width for the Panel
- <Track>: Details information about every track in a session
  - color= The default color for the data in the track
  - expand= Whether the track is expanded or not
  - height= The default height of the track
  - id= The id assigned by IGV to this track
  - name= The display name for the track

4.2 Some of the main utilizations

(I will not write down all the passages needed to obtain the figures represented below, as they are included in the exercise file delivered by the professor)

4.2.1 RNA-seq alignments

Figure 4.7: the height depends on the quantity of reads connecting the different exons.

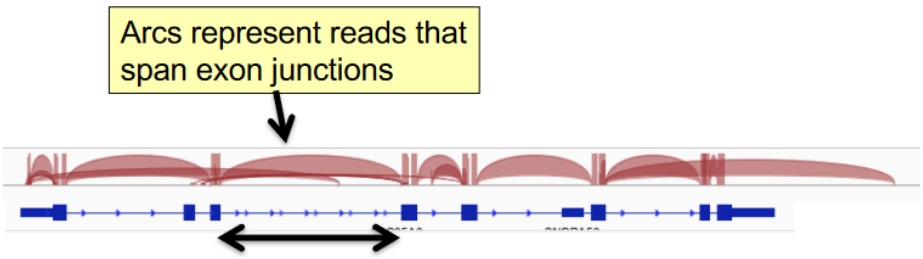
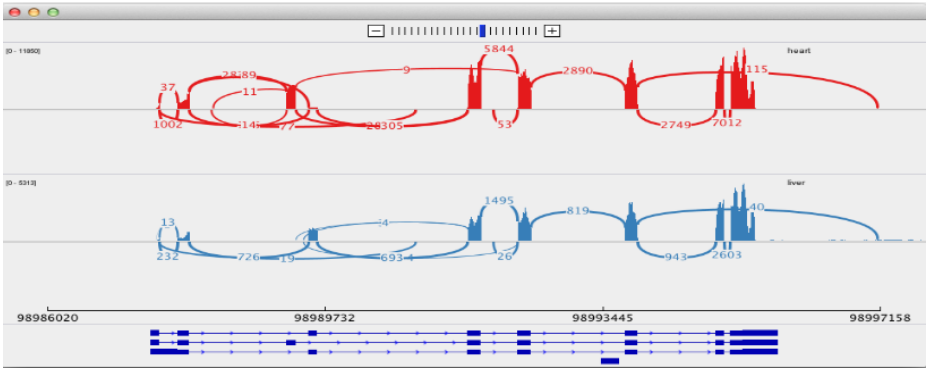
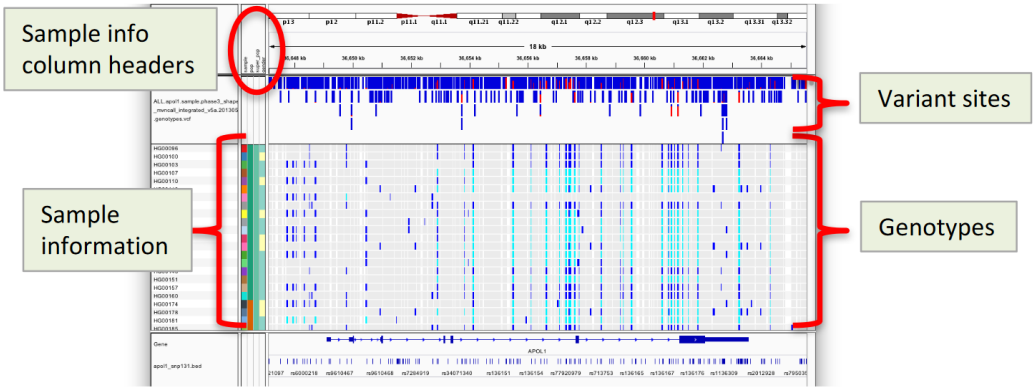


Figure 4.8: Sashami plots: The number of reads connecting exosomes are represented here on the curved lines. The peaks represent coverage within exons.



4.2.2 Study of variants

It is possible to study variants from different samples.



It is also possible to sort the samples in different ways and to group them considering different characteristics.

4.3 Exercise

The goal was to read pairs/end order/coverage/insert sizes at following coordinates (hg19). Interpret, if possible, as inversion, inverted duplication, tandem duplication, or deletion.

Figure 4.9: Tasks performed

Task A (1:5)
Refer to IGV_Exercises.pdf and finalize exercises (typo CAP9->CAPN9)

Task B
Upload from folder
NA12878.mapped.ILLUMINA.bwa.CEU.high_coverage_pcr_free.RR.bam

Inspect read pairs/end order/coverage/insert sizes at following coordinates (hg19). Interpret, if possible, as inversion, inverted duplication, tandem duplication, or deletion. (Color by pair orientation)

Region 1 chr1:11,043,245-11,061,901
Region 2 chr5:9,410,315-9,413,699
Region 3 chr7:31,576,117-31,599,940 (tricky)
Region 4 chr12:12,540,452-12,550,470
Region 5 chr5:79,041,411-79,054,952

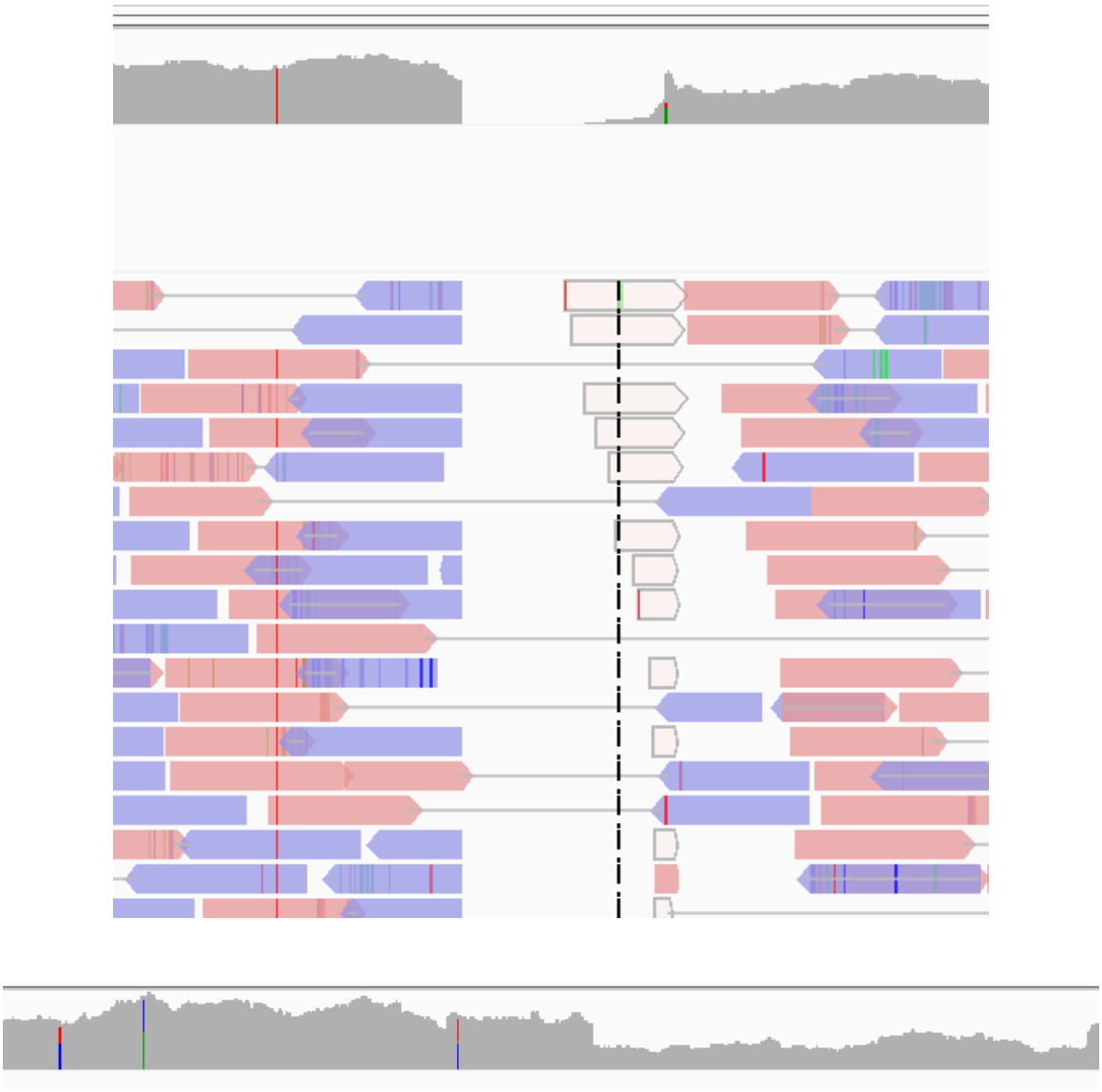
Task C
Save a session. Inspect xml file using a text editor.
[igv]
1. load BAM file
2. go to first genomic region
3. save session (*.xlm)
4. open *.xlm with text editor

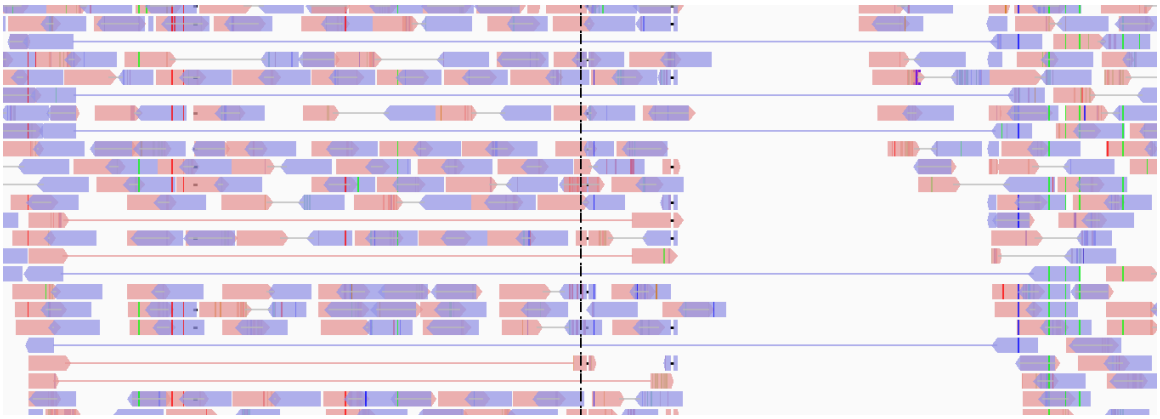
Genomics 2020-2021, Demichelis 14

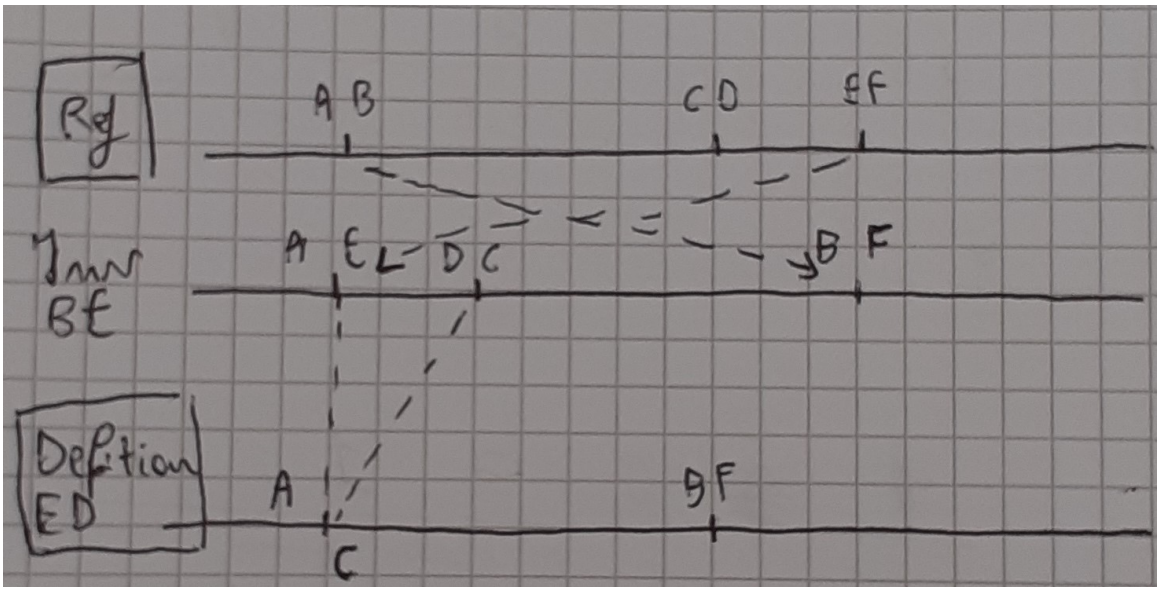
4.3.1 Task B

Figure 4.10: ***chr1:11,050,009-11,055,137***: It could be a tandem duplication on one of the two alleles and a deletion on the other allele. The reason why I would suggest the presence of a deletion is due to the fact that the coverage remains quite constant, despite of the duplication.

Figure 4.11: ***chr5:9,410,315-9,413,699***: it is quite clear that both the alleles were deleted in that region, because of the decrease in coverage






Ref
A B
C D
E F
A E L D C
B F
Deletion
E D
A
C
F

Chapter 5

Tumor Evolution Studies via NGS data

Tumor board: organism research oriented (and not) hospitals, patient not strictly assigned to one doctor but many specialist. E.g. oncologists, pathologists, geneticists...
TUmor boards teach young doctors how to manege difficult events.

5.1 Tumor evolution

What are the somatic events that occur during tumor genesis/evolution and when do they arise?
First of all, remember that every cancer cell was once a healthy celss that underwent some stress (UV light, radiation...).
Typical traits of cancer are:

- Cancer is a dynamic disease, that's why evolution of the desease is so important to track.
- During the course of disease, cancers generally become more heterogeneous, which is often related to treatment resistance.
- The bulk tumor includes a diverse collection of cells harboring distinct molecular signatures with differential levels of sensitivity to treatment.
- This heterogeneity might result in a non-uniform distribution of genetically
- distinct tumour-cell subpopulations across and within disease sites (spatial heterogeneity) or temporal variations in the molecular makeup of cancer cells (temporal heterogeneity)

Heterogeneity is a big obstacle in cancer treatment. Two characteristics to keep in mind are:

- Heterogeneity provides the fuel for resistance; therefore, an accurate assessment of tumor heterogeneity is essential for the development of effective therapies.
- Multiregion sequencing, single-cell sequencing, analysis of autopsy samples, and longitudinal analysis of liquid biopsy samples are all emerging technologies with considerable potential to dissect the complex clonal architecture of cancers.

However, techniques to study tumor heterogeneity are hindered by intra-patient heterogeneity: spatial and temporal.

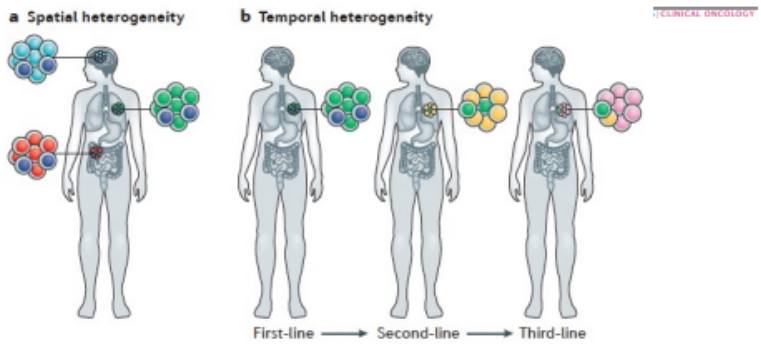


Figure 5.1: a) Spatial heterogeneity denoes an uneven distribution of cancer subclones across different regions of the primary tumor and/or metastatic sites. b) Temporal heterogeneity refers to variations in the molecular makeup of a single lesion over time, either as a result of natural progression of the tumor or as a result of exposure to selective pressures created by clinical interventions. Colours denote the presence of subclones with different genetic features.

Let's focus for the moment only on spatial, and not temporal, tumor heterogeneity.
Certain cells respond to treatment and some don't. Red = cell resistant to drug, blue = cell sensitive to drug.

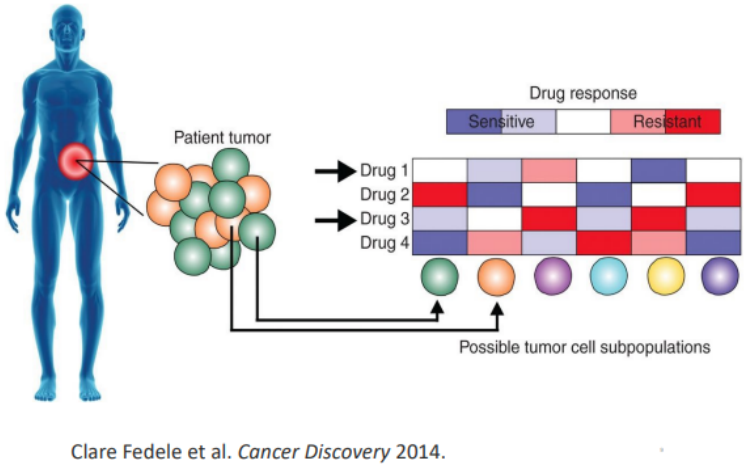


Figure 5.2: Schematic view of what a heterogeneous cancer may look like.

Linear evolution vs branching evolution.
The feature of this set of cells over time and for some reason either the new population replaces the older, or there's a branching and the tumor mass becomes heterogeneous.

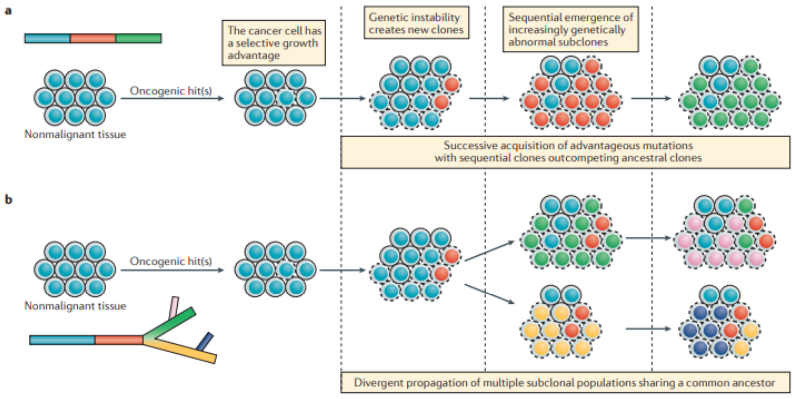


Figure 5.3: Schematic view of what a heterogeneous cancer may look like.

a) everything branches out from the monoclonal origin, but b) polyclonal origin, indipendent metastatic processes. Cells from indipendent lesions meet and form a highly diverse metastatic tumor.

