Computational Human genomics

Maurizio Gilioli

May 28, 2022

Contents

Chapter 1

Introduction

1.1

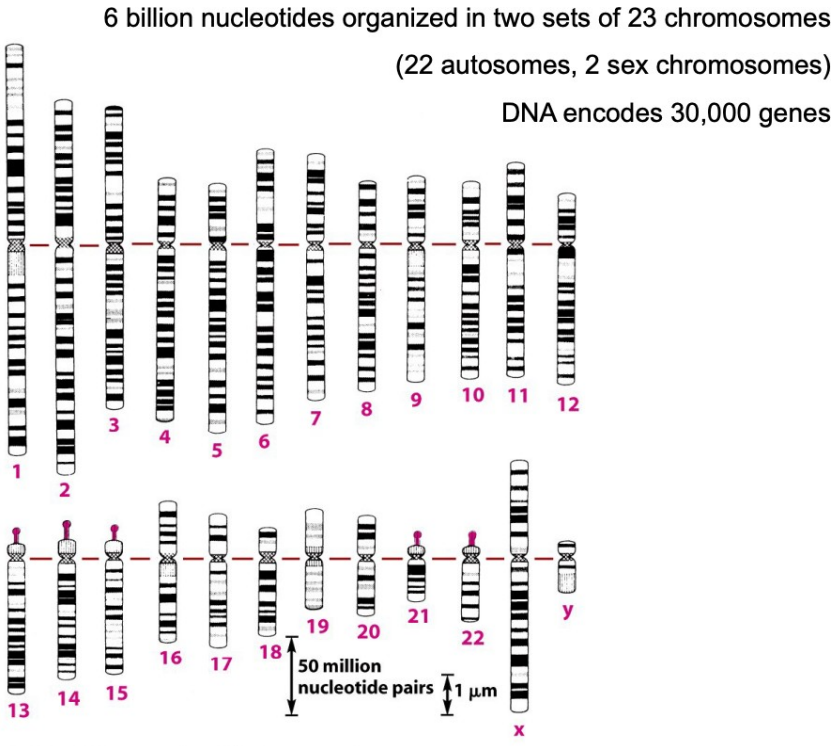


Figure 1.1

The words variations, aberrations and lesions are often interchanged. Aberrations and lesions are mainly used for acquired lesions, instead variations are mainly used for the inherited ones.

1.1.1 Genetic Make-Up

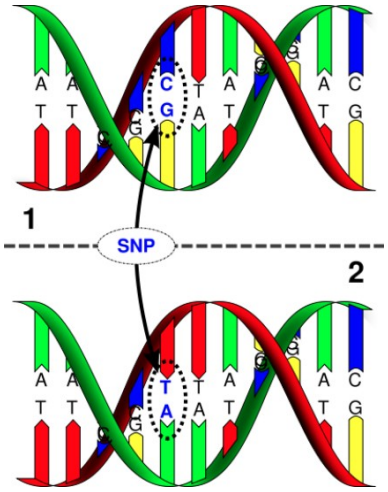
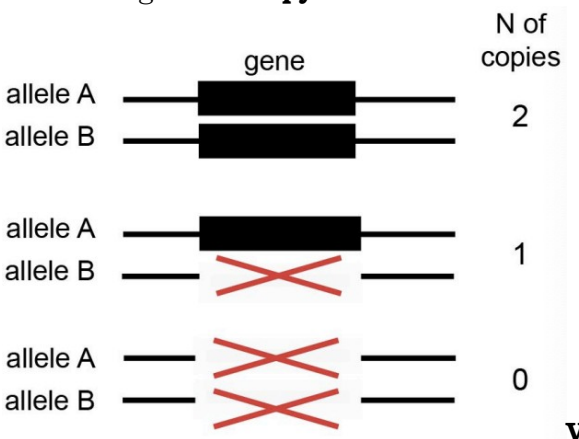


Figure 1.2

Single Nucleotide Polymorphism (SNP) is a sequence variation affecting single amino acid → point mutation (figure 1.2).

The genomes of two unrelated individuals have about 1% of different bases → that percentage corresponds to the SNPs.

But looking at the **Copy Number**



Variants (CNV), that difference will be way higher than only 1%

DNA not present in only two copies, but in multiple, single or even zero copies (hemizygous loss, homozygous loss)

They are less known as inherited type of variants because they are harder to detect and identify, but they provide a lot of uniqueness in each of us.

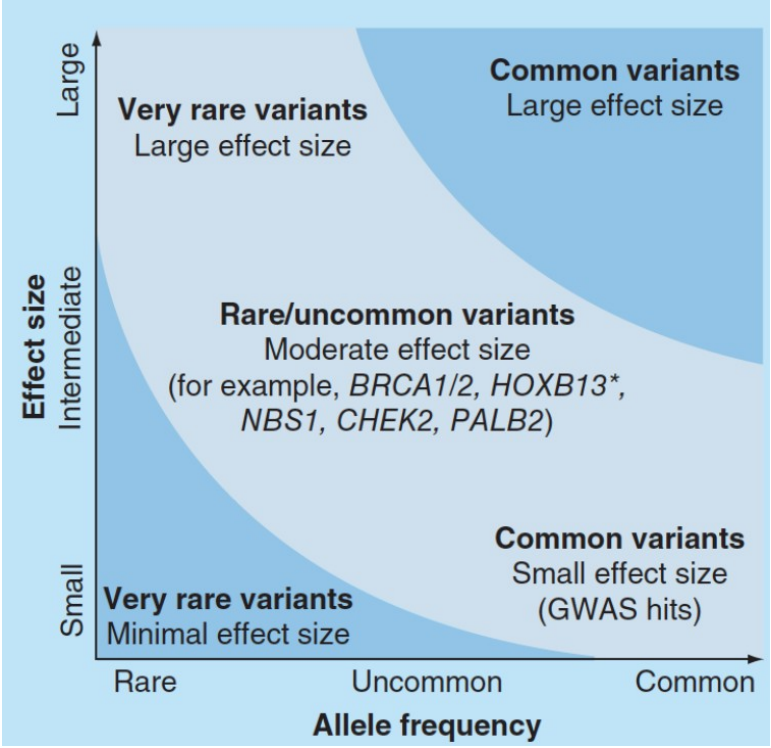
Why are SNPs and CNVs so important?

They are responsible of human diversity → genetic changes

Hundreds of CNVs per individual and 20% of them potentially affect protein- coding genes

Differences in Genetic Make-up:

Very common variants are variants that are distributed in the population as the common allele, so that 1/2 of the population has an heterozygous genotype at that position, 1/4 has an homozygous genotype for one allele and 1/4 has an homozygous genotype for the other allele.



The **penetrance** is the proportion of individuals carrying an allele (or a genotype) that also expresses the trait (phenotype) associated with it. Obviously, penetrance is directly associated with the size of the effect produced by the variant.

The **allele frequency** is calculated by dividing the number of times the allele of interest is observed in a population by the total number of copies of all the alleles at that particular genetic locus in the population.

The allele frequency is low with very rare variants

Well known variants: BRCA1/2, HOXB13, NBS1, PALB2, CHEK2 → they have moderate size effects, meaning that all the people who have the variants, have the disease

1.1.1.1 Differences in genetic Make-Up, example

Absorption, distribution, metabolism and elimination (ADME) genetic variants determine pharmacokinetic variability of certain compounds, influencing the patients' 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.

Three main ways to study genetic variants: GWAS (genome-wide association study) SNPs panel Candidate SNP

For example, in terms of hypothesis, if I study all the variants in the human genome and I query them in a large population, I generate data without specifying SNP to search. Instead, if I have a very specific hypothesis, for example I want to query if a SNPs in the CYP gene relates to the conversion of androgens to estrogens, I don't need to run an GWAS or a wide SNPs panel. I can query those SNPs because I have an *a priori* hypothesis and I want to test them.

This type of differential design for an experiment it is not only true for inherited variants and ADME genes, but also to predisposition to diseases and to study human tumors.

Precision medicine → treatment (or dosage) of a patient based on their individual traits: takes into consideration genetic and genomic of the individual and tumor/disease cells

Starch rich diet → CNV in the genome Drink beer and turn red → ADME gene

Athletes with a deletion of a gene, the steroids were not found in the anti-doping tests

1.1.2 Acquired DNA aberrations

Somatic variants are the variants NOT inherited from parents and not transmitted to offspring. They are:

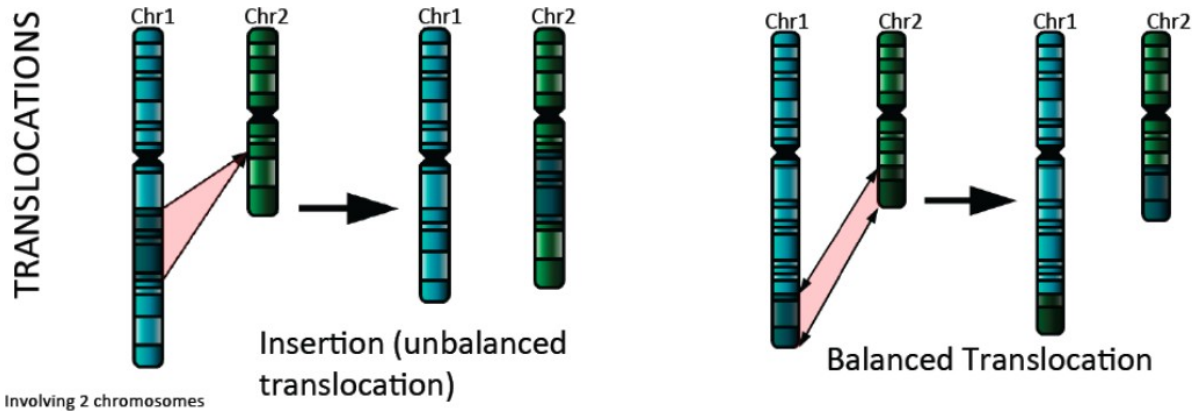
Single Nucleotide Variants (SNV) are somatic changes of single nucleotides present in only certain cells, instead of SNPs that are present in all cells of our body.

Indels are changes that involve few nucleotides by INsertion and DEletion

Rearrangements are mutations that can involve events like translocations, inversion, chromothripsis,.. usually these events are caused by breakage in the DNA double helices a two different locations, followed by a rejoining of the broken ends to produce a new chromosomal arrangement of genes, different from the beginning

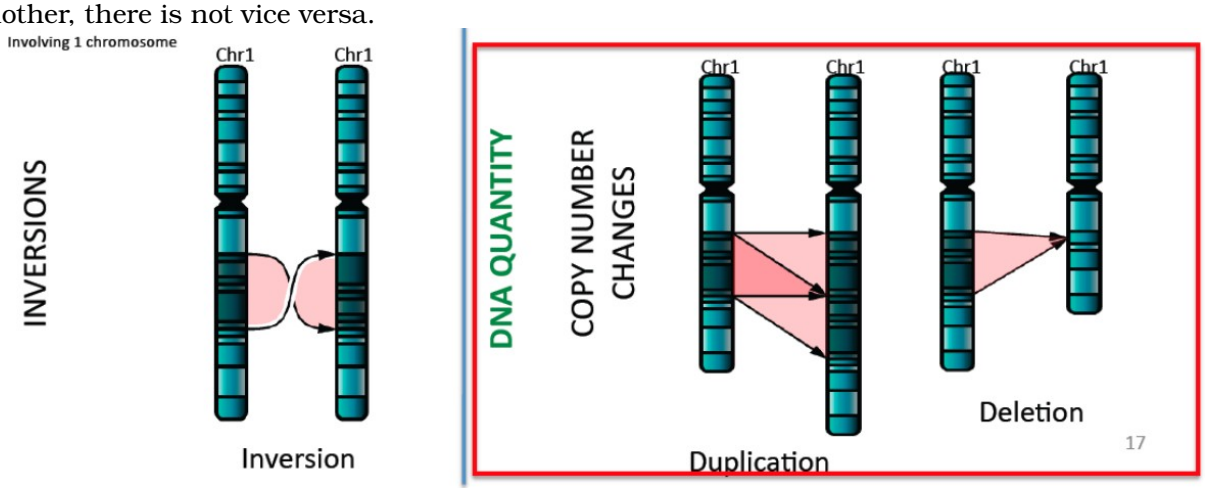
Somatic copy number aberrations (**SCNA**) are somatic changes similar to CNVs. They can be every change related to the number of copies like loss of a portion of a genome, loss of both alleles, extra copies...

Examples of acquired DNA aberrations:



Balanced translocation: you conserve the quantity of DNA, there isn't any loss or gain

Unbalanced translocation: A genomic portion is translocated from a chromosome to another, there is not vice versa.



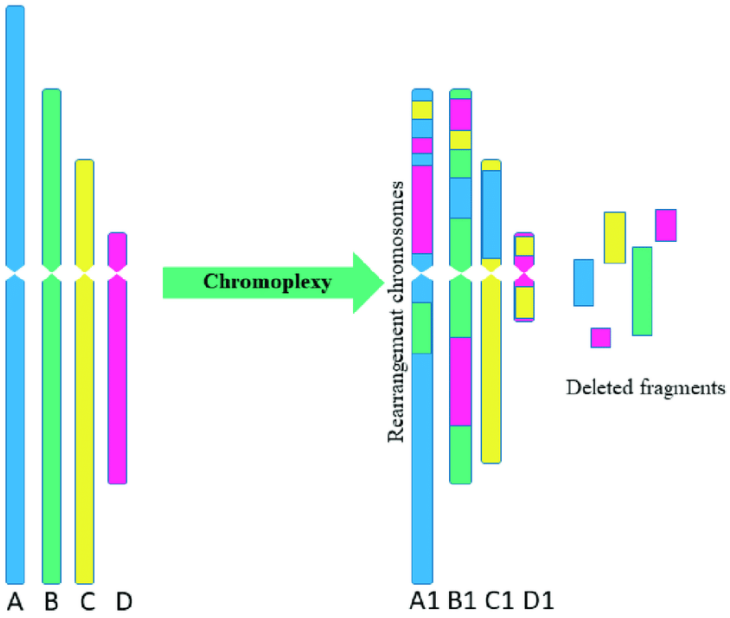
Inversions in only ONE chromosome: everything is normal instead in the break points

Copy number changes: duplication or deletion

It could happen in the same chromosome but also in different chromosomes

Other types of complex somatic events include:

Chromoplexy: 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: a clustered chromosomal rearrangement in confined genomic regions that results from a single catastrophic event, usually limited to one chromosome

Kataegis: a phenomenon that is characterized by large cluster of mutations (hypermutation) in the genome of cancer cells. An APOBEC family enzyme might be responsible fo the kataegis process

when an aberration (clonal) occurs, all the cells will harbour the aberration and at some point another aberration (subclonal of the other) could appear in just one cell line. The **clonal** aberration is present in all the cells, the **subclonal** aberration is inherited in just one cell line. Clonality is an important information that allow us to study evolution.