A major concept: is treatment resistance encoded in the original sets or is driven by the treatment itself?
Selection of clones tat provide resistance, or transofmration of clones under teatment pressure.

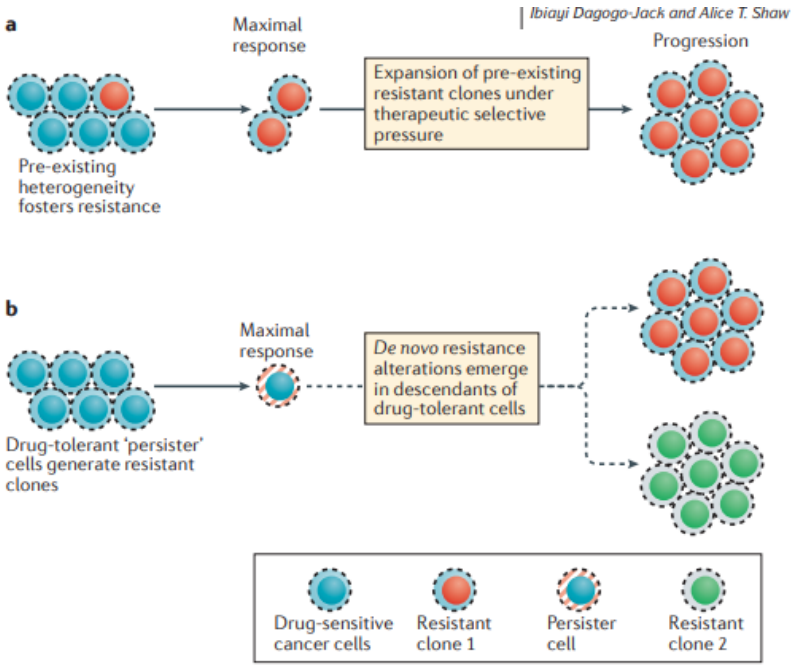


Figure 5.4: Schematic view of what a heterogeneous cancer may look like.

5.2 From sketches to sequencing data evolution information

Multiple individuals and on the left side there's evollution inferred frolm the data. On the right, assessment that suggests that it's compatible with monoclonal seeding.
What lesions ocured first and which is the model that fits better the data.

5.2.1 Tumor evolution and heterogeneity

To summarize, there are the difficult tasks one has to deal with when studying tumor data:

- intra tumor heterogeneity;
- inter tumor/intra patient heterogeneity;
- Inter-patient heterogeneity;
- clinical/treatment relevance;
- time dependency;
- admixture DNA (tumor purity);

However, if recognized, they can also provide for insightful hints int he analysis.

5.2.1.1 Admixture

Tissue from any source we have multiple cell types, each present with a given ratio.
Deconvultion looking at NGS data. Lesion 100% pure if the contamination of the adm of tumor cells is very low.
This info (purity = 1- adm) is improtant for

- agressvness
- every interpetation of somatic data needs to be interpreted in the contex of tumor purity. Lesion is clonal if all tumor cells have it, a less present lesion instead might be subclonal.

5.2.2 Useful feature from NGS data

- Polymorphic information that is presentin the genetics of every individual. SNPs are very helpful in all somatic analysis.
- MAF= minor allele frequency, freq a ehich the allele at a polymorphic site is present in a population. Used for SNPs in general.
- AF= allelic fraction, how many times ina specific locus a base is represented. How mnay reads represent the alternative allele, how muh support.

5.2.3 Allelic Fraction (AF) properties

Most important algorithms: how to exploit informative SNPs for interpetation tumor studies.

- Informative SNPs are sites in which the individual has a heterozygous base. For each SNPs we can calculate the AF.
- Neutral Reads
- Beta measure that No lesion and the two allelic equally represented then beta is one. With beta moving towards zero we have a measure
- Nref

AF: an info SNPs with loss of heterogeneity due to monoallelic los sis either 0 or 1, anything in the middle is an indication of either admixture or subclonality of the lesion.

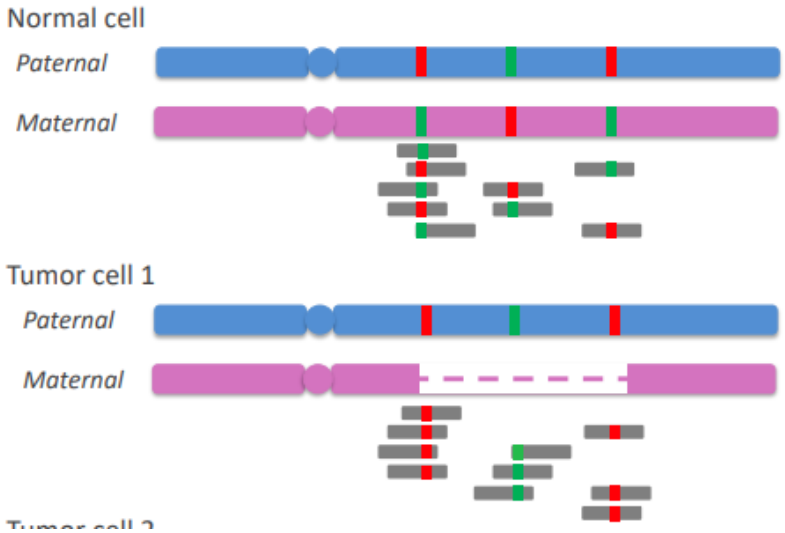


Figure 5.5: Schematic view of what a heterogeneous cancer may look like.

An example of the use of AF and beta measures for tumor data analysis in depicted in

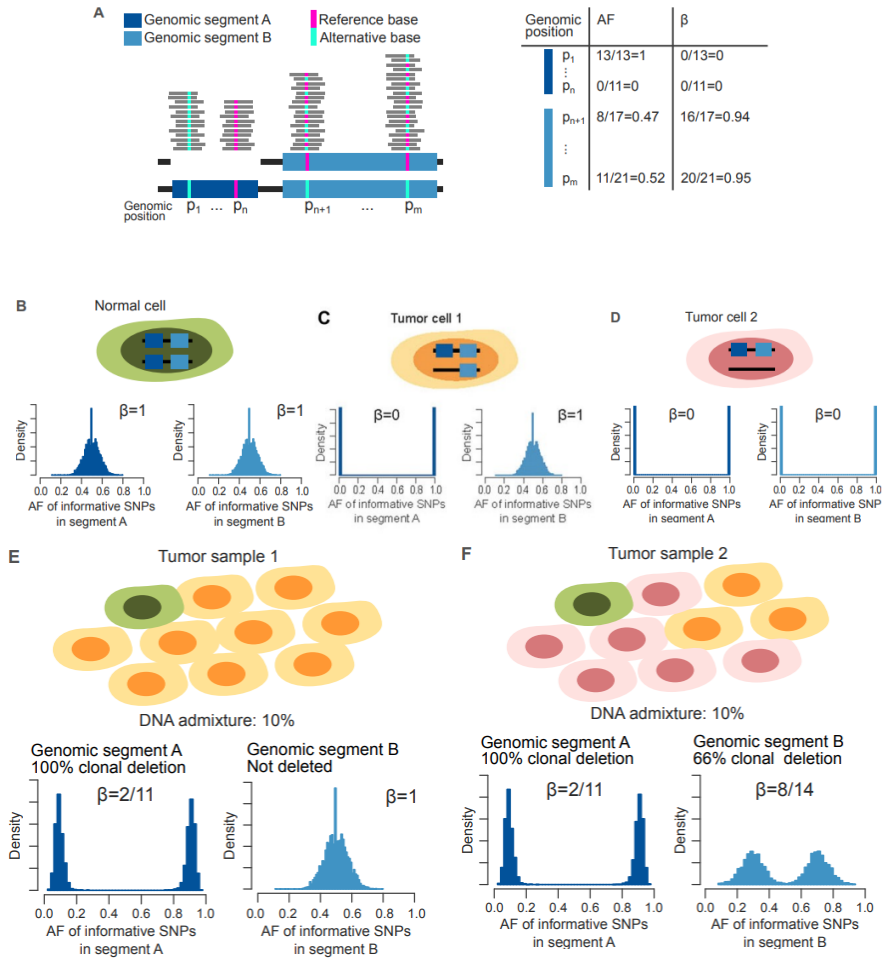


Figure 5.6: **A)**Example of the allelic fraction (AF) and beta (β) in computed in five genomic positions (p_1 to p_m). Positions p_1 to p_n are within a hemizygous deleted genomic segment A, while genomic positions p_{n+1} to pm lie within a wild type genomic segment B.
(B-D) Examples of a normal cell and two different tumor cells. Tumor cells 1 and 2 differ for the status of genomic segment B. Histograms below cell cartoons report the expected distribution of the allelic fraction of SNPs in genomic segments A and B together with the associated beta values.
(E-F) Examples of two different tumor samples. Tumor sample 1 includes one normal cell and nine tumor cells with deleted genomic segment A and wild type genomic segment B. Tumor sample 2 differs from tumor sample 1 in the presence of six tumor cells with a hemizygous deletion of genomic segment B. Expected distribution of the AF of informative SNPs together with estimated beta are depicted below each tumor sample cartoon.

In each genomic segments there are many info SNPs, but if there is not any, there's nothing we can say.

5.3 Coverage and AF properties

Does the mean coverage of the experiment impact on the ability to use the AF properties? Intuitively, the deeper the sequencing the more likely it is to distinguish distribution that are not so close to each other. It is especially important when β is close to zero. An example is shown in figure **??**.

ccc

5.4 Computing Beta

A Beta for each genomic segment S:

- compute the observed distribution of the AF of informative SNPs in S
- find the values of Beta and Nref such that the expected distribution of the AF matches the observed AF
- compute uncertainty around Beta as a function of:
  - the mean coverage of S
  - the number of informative SNPs in S

5.5 Global vs Local Estimates of admixture

We'll discuss two types of sample and how to determine cell population (admixture).

In sample one depicted in figure 5.7 there's a clonal cell population, meaning no heterogeneity. On the x axis we have genomic coordinates indexed by informative SNPs for that individual.
For all the info SNPs present in this chunk of DNA, we see drops in AF that are smaller or wider, but more or less for each one of those lesions (drops), representing a drop or a gain in the amount of DNA is almost identical in all of these chunks. The representation of the lesions is supported by the same data along the stretch of DNA. Thinking in terms of how much the AF distribution from 0 and 1 and the center is basically the same across all of them. Meaning, the level of admixture, both globally and locally, is the same.
The amount of cells that have the first, second, third lesion etc. is the same.

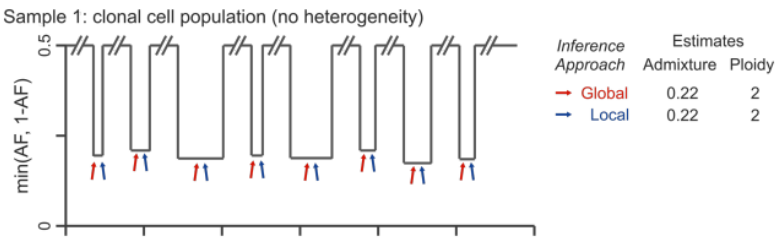


Figure 5.7

In the picture 5.8 below we can see another sample. We can clearly discover multiclonal cell population. The depth of the lesion is proportional to the number of cells that carry the lesion. In

this case the definitions of local and global admixture change: a **global** value is a global value of tumor purity, a local value is a value of the clonality of the lesion in the diseased cell population.

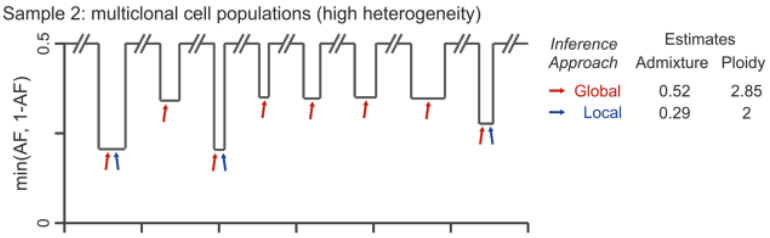


Figure 5.8

5.5.0.1 Estimate of DNA admixture (1-Purity)

We will now translate the concepts expressed in the previous sections in a 2-dimensional space, as shown in figure 5.1.

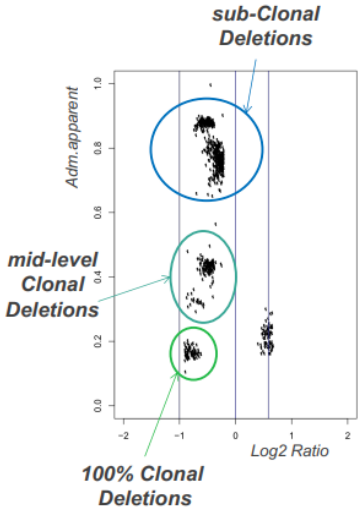


Figure 5.9: Every dot is one genomic segment, they form clusters.

On the x axis we have the Log2 ration measure and on the y axis the admixture apparent, which is proportional to the beta value.
The Log2 Ratio is basically tumor over normal in the log2 space. It allows to interpret info about every segment int he genome when coupled with the beta measure.
The admixture apparent is calculated as

$$Adm.\ apparent = \frac{\beta}{2-\beta} \tag{5.1}$$

This measure associates an apparent DNA admxiture to each monoallelic deletion. It is useful to calculate the clonality values, for which the formula is:

$$Clonality: \frac{1 - Adm.\ apparent}{1 - Adm.\ global} \tag{5.2}$$

The lowest cluster is the one used to assess admixture and the other ones are subclonal lesions.
Moreover, the closest the points are, the most probable it is that the lesions happened close in time, and viceversa.

5.6 A challenging case (PR-2741

Data of a real case of prostate cancer in which we can see a clear drop in coverage in region 2 of the 5th chromosome. The DNA present in region 2 could come either from admixting cells or from cells that do not have the deletion.
Looking at the AF of region 1, 2, 3 both from the tumor and the match normal normal sample we

observe more or less the same two modes of distribution.
In the tumor sample in fact we do not see two peaks in 0 and 1 as expected. It could be signal of intervening normal cells, that bring the modes to the center, or subclonality event.

Chapter 6

Tumor evolution studies (continued)

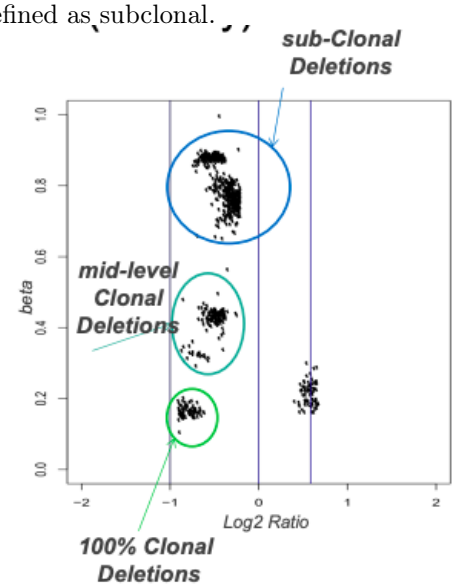
Written by Giorgia Bucciarelli

6.1 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.

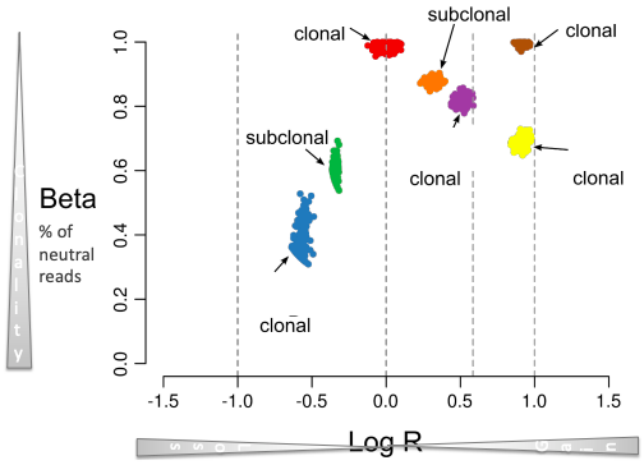
In the figure each dot is a genomic segment or a gene that clusterize in the space and when dots are in a same cluster it means that they very likely share the same copy number status and also the same level of clonality.

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.

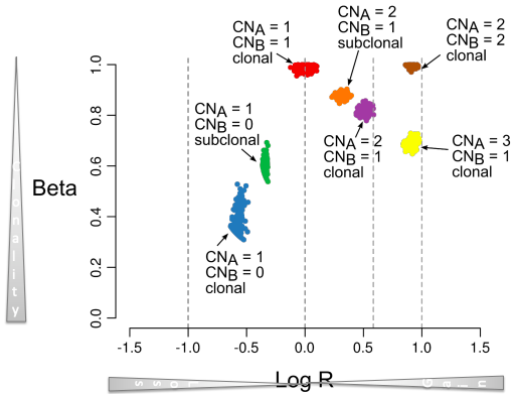
So the graph in the figure puts in relation the copy number status (log2 ratio) and the purity/-clonality of the sample (Beta); the more we go towards the left the fewer number of copies, the lower on the y axis the higher the clonality.

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.

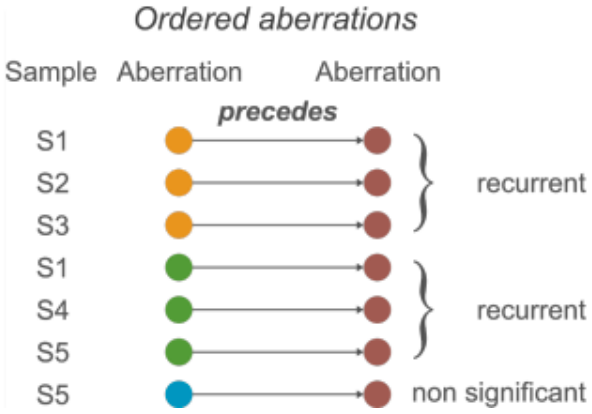
- Blue one: one copy of DNA, so we have a deletion
- Green one: also one copy of DNA but with subclonality

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:

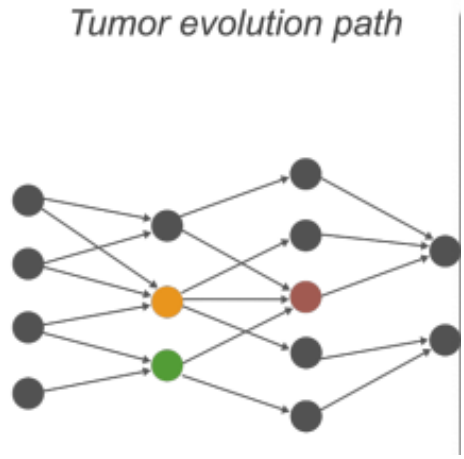
- In sample 1 the brown lesion is subclonal to the orange one, and that same lesion is also subclonal to the green one.
- In sample 2 we have again the support of the relation between the brown and orange lesion with the same level of subclonality (brown subclonal to orange).
- In sample 3 is the same 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.

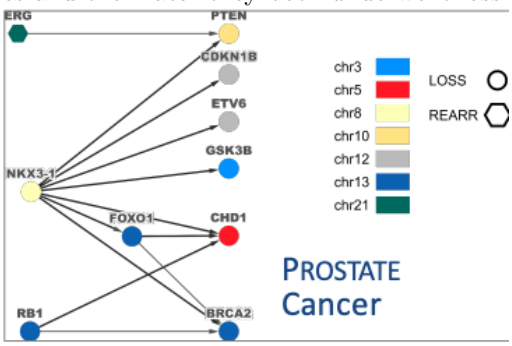
So having multiple observation supporting that aberration x precedes aberration y (i.e. aberration y is subclonal to aberration x) we can build an evolution chart.



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.

6.2 Ploidy and purity correction on $\log_2(\frac{T}{N})$ 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?

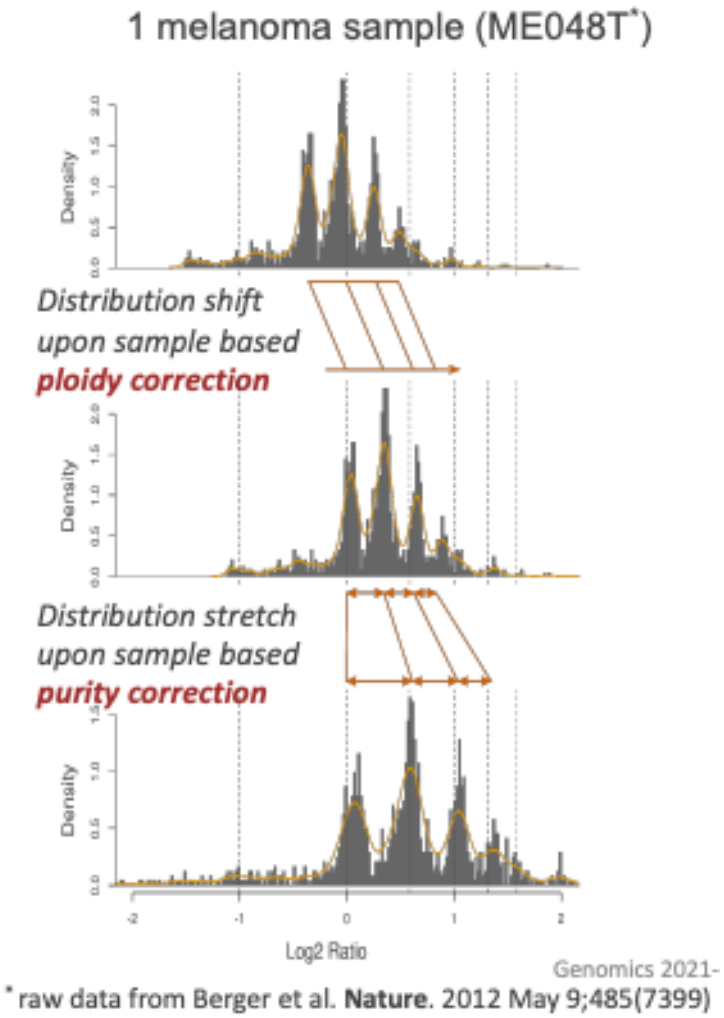
E.g.: we have a sample a 100% pure and with 50% of clonality (a lesion present in 50% of the cells) and a second sample with a tumor purity of 10% and a clonality of 100% (a lesion present in 100% of the cells), we need a way that allows us to compare numbers without having to convert everytime for every 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

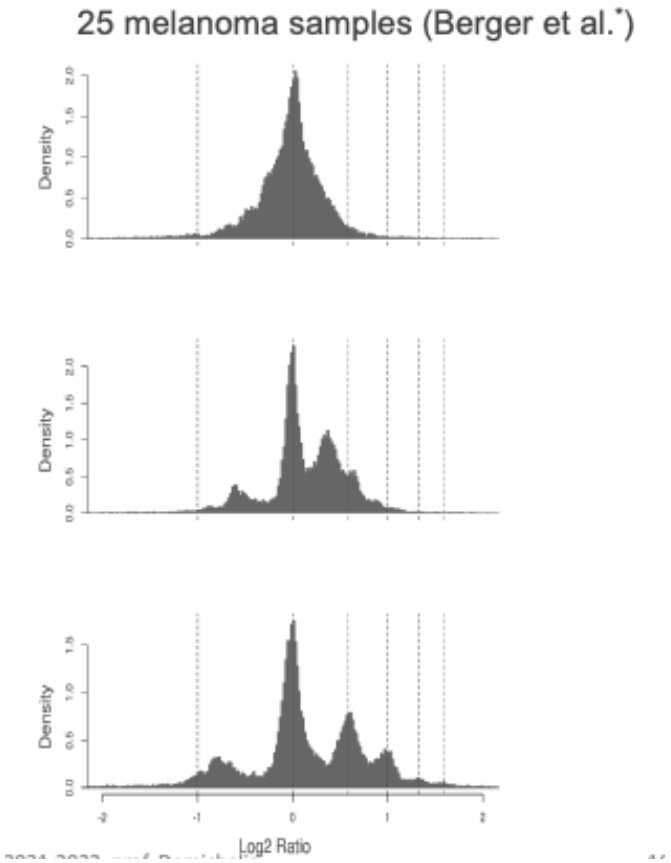
We correct our data and the *purity correction causes a stretch between the peaks*, since tumor

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



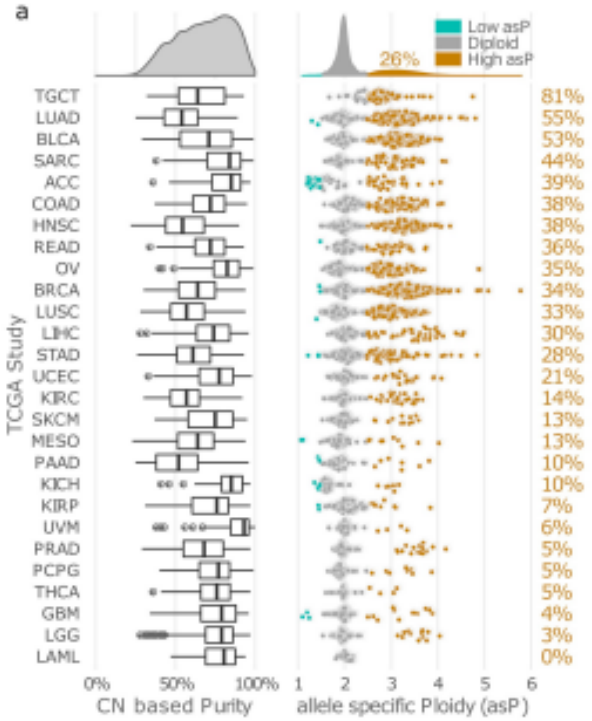
- 1st graph: The distribution of the log2 data of uncorrected signal, every melanoma sample is highly aberrant with a ploidy that is different between different individuals and a purity that is also different between different individuals. But we do have the tumor ploidy and purity so we can correct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

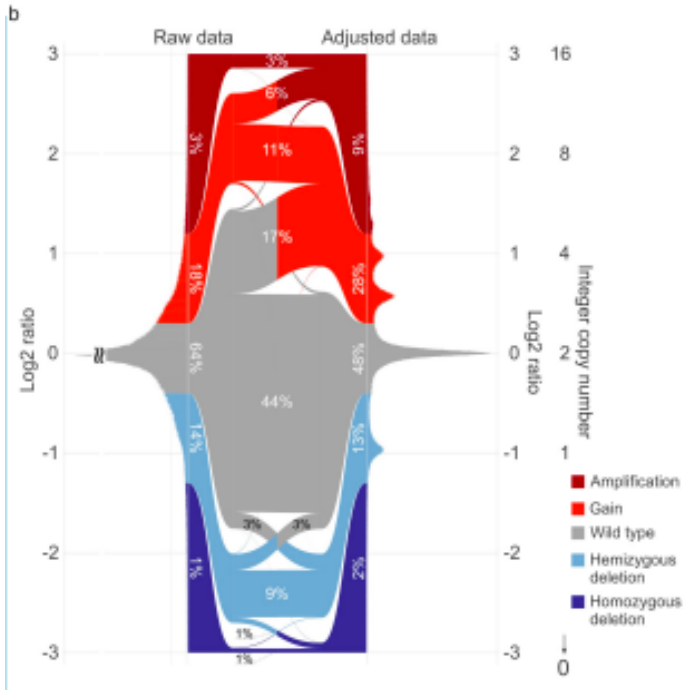
- The majority of the median distributions are above 50 %.
- The overall tumor cellularity was almost 70%.

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

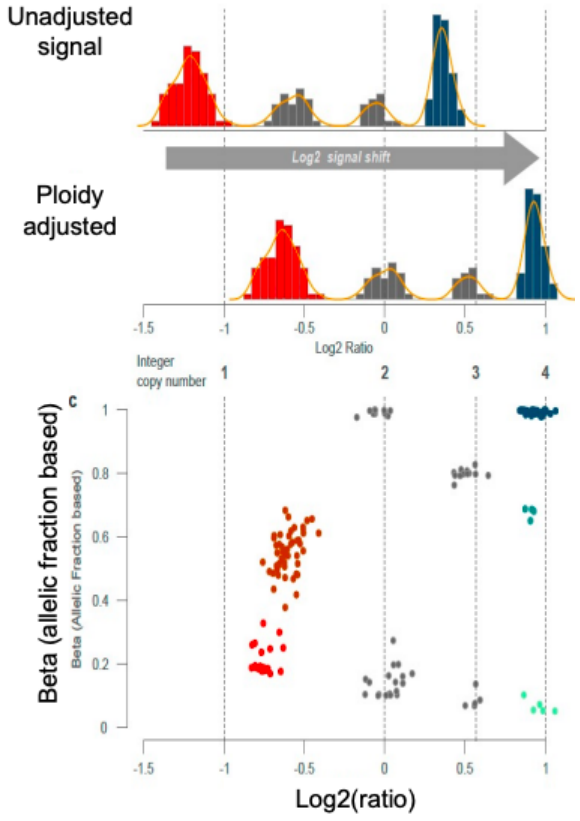


- On the y axis we have the log2 ratio
- On the left side we have the raw data

- 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

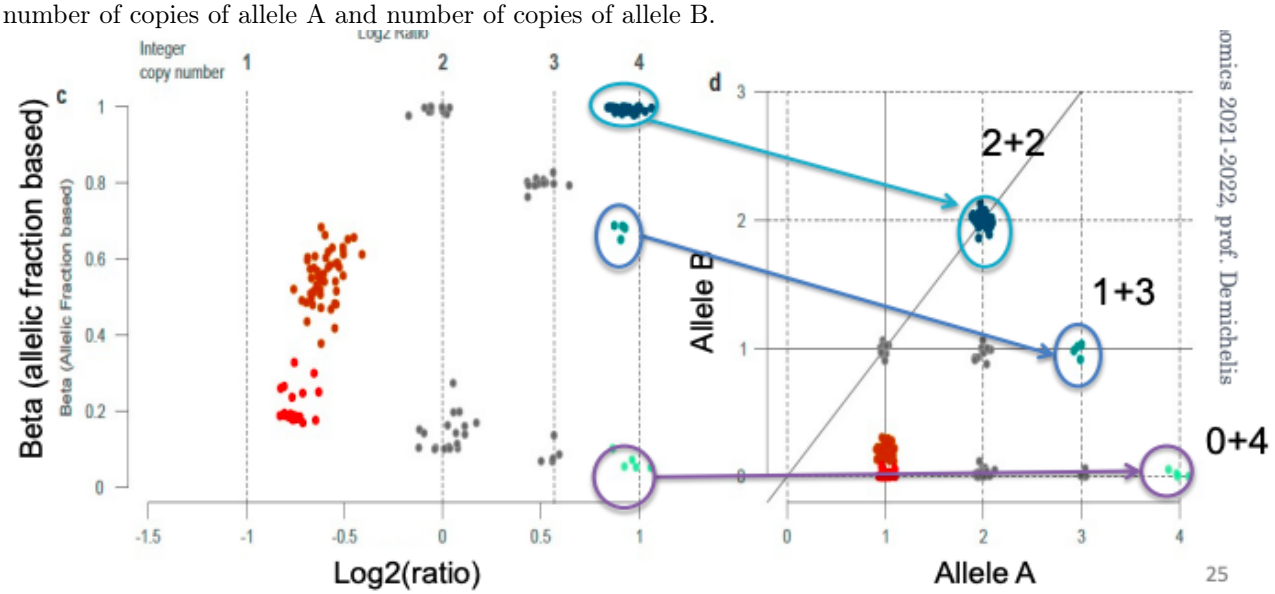
So, in this situation, the gene x will have two copies but both of them coming from the same allele (B).

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

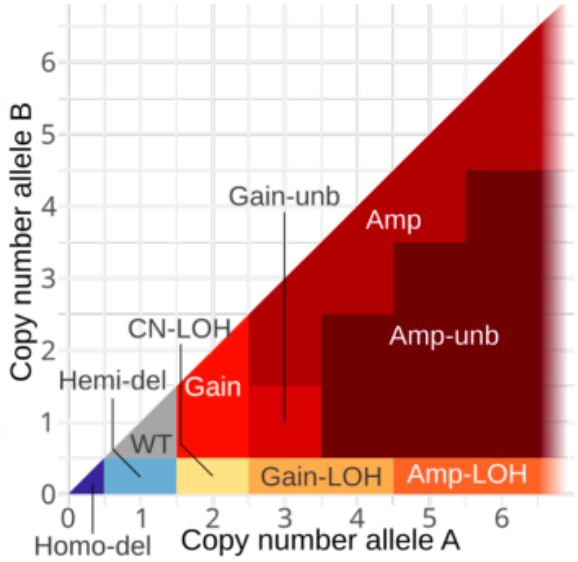
The equations are not important, what's important is that once we have corrected the data then we can shift our analysis up to the level of number of copies of each allele for each gene.

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.

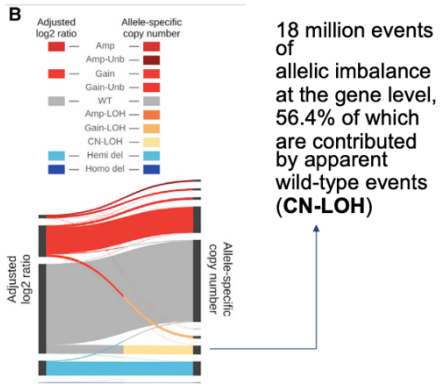
So, this whole computation allows us:

- 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).

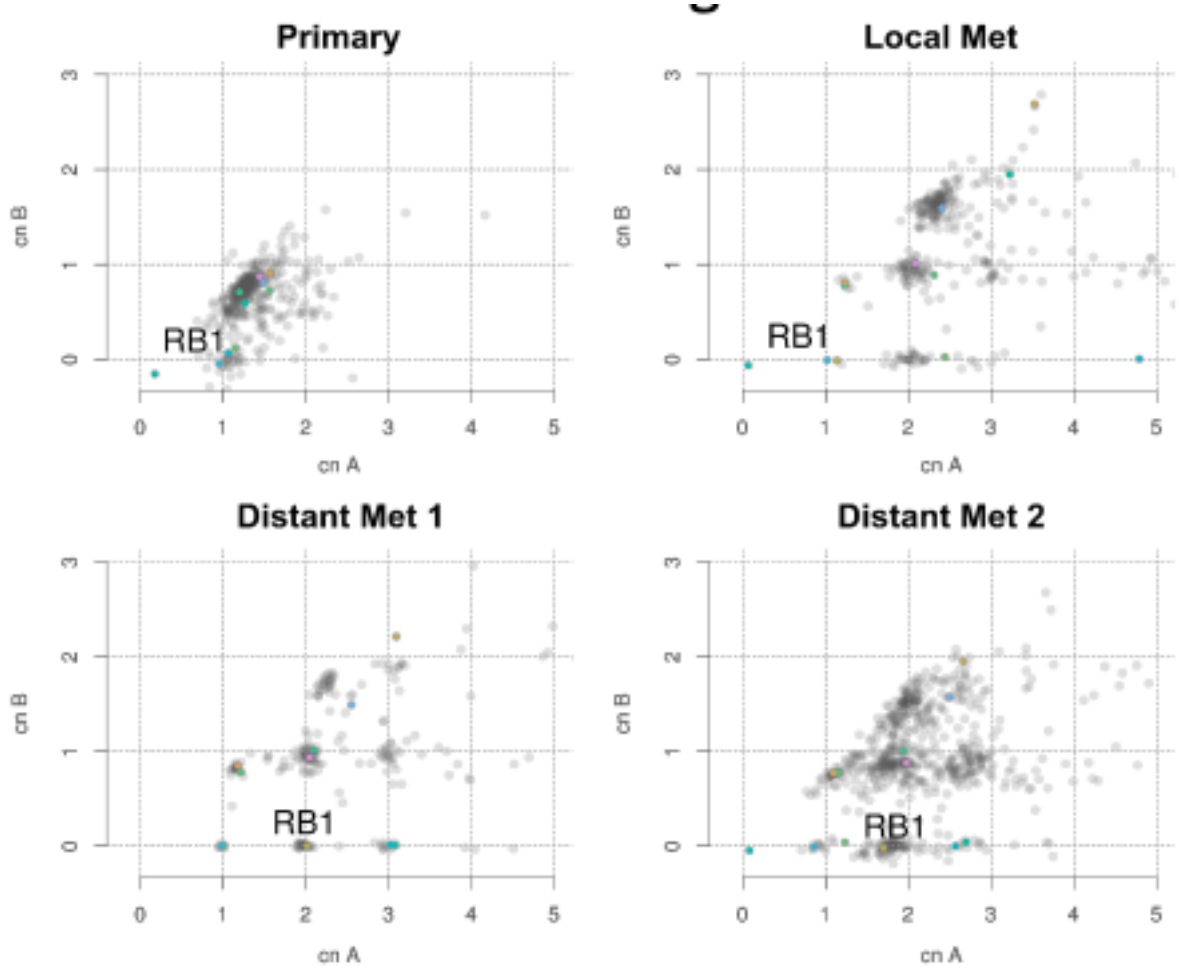
Copy-neutral because the number of copies doesn't change but there's been loss of heterozygosity.