1.2 Experimental approaches

Experimental techniques to detect variants/aberrations **prior to NGS**: a failure because it was very hard to determine the starting points of the aberrations.

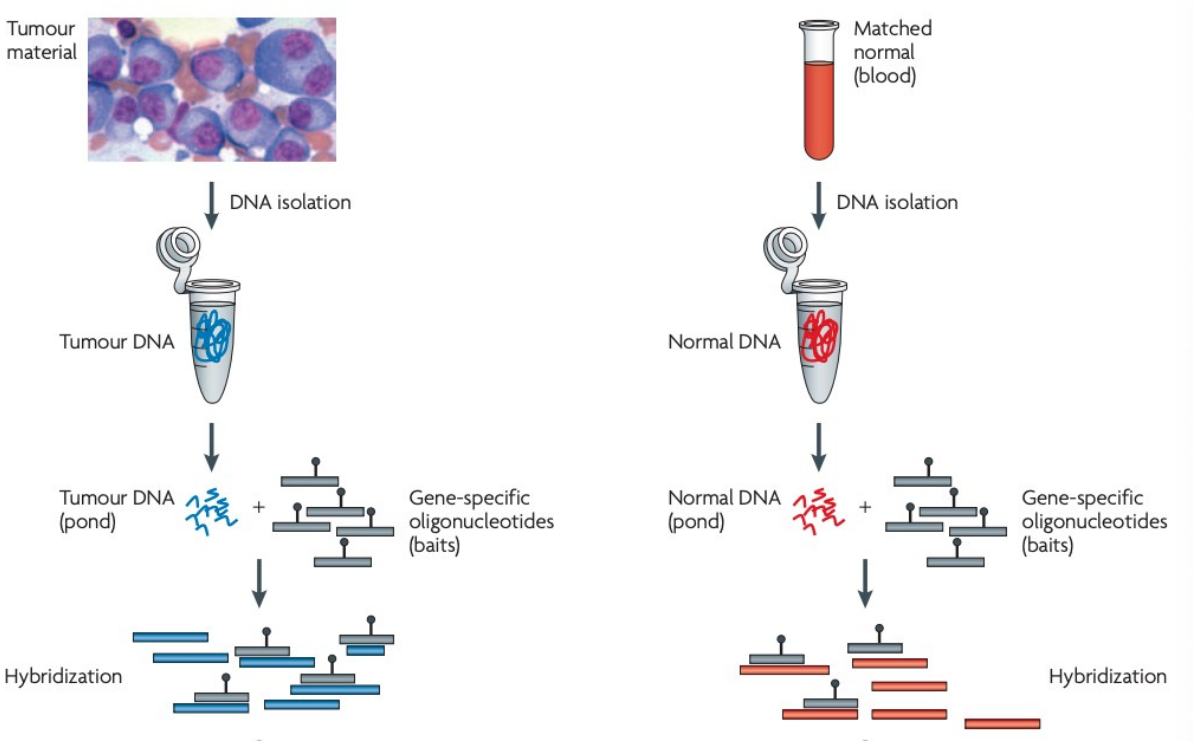


Figure 1.3: Meyerson et al. 2010, "Advances in Understanding Cancer Genomes through Second-Generation Sequencing.", Nature Reviews Genetics, https://doi.org/10.1038/nrg2841

Bulk of tumor tissue/cells from the blood procedure (figure **??**):

1) DNA isolation
2) Gene-specific oligonucleotides (**baits**) that get hybridized onto the tumor DNA → the baits have a tag that allows them to be isolated
3) The DNA does get fragmented
4) The captured DNA is eluted and prepared into sequencing libraries
5) Sequencing
6) Aligned to the bait sequences

We repeat the procedure for healthy cells of the same individual in order to **detect somatic mutations**.

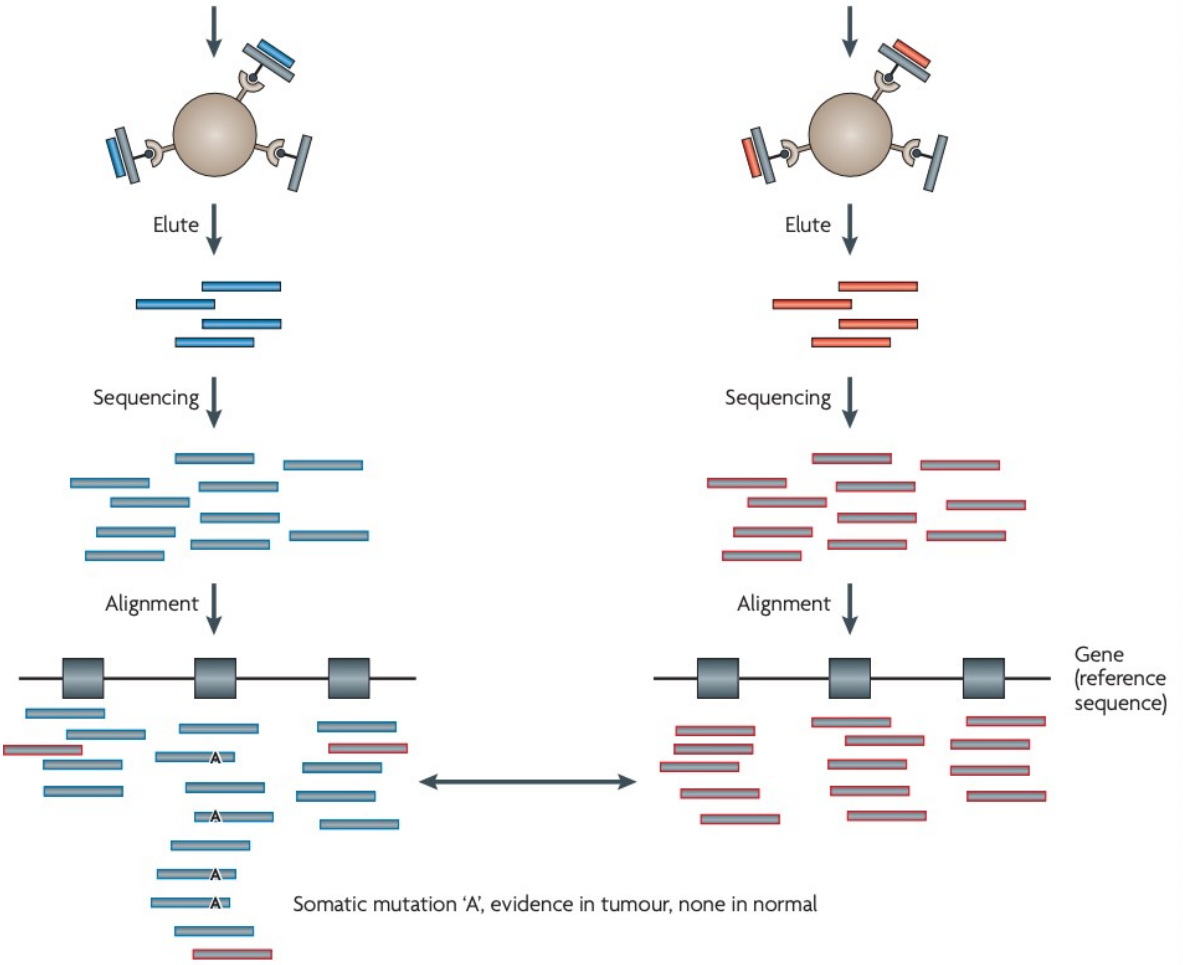
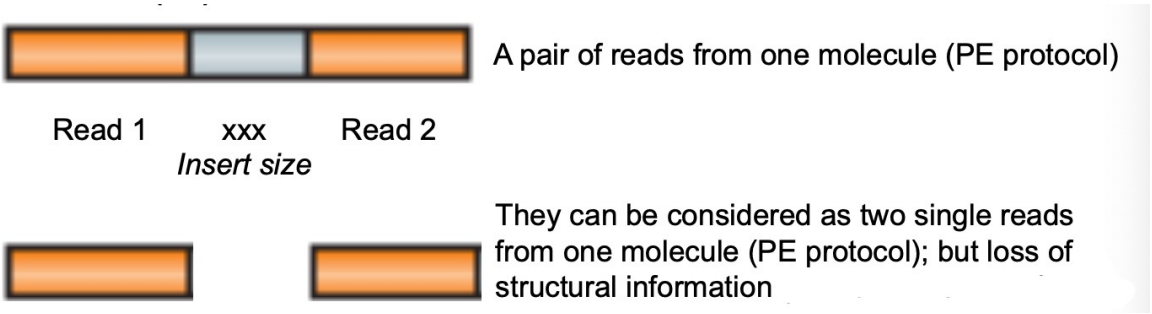


Figure 1.4: Beads capture

We sequence baits because is way cheaper (exons of 50 bases instead the whole genome)

After fragmentation procedure, before adding the adapters, we can choose between two different sequencing approaches 1.2:

One read from one molecule (SE protocol)



- ***Paired End (PE) sequencing***

  You will sequence only one part of a molecule (length of 150 bp → based on the power of the sequencing machine we are using). You will know exactly 150 bp for every molecule you sequence, but you lose information (the second end of the pair).

- ***Single End (SE) sequencing***

  You information about the length of the DNA portion between the ends. It's more expensive, but:

  - it gives information about the localization of the molecule
  - you can treat each end as single read

1.2.1 Information after reads mapping over reference genome

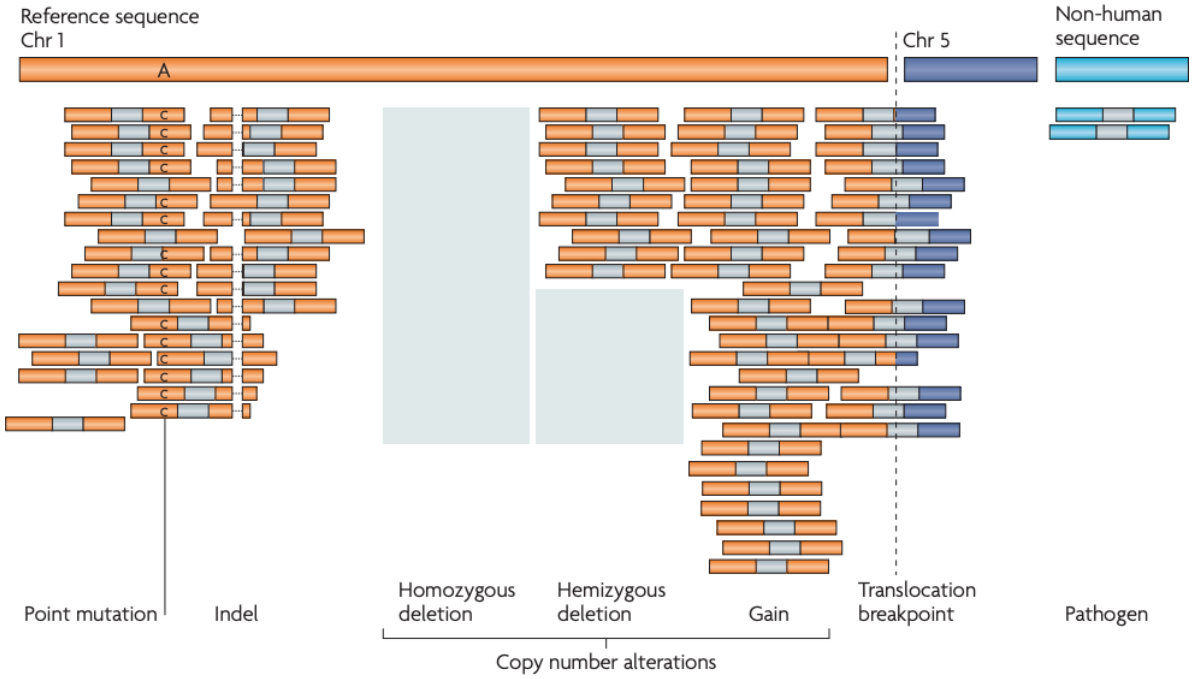


Figure 1.5: In the following picture: a view of reads that are mapped against reference genome and what we would look if we have any of the variations that we mentioned

Following the mapping of the reads over the reference genome, different types of genomic alteration/information can be detected:

- You can clearly identify **point mutations**. If a point mutation is present in the molecule that you sequenced and present on both alleles of the genome, it can be seen in all the reads very clearly.
- You might see **indels** (shown here by a dashed line). You will see a little space because the reference genome has more nucleotides than the sequenced molecule.
- If you have **homozygous deletion**, you don't see anything mapped in that portion: there's no DNA. Doesn't matter if SE or PE.
- If you have **hemizygous deletion**, you see the see the read mapped to that portion where the hemizygous deletion is sitting, that is more or less proportional to half of the reads that you have in regions where you don't have a copy number change. Doesn't matter if SE or PE.
- If you have **gain**, what you get is higher number of reads aligned against that part of DNA, underlying the fact that the molecule you sequenced has extra DNA for that portion of reference genome. Doesn't matter if SE or PE.

- **Translocation breakpoint** are very important!! You will have one end mapping the chr1 and the other end mapping the chr5. Those two ends come from the same molecule of the *target cell* (the cell we sequenced), it means the cell has a translocation between chr1 and chr5. Without the PE protocol you cannot have this result.

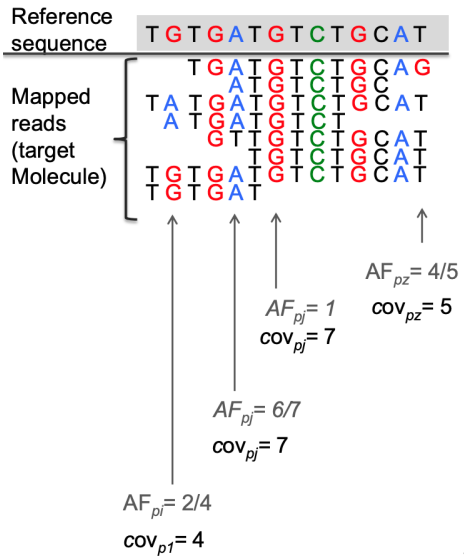


Figure 1.6: View of sequence alignment

The **local coverage** (cov) as shown in figure 1.6 at position i is the number of reads that span p_i.

The **allelic fraction** (AF) as shown in figure 1.6 at position i is the proportion of reads that supports the reference base in p_i (= the reference or the alternative allele).

1.2.2 Whole Genome Sequencing Coverage

$$cov = \frac{L\dot{N}}{G} \tag{1.1}$$

where:

- **L** is the read length.
- **N** is the number of mapped reads.
- **G** is the haploid human genome length.

This is super important because it saves us time and money when we design an experiment. When you design an NGS experiment, you should know before what is the type of coverage you need to answer the question you wanna ask with your experiment. For example, if you want to look at the genotype of SNPs (inherited polymorphisms at single side), you don't really need a coverage which is above 10 or 15. So you can design your experiment in order to have an average coverage equal to 10 or 15. To do that, you reverse

the equation and count how many reads you need to generate to achieve that goal.

N.B.: The number of mapped reads will be always lower of the number of reads generated by the machine (than the expected). There might be duplicates that you might not be able to use because there might be reads that have a quality below the threshold you intend to use.

1.2.2.1 Difference between sequence coverage and physical coverage

A graphic view of how SE (Single End Sequencing) or PE (Paired End Sequencing) can be used:

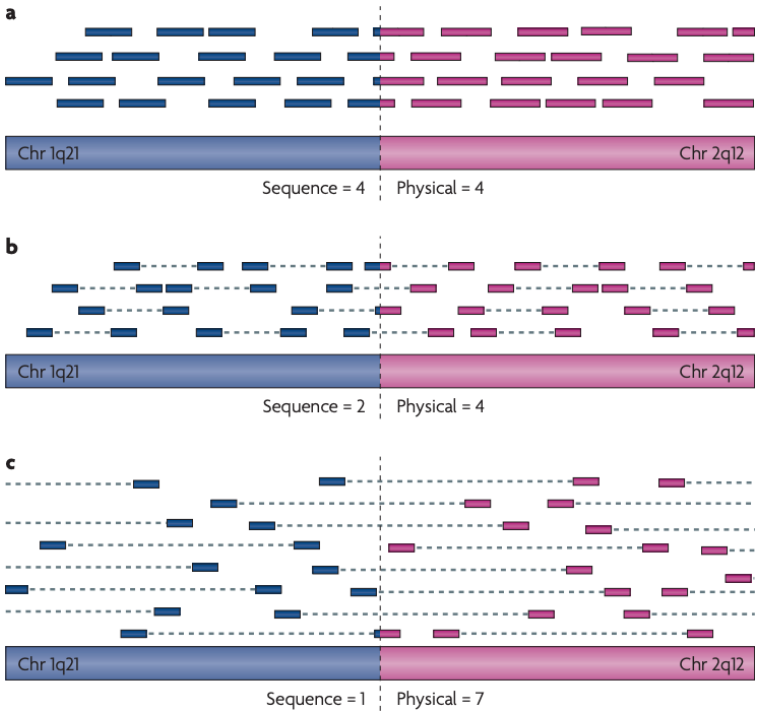


Figure 1.7: Panel A - SE protocol; Panel B - PE protocol; Panel C - PE protocol

Three different scenario are depicted that vary in the length of the DNA fragments that are sequenced. **Sequence coverage** represents the number of sequenced reads that cover the site; this affects the ability to detect point mutations. **Physical coverage** measures the number of fragments that span the site; this affects the ability to detect the rearrangements, based on paired reads that map to different chromosomes. It is a way informative type of coverage: for instance for translocations, deletions ...

In Paired End sequencing protocols, the physical coverage is always higher than the sequence coverage. Choosing the method illustrated in panel 3 (figure 1.7).

Making estimation of intended coverage and observed coverage is very important. Below I will report an example:

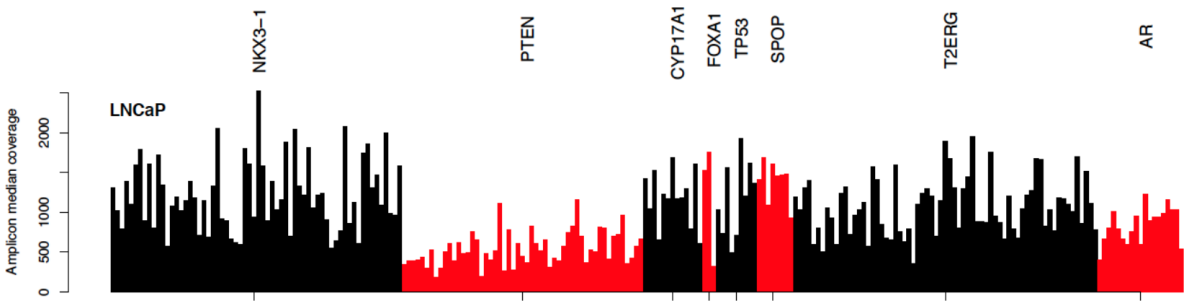
1.2.2.2 Example coverage observation

In these panels were designed to sequence a set of 10 genes that the researchers were interested in for prostate cancer. They designed this panel, sequenced cell lines on this panel and observed the following points

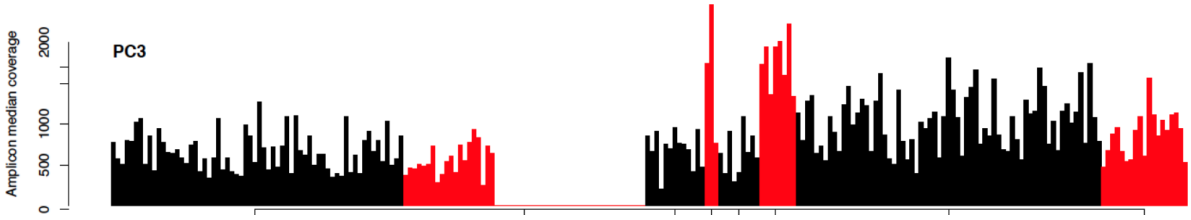
- On x-axis: the genomic location
- On y-axis: the local coverage (amplicon median coverage = each bar represents the local coverage of about 30 bp)

The different colors represent the different genes

- 1^{st} **panel**: Local coverage (pile-up) of selected areas (targeted sequencing assay): 7 genes
  - + 1 multi-gene region (T2ERG). Alternate colors indicate targeted areas The barplot show a single sample (LnCaP cell line; cancer cell line) data.
  - Apparent **deletion** of PTEN (monoallelic deletion) because the local coverage of PTEN is significantly lower than the one from other genes.

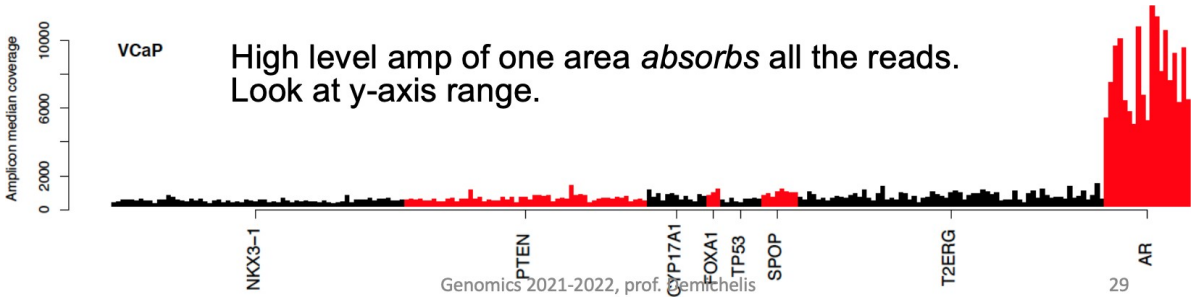


- 2^{nd} **panel**: Local coverage (pile-up) of selected areas (targeted sequencing assay): 7 genes
  - + 1 multi-gene region.
  - Monoallelic deletion and partial biallelic deletion of PTEN because one portion is deleted and one not. PTEN has a **partial homozygous deletion**.
  - The PC3 cell line shows a little bit of gain in the gene SPOP and FOXA1.
  - The average coverage for the PC3 cells is approximately the same as the previous sample.



- 3**rd panel**:

- There's no homozygous deletion but has a high level amplification of one area *absorbs* all the reads. Massive amplification of the Androgen Receptor (AR) → error: because it inhibits the sensitivity of detecting copy number changes in any other gene.
- When designing a panel we must pay attention and make sure that we don't have potential aberration that basically will draw all the attention of your experiment and leave you without information or sensitivity in all other regions.
- It's easy to increase the experimental coverage (i.e. the sequence depth) at later point. Provided tour original sample/library is still available, you can perform another run of sequencing and then combine the output from different runs



1.2.2.3 Note that this isn't possible with array-based technologies.

What are the limiting factors of NGS DNA-seq experiment, in any?
Repeated regions due to **short reads**
What is the problem of short sequencing on long genome?
Complexity regions CG content

1.3 The reference sequence of the human genome

Many years ago, some people claimed that the entire human genome was sequenced but it wasn't true at all. There were still unknown or missing regions. In 2022 we finally have the complete human reference genome sequence.

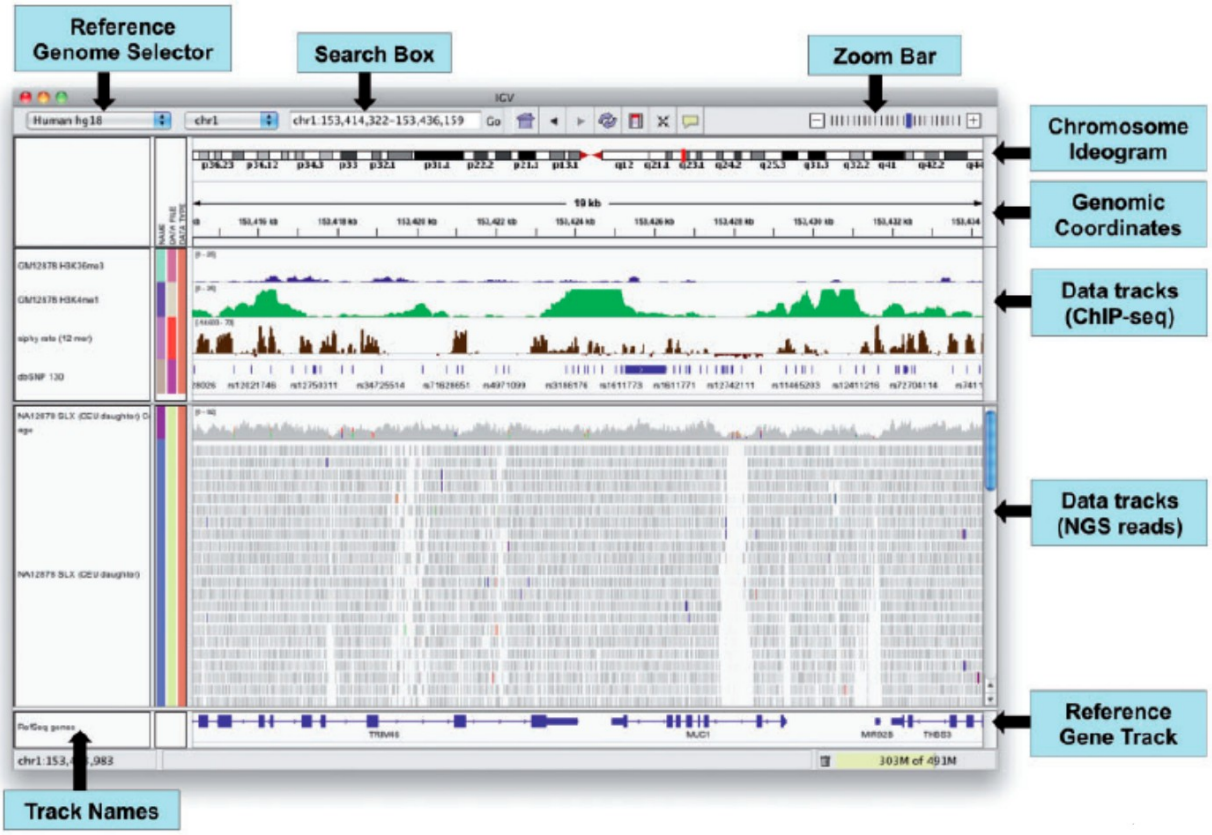
But we need to consider the polymorphisms, there is no **unique** genome. How to integrate them into a single reference genome? There is a consortium that deals with these problems. They assemble a reference genome that reflects the more common (in the whole population) sequences at each position of the human genome, but also tracks information of everything that is polymorphic. So that we can use the latest release of what they built as reference genome and then use databases to learn about all the polymorphic sites and all the features of every polymorphic variants