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.

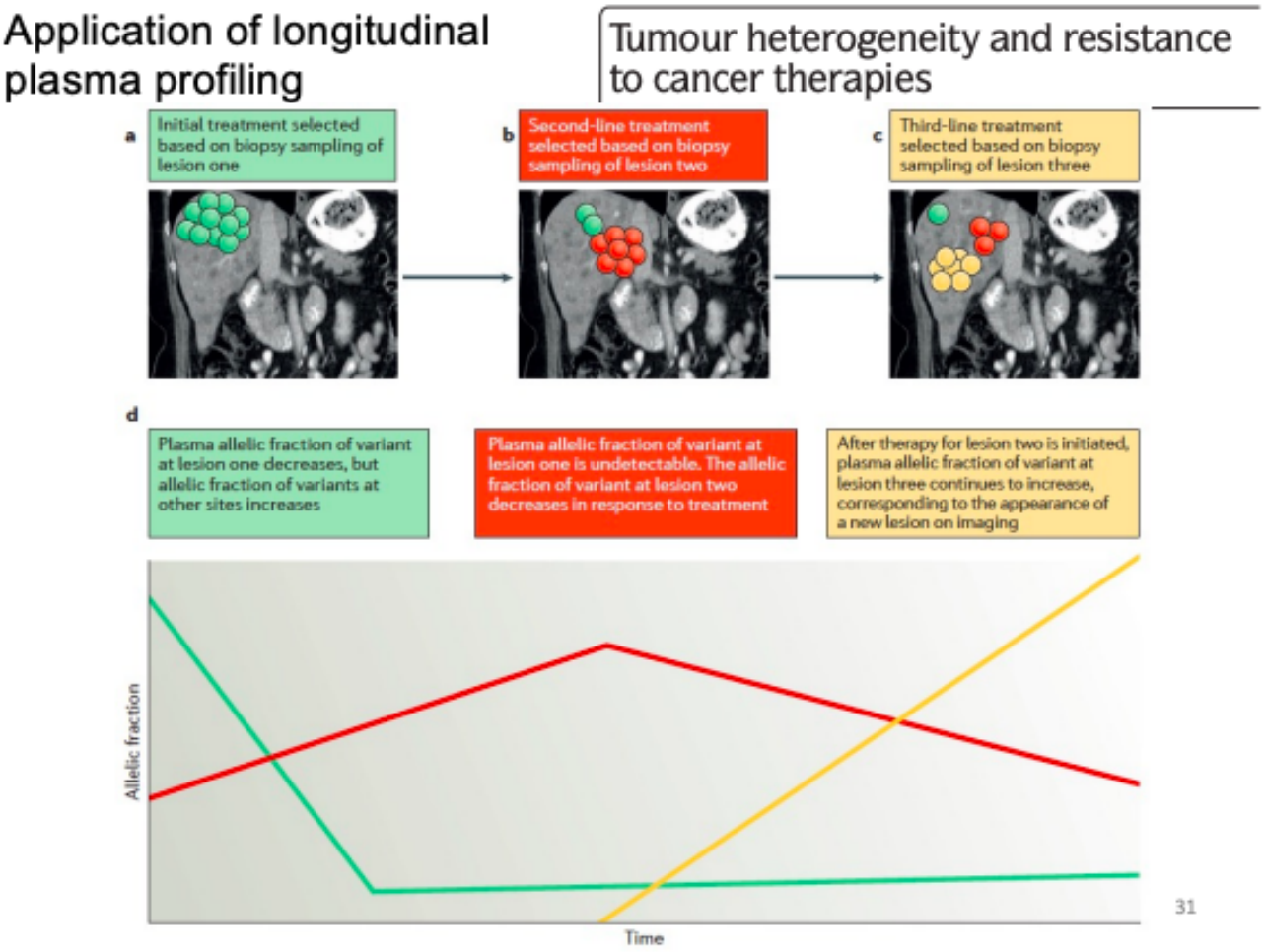
What we see:

- 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.

Chapter 7

Tumor evolution studies via NGS data: SNVs-based methods

Written by Linda Cova

There is a large number of tumors where copy-number aberrations are minimal. Consequently, it is difficult to use copy number based approaches for these kinds of tumors. It is estimated that about 3% of primary tumors present flat genomes, meaning that they display very few copy number changes. These types of tumors are correlated to a better prognosis both in overall survival and progression-free interval, but relapses are still present so the assessment of these tumors is important.

In order to address this issue, some tools were developed to detect tumor purity via SNVs.

7.1 Rationale of somatic point mutation based assays

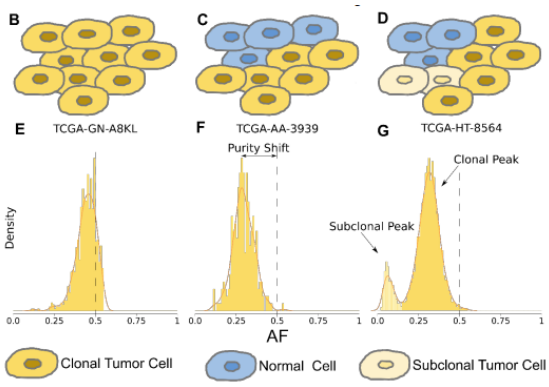


Figure 7.1: Peaks shift for a clonal tumor cell population and some mixed populations. Considering two genomic locations, healthy cells have genotype AA-AA, clonal tumor cells AB-AA and subclonal tumor cells AB-AB, where B is the alternative allele associated with a somatic point mutation

The distribution of allelic fractions of the clonal population only is symmetric, with the main peak around 0.5. A mixed population of clonal tumor and normal healthy cells shows a shifted peak.

The distance from 0.5 to the peak is proportional to the fraction of normal cells, because normal cells contribution moves the peak towards the side from the center (purity shift displayed in 7.1. A subclonal point mutation is identified with a second peak towards 0, because its allelic fraction is probably far distant from 0.5.

7.2 TPES (Tumor Purity Estimation)

Alessio Locallo (Demichelis' student, 2019)

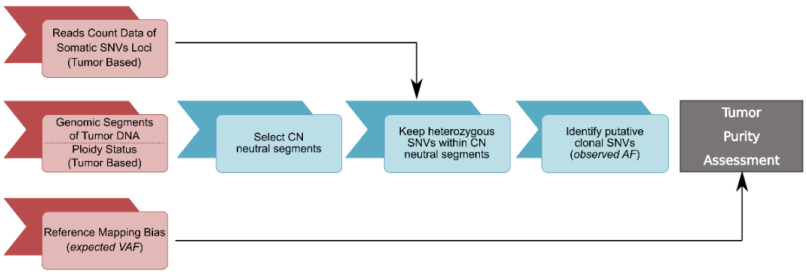


Figure 7.2: Workflow of the TPES algorithm

It is important to consider the **Reference Mapping Bias**: a polymorphic locus carrying a non-reference base is less likely to be mapped during the alignment process. With a perfect SNV (clonal, monoallelic, in highly pure tumor), allelic fraction will not be 0.5 because the aligner considers the variation as an error and sometimes discards the read containing it: some signal is lost.

TPES steps:

- **Selection of CN-neutral segments**: point mutations that are flat in terms of copy-number are perfect for flat genomes and easier to deal with. This is the first filter implemented by this tool: a threshold is set on the log2 of the tumor over the normal.
- When considering the **allelic fractions** of all the somatic mutations of whole genome, a major peak is expected around 0.5 (expected VAF). Other peaks can be originated from things that escaped the previous filter or from monoallelic mutations with copy-neutral LOH (loss of heterozygosity): in this case the allelic fraction results doubled. So another threshold on allelic fraction is needed (maxAF=0.55)
- Identification of **putative clonal SNVs**: the peak closer to 0.5 is the most useful to determine tumor purity. The others are related to subclonal events.

With enough point mutations and after peaks identification, purity is assessed with the following equation:

$$1 - purity = admixture = 1 - \frac{observedVAF}{expectedVAF}$$

7.3 How many SNVs are needed to assess tumor purity?

The number of SNVs changes for each tumor type, so not all tumor types guarantee enough SNVs. The minimum number of SNVs needed to obtain reliable results can be assessed with a **comparative**

analysis. The Spearman's correlations between the results of two different purity calling algorithms uding decreasing number of SNVs are computed. The subsampling approach (which SNVs to consider?) is to subsample the SNVs as many time as possible to have higher confidence on the results. At each iteration, as many samples as possible are used, but the number decreases when the number of SNVs increases.

The computations determined 10 as the minimum number of SNVs needed to infer tumor purity. With this number, tumor was detected in 80% of samples by combining TPES and CLONET (CN-based). The 20% could be tumor-free or not detected samples. Since both SNVs and CN based methods failed, this 20% could be possibly detected with methylation.

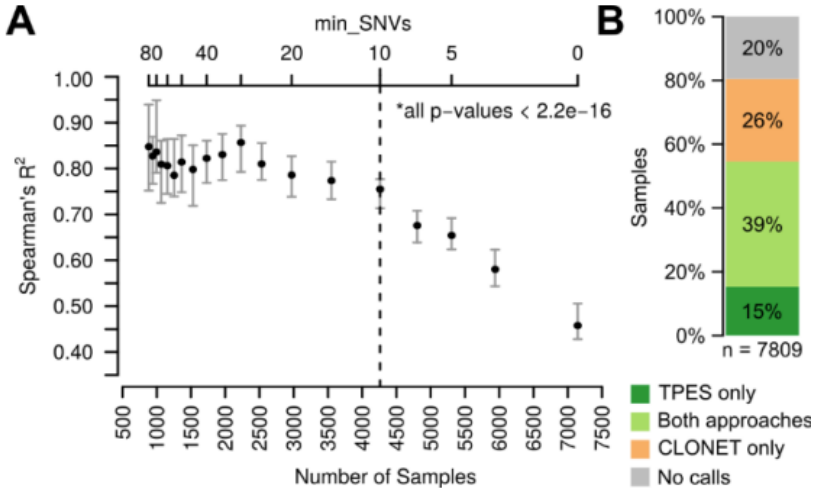


Figure 7.3: **A**) Correlation between two purity estimating algorithms (TPES and CLONET) with decreasing number of SNVs considered. **B**) Percentages of samples where tumor purity was assessed by the two tools (TPES considering with 10 SNVs)

7.4 Comparison between purity callers

TPES was compared to other tools that do the same thing but with a range of different methodologies: good correlation between the results was found, in particular with the CN-based algorithms. This shows that genomics is more reproducible in general to assess purity, while methods relying for example on image analysis give different results.

The best solution to assess tumor purity is to couple and CN-based and a SNV-based approach: some samples are only detected by one of the two so a combination gives the best results globally.

7.5 Pros and Cons of SNVs-based tumor purity assessment

Pros:

- Best-suited for CN neutral tumor genomes
- Applicable to a range of NGS techniques
- Fast and low demanding in terms of computational resources

- TPES is available as R package on CRAN

Limitations:

- Needs a reasonable number of putative clonal somatic heterozygous SNVs per sample
- Sensible to subclonal cell populations which could influence clonal peak detection

Chapter 8

Liquid biopsies in oncology

Written by Linda Cova

8.1 Liquid vs Tissue biopsies

Tissue Biopsy	**Liquid Biopsy**
Accurate and detailed view of one tissue only	Landscape overview, with resolution depending on tumor burden, releasing rates, metastases and tumor heterogeneity. It is possible to get an aggregated signal of different tumor cell populations
Single tumor	Possibility of getting signal from multiple tumor masses
Signal relative to a specific point in time	At a certain point in time, but multiple serial samples can be collected
Invasive and painful for the patient, not feasible for all the tissues	Minimally invasive (so it is possible to collect samples multiple times) and can be coupled with a routine blood draw
	It is possible to design specific assays to detect minimal quantities of tumor cells, for example the ones left behind after surgery. This is useful to detect minimal residual disease (MRD) and avoid tumor recurrence
	The collection of serial samples allows for example to track clonal evolution of the tumor over time, to catch treatment resistances early on and to monitor the patient's response to the treatment
	It can be used for early detection of cancer, many studies are trying to reach this objective

Material availability	
From needle biopsies, biopsies, surgical resections (if some material is left after the clinical protocol and the patient agrees to a research protocol)	From circulating tumor cells, extracellular vesicles, cell-free DNA (the most interesting). In healthy donors there is 4ng/ml cell-free DNA (below 10 anyway), in tumor patients 100s ng/ml (but the range is really wide). The numbers are higher if the tumor is metastatic and the treatment is also very influential on the quantity of cfDNA. Tumor patients under treatment have cfDNA quantities comparable to healthy people. Anyway, cfDNA quantity is influenced by a number of factors in addition to cancers so it is not a good diagnostic feature by itself
Tumor content	
Tumor content can be assessed with a microscope: the proportion of tumor cells compared to healthy cells is measured based on morphology with a simple staining of the tissue slide. So tumor content is assessed by counting cells and considering the magnification of the image. If subtyping is needed, a staining for markers is performed. Computational methods are also available	The fraction of circulating tumor DNA (ctDNA) is inferred with methods based on genomics (or possibly also methylation)
Tumor ploidy/aneuploidy	
Inferred with cytogenetics, FISH, or from NGS data	Inferred based on genomics but it is quite tricky

8.2 Issues in the interpretation of cfDNA data

8.2.1 Normalization on tumor content

When interpreting data from liquid biopsies, it is fundamental to contextualize a mutation after observing it. In order to associate a particular mutation to a particular diagnosis the signal has to be normalized based on tumor content. Without normalization, tumor content is the most influential variable on the patient's prognosis and this can be misleading. For example, one mutation could look like it is linked to a specific type of tumor when it is actually present in other types too but it is not detected due to the low tumor content of some samples 8.1. For this reason not all the literature available about liquid biopsies is reliable: lack of normalization leads to completely wrong conclusions. This applies to all kinds of assays: from microarrays to the sequencing of extracellular vesicles.

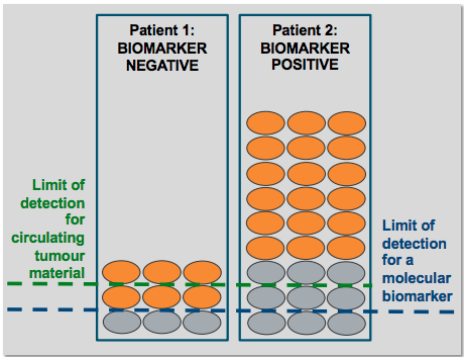


Figure 8.1: The two samples have the same percentage of tumor cells, however the first one results negative for the marker because of the low tumor content

8.2.2 Quantity of input material

Another source of errors in the interpretation of cfDNA data is the amount of input material: if the patient's tumor content is high the results could be obtained with a limited amount of extracted nucleic acid, but if the tumor content is low, too little material can lower the chances of detecting tumor cells **??**. The problem is that in most cases the tumor content is unknown before the analysis and this must be considered when designing an experiment. Usually the standard procedure is to begin with 2ml of plasma. If no tumor is detected, one should repeat the assay with more material (or sequence another vial and combine the results) to be sure that the tumor is not present and not just undetected. In some cases some information about the state of the patient is available: for example if a patient is in remission more material is required.

Keep in mind that if the sample is pure, 10 ng of DNA should correspond to around 1500 diploid tumor genomes.