Genome Reference Consortium: where you can find different versions of human reference genome

UCSC Genome Browser on Human:
where you can upload different versions of the reference

1.3.1 Interpreting pair orientation

Using IGV (Integrative Genomics Viewer)



The main characteristic of IGV is that it is a main view viewer: all the information are in one window

Every vertical bar is a read

On the x-axis there's the genome coordinates at the top, the reference genome at the bottom (we can select the reference genome we prefer)

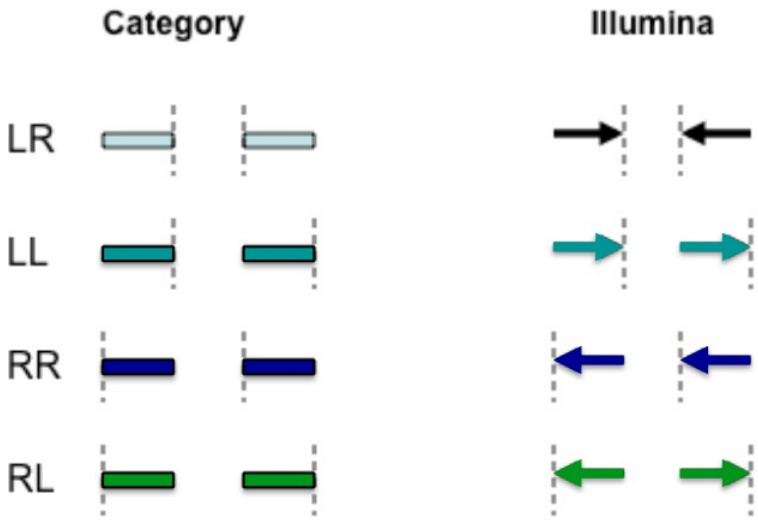
Along with the data tracks there is the local coverage of the kb shown in the window (of the sample we are looking at)

You can get any information you want of any single read that you are uploading, very useful to see difference from the reference genome because every aberration or whatsoever is highlighted by a different color in the local coverage of a nucleotide base. Moreover, it gives information about the quality of the read and the bases, if you have a PE protocol, it tells you also information about the PE for each of them

The **orientation** of paired end can be used to detect structural events, including: inversions

duplications translocations

According to the Illumina protocol, the two ends are LR oriented, but we could also obtain other orientations, like LL RR RL, if we come up against the events mentioned above.

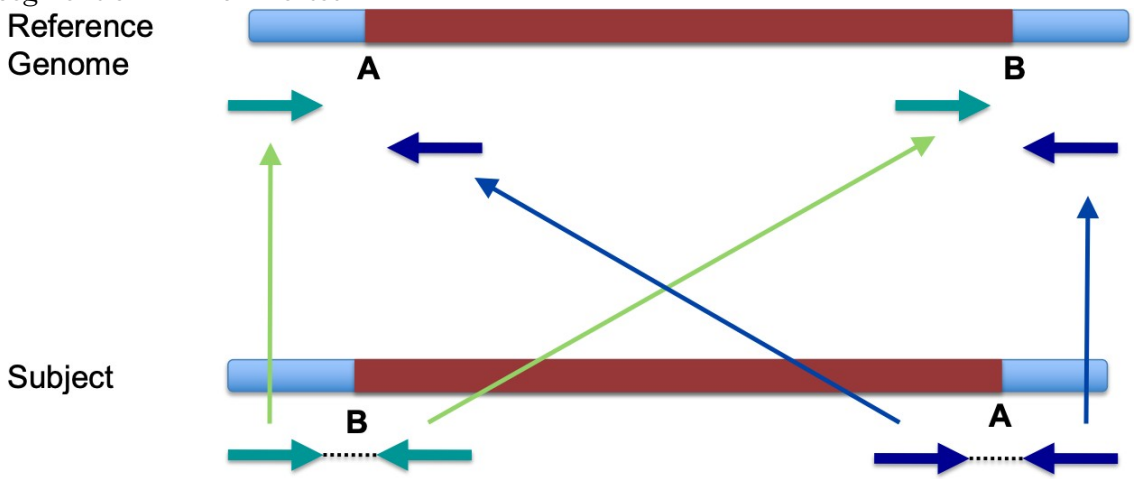


LR normal reads. They are left and right (respectively) of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome

LL, RR implies inversion in sequenced DNA with respect to reference RL implies duplication or translocation with respect to reference

1.3.2 Inversion

A segment of DNA is inverted



The most important pairs are the ones that stand between junctions because they are the most informative ones.

Here one end mapped where it was on the reference genome while the other end reversed its orientation

In IGV:

Information that help us:

The insert size from the target molecule (= the subject) is way longer. For all the pairs that are at the breakpoint, the insert size is different from the expected

The orientation is different

If you look at the local coverage, you can see a **drop** in two points: at the breakpoints. The reads that are mapping the junctions cannot map the reference genome because the breakpoint sequence does not exist in the reference genome. So, if we have an inversion in only one of the two alleles, then the reads coming from the allele with the inversion will not contribute to the local coverage at the breakpoint. The sequence in your target molecule exists only in one allele, so at the breakpoints you will only have reads contributing to the local coverage coming from one allele. The allele with the inversion will not have the

AB sequence, but only the BA sequence. That's the why of the drop in the local coverage.

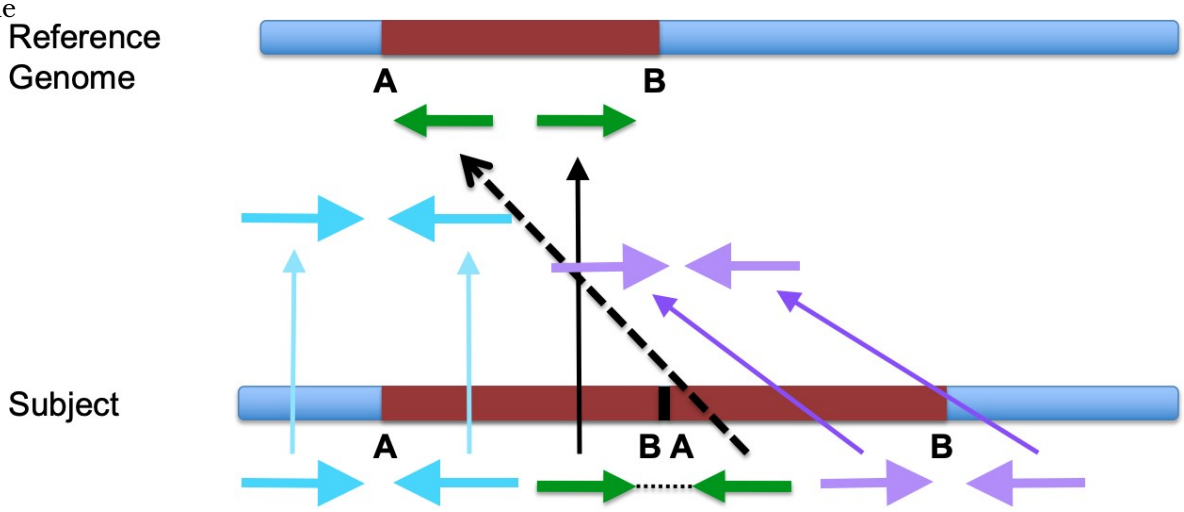
Moreover, we can notice that the coverage on the middle part does not change significantly from the coverage on the sides. That suggest this is not either a gain or a deletion, the only thing that might have happened is an inversion. Therefore, the inversion is not biallelic (because we see DNA, the drop doesn't go to 0)

When you align reads against a genome, you can allow for a certain mismatches or partial alignment. So, if you impose certain thresholds to your aligner, you can also say that if there are reads that align for 80% and have 20% of sequences misaligned, you align them in any case. So you will have reads that are correct up to the breakpoints and the browser will shows the mismatches beyond the breakpoint. So, you can have a partial

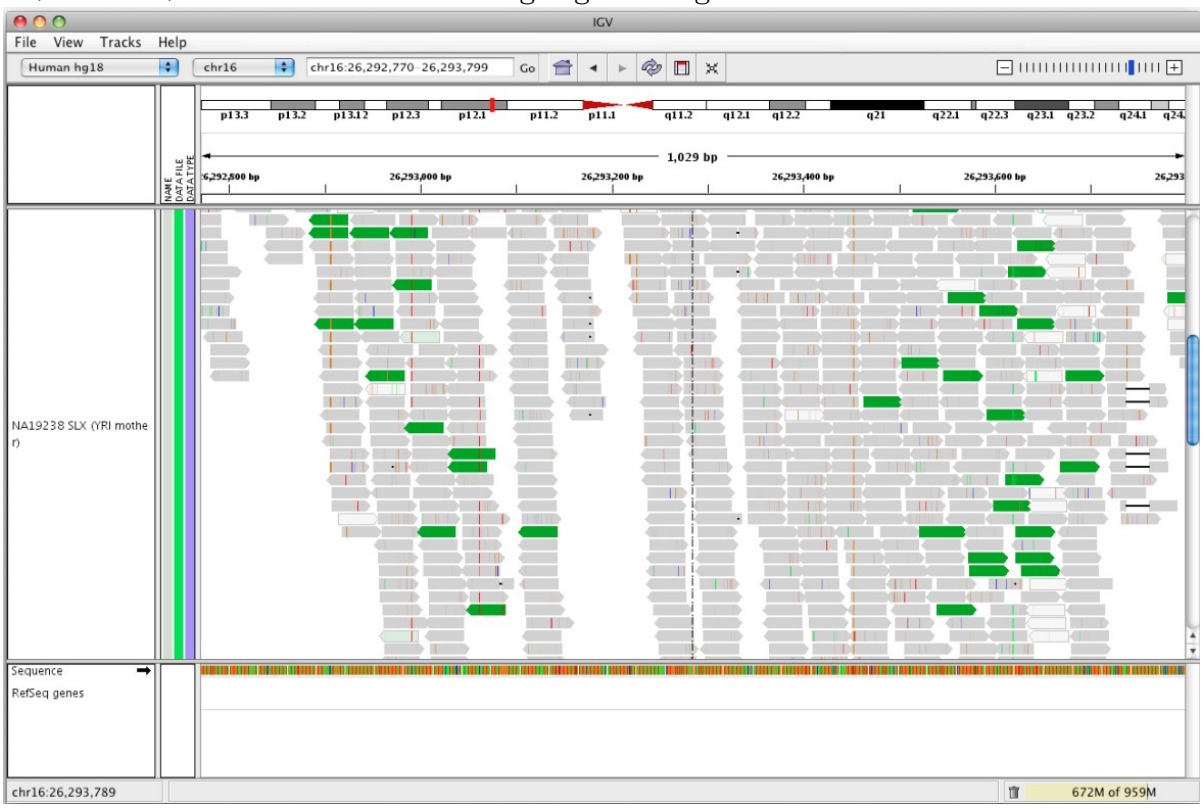
drop of coverage because you allow mismatches in your alignment.

1.3.3 Tandem duplication

A segment of DNA is duplicated and inserted in the target molecule adjacent to the original one



So, as result, the orientation instead of going inward goes outward.



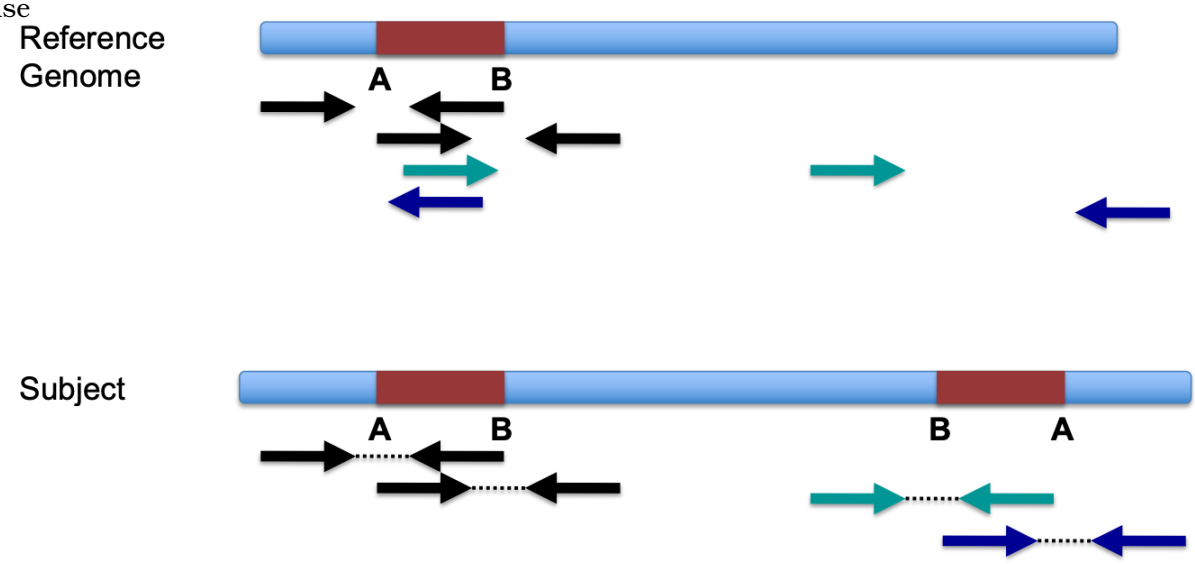
What do you expect to see from coverage? We will have a gain in coverage that is proportional to the extra copy. We need to pay attention to the double because it is a double contribution of that allele, but if a tandem duplication happens only in one allele and the other allele has his own one copy, then the local coverage corresponding to the tandem duplication will be 3/2 of the expected coverage.