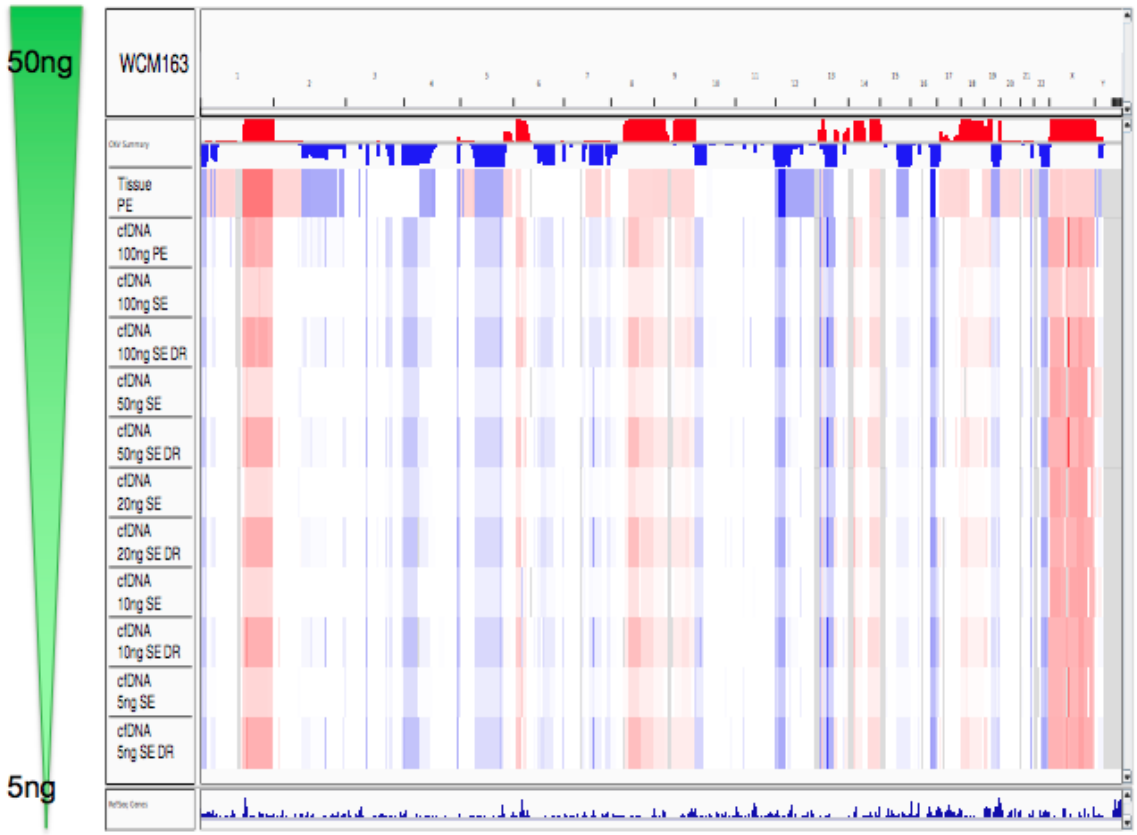


Figure 8.2: The same analyses are repeated with different quantities of starting material: for a patient with high tumor content the results do not change even with as little as 5ng of cfDNA

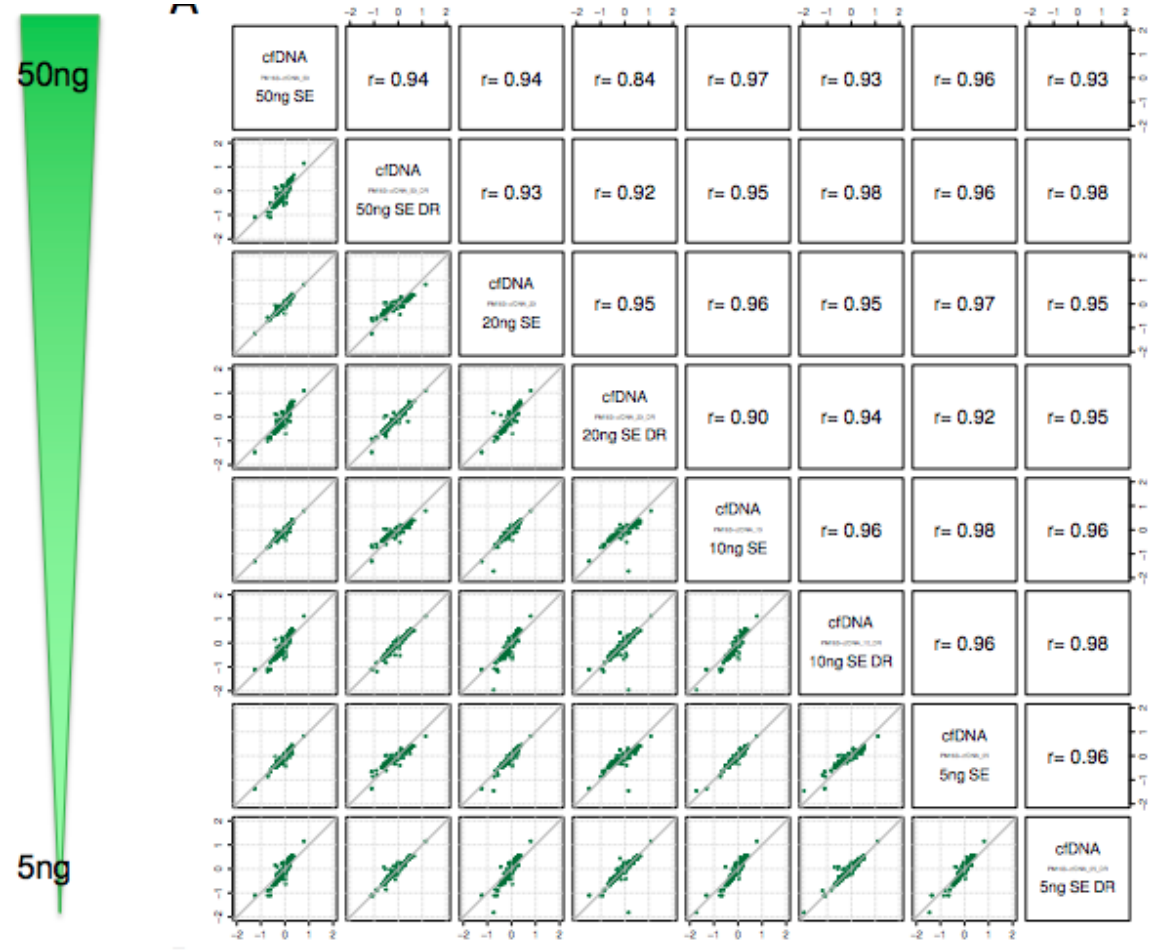


Figure 8.3: Copy number signal correlates very well between different initial amounts of cfDNA for a sample with high tumor content

8.3 SNV detection in liquid biopsies

Technical problems:

- PCR artifacts
- Sequencing errors: one mutation should be validated by multiple reads to be confirmed
- Problems related to the depth of coverage: the required coverage should be estimated considering the expected tumor content of the sample and deeper sequencing may be required

Biological problems:

- Low tumor content: ctDNA/cfDNA ratio
- Clonal hematopoiesis (when a hematopoietic stem cell starts making cells with the same genetic mutation): to distinguish the signal coming from clonal hematopoiesis, compare it with what

has been sequenced before from solid tumors. It is rare to observe something in liquid biopsy that has never been noticed in solid ones.

- Copy number variations and ploidy: with a whole genome duplication and a SNV only present on one allele, the signal corresponding to the mutation is only 25% and has to be correctly interpreted.
- Intra-patient tumor heterogeneity: very low allelic fractions for SNVs that are not clonal can be difficult to observe

Multiple **tools** are available to detect SNVs. Each tool will probably give different results (or partially concordant ones). Each tool can be tuned to favour some types of calls, so the tuning parameters should be carefully selected.

8.4 Requirements depend on the application

Application	**Requirements**
• Early tumor detection • MRD detection • Recurrence detection	Tumor quantity is low so a low signal is expected: higher quantity of starting material is required but there needs to be a balance between the number of false positives (with too much material) and false negatives (with too little) that can be produced
• Tumor dynamics • Treatment response • Mechanisms of resistance	The assay should be designed in order to be able to distinguish between different clones (sub clonality analysis)
Single biomarker assessment	The only important thing is to detect whether one point mutation is present or not, so in this case tumor content is not important. A targeted assay is used and specific locations associated with the SNV are sequenced as deep as possible to detect the mutation

8.5 Whole genome vs targeted sequencing

Whole genome sequencing has higher computational cost, while targeted assays have higher sample preparation time. The sequencing cost is higher for whole genomes but it does not decrease evenly 8.4.

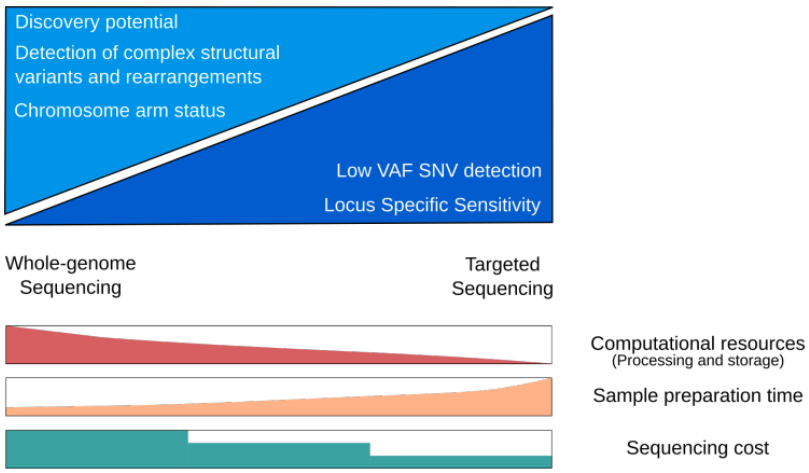


Figure 8.4: Whole genome vs targeted sequencing

8.6 Take-home message

Possible exam question: what type of assay should be run and which are the requirements for a specific situation.

Chapter 9

Epigenetic profiling of cell-free DNA

Gian Marco Franceschini - *gian.franceschini@unitn.it*
Written by Linda Cova

9.1 Introduction

All the cells of the human organism present the same genetic information but they give rise to different types of tissues and cells. This happens mostly thanks to epigenetics. The main epigenetic modifications are:

- **DNA methylation**: in humans they are mainly found on CpG islands (genomic regions with high CG content)
- **Histone post translational modifications** (PTMs)
- The **chromatine architecture**
- ...and many others

All levels of epigenetic controls are often dysregulated in cancer: these variations usually go in favor of cancer cells survival. For this reason, epigenetic reprogramming has recently been added to the hallmarks of cancer.

The epigenetic landscape is very different from the genetic one. DNA mutations are directional: they cannot be reverted so they accumulate with subsequent cells generations. The epigenome is plastic, so it can be reverted (possibly through therapy but this can happen physiologically). Moreover, the human epigenome is tissue/cell specific while the genome is unique.

9.2 DNA methylation

DNA methylation is the addition of methyl-groups to cytosines in CpG islands. It is regulated by enzymes that are responsible for regulating the cell-specific transcriptional state. These enzymes can be:

- Cis-factors: local control
- Trans-factors: genome-wise control

CpG islands are spread through the genome and when they are in a promoter they regulate gene expression through transcriptional silencing of the corresponding gene if they are methylated. The mechanisms are multiple and still not completely clear: DNA methylation could for example impair the binding of transcription factors or recruite repressing proteins. This methylation landscape is highly regulated and tissue-specific.

In cancer tissue, hypomethylated and hypermethylated regions are often observed, leading to an abnormal regulation of gene expression. In addition to that, hypermethylation of pericentrometric heterochromatin in cancer can lead to mitotic recombination and thus genomic instability 9.1.

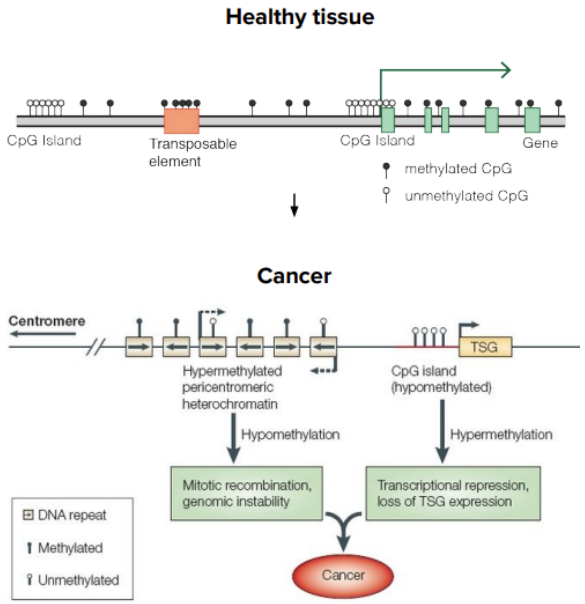


Figure 9.1: Methylation patterns altered in cancer

This landscape of regulation is very complex: DNA methylation can regulate gene expression but it is not the only regulating factor, some histone modifications also contribute for example.

DNA methylations are not inherited across generations, so there is no accumulation of methylation variants, as it happens with regular DNA mutations. Each individual is born with a brand new methylation landscape that is then disrupted during life (not only due to cancer or disease). Interestingly, i could be possible to exploit variations in the DNA methylome to measure age by computing how many cell divisions led to that specific methylation state.

9.3 How is DNA methylation measured?

The fist step is the **bisulfite conversion**: thanks to bisulfate ions, unmethylated Cs are converted into Us. With some particular alignment algorithms that are aware of these modifications one can detect the errors and thus methylations. Both array-based and shotgun-sequencing-based assays are used to this aim. The result of such an assay is a series of **beta values**: the fraction of reads corresponding to one genome site that is methylated.

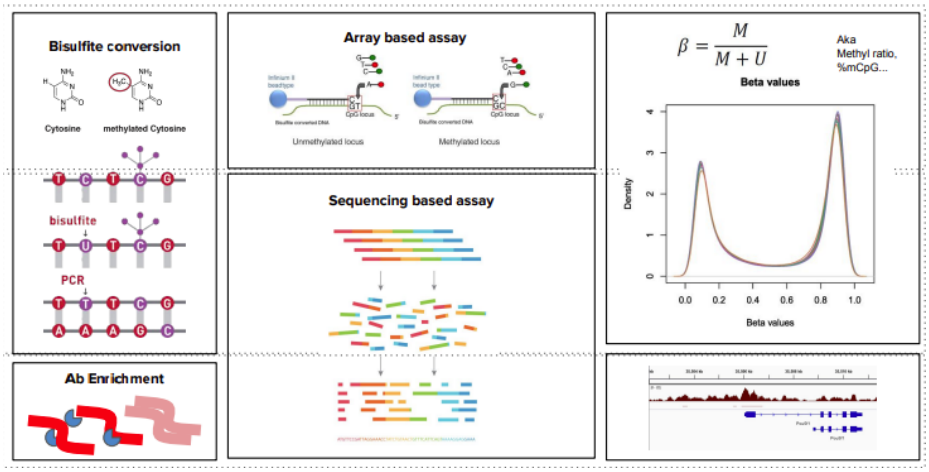


Figure 9.2: Main methods for DNA methylation measurement

Immunoprecipitation-based methods are also available, but the most frequently used methods are based on whole genome profiling.

It is useful to analyze the sequencing result at the **single-read level**: different methylation configurations can lead to the same global methylation level but have different biological interpretations. For example one methylation level of 0.5 can be the result of one completely methylated allele with the other one unmethylated or two half-methylated alleles. This kind of information is important in order to determine, for example, if the sample contains different types of cells or if it there is some disrupting pathological situation.

9.4 Tissue-specific vs disease-specific DNA markers

- **Aspecific** → most of the genome
- **Tissue-specific** → methylations that regulate gene expression to activate the tissue-specific functions of cells
- **Disease-specific** → CpG hypermethylation, genome-wide hypomethylation and other modifications usually correlated with cancer
- **Tissue+cancer-specific** → methylation patterns specific of cancer in a certain tissue. These markers allow to discriminate between different tumor types.

9.5 DNA methylation based liquid biopsy

When a cancer cell dies, its DNA is released in circulation and it is potentially possible to get it with a liquid biopsy. The goal is to analyze methylations of cfDNA to retrieve information about the state of the patient, and possibly detect early-stage tumors.

For this purpose, when compared to genomic DNA, the analysis of the methylation landscape has some positive and some negative aspects. For genomic DNA, the percentage of actually informative

signal on the whole information that is obtained can be small and difficult to observe, on the other hand, for DNA methylations it is difficult to discriminate between what is aberrant and what is not because the modifications are tissue-specific and it is difficult to obtain clear background references to make a comparison.

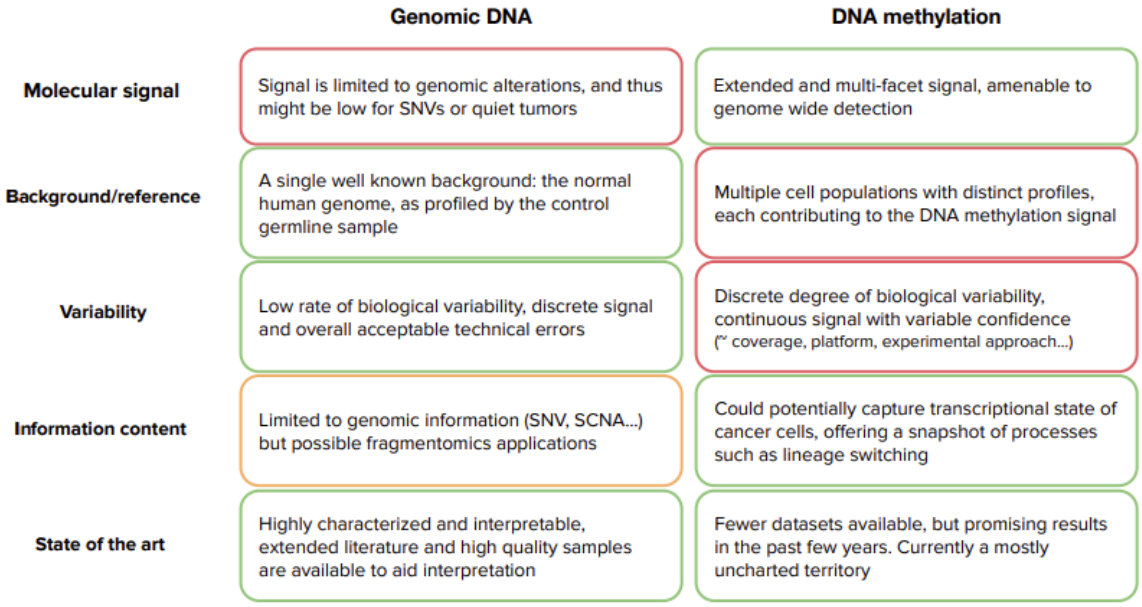


Figure 9.3: Comparison of genomics vs methylation for cancer detection

9.5.1 Workflow

First, the data is sequenced from solid and liquid samples: the methylation profiles from solid samples are needed as reference. The reference profiles for liquid biopsies analysis are derived from white blood cells and from the cancer type of interest. White blood cells are the background reference for cfDNA, since the most frequent genomic material in circulation originates from this type of cells. If a methylation pattern different from the one of blood cells is found in cfDNA it means that cells of some other tissue are dying and their material is going into circulation and it is not a positive signal. With these pattern as reference, the goal is to discover biomarkers and perform feature selection. Subsequently, a model is fitted and optimized to perform predictions on new data. The last step is performance evaluation.

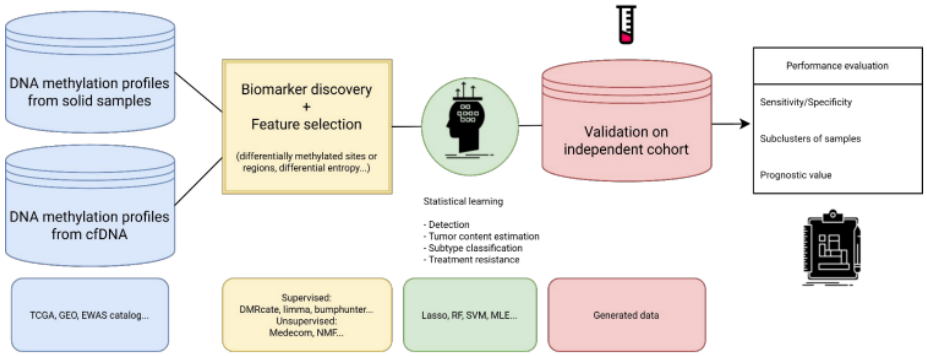


Figure 9.4: Common analysis workflow

9.5.2 CCGA study

The Circulating Cell-free Genome Atlas (CCGA) is a study conducted by Grail designed to characterize the landscape of genomic cancer signals in the blood of people with and without cancer. The study enrolled approximately 15,000 participants. Their goal is early and simple detection of cancer from analysis of methylations on cell-free DNA.

They performed whole genome methylation profiles after choosing between three different independent methods (the other two were targeted sequencing and whole genome sequencing for CNVs but they decided to further develop the methylation path). In the second phase they developed an assay for a targeted methylation study: the best features to discriminate between the two classes (cancer vs non-cancer) were selected in order to sequence the areas with these modifications without whole genome analysis. A model for this classification was developed, trained and validated. The last step is a large-scale clinical validation with a 5 years follow-up that is still in progress.

The results are great but not for all cancer types: sensitivity is better for cancers of highly-vasculated tissues and metastatic tumors, while some types of cancer prduce a lot of false negative results. Moreover, detection is obviously better when cancer progresses but the goal is early detection.

9.5.3 Deconvolution approaches

Deconvolution of cell-free DNA is another task to be performed on DNA methylation other than classification. The goal is to explain the observed signal with a combination of pure signals: discover the main contribution that led to a specific methylation landscape, one example is tumor profiling.

From liquid biopsies, it is possible to detect which are the main contributors to the cfDNA. These results can be compared with the ones obtained from cancer patients to determine which are the contributors to the difference in cfDNA that is observed and to infer data for tumor disgnosis or treatment resistance detection.

In order to perform deconvolution, **high-quality reference atlases** are needed: one was built with the contribution of Grail. They sorted healthy donors cells with FACS and profiled them. Cell type specific metylation profiles were built, so it is possible to use this atlas to select biomarkers, like a reference genome. They generated specific methylation patterns for 39 human cell types from 207 methylomes.

9.6 Targeted panel approaches for tumor content estimation

Demichelis'group study

Their interest is detection of treatment resistance in prostate cancer. The goal is to know when the tumor becomes resistant, in order to be able to change or calibrate the therapy. A sequencing panel was developed to detect the amount of cancer-derived DNA in circulation, and interestingly only 50 regions are sufficient to get a satisfying estimation. A model is built to know how much ctDNA is expected after treatment and it is possible to get a score that estimates the level of resistance.

Part II

Papers

Chapter 10

Role of non-coding sequence variants in cancer

10.1 Abstract

Patients with cancer carry somatic sequence variants in their tumour in addition to germine variants in their inherited genome. Numerous studies have noted the importance of non-coding variants in cancer. The overwhelming majority of variants occur in non-coding portion of the genome.

10.1.1 Introduction

One of the most important benefit of whole genome sequencing is the identification of variants in non-coding regions of the genome, with most of them lying in such regions. One of the biggest challenges is to identify driver mutation and distinguish them from passenger mutations.

10.2 Genomic sequence variants

The general properties of sequence variants are applicable to non-coding variants. They range from single nucleotide variants to small insertion and deletion of less than $50bp$ or indels, to larger structural variants. The latter can be copy number variants CNV or copy-number neutral. An average human genome contains 4 million germline sequence variants, whereas a tumour genome contains thousands of variants relative to the same individual germline DNA. Somatic variants are rarer in healthy tissues. Somatic mutation frequency varies across different cancer types. Some germline variants may be responsible for tumorigenesis (high penetrance) or modulate the effect of somatic variants (low penetrance). The germline variants associated with increased cancer susceptibility do not have a fitness effect at reproductive age, which can be the reason for the continued prevalence of such variants in the population. Germline variants show LD that increase the difficulty in disentangling the causal disease variants. A much higher fraction of somatic variants consist of structural variants and unlike germline variants they happen on a specific tissue. However germline variants can have a functional effect in specific tissue if they occur in regions of closed chromatine or if they disrupt a binding site of a tissue-specific transcription factor. Kataegis is characteristic only of somatic variants. Moreover somatic variants are not inherited and so they are not subject to meiosis and do not show LD.

10.3 Non-coding element annotation

Non coding elements can have diverse roles in the regulation of protein-coding genes. They consist of cis-regulatory regions and ncRNAs. They are identified by functional genomics approaches or sequence conservation and display cell and tissue specificity.

10.3.1 Cis regulatory regions

Cis-regulatory regions include promoters and distal elements which regulate gene expression following binding by TFs. TFs bind to specific DA sequences within thier larger regions of occupancy which can be identified using chromatin immunoprecipitation followed by sequencing assays. They bind DNA in regions of open chromatin identified using DNAase I hypersensitivity assays and DNAase I footprinting. DNA methylation and other histone modification can modulate TF accessibility. Several histone marks are ssociated with specific putative functions. Most of sequence-specific TF and chromatin maks lead to highly localized ChIP-seq signals, other marks are associated with large genomic domains. Epigenetic hcanges can alter FT accessibility in different cellular states and can change the activity or regulatory elements.

10.3.2 Distal regulatory elements

Distal regulatory elements may regulate gene expression by interacting with promoters in the 3D structure of the genome. Linking them to their target region is crucial to understand the effects of sequence variants in them. Multiple approaches have been used linke chromosome conformation capture: regulatory sequences can control transcription by looping to and physically contacting target coding genes that are located tens or hundreds of kilobases away. It probes one-versus-one contacts in the 3D space of the genome. Other variation control one-versus-all, many-versus-many and all-versus-all contacts. Other approches include correlation of histone marks at enhancer regions and target gene expression across multiple cell lines. Links between epression quantitative trait loci and associated genes. The resulting linkages can be studied as a comprehensive network.

10.3.3 RNA-seq

RNA-seq reveals non-coding transcripts, which can be confirmed to not have protein-coding ability by the absence of open reading frames or proteomic analysis. Certain histone modification can also indicate ncRNA activity. ncRNA can be divided into categories and they act through different mechanisms to modulate genexpressions. In particular miRNA and lncRNA are important in cancer biology. miRNA inhibit target gene expression by binding to the 3'-UTR and causing mRNA degradation or repression of translations. The mechanisms of action of lncRNA remain unclear, but a number of lncRNA have been shown to act as molecular scaffolds that bind proteins, DNA or other RNA molecules and are able to modulate gene expression.

10.3.4 Transcribed pseudogenes

Transcribed pseudogenes are a type of ncRNA tha bear a clear resemblance to functioning protein coding genes. They are copies of coding genes that have lost their ability to code for proteins owind to disabling mutations. They can be divided into duplicated and processed based on their formation from duplication or retrotransposition of the parent gene. Processed pseudogenes lack the promoter sequence and intronic strucutre and contain a 3'-poly(A) tail. These pseudogenes can be transcribed

and regulated the expression of their parent genes, generating endo-siRNA and partecipating in the RNA interference pathway or by acting as molecular sponges.

10.3.5 Evolutionary conservation

Evolutionary conservation of genomic sequence across multiple species is used to annotate non-coding regions. Comparative analysis allowed the discovery of these ultra-conserved elements, the majority of which do not overlap protein-coding exons. Analysis of these sequence is important because they have been show to have a role in cancer biology. Non-coding elements exhibit conservation among humans. Negative selection within the population can be estimated using enrichment of rare alleles and reduced density of single nucleotide polymorphisms. These can be important to identify elements that show human-specific conservation in functional non-coding categories. The ultra sensitive elements and have strong depletion of common polymorphisms and enrichment of known disease-causing mutation Negative selection can be used to identify candidate cancer driving mutations.

10.4 Roles for somatic variants in cancer

Because cancer genomes contain a higher fraction of structural variants than germline genomes, variant detection becomes challenging. The depth of coverage needs to be more than typically used to account for the decreased purity and increased ploidy.

10.4.1 Gain of TF-binding sites

TERT encodes the catalytic subunit of the enzyme telomerase. This allow to lenghtens telomeres, allowing cells to escape apoptosis and become cancerous. TERT expression is typically repressed, but it can be overexpressed in cancer. Recurrent mutation in the promoter of TERT in many different cancer types have been found. These mutations create binding motifs for the ETS family like TCF leading to their binding to TERT and upregulation of its expression. Tumours in tissues with low rates of self-renewal tend to exhibit higher frequencies of TERT promoter mutations. Gain of TF-binding site has been observed for enhancers, an important distal cis-regulatory elements that play a major part in gene transcription.

10.4.2 Fusion events due to genomic rearrangements

Genomic rearrangments can lead to fusion of active regulatory elements with oncogenes. Moreover somatic structural variants juxtapose coding sequences proximal to active enhancers during enhancer hijacking. So in these genomic rearrangements bring oncogenes adjacent to active promoters or enhancers.

10.4.3 ncRNAs and their binding sites

Disregulation of ncRNAs is a cancer signature and it can be due to the presence of somatic variants in them. MALATI or metastasis-associated lung adenocarcinoma transcript 1 is an example of this. Mutation of MALATI might be under positive selection in the tumour. In another example copy number amplification of a lncRNA is thought to contribute to neuroblastoma progression. Mutation in the binding sites of ncRNA are linked to cancer.

10.4.4 Role of pseudogenes in modulating the expression of a parental gene

becase of their resemblance to their parental protein-coding genes, transcribed pseudogenes are thought to have a natura way of affecting and regulating their parental counterparts. Pseudogene deletion or amplification can affect competition ofr miRNA binding.

10.5 Roles for germline variants in cancer

Most of the non-coding germline variants associated with cancer susceptibility can be analyzed through WGS data from healthy and ill individual. Germline-non coding variants can affect gene expression in many different ways: point mutation can distrupt binding motifs. GWAS SNPs and the one in LD with them might help to identify the causal variants and shed light on their mechanism of actions.

10.5.1 Promoter mutations

Germline mutation can create binding motifs with functional effects in the tissues where the TF is expressed. Moreover they can upregulate the binding.

10.5.2 SNPs in enhancers

Multiple SNPs in a gene desert can increase the risk of cancer: this can be due to the fact that they happens in regions that act as enhancers . Tissue specificity might be the reason why they are associated with specific cancers. Hormone-regulated cancers have mutation in TF-binding sites that vary with age owing to a differential TF activity during a person lifetime.

10.5.3 Variants in introns

Variants in introns can affect splice sites and cause loss of regulatory repressor elements. Germline CNV spanning intronic inhibitor regulatory elements can lead to the overexpression of target transcripts, modulating cell proliferation or migration.

10.5.4 SNPs in ncRNA and their binding sites

Most cancer-associated polymorphisms are related to increased risk, some of them can be beneficial.

10.5.5 Others

Other methods to identify variants with functional consequences such as ECTS and allele-specific expression analysis have been used to interpret cancer-associated loci identified through GWAS. These reveal germline determinants of gene expression in tumours and help to establish a link between non-coding risk loci and their target coding genes.

10.6 Interplay between germline and somatic variants

Cancer results from a complex interplay of inherited germline and acquired somatic variants. Loss of heterozygosity events affecting non-coding element have been observed. Somatic variants disrupt

the only functioning copy of the non-coding element. One example is the loss of miRNA or lncRNA. However some mutation can weaken the effect of a somatic variant.

10.7 Computational methods for identifying variants

Computational prediction of drivers is a challenging task. Driver identification uses detection of signals of positive selection or prediction of mutations with high functional impact. Analysis of the recurrence of somatic variants from tumour samples in functional elements to identify regions under positive selection is similar to the burden test strategy. Such analysis can be done in a specific cancer type or across multiple cancers. In addition tools that try to do this need to account for genomic mutation rate covariats that lead to mutational heterogeneity across the genome. Computational identification of non-coding drivers is more challenging than the coding one because of thei complex and varied modes of action. Non-coding mutation are also more abundant and the key mutations have to be distinguished from a larger set of passenger events. Some methods analyse the recurrence of somatic variants from tumour samples in functional elements. Tools exits to annotate and prioritize potentially functional non-coding variants with high impact. These tools can interpet SNV and indels or some structural variants. Some of them try to interpret the effect of cis-regulatory mutations at a nucleotide level of resolution by computing whether they create new TF-binding motifs. Biological networks can provide information about the connectivities of the target genes of non-coding variants. High inter and intra-species conservation tend to be an indicator of function.

10.8 Experimental approaches for functional validation

Experimental approaches to understand the effects of cis-regulatory mutations in promoters and enhancers on cellular functions have main strategies. First they require introducing the sequence variants,determining the resulting molecular level effects on transcription using high and low throughput functional assays and demonstrating direct biological significance. One way to intrduce sequence variants involves the use of CRISPSR-Cas9 systems. Then the effect evaluated through sequencing screening or luciferase reporter assays. Analysis of the mutation in a high-throughput manner can be achieved using a modification of cis-regulatory eleemnt analysis by sequencing. Synthetic promoter libraries drive the expression of a common reporter gene and a downstream unique barcod sequence that identifies the upstream promoter. RNA-seq reveals the effects of promoter variants on the expression levels of their paired barcode sequence. The activity of enhancers it independent of their location, so they can be incorporated into high-throughout reporter assays using different reporter construct arrangements. In CRE-seq approaches the enhancer is placed upstream of the reporte gene and the barcode. The cloned libraries can be transfected into eukaryotic cells in pooled format and RNA-seq is used to asses the resulting expression level of the reporter driven by each variatn element. Visible reporte assays using synthetic transcription reporter construct that contain the regulatory sequence of the reporter gene enable direct validation. Other approaches are needed to validate variants in ncRNA, UTR and intros. Monogene assays can be used to test the effects of intronic variants: the variant sequence is cloned into transcription-competent minigene vectors and transfected into mammalian cells. This is followed by examination of the splicing patterns of the transcripts. Functional screening help identify the best candidates but still needs tumour type specific validation. Functional valdation requires demonstrating oncogenic properties that are increased owing to the variant in question. Wild type and mutants are compared in vitro and in vivo. Overall functional validation of non-coding variants is important to understand their biological consequence.

High-throughput prioritization of putative functional mutations is crucial before testing of the most promising candidates in in vivo systems.

Chapter 11

Advances in understanding cancer genomics through second-generation sequencing

11.1 Abstract

The application of second generation DNA sequencing technologies is allowing substantial advances in cancer genomics. These methods are increasing the efficiency and resolution of detection of each of the principal types of somatic cancer genome alteration.

11.1.1 Introduction

A near term medical impact is the elucidation of mechanisms of cancer pathogenesis, leading to improvements in the diagnosis of cancer and the selection of cancer treatment. It has become feasible to sequence expressed genes, known exons and complete genomes of cancer samples. Most of the genomic alteration that cause cancer are somatic. Studying these alteration can improve therapies targeted against the production of these alterations. Comprehensive genome based diagnosis of cancer is increasingly crucial for therapeutic decisions. Some genomic alterations in cancer are prevalent at a low frequency in clincal samples, owing to substantial admixture with non-malignant cells. These methods makes it feasible to discover novel chromosomal rearrangements and microbial infections and to resolve copy number alterations at very high resolution. The data generated from second-generation sequencing provides a statistical and computational challenge. This will be partly solved by systematic analysis of large cancer genome data sets.

11.2 Cancer-specific consideration

Cancer samples and genomes have general distinct characteristics from other tissue samples that require particular consideration.

11.2.1 Characteristics of cancer samples for genomic analysis

Cancer samples differ in their quantity, quality and purity from the peripheral blood samples. Diagnostic biopsies from patients with disseminated disease tend to contain few cells, therefore the quantity of nucleic acid available may be limiting. An alternative approach to deal with small sample is whole-genome amplification, but it does not preserve genome structure and can crate artefactual alteration. Nucleic acids from cancer are of lower quality due to formalin fixation and paraffin embedding necessary for microscopi histology. They will have undergone cross-linking and be degraded. Special experimental and computational methods are required. Moreover cancer specimens can include substantial fraction of necrotic and apoptotic cells. Moreover a cancer specimen will have a mixture of cancer and normal genomes and the cancer themselves can be highly heterogeneous and composed of different clones.

11.2.2 Structural variability of cancer genomes

Cancer genomes vary in their sequence and structure compared to normal genomes and among themselves. Cancer genomes vary in their mutation frequency, in global copy number or ploidy and in genome structure. The presence of a somatic mutation is not enough to establish statistical significance: it must be evaluated in terms of the sample-specific background mutation rate. The analysis of mutations must be adjusted for the ploidy and purity of each sample and copy number at each region. To identify somatic alteration, comparison with matched normal DNA from the same individual is essential.

11.3 Experimental approaches

The application of second-generation sequencing has allowed cancer genomics to move from focused approaches to comprehensive genome-wide approaches.

11.3.1 Whole genome sequencing

Complete sequencing of the genome of cancer tissue to high redundancy, using germline DNA sequence from the same individual as a comparison has the power to discover the full range of genomic alterations using a single approach. So it is the most comprehensive characterization of the cancer genome and the most costly. The major potential is the discovery of chromosomal rearrangements. It also may be able to detect other types of genomic alterations like somatic mutations of non-coding regions as well as non annotated regions. The two main parameters to consider when performing WGS are depth of coverage and physical coverage. Sequence depth is measured by the amount of over-sampling, typically at least a 30 fold average coverage is needed. Physical coverage is important for detecting rearrangements. This is helped by paired-end sequencing. The expected distance between paired reads is used to place the reads on the reference genome. The distance between the paired reads can be increased creating jumping libraries by circularization. This has two limitation: the coverage is lower and point mutation resolution is lower. Second it requires large high-quality DNA, which may not be possible with all clinical cancer samples.

11.3.2 Exome sequencing

Target sequencing approaches have an increased sequence coverage of regions of interests at lower costs and higher throughput. Any subset of the genome can be targeted. Capillary-based sequencing

has been proben powerful to focus sequencing efforts on the coding genes of interest. Uneven capture efficiency across exons can mean that not all exons are sequenced and some off-targed hybridization can occur. The higher coverage make WES suitable for mutation discovery in cancer samples of mixed purity.

11.3.3 Transcriptome sequencing

RNA-seq is a powerful approach for understanding cancer. Transcriptome sequencing is sensitive and efficient in detecting intragenic fusions like in-frame fusion events that lead to oncogene activation. Transcriptome sequencing can be used to detect somatic mutations by finding a matched normal sample. Mutation detection is hampering due to a lack of statistical power. RNA-seq allows analysis of gene expression profiles and is powerful for identifying transcripts with low-level expression. It can also detect novel transcripts, alternative splice forms and non-human transcripts.

11.4 Detecting classes of genome alterations

Second-generation sequencing can provide a comprehensive picture of the cancer genome detecting each of the major alterations in the cancer genome.