If you have a read that maps BA, do you expect to see it in the mapped reads? Partial mapping. As we said before, if you allow your mapper to have some mismatches of a certain percentage of bases from your reads, you can still see some coverage contributed on one end of the segment and mismatches on the other side.

For what concern the junctions, you shouldn't see any difference of coverage because that sequence exists only once in the target molecule. The local coverage increases only in correspondence of the segment AB.

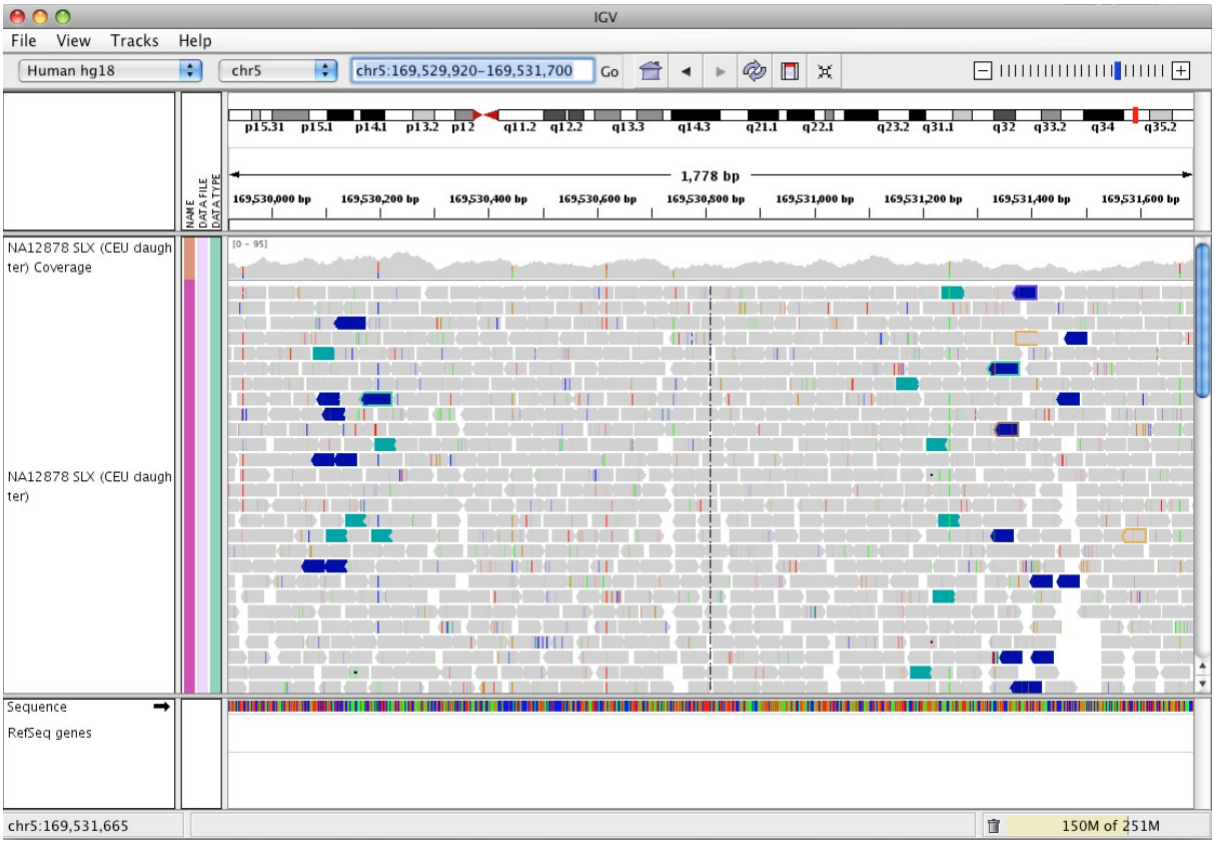
1.3.4 Inverted duplication

The duplication is inverted but it's not located near the original fragment, but somewhere else



Take into consideration:

- overlapping of "left" and "right" reads on the reference genome
- coverage depth (copy number)



There is a gain of coverage in the duplicated region and a tiny drop in the break points where the sequence exists in only one allele

1.3.5 Deletion

Deletion of a segment of DNA

If the deletion is larger than the size of the reads, we should see half of the coverage in the deleted regions

If the deletion is larger that the size of the reads, we should see a tiny little space corresponding to the missing nucleotides

Elements to consider:

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

Ask yourself if the sequence exists and where it is

Chapter 2

Genetic Figerprinting

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.

2.0.1 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.

2.1 SNPs features

2.1.1 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.

2.1.2 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`

2.1.3 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. In these

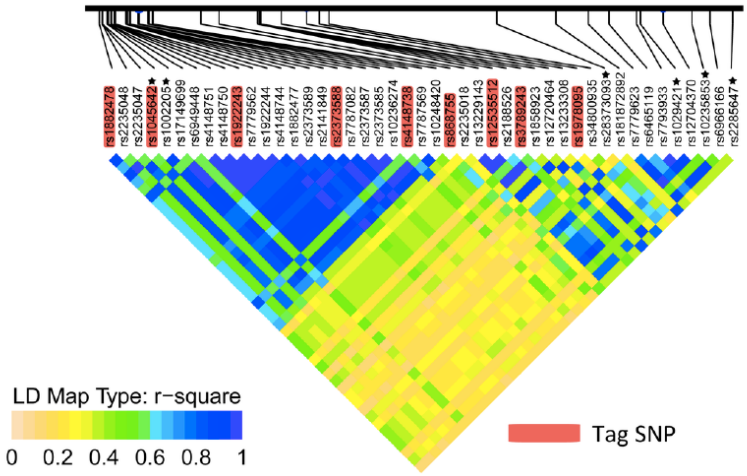


Figure 2.1: Linkage disequilibrium plot

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* 2.1. 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).

The colors indicate the strenght of pairwise linkage disequilibrium (LD) according to r2 metrics. 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).

2.1.4 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.

2.1.5 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.

2.1.5.1 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.

2.1.5.1.1 Some examples: In each example in figure 2.2 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.

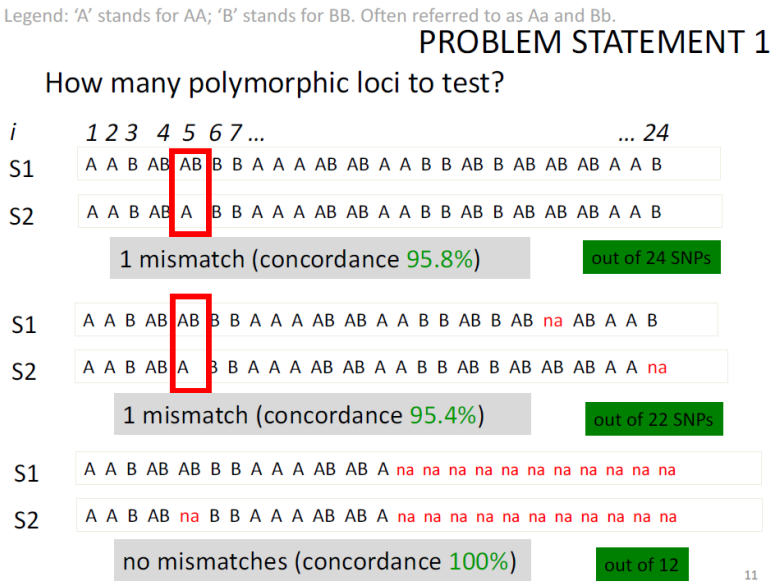


Figure 2.2

2.1.5.2 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.

2.2 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.

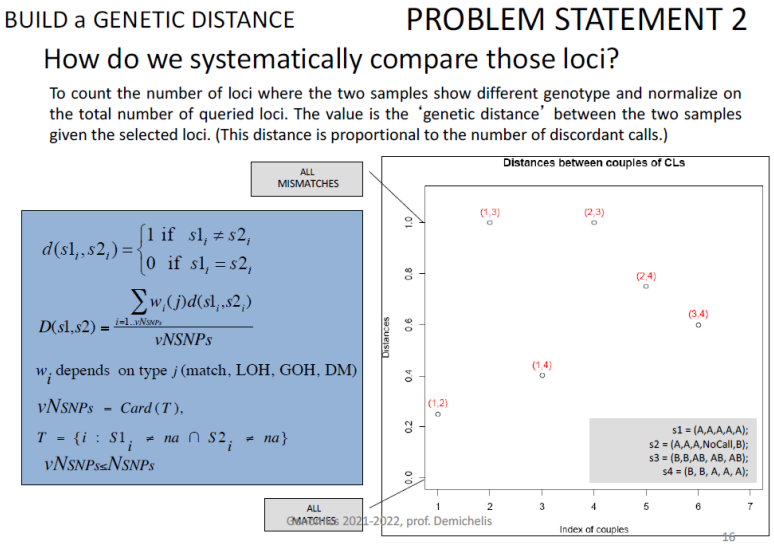


Figure 2.3: Genetic Distance graph with 4 samples

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 2.3 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.

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.

This other example at figure 2.4 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

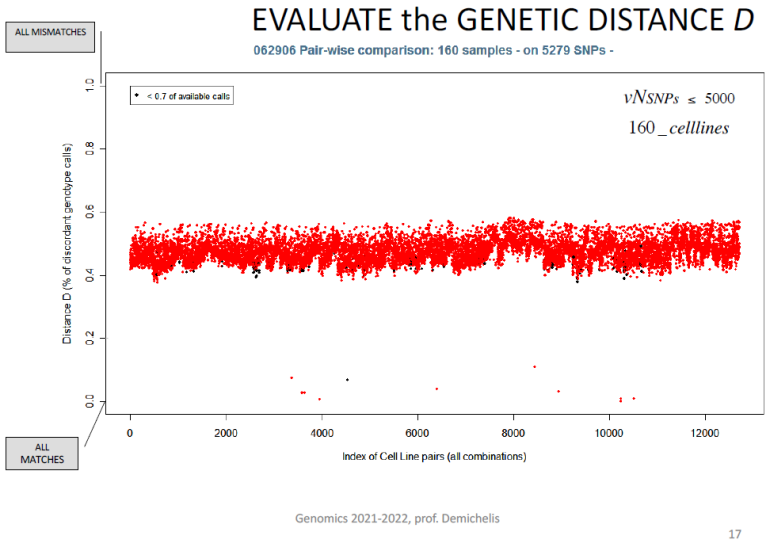


Figure 2.4: Genetic Distance graph with 160 samples

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).

In this last example at figure 2.5 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.

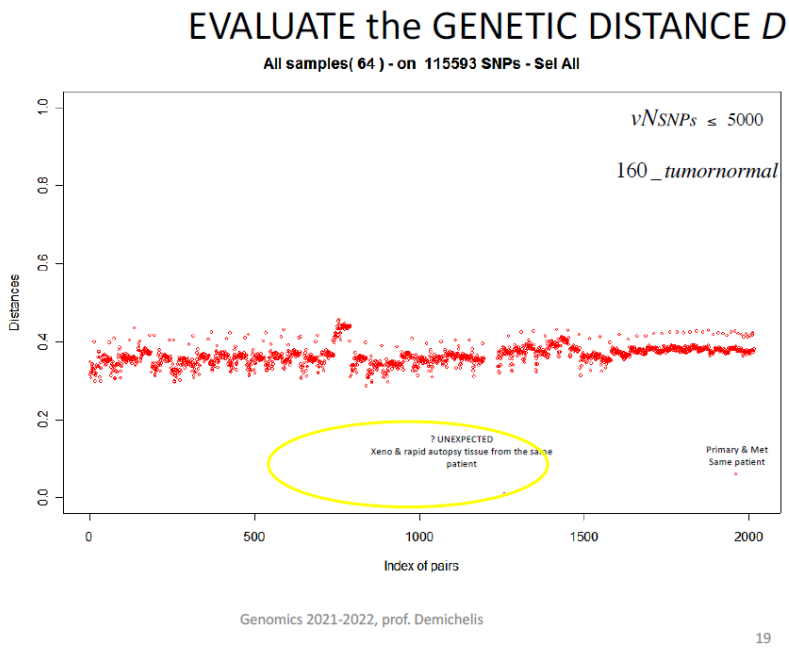


Figure 2.5: Genetic Distance graph with tumor samples

2.2.1 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.

2.2.2 Further considerations

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

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) 2.6.

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

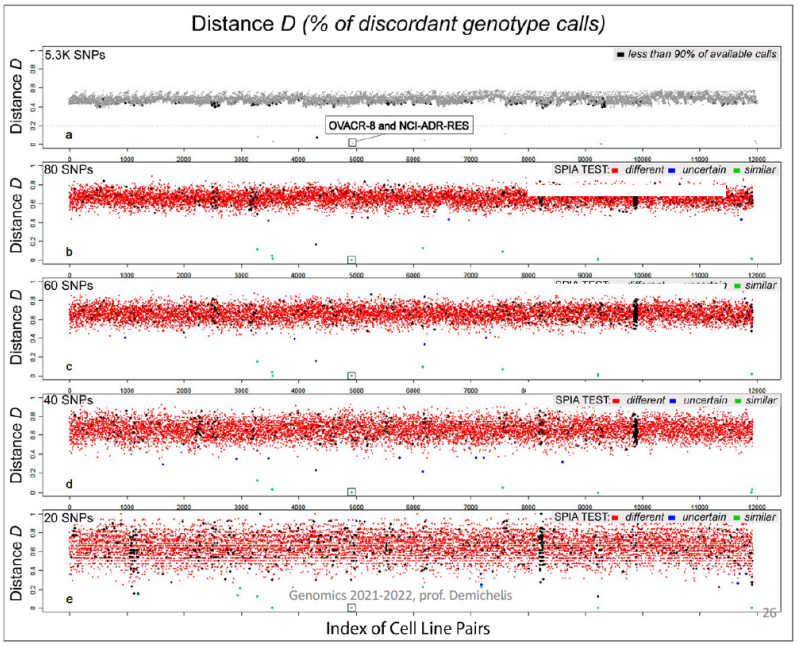


Figure 2.6: Genetic Distance graph at deacreasing number of selected SNPs

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.

2.3 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.

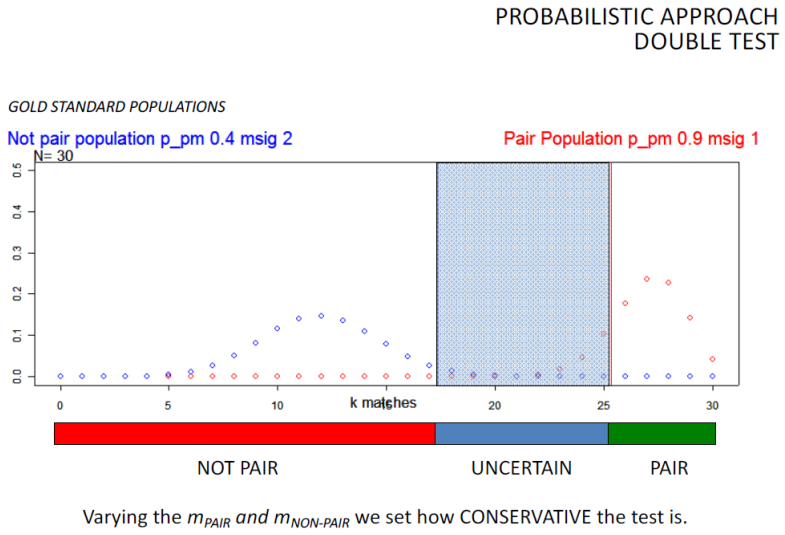


Figure 2.7

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

2.3.1 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 $K_{mean} \pm msd$ where m is the number of standard deviations used to define the thresholds which will lead to have a smaller or wider confidence area.

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 2.7 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.

2.4 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.

2.4.1 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.

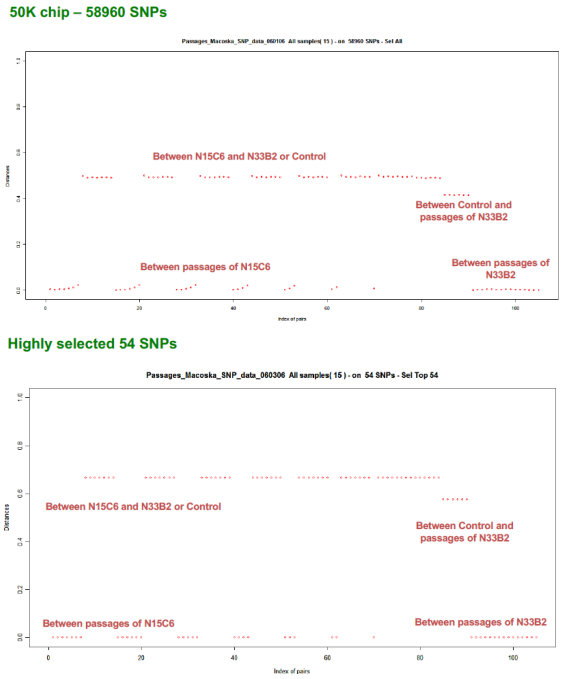


Figure 2.8

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.

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).

2.4.2 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.

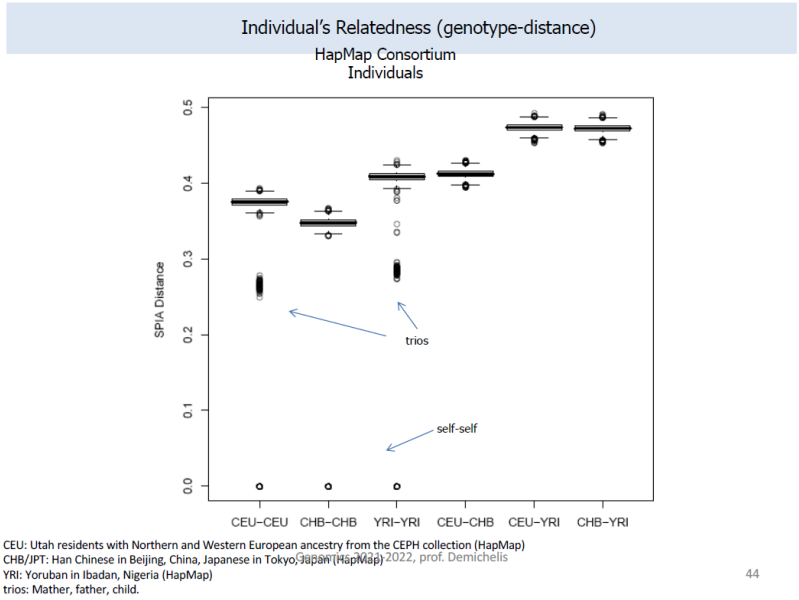


Figure 2.9

By looking at the data based on SPIA Assay (a genotype base assay which measures distance) at figure 2.9 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.

2.4.3 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.

2.4.4 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?

2.4.4.1 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 2.10.

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,

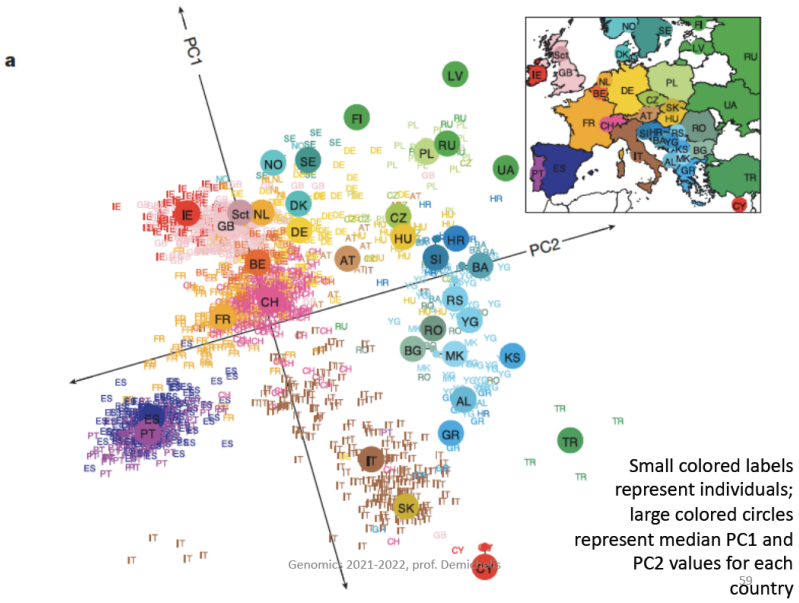


Figure 2.10

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').

Focusing on Switzerland, they could even make inference on the linguistic canton 2.11. 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.

2.4.4.2 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.

2.4.5 SPIA Assay

In a hand on lesson we performed ourselves a SNP-based genetic distance test using the R package 'SPIAssay'. You can find an R Markdown of that lesson in the folder 'Additional material'.

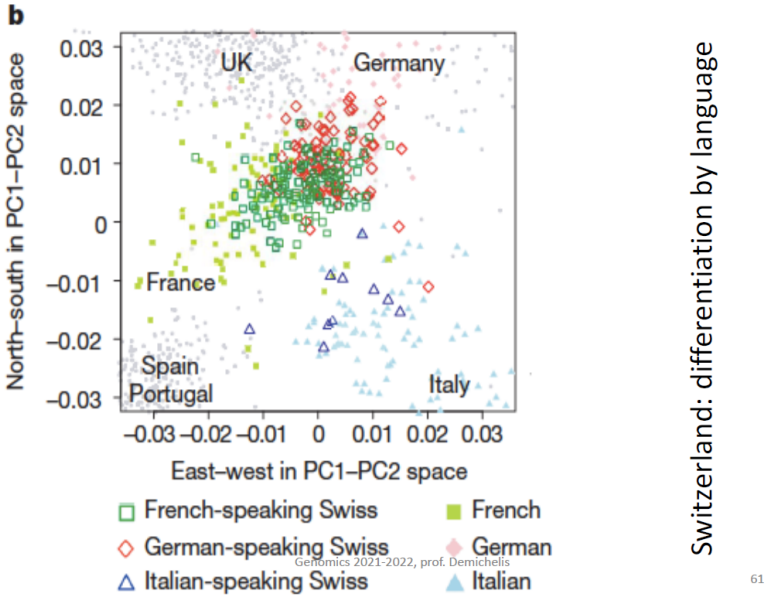

b
North–south in PC1–PC2 space
0.03
0.02
0.01
0
−0.01
−0.02
−0.03
UK
Germany
France
Spain
Portugal
Italy
−0.03 −0.02 −0.01 0 0.01 0.02 0.03
East–west in PC1–PC2 space
French-speaking Swiss
French
German-speaking Swiss
German
Italian-speaking Swiss
Italian
Switzerland: differentiation by language
Genomics 2021-2022, prof. Demichelis
61

Figure 2.11

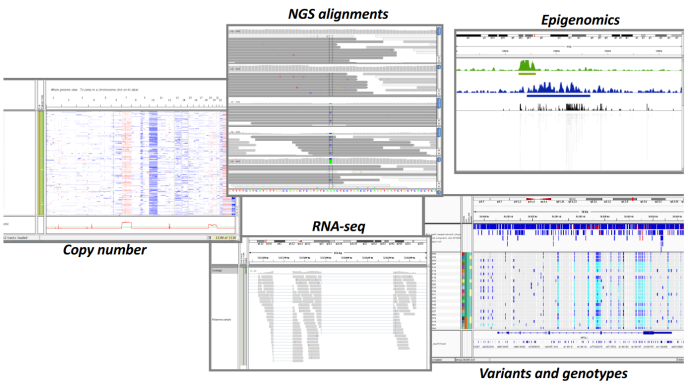
Chapter 3

IGV (Integrative Genomics Viewer)

3.1 Main characteristics

The human genome nowadays is being explored extensively thanks to of 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**. This is an approach for desease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person. The right drug, at the right time and at the right dose for each individual.

Figure 3.1: 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, and also to jump

in precise positions of the sequence. It is possible to search for genomic coordinates or gene names. 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 are present in the Supplementaty information - Integrative Genomics Viewer pdf file.

Figure 3.2: All the important elements to navigate into IGV are reported in the figure

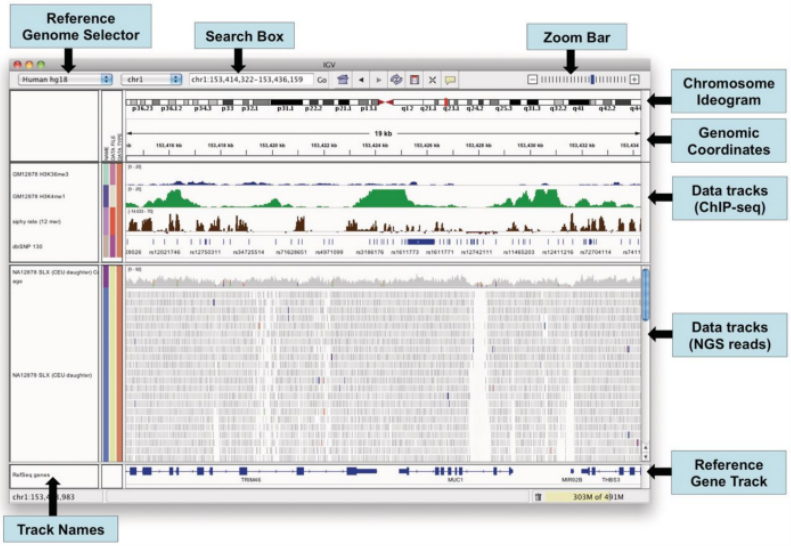


Figure 3.3

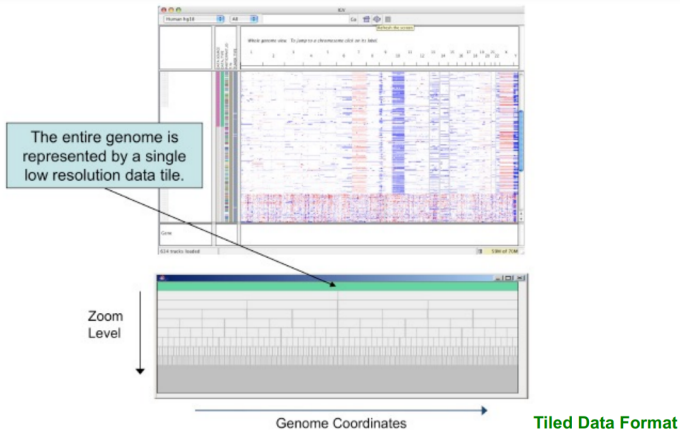


Figure 3.4

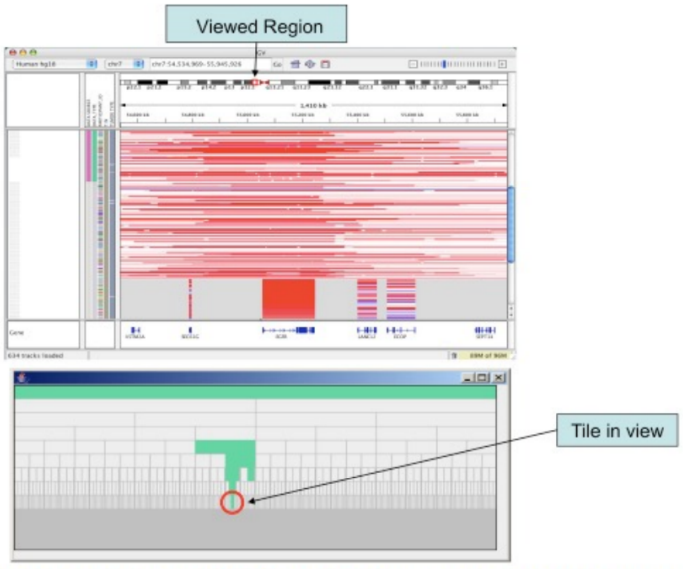
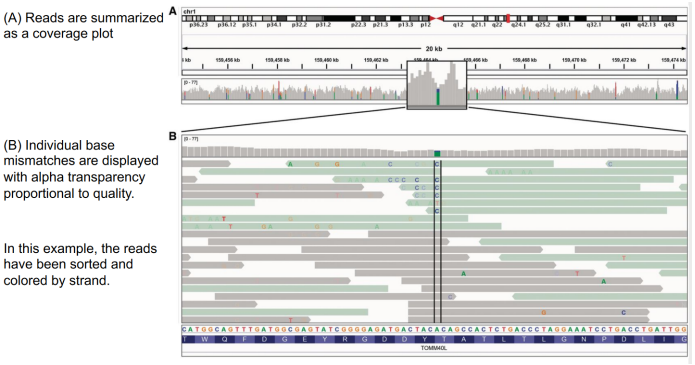


Figure 3.5



3.1.1 Igvtools

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

Figure 3.6: igvtools possible operations, the "count" function allows to generate coverage data, and it takes in input a BAM file. The obtained data 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

3.1.2 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 3.7: 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

3.2 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)

3.2.1 RNA-seq alignments

Figure 3.8: the height depends on the quantity of reads connecting the different exons.