11.4.1 Somatic nucleotide substitutions and small insertion and deletion mutations

Nucleotide substitutions are the most comommon somatic genomic alteration occurring at a frequency of one in a million. Insertion and deletions are tenfold less common. The rate of mutations varies greatly between cancer specimens. Detection of somatic mutations in cancer requires mutation calling on the tumour DNA and the matched normal DNA, coupled with comparison to a reference genome. False positive are inaccurate detection of an event in the tumour and detection of a germline event in the tumour but failure to detect it in the normal. Noise can be due to machine-sequencing errors, incorrect local alignment and discordant alignment of pairs. Moreover it can be caused by failures to detect the germline alleles that differ from the reference sequence in the normal sample. False negative si often due to insufficient coverage. Statistical significance of an alteration can be assessed by comparison to the sample-specific background mutation rates in the specific nucleotide context and correcting for multiple hypothesis testing. Computational tools predict the effect of an amino acid change on the protein structure and function, and some tools aim to distinguish driver from passing alterations. Experimental validation is the most powerful method.

11.4.2 Copy number

Second generation sequencing methods offer substantial benefits for copy number analysis, including higher resolution and precise delineation of the breakpoints of copy number changes/ The digital nature allow to estimate the tumour to normal copy number ration at a genomic locus counting the number of reads in both tumour and normal samples in the locus.

11.4.3 Chromosomal rearrangements

Second-generation sequencing has been shown to allow systematic description of the rearrangements in a givencancer sample. Extension of these approaches to large numbers of samples should lead to the discovery of the makor recurrent translocations in cancer. Intrachromosomal rearrangements,

inversions, tandem duplications and deletions, insertions of non-endogenous sequences like viral ones, reciprocal and non-reciprocal interchromosomal rearrangemnts and complex rearrangements like combinations of these various events can be detected through second-generation sequencing.

11.4.4 Microbe-discovery methods

In addition to somatic alterations many cancers are caused by microbial infections. Neither array methods nor directed sequencing approaches can identify new examples of microbial genomes that have inserted themselves into the human genome. COmputational subtraction of the sequence from a sample from the human reference genome can detect non-human sequences and identify novel microbial infections associated with human disease. Challenges include low concentration of the microbial agent ,hit and run mechanisms, quality issue that cause artefacts and incompleteness of human genome reference samples.

11.5 Computational issues

The three main challenges in developing computational solutions are the need to simultaneously analyse data from tumour and patient to identify rare somatic events, ability to analyse very different and highly rearranged genomes and to handle samples with unknown levels of non-tumour contaminations and heterogeneity within the tumour.

11.5.1 Alignment and assembly

Reads must be aligned to the specific chromosome, position and DNA strand from which they are most likely to have originated. Thiese are perfomed against reference human genomes using methods developed for normal samples. The uniqueness of every cancer genome and the difficulty of correctly assigning rearranged sequences from homologous regions mean that de novo assembly of cancer genomes is likely to become the most powerful approach.

11.5.2 mutations detection

As somatic genome alteration are rare, any method that detects mutations in cancer must do so with low false positive rates. The first report of a method specific for somatic mutation calling or SNVMix. Systematic analysis of false-positive and false-negative rates of the methods based on real cacner data is yet to be performed. A naive somatic mutation caller can be built by applying a germline single-sample mutation caller to thetumour and normal data sets: somatic events are those detected only in the tumour. Somatic mutation calling is more complex because cancer samples vary in purity and ploidy. A key parameter for each mutation is its allelic fraction: the expected fraction of reads in the tumour that harbour the mutation among all reads that map to the same genomic location. The allelic fraction captures the local complexity of the tumour genome, the non-tumour contamination levels and any mutation-dependent experimental or alignment bias.

11.5.3 Validation of mutation and rearrangement calls

Accurate estimation of false positive and false-negative rates is a challenge. The second can be estimated by validation of the event using an orthogonal technology: a genotyping assay such as mass spectrometric analysis. This is not sufficiently sensitive to validate mutations with low allelic fracitons. Current efforts are focused on applying deep targeted second generation sequencing to

validate the events. For validating rearrangments the current methods require PCR amplification of the region surrounding the event follwed by sequencing of this region. They are not high-throughput. A deveolping concept is to capture the rearranged sites using a similar protocol to the exon capture approach and apply deep sequencing.

Chapter 12

Integrative genomics viewer

12.1 Introduction

Experienced human review is essential in analysis of the datasets generated during genomic studies. The integrative genomics viewer or IGV is a visualization tool that enables intuitive real-time exploration of diverse, large scale genomic data sets. It supports integration of aligned sequence reads, mutations, copy number, RNA interference screens, gene expression, methylation and genomic annotations. IGV makes use of efficient, multi-resolution file formats to enable real-time exploration of arbitrarily large data sets over all resolution scales. The user can zoom and pan across the genome at any level of detail, from whole genome to base pair. Sample annotations can be defined and data divided into tracks. Annotations are displayed as a heatmap. Its scalable architecture makes it well suited for genome-wide exploration of NGS datasets, both basic aligned read and its derived results. As the user zooms below the $50kb$ range individual aligned reads become visible and putative SNPs are highlighted as allele counts in the coverage plot. Zooming in further individual base mismatches become visible, highlighted by color and intensity according to base call and quality. Reads can be sorted by quality, strand, sample and other attributes. IGV use paired ends reads to color-code paired ends if their insert sizes are larger than expected, fall on different chromosomes or have unexpected pair orientations. Intra and inter chromosomal events are readily distinguished by color-coding.

Chapter 13

Tumour heterogeneity and resistance to cancer therapies

13.1 Abstract

As a result of cancer heterogeneity, the bulk tumour might include a diverse collection of cells harbouring distinct molecular signatures with differential levels of sensitivity to treatment. This might reult in a non-uniform distribution of distinct subpopulations across and within disease sites or temporal variations. This provides the fuel for resistance.

13.1.1 Introduction

The stochastic nature of cancer initiation reinforces the notion that the development and progression of cancer does not follow a fixed course. The ongoing evolution of cancer migh generate a molecularly heterogeneous bulk tumour consisting of cancer cells harbouring distinct molecular signatures with differential levels of sensitivity. Intertumoural heterogeneity is the heterogeneity between patients harbouring tumours of the same histological type. Intratumoral heterogeneity is spatial or temporal heterogeneity: dynamic variations in the genetic diversity of an individual tumour over time. Oncogenic drivers can be exploited to treat cancer, but almost all of them develop resistance to targeted therapies. Intratumoural heterogeneity drives the evolution of cancers and fosters drug resistance. A comprehensive understanding of tumour dynamics is essential for the development of effective and durable therapeutic strategies.

13.2 Causes of intratumoral heterogeneity

13.2.1 Genomic instability

Instability might result from exposure to exogenous mutagens and aberrations in endogenous processes. Characteristic genetic signatures associated with some of these mutagenic processes have been identified by large-sacle genomic sequencing. Exposure to chemotherapy might increase the mutational spectrum of a tumour and create genomic instability. Genomic instability can also result from chromosome-level changes that lead to gains or losses of whole-genome segments rather than point mutations.

13.2.2 The clonal evolution and selection hypothesis

Genomic instability fosters genetic diversity by providing the raw material needed for the generation of tumour heterogeneity. Dynamic chromosomal instability can lead to copy-number imbalances and non-uniform loss of chromosomal segments harbouring specific alterations that can contribute to mutational heterogeneity across different regions. Increased levels of genomic instability promote the emergence of more competitive subclones. Genomic instability cooperates with other factors to promote the development of tumour heterogeneity. The clonal evolution method and or the selection framework are used to explain how clonal diversity is generated and maintained. This model is based on the hypothesis that tumour initiation occurs in a stochastic manner, beginning with an induced change that confers a selective growth advantage and leads to neoplastic proliferation. The genomic instability creates additional genetic diversity subjected to evolutionary selection pressures, resulting in the sequential emergence of increasingly genetically abnormal and heterogeneous subpopulations. Linear evolution describes evolution owing to the successive acquisition of mutations that confer a growth and survival advangage. Sequential clones have advantageous mutations and out-compete ancestral clones. Alternatively branching evolution denotes the emergence and divergent propagation of multiple sub clonal tumour cell populations that share a common ancestor. Branched evolution has a greater opportunity to create a more heterogeneous tumour. Moreover different sub clones might cooperate for tumour propagation in cancer.

13.3 The spectrum of tumour heterogeneity

13.3.1 Spatial heterogeneity

Cancer can ignore growth suppression signals, invade local tissues and metastasize to distant organs. The molecular make-up of cancer cells in different sites can be different, owing to the variable influences of micro-environment related factor. Heterogeneity might exist among the cell present within the parent tumour. The uneven distribution og diverse tumour subpopulations across different sites and within a tumours is termed spatial heterogeneity.

13.3.1.1 Heterogeneity at a single disease site

Primary tumours contain multiple geographically separated and molecularly distinct cellular subpopulations. This can result in an uneven distribution of key molecular alterations across different regions. It might manifest as the ubiquitous presence of key molecular driver alterations, with an unequal distribution of additional molecular alterations. The pattern of spatial heterogeneity observed is reflective of the specific evolutionary context. Multiragion sampling is an informative investigational strategy that improves the ability to determine the extent of spatial heterogeneity within an individual tumour. Many of the unevenly distributed passenger mutations are not expressed. Markers of different impact can be present in geographically distinct regions within the same tumour. Genomic instability is a better biomarker than the alterations detected. A substantial level of genetic diversity exists between individual cancer cells. Multifocal tumours (multiple histologically similar cancers within a single organ) pose a unique challenge because genetic homogeneity cannot be assumed. Moreover the potential exists for divergence.

13.3.1.2 Comparison of spatially distinct disease sites

The genetic makeup of cancer cells at a specific metastatic site might differ from that of the parent tumour. The degree of genetic discordance might reflect whether the metastases occurred as late

events or arose through dissemination early in the course of tumour development. Comparison of the genetic make-up of different metastases reveal substantial levels of heterogeneity. In the simplest scenario, seeding of multiple metastatic sites by identical clones, all metastatic sites would have the same genetic signature. This uni-directional flow might not be a universal scenario: tumour self-seeding and exchange of tumour material between different metastatic sites can occur. Moreover polyclonal seeding can happen. In some cases distant metastases and arise from independent seeding by genetically distinct subclones originating from the primary tumour. Moreover site specific factors could promote genetic divergence after initial colonization.

13.3.2 Temporal heterogeneity

Temporal heterogeneity refers to the dynamic variation in the genetic diversity of a tumour over time. Chemotherapy can alter the molecular make-up of tumours over time by creating shifts in the mutational spectrum. Mutations in genes that are fundamental to replication and cell-cycle regulation can contribute to genomic instability. Targeted therapies can extort selective pressuresoncogene-driven cancer cells.

13.3.2.1 Genomic complexity might increase with exposure to targeted therapies

The efficacy of targeted therapies reflects therapeutic vulnerabilities resulting from a dependence on specific growth signals and the trincar location of the driver alteration. Resistance can arise through mutations, activation of bypass signalling pathways and cell-lineage changes. De novo resistance alterations can be present at low variant allele frequencies in pretreatment tumour specimens. Resistant clones merge from the selective expansion of pre-existing populations during treatment with targeted agents. The genomic complexity increases with exposure to sequencing systemic therapies: the single genetic snapshot depicted in a diagnosti biopsy sample might become outdated during the clinical course. Serial characterization of tumours at multiple time points is necessary in order to accurately capture the various temporal shifts that take place during clonal evolution.

13.3.2.2 Longitudinal sampling provides insight into temporal heterogeneity

Longitudinal profiling has the potential to decipher the role of clonal evolution. Repeat biopsy sampling enables the tailored use of sequential therapies. Clonal evolution that arises from the selective pressures created by targeted agents is dynamic. Clonal dynamics are not always easily manipulated by treatment interruption. Longitudinal sampling might be most clinically relevant when used as a tool to enable the selection of subsequent treatment strategies.

13.3.2.3 Residual drug-tolerant cells can foster temporal heterogeneity

A reliance on biopsy samples might fail to detect cancers at the early or intermediate stages of resistance. The residual disease left to therapies could harbour a small population of quiescent drug-tolerant cells that have survived owing to adaptive activation. Acquired resistance is attributed to selective expansion of pre-existing subclonal population. Data from some studies suggest that the ongoing evolution of drug-tolerant cells leads to de novo generation of resistance alterations. These can emerge from single-cells clones derived from drug-tolerant cells. This hemphasizes the necessity of developing sensitive technologies that enable the early detection of resistance. The emergence of resistance highlights the need to develop therapeutic strategies that target the minimal residual disease state.

13.4 Noninvasive monitoring of heterogeneity

Analysis involving single-site biopsy sampling might result in underestimation of the degree of spatial heterogeneity, and sampling intervals tolerable by the patient might not enable the true extent of temporal heterogeneity to be captured. Liquid biopsies that facilitate longitudinal analysis of tumour-derived genetic material are a promising strategy for addressing the shortcomings of tissue sampling. Genotyping of circulating tumour cells, circulating exosomes and circulating cell-free tumour DNA or ctDNA had promising results.

13.4.1 Analysis of ctDNA

Analysis of ctDNA is a sensitive and highly informative method of identifying clinically relevant genomic alterations with a high degree of concordance with tissue biopsy. ctDNA might enable the identification of alterations not detected by tissue genotyping. Optimizing ctDNA platforms to increase sensitivity for very-low-frequency mutations might enable the early detection of resistance and relapse. This is because the detection of variants associated with treatment resistance in plasma can precede the emergence of evidence of radiographic progression by 10 months in some patients. Longitudinal plasma analysis is an effective tool for gauging the influence of treatment on the molecular and genetic makeup of a patient's cancer over time. Plasma clearance might be predictive of a clinical response. Serial plasma analysis can enable the kintetics of dominant alterations present before treatment to be monitored and to capture clonal shifts occurring during therapy. Several studies found that genotyping of plasma samples enables the kinetics of intratumoural heterogeneity to be captured in a timeframe that is potentially conducive to guiding clinical decision making. Plasma samples contain ctDNA from multiple metastatic sites so it can enable the detection of clinically relevant alterations that are not identified through analysis of tissue biopsy samples. Reliance on tissue sampling alone often underestimates the degree of overlap between distinct driver alterations. Heterogeneity of alterations associated with resistance in plasma samples correlates with shorter PGS durations. Analysis of pretreatment plasma samples can provide some insight into the probable disease outcomes of patients receiving treatment. Sampling of multiple lesions during autopsy can improve upon the ability of tissue sampling to capture the extent of molecular heterogeneity present in cancers. Plasma genotyping might provide a more-comprehensive readout of tumour heterogeneity.

13.5 Overcoming heterogeneity

Higher level of intratumoural heterogeneity predispose patients to inferior responses to anticancer therapies, including to targeted agents. The degree to which subpopulations coexist will affeect clinical outcomes. Cancers become more heterogeneous and complex with successive exposure to systemic agents: responses to subsequent lines of therapy are often not as robust as responses to initial treatments. The current paradigm of sequential treatment is suboptimal: it fails to address the heterogeneity that might underlie incomplete responses to treatment. A bulk solid tumour is a heterogeneous entity that predominantly consists of drug-sensitive cells: mathematical modelling might enable the design of dosing schedules that account for this inherent heterogeneity. Withdrawal of targeted therapy can negate the selective advantage conferred upon drug-resistant celss and enable reporpulation of the tumour with drug-sensitive cells. Intermitting dose scheduling can temporarily suppress clonal outgrowth, although the effect is to subdue heterogeneity rather than to eliminate it. Combination approaches that target heterogeneous tumour populations have proven successful in preclinical studies. Plasticity between different signalling pathways is a potential manifestation

of temporal heterogenity under therapeutic selective pressure. Drug combinations targeting multiple singalling pathways could provide another means of addressing intratumoural heterogeneity. The characteristics of the tumour before treatment could be used to design therapeutic approaches. Drug tolerant cells can develop a wide ragne of resistance mechanisms and targeting this population offers another opportunity to curtail intratumour heterogeneity. Combinations are likely to be more effectiev than monotherapy. Many tumour lack actionable genetic alterations. In these cases strategies that target more ubiquitous sources of heterogeneity are likely to be most applicable. Genomic instability is a pervasive and ideal target. Countering the development of genomic instability is a more daunting task than silencing a dominant signalling pathway. This is likely to be most effective in patients with cancers that are prone to mutagenic stress.

Chapter 14

Unravelling the clonal hierarchy of somatic genomic aberrations

14.1 Introduction

14.1.1 Abstract

Defining the chronology of molecular alterations may identify milestones in carcinogenesis. The analyses highlight the diversity of clonal evolution within and across tumour types that might be informative for risk stratification and patient selection for targeted therapies.

14.1.2 Background

Cancer arises from clones that undergo intense evolutionary selection during disease progression. THis process may lead to subclonal divergence resulting in genetic and molecular heterogeneity. Several methods have been developed to quantify DNA admixture and ploidy from SNP array data that use the relative abundance of specific allele signal (B allele frequency) and the tumour over normal signal ratio or Log R to measure the complexity of the cellular population. Using gerline heterozyfous SNP loci or informative SNPs tumour purity and ploidy are estimated analyzing allelic fraction values. Subclonal alterations will appear as outliers from the computed admixture and ploidy. Global methods are well-suited fro tumour samples with homogenous genomic aberrations. This approaches are suboptimal with tumour samples with high heterogeneity. Local optimization uses creates estimates of purity and ploidy from few clonal events. The AF values of informative SNPs in a somatic deletion result from the composition of signal from non-tumour cells, tumour cells without the deletion and tumour cells harbouring the deletion. Modelling the probability distribution of the observerd AF, a local estimate of the DNA admixture is computed, accounting for both normal cell admixture and subclonal tumour cell population. After the estimation for all deletions across the genome only selected regions contribute to the computation of the tumour sample global admixture.