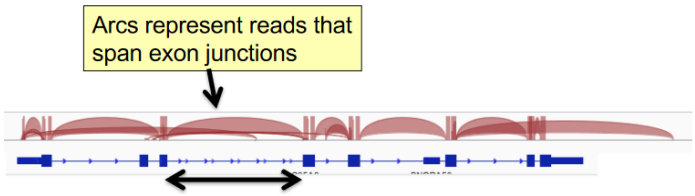
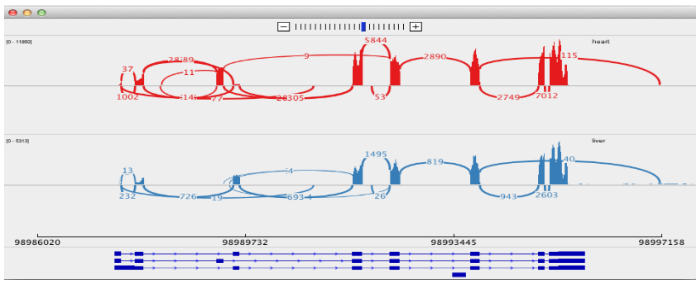
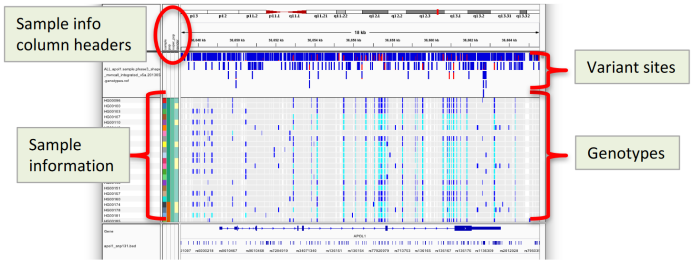


Figure 3.9: Sashami plots: The number of reads connecting exosomes are represented here on the curved lines. The peaks represent coverage within exons.



3.2.2 Study of variants

It is possible to study variants from different samples.



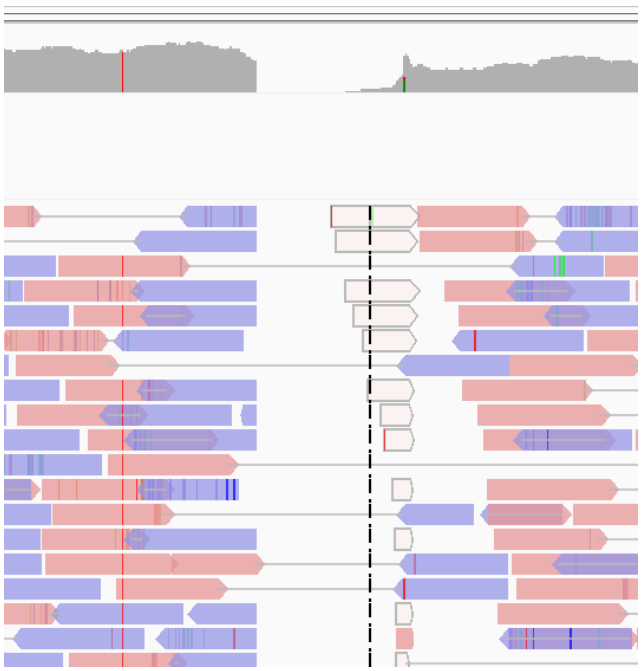
It is also possible to sort the samples in different ways and to group them considering different characteristics.

3.3 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 3.10: ***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 3.11: ***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



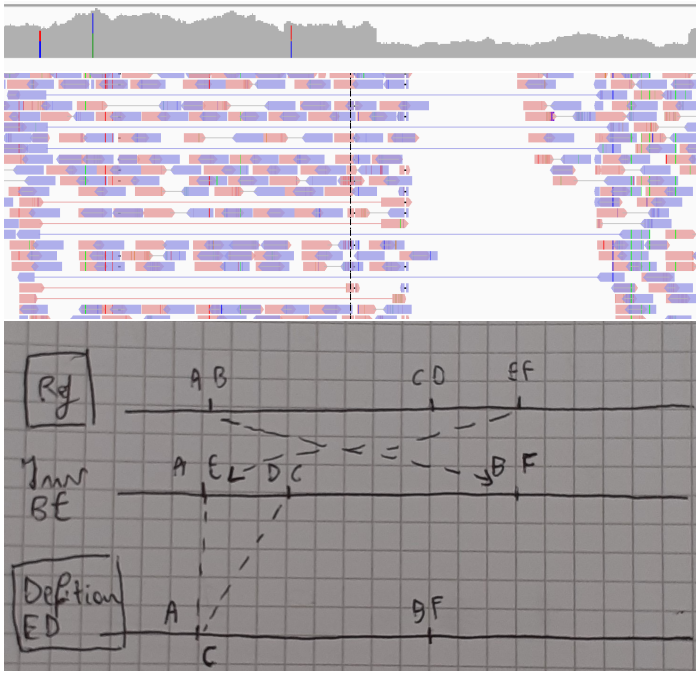


Figure 3.12: ***chr7:31,576,117-31,599,940***: Basically you have an inversion between B and F, and after the deletion of the ED portion, the other allele remains normal.

Chapter 4

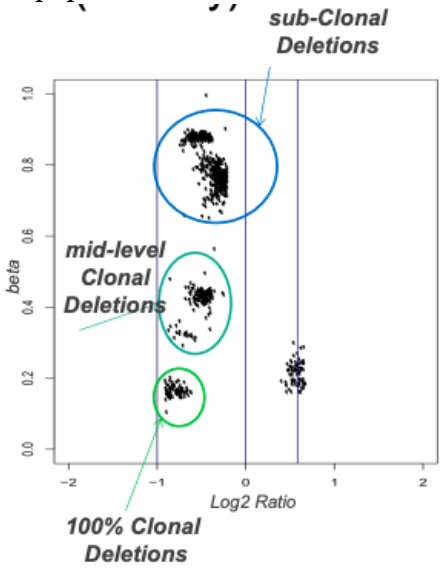
Tumor evolution studies (continued)

4.1 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.

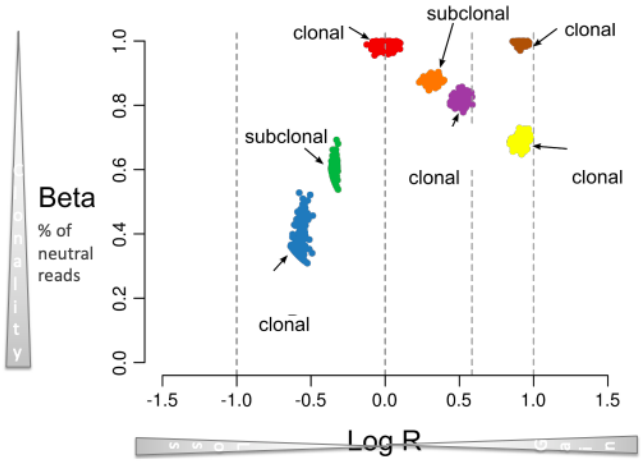
In the figure each dot is a genomic segment or a gene that clusterize in the space and when dots are in a same cluster it means that they very likely share the same copy number status and also the same level of clonality.

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.

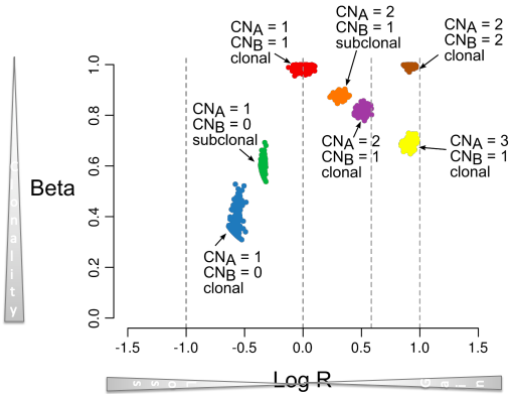


We have losses and gain of DNA copies, moving on the x axis.

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.

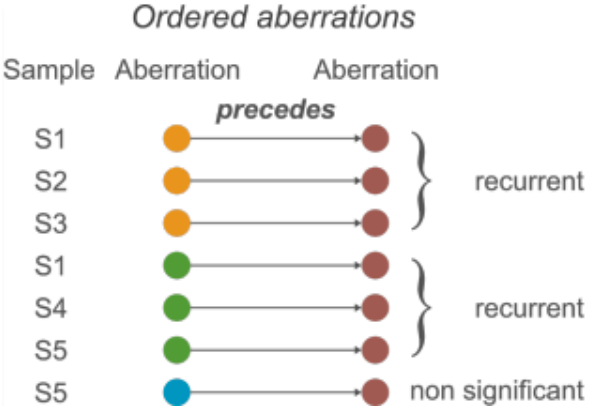
- Blue one: one copy of DNA, so we have a deletion
- Green one: also one copy of DNA but with subclonality

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:

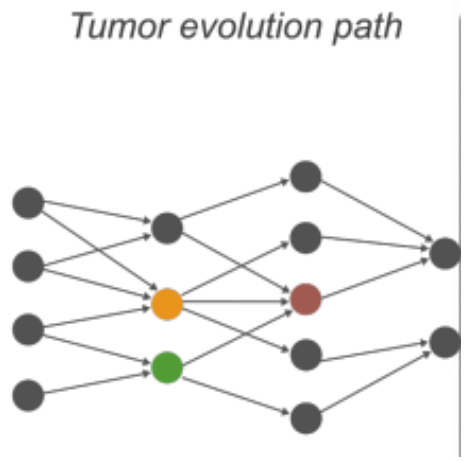
- In sample 1 the brown lesion is subclonal to the orange one, and that same lesion is also subclonal to the green one.
- In sample 2 we have again the support of the relation between the brown and orange lesion with the same level of subclonality (brown subclonal to orange).
- In sample 3 is the same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

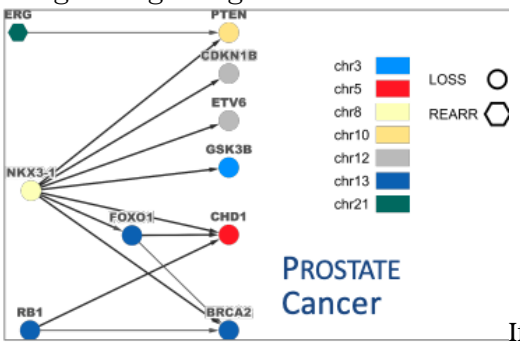
So having multiple observation supporting that aberration x precedes aberration y (i.e. aberration y is subclonal to aberration x) we can build an evolution chart.



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.

4.2 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?

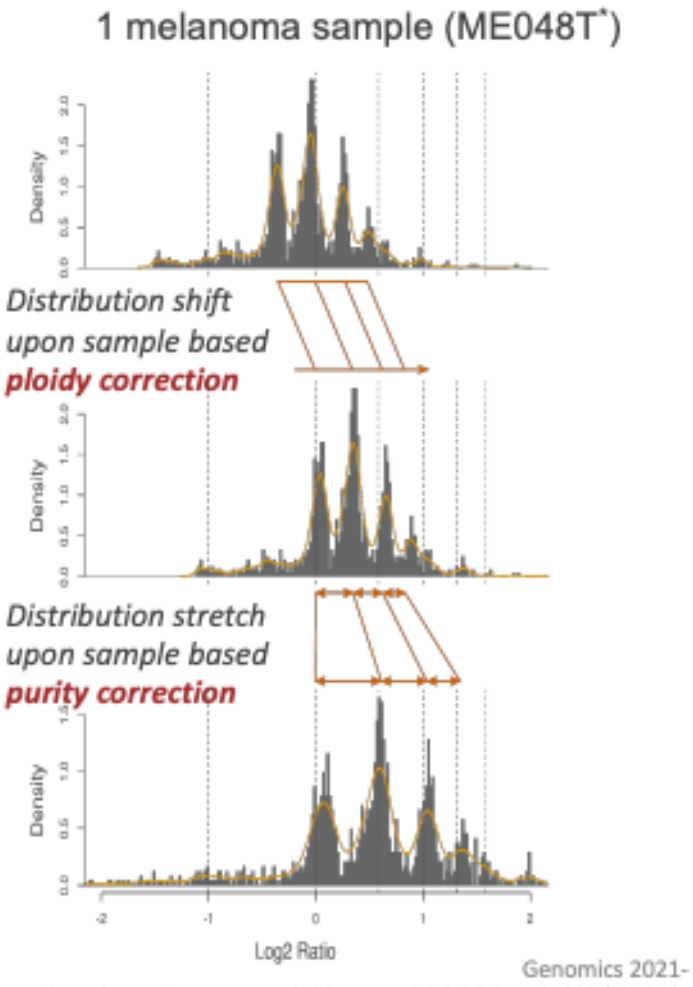
How can we compare two different samples for which we quantify completely different levels of tumor content?