14.2 Results

14.2.1 Clonality assessment of aberrations from sequencing reads

The reads mapped into a genomic window can be partitioned into a set containing reads that equally represent parental chromosomes and a set containing reads from only one parent chromosome. There are four steps that from neutral read counts, allow inference of clonality of any genomic window. First the percentage of neutral reads within a genomic segment are estimated independently of its Log R value. Then the Log R value is used to relate the neutral reads with a local estimate of DNA admixture. Local estimates are aggregated to estimate global admixture and clonality of somatic copy number aberrations. Aneuploidy genomes are identified and the analysis corrected accordingly. The analysis is then extended to point mutations and structural rearrangements. For each genomic semgent Seg the expected AF of the informative SNP has a bimodal distribution that relates to the composition of the DNA sample. The distance between the two modes is proportional to the percentage of neutral reads β. The expected distribution of the AF varies accordingly with β and N_{ref}, the proportion of reference base reads in the allele represented by active reads. For each input segment Seg, optimization based on swarm intelligence finds a β that minimizes the difference between the expected and the observed AF distribution. Then the Log R of Seg allows computing a local estimate of the admixture. If Seg defines a mono-allelic deletion, β corresponds to the percentage of reads deriving from cells that do not Harbor the deletion and relates to a local estimate of the percentage of admixed cells:

$$Adm.local = \frac{\beta}{2 - \beta}$$

Local admixture values are clustered and the lowest median determines the global admixture of the sample. The more the local admixture value differs from the gloval the more Seg is subclonal. The clonality of Seg or Cl_{Seg} is computed as the percentage of tumour cells in a sample harbouring Seg. If Seg is a gain $Adm.local$ extends by rescaling the percentage of neutral reads β to recover the percentage of reads sequenced from cell that does not harbour the gain of Seg. Bi-allelic deletions are treated separately. If the deletion is clonal its AF has binomial distribution $\beta = 1$ and represents DNA admixture. In case of subclonality β is proportional to the percentage of tumour cells that do not harbour the deletion. Aneuplidy causes a shift in the Log R vs β space. In any segment with an empty active reads set each allele has the same number of copies and $\beta = 1$. The ploidy of a sample is the shift in the Log R values of the neutral segment that best accounts for the observed Log R values. Log R data are corrected for ploidy and Adm.global to achieve better estimates of the segment copy number. Clonality estimates build on the assumption that reads supporting the alternative allele are representative of the amount of tumour DNA harbouring the mutation. The proportion of reads supporting the alternative allele of a pure and clonal hemizygous PM has symmetric binomial distribution. Adm.global represents the percentage of reads from admixed cells that have to be ignored to compute the correct value of AP. A PM is subclonal when its corrected AP has a low probability to be clonal. The same principle applies to REARRs. The total numberof reads that span both sides of a breakpoint defining a REARR is a proxy of the number of cells harbouring the rearrangement. The difference between the expected and observed proportion of reads supporting the alternative allele is proportional to the subclonality of the considered REARR.

14.2.2 Inferring the order of mutations in a tumour sample

The assessment of the clonality of each somatic aberration enables the deconvolution of the sequence of oncogenic events that occur during tumour initiation and progression. Assuming that clonal

alterations pre-dates subclonal alterations within the same tumour, pairs of genes aberrant in the sample sample and across multiple tumours are considered to determine the directionality of the clonal-subclonal hierarchy. To minimize the number of of false posistives (clonal called subclonal) the estimation uncertaninty around β is computed and propagated to clonality values. This enables robust comparison of aberration clonality across different tumour sample data. If a clonal aberration A_1 and a subclonal A_2 occur within the same sample S, A_1 has been acquired before A_2 in S and A_1 precedes A_2 in S. The same dependency has to be found consistently across samples to derive the rule that links A_1 and A_2. This can produce an evolution path draft and in the presence of adequate sample size and frequencies of co-occurring aberrations, the statistical significance of the relation between A_1 and A_2 can be assessed by testing the null hypothesis that the two aberrations are independent and consider a binomial distribution with number of trials n equals the number of samples where A_1 is clonal and A_2 is subclonal or vice versa. With

14.2.3 In silico and in situ experimental validation

To assess if the coverage depth typical for large scale sequencing experiments has an effect on clonality estimates miSeq ultra-deep sequencing data was queried. Excellen agreement in downstram clonality calls for deletion was observed. CLONET did not assign clonality values to aberrations in which MiSeq does not confirm AP values. Next studying PMs and assessing high correlation of AP values between WGS adn MiSeq data, suggesting that the study coverage duse not significantly impact the ability to assess aberration clonality. In order to validate the clinality statuse of complex structural genomic aberrations, in situ tests was used. The ability to assess rearrangemnt clonality was deminstrated focusing on well-characterized REARRs. Perfect agreement was demistrantion. Also subclonal bi-allelic deletion was validated by fluorescence. The prediction highlights a small subclonal bi-allelic deletion within a larger clonal mono-allelic deletion, suggesting that selective evolutionary pressure is acting on the genomic region.

14.2.4 Comparative analysis reveals different mechanisms of tumour deregulation

THe mean number of events classified as clonal or subclonal by means of the proportion test with FDR correction. Deletions are more heterogeneous than gains in prostate and lung cancer, while melanoma had the opposite behaviour. Comparing the proportion of clonal/subclonal losses and gains the prostate and lung samples are statistically indistinguishable. This suggests that temporally distinct mechanisms lead to loss and gain across the three tumour types. Prostate cancer in terms of PMs exhibits more subclonal events than melanoma, suggesting a more central role of PMs in melanoma oncogenesis compared with prostate cancer. Aggregated values reflect only part of the story: great variability in the percentage of clonal events within a single combination of tumour and aberration is observed. Then the distribution along the genome of the variability in the clonality status of aberrations was assessed. Commonality beween the three tumour types in some regions can be observed. Then the capability of clonality analysis to highlight tumour specific mechanism of deregulation was investigated. Considering PTEN deletion, which is involved in many cancer types, it was seen how the timing of the alteration is different and may point to differential roles for pathway inactivation. The focal and subclonal deletion in prostate samples suggests that evolutionary pressure is acting later and may promote cancer progression at a later stage. PTEN is homogenously lost in metastatic melanoma. In lung cancer this loss is more rare. CLONET can identify tumour lineage specific subclonality.

14.2.5 Clonal hierarchy of genomic aberrations

The temporal evolution of driver aberrations was analysed to build evolution maps capitalizing on the information from multiple individuals' samples in the absence of multiregion samples. Given the sample size and the mutation frequencies, drafts of evolution maps were built by implementing the following rule. In particular, an arrow from A_1 to A_2 is drawn if:

- A_1 and A_2 co-occur in at least two samples.
- A_1 preceded A_2 in at least one sample.
- A_2 does not precede A_1 in the considered dataset.

The sensitivity of CLONET allowed the identification of additional genes whose loss precedes the homozygous deletion. No contradictory relations were detected in independent datasets. In order to investigate common patterns of progression across tumour types, a large set of putative cancer genes was interrogated and applied pairwise intersections of identified paths. The evolution of known cancer signalling pathways was explored: both ommon themes across tumour types and tissue-specific patterns emerged. Recurrently deregulated pathways where detected as early drivers. THe timing of dysregulation along the evolutionary paths can be independent across tumour types.

14.3 Materials and methods

14.3.1 CLONET pipeline

SNPs have been extracted from BAM files using an in-hous procedure, SCNAs were detected using SegSeq from tumour and normal sequencing-based data, PM coordinates were as in original corresponding manuscripts and REARRs were identified by means of dRanger and Breakpointer. To avoid germline background effects, genes that intersect significant with known germline copy number variants were filtered out.

14.3.2 CLONET on exome and targeted sequencing data

The analysis of samples with few SCNAs provided that informative SNPs read counts and Log R values are available is enabled. Individual specific informative SNPs can be identified from matched normal DNA samples. Appropriate Log R values can be optained for exome genomi segments with platform specific strategies and provided to CLONET as input. Array-based segemnted data or SCNA segments directly inferred from eome data with recent well-performing tools. CLONET combines segment input with exome-derived read counts to estimate purity and ploidy. Then subclonal aberrations are called based on sequencing data. Copy number calls derived using custom control regions and very hig-coverage allowed for CLONET based clonality estimation even in the case of low tumour content.

14.3.3 Expected distribution of the allelic fraction of a genomic segment

Consider a genomic segment that spans a set of informative SNPs for the individual of intersect. For any of them with coverage *cov* the total number of reads r supporting the reference base is the sum of the neutral reads r_n and the active reads r_a supporting the reference base. β is the ration between neutral reads and the total number of reads spanning the SNP of interest. The probability of having k reference reads is the convolution of the probability of observing $\beta \cdot k$ neutral reads and $(1-\beta) \cdot k$ active reads:

$$P(r = k, 0 \leq k \leq cov) = Conv(P(r_n = \beta \cdot k), P(r_a = (1 - \beta) \cdot k))$$

$P(r_n = \beta \cdot k)$ is assumed to follow a binomial distribution with trials $\beta \cdot cov$ and probability of success ps. All active reads support the reference or the alternative base. N_{ref} is the proportion of informative SNPs within the aberration that carry the SNP reference base in the allele represented by the active reads. $P(r_a = (1 - \beta) \cdot k)$ follows a categorical distribution iwth values equal to N_{ref}.

$$P(r = k|cov, \beta, N_{ref}, ps) = (1 - N_{ref}) \cdot B(k|\beta \cdot cov, ps) + N_{ref} \cdot B(k - (1 - \beta) \cdot cov|\beta \cdot cov, ps)$$

Where $B(m|n, p)$ is the probability mass function of a binomial distribution, the probability of m successes in n trials with success probability P.

14.3.4 Estimated proportion of neutral reads for a genomic segment

β and N_{ref} can be inferred from the sequencing coverage at informative SNPs within the segment. Given a segment Seg and a set I of informative SNPs in Seg, each SNP in I is a sample from the distribution described earlier. Optimization can allow for the identification of values β and N_{ref} for each segment using the Kolmogorov-Smirnov for the likelihood that I are a sample of the distribution and a particle swarm optimization finds a candidate $\hat{\beta}$ and $\hat{N}_{ref}$ that best represents the distribution of the allelic fraction of the SNPs in I.

14.3.5 From neutral to non-aberrant reads

Consider a Seg if the $Log\ R$ value of Seg support a SCNA C, reads that cover Seg from cells harbouring C are considered aberrant. If Seg is a candidate mono-allelic deletion β corresponds to the percentage of reads that cover Seg and are sequenced from cells harbouring both alleles. If the Log R value supports a gain with $cn > 2$, β has to be rescaled to obtain the percentage of sequenced cells that have copy number cn. If cn is odd, the number of neutral reads is the sum of the neutral from admixed plus the neutral of the gain. β_{cn} of reads from cells with cn is computed from β by removing neutral reads due to the gain:

$$\beta_{cn} = 1 - cn_G \cdot (1 - \beta)$$

If cn is even and one copy difference between allele is allowed β is closed to one.

14.3.6 From aberrant reads to aberrant cells

Given a somatic mono-allelic deletion M the local admixture $Adm.local$ is the proportion of cells not harbouring M over the total number of cells. Let a define the total number of reads supporting the alternative allele,, as the sum of neutral a_n and active a_a reads. For any informative SNP within M, the local admixture is:

$$Adm.local_M = \frac{\frac{r_n + a_n}{2}}{\frac{r_n + a_n}{2} + (r_a + a_a)}$$

The proportion of non-aberrant reads covering M is:

$$\beta_M = \frac{r_n + a_n}{r_n + a_n + r_a + a_a}$$

14.3.7 Uncertainty assessment and its propagation to clonality estimates

To optimize sensitivity and specificity the estimation uncertainty ϵ around β is computed. The value of ϵ varies upon the mean coverage and the number of informative SNPs. The mean coverage controls the ability to discern the two modes of the AF distribution. Higher β requires higher coverage. The procedure to infer the value of β is independent from its Log R value. Segments aggregate into cluster corresponding to copy number and define a clonality status. Restricting to putative somatic mono-allelic deletions, B_{min} with the lowest median of β would represent 100% clonal deletions. B is the set of β values of all the putative somatic mono-allelic deletions. B_{min} is the smallest subset of B such that $min(B)$ in B_{min} and for all β' in B and not not in B_{min}, $max(B_{min})+err(max(B_{min})) < \beta'-err(\beta')$. The median value is selected as candidate $Adm.global$. Given a somatic copy number C in a sample, the local and global admixtures are computed. The clonality Cl_C of C is the percentage of timor cells in a sample harbouring C:

$$Cl_C = \frac{1 - Adm.local_C}{1 - Adm.global}$$

The more the value local differ from the global the more C is subclonal.

14.3.8 Clonality of bi-allelic deletion

For subclonal bi-allelic deletion the allelic fraction signal comes from cells with two or one allele. Consider a subclonal bi-allelic deletion where n, m and b denote the proportion of cells with two, one and zero alleles. The local estimate of the admixture can be computed. This is the proportion of cells with two alleles in the subpopulation of cells with one or two alleles, $n = Adm.local(n + m)$. The proportion of normal cells in the sum is equal to the global DNA admixture. The clonality of a bi-allelic deletion Cl_B is the percentage of cells harbouring the bi-allelic deletion over the number of cells with a mono- or a bi-allelic deletion $\frac{b}{m+b}$.

$$Cl_B = \frac{Adm.global - Adm.local \cdot Adm.global}{Adm.local \cdot (1 - Adm.global)}$$

Chapter 15

TPES: timor purity estimation from SNVs

15.1 Abstract

Tumour purity is the proportion of cancer cells in a tumour sample. It impacts on the accurate assessment of molecular and genomics features.

15.1.1 Introduction

Genomic and molecular analysis of tumour samples require the quantification of tumour and admixed normal cells proportion. In ored to asses the somatic lesion detection boundaries and to perform comparative analyses several tools were built to quantify TP from NGS data. The approaches based on SCNA fall short for samples with quiet genomes. To solve this purity can be estimated through the distribution of variant allelic fractions within copy number neutral tumour segments.

15.2 Materials and methods

The VAF distribution of a set of clonal monoallelic SNVs from a pure tumour sample should be centred in 0.5. Technical and cancer specific factors may influence the VAF value as reference mapping bias. Moreover in the case of subclonal events the VAF is altered. Clonal monoallelic SNVs in a diploid segment are suited for TP estimation and are named p-SNV. Given a set of p-SNVs, TP could be computed as:

$$\frac{observerd\ VAF(pSNP)}{expected\ VAF}$$

Where *ovserved* VAF is computed from the tumour data while *expected* VAF is the value expected from a pure tumour sample accounting for reference mapping bias. p-SNVs are selected with a conservative procedure. To minimize the number of false positive p-SNVs for each sample, TPES introduces two main filtering steps. In the first SNVs are selected from copy-number neutral segments applying a conservative filter on the Log R value of each genomic segment. Moreover the log R is adjusted for ploidy and SNV are retained only with a number of reads mapping the alterantive base and AG above and below threshold. Chromosome X and Y are excluded to avoid

gender stratification. This nominates a set of heterzygous copu-number neutral SNPv. The second filter TPES removes putative subclonal mutations. Observed VAF distribution is smoothed by kernel density estimation. Local maxima of the underlying distribution can be observed. The peak with the highest VAF value is the candidate observed VAF.

Chapter 16

SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines

16.1 Abstract

Experiments reported in the scientific literature may contain pre-analytic errors due to inaccurate identities of the cell lines employed. To address this a simple approach to enable accurate determination of cell line identify has been developed by genotyping SNP.

16.1.1 Introduction

Cell lines are important in the identification of therapeutic targets and in understanding molecular pathways related to drug-tumour interactions. One recognized risk in cell line maintenance is human error, either by mislabelling or cross-contamination. In an effort to identify latent cross-contamination or other errors SNPs are considered. SNPs as DNA markers have been shown to be well suited for different purposes such as animal identification, identification of population ancestry and for forensic purposes. The ability of high-density oligonucleotide arrays to genotype hundreds of thousands of SNP loci in parallel provides a molecular fingerprint of each sample. To this end an assay that emplys 30 to 50 single loci and capable of distinguishing any two DNA sample has been developed. This assay can identify a given sample comparing its genotype with a reference dataset.

16.2 Material and methods

16.2.1 Genotype distance

To evaluate the similarity of two DNA sample the similarity measure D is introduced. D is proportional to the number of genotype mismatches between the samples and is normalized to the number of genotype calls available for both samples. Given a set of N_{SNPs} of individual SNPs, $CL1$ and $CL2$ are ordered sets of genotype calls of two samples and $vN_{SNPs} - Card(T)$, where $T = \{i : cl1_i \neq NoCall \cap cl2_i \neq NoCall\}$. For $vN_{SNPs} > 0$, D is defined as:

$$D(CL1, CL2) = \frac{1}{vN_{SNPs}} \sum_{i=1,\ldots,N_{SNPs}} d(cl1_i, cl2_i)$$

Where:

$$d(cl1_i, cl2_i) = \begin{cases} 1 & if\ cl1_i \neq cl2_i \\ 0 & if\ cl1_i = cl2_i \vee cl_i = NoCall \end{cases}$$

The distance is normalized over the number of available calls. Moreover the algorithm evaluates:

- The count of mismatches where the two samples are homozygous for different alleles.
- The count of mismatches where one is homozygous and the other is heterozygous.
- The count of homozygous matches and the count of heterozygous matches.

For each mismatch the algorithms reports the identifier of the sample with largest number of heterozygous calls. The implementation of the distance can be modified to weight different types of mismatch.

16.2.2 SNP panel selection procedure

Using a small number of SNPs samples can still be accurately distinguished. Initial filters on the selection of SNPs where:

- SNPS with the rs identifier.
- SNPs not in intronic regions.
- SNPS represented on the 10K Affymetrix oligonucleotide array.

On the training set the minor allele frequency, the heterozygosity rate and the call rate for each SNP across all sample have been computed. Then SNP satisfying the Hardy-Weinberd equilibrium applying elastic boundaries and having SNP call rates than 80% have been filtered. Then on the test set, the heterzygosity rate of the identified SNAP has been computed. At each iteration a variable number of SNPs has been computed. SNPs have been ranked according to the selection rate.

16.2.3 SPIA probabilistic test on cell line genotype distance

A double probabilistic test to apply on the genotype distance is applied to discern when two cell lines are close enough to be called similar. The test score depends on the number of matches and on the total number of SNPs evaluated. The test relies on the probability of the evaluated distance belonging to the population of real matched pairs or to the population of real non-pairs. If the output is not clear a second panel of SNPs would need to be investigated. If the SNPs are independent and the genotype call probability being the same at each SNP, the probability of habing k matches out of N SNPs follows the binomial distribution:

$$P_k = \binom{N}{k} P^k Q^{N-k} = \frac{N!}{k!(N-k)!} P^k Q^{N=k}$$

Where P and Q are the probability of match and mismatch and N is vN_{SNPs}. By knowing the probability of match at a single SNP for a real matched pair P_M and for a non-matched pair

P_{non-M} the distribution of real matched pair and non-pair can be drawn. For a given vN_{SNPs} then areas corresponding to "different", "uncertain" and "similar" can be defined. The area limits depend on the level of confidence needed. The mean number of successes k_{mean} is equal to NP_M and the standard deviation $sd_{kmean} = \sqrt{NP_M(1-P_M)}$. The probability that a distance measurement falls within M standard deviations from the mean is given by the integral of the distribution function. By setting m the area limits can be defined. The smaller the number of SNPs the narrower the region of uncertainty and the higher the probability of making an incorrect call.