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).

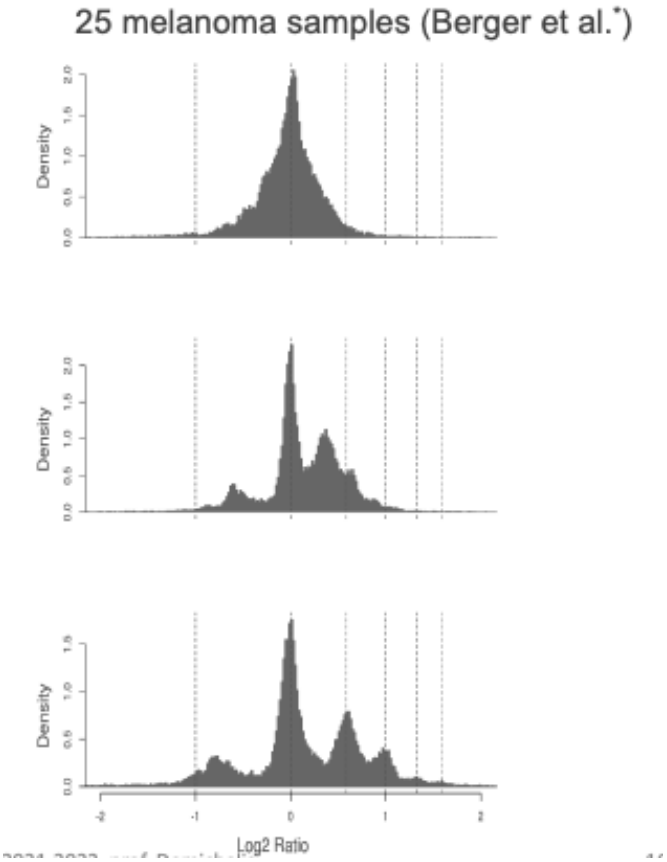
Purity correction

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



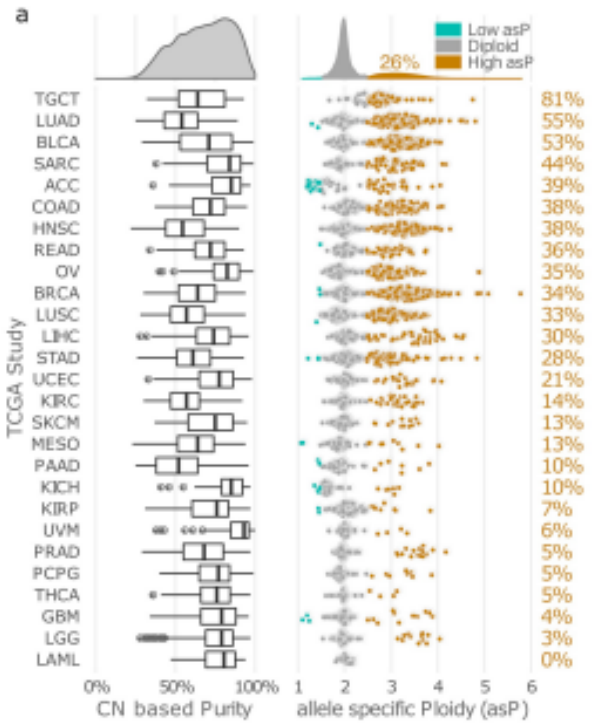
- 1st graph: The distribution of the log2 data of uncorrected signal, every melanoma sample is highly aberrant with a ploidy that is different between different individuals and a purity that is also different between different individuals. But we do have the tumor ploidy and purity so we can correct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

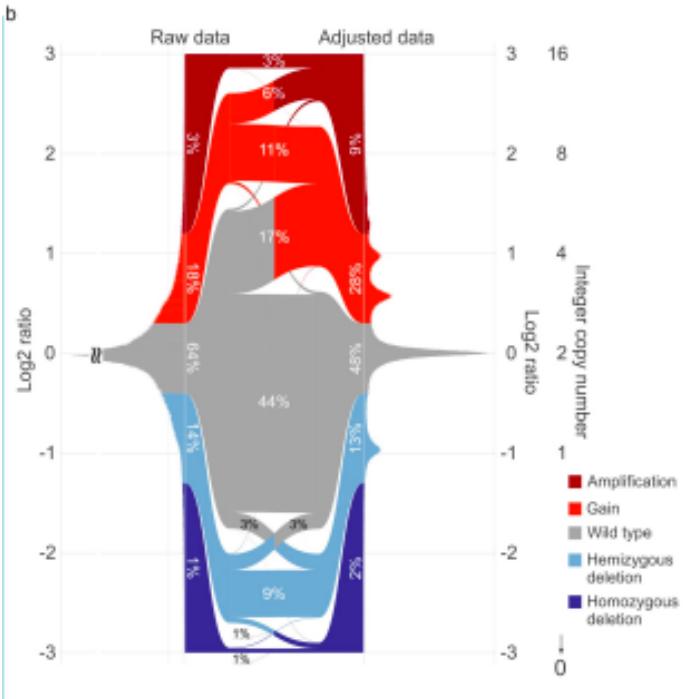
- The majority of the median distributions are above 50 %.
- The overall tumor cellularity was almost 70%.

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.

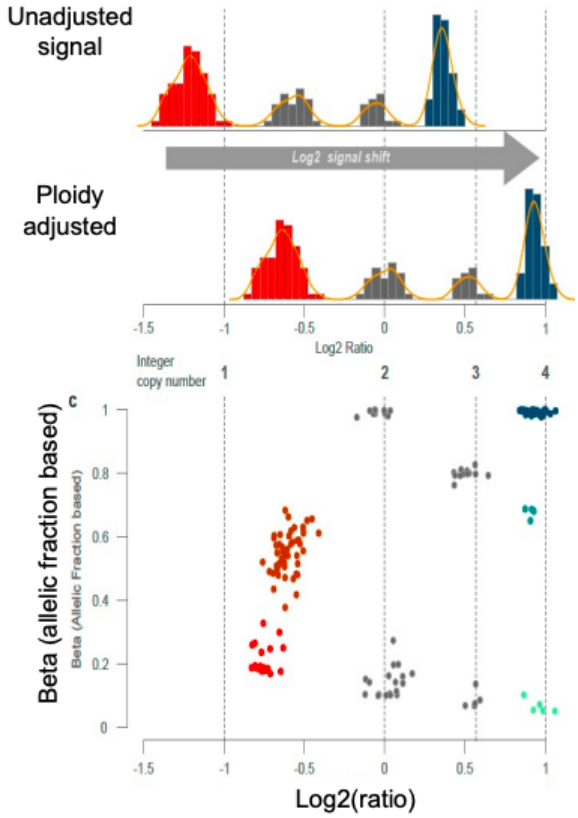
The figure shows what happens to data when we correct for ploidy and purity



- 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)



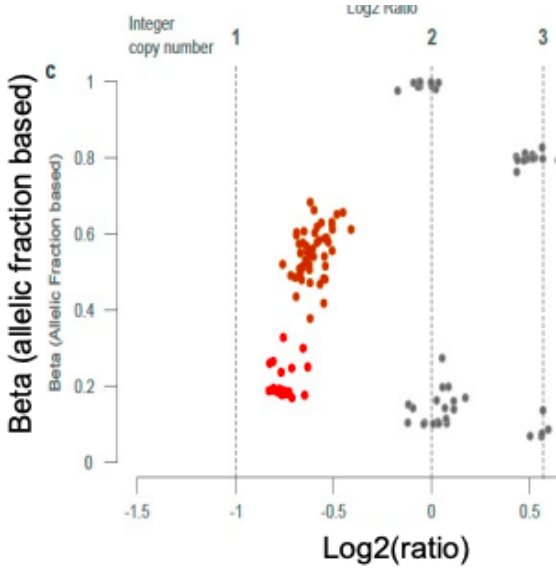
Thinking in terms of allele specific data:

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



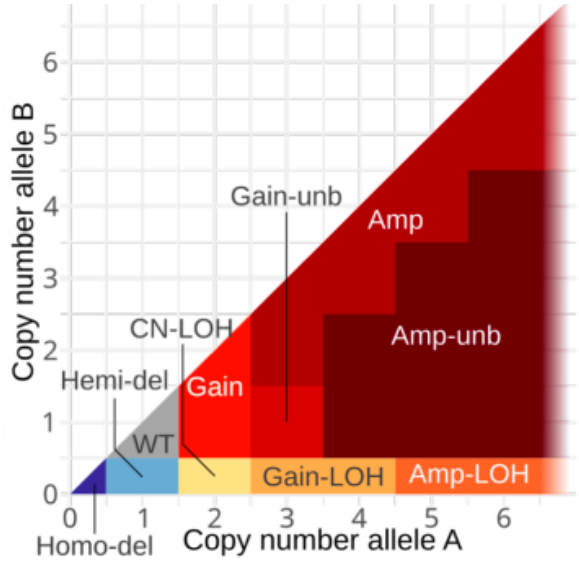
the number of copies of allele A and number of copies of allele B.
four copies we can have different combinations:

- 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.

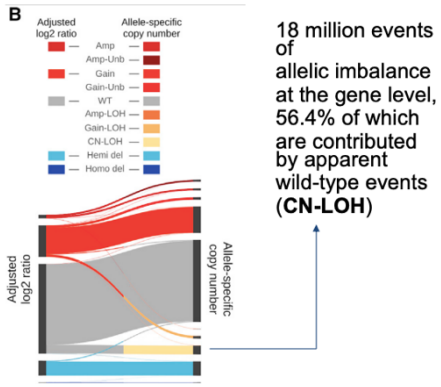
So, this whole computation allows us:

- 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).

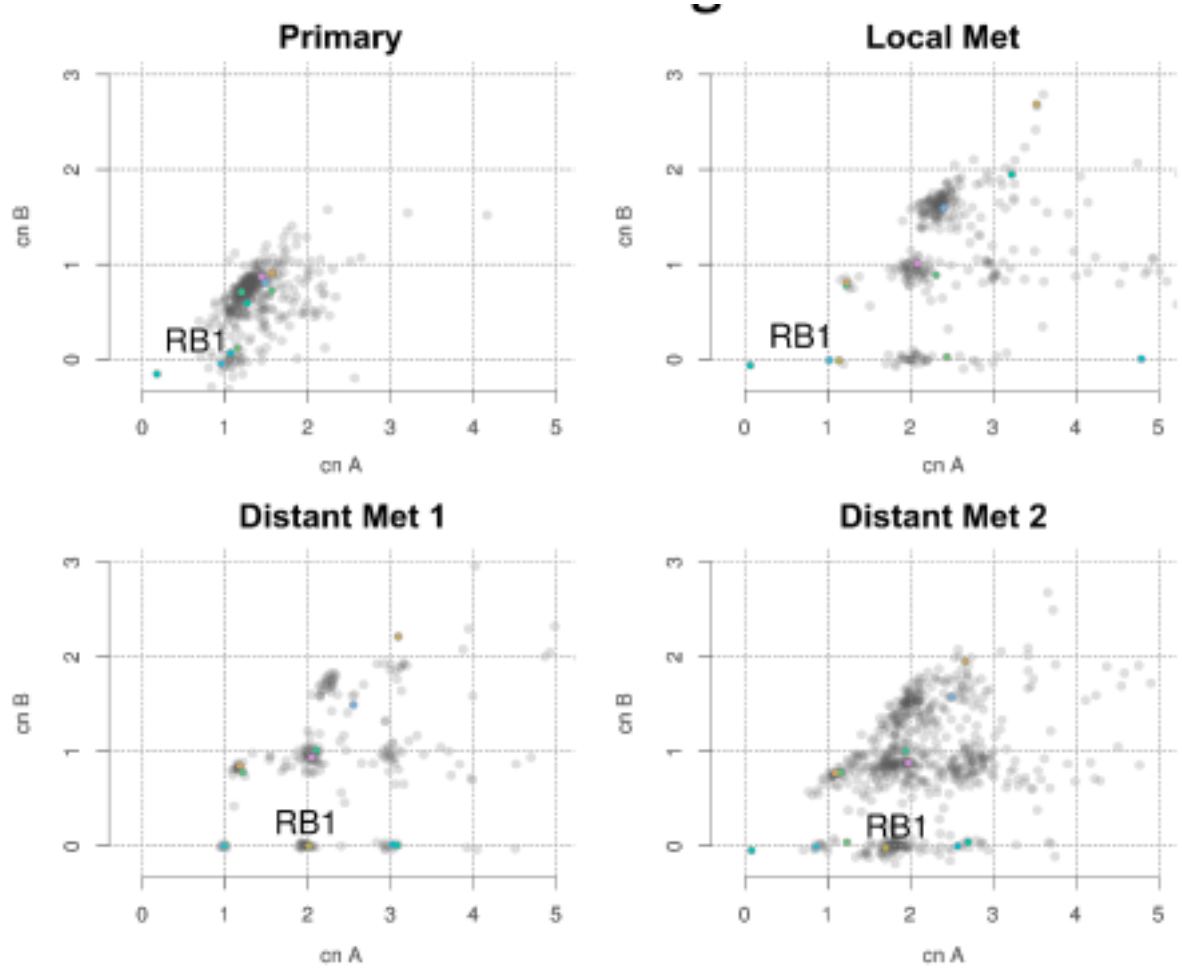
Copy-neutral because the number of copies doesn't change but there's been loss of heterozygosity.

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.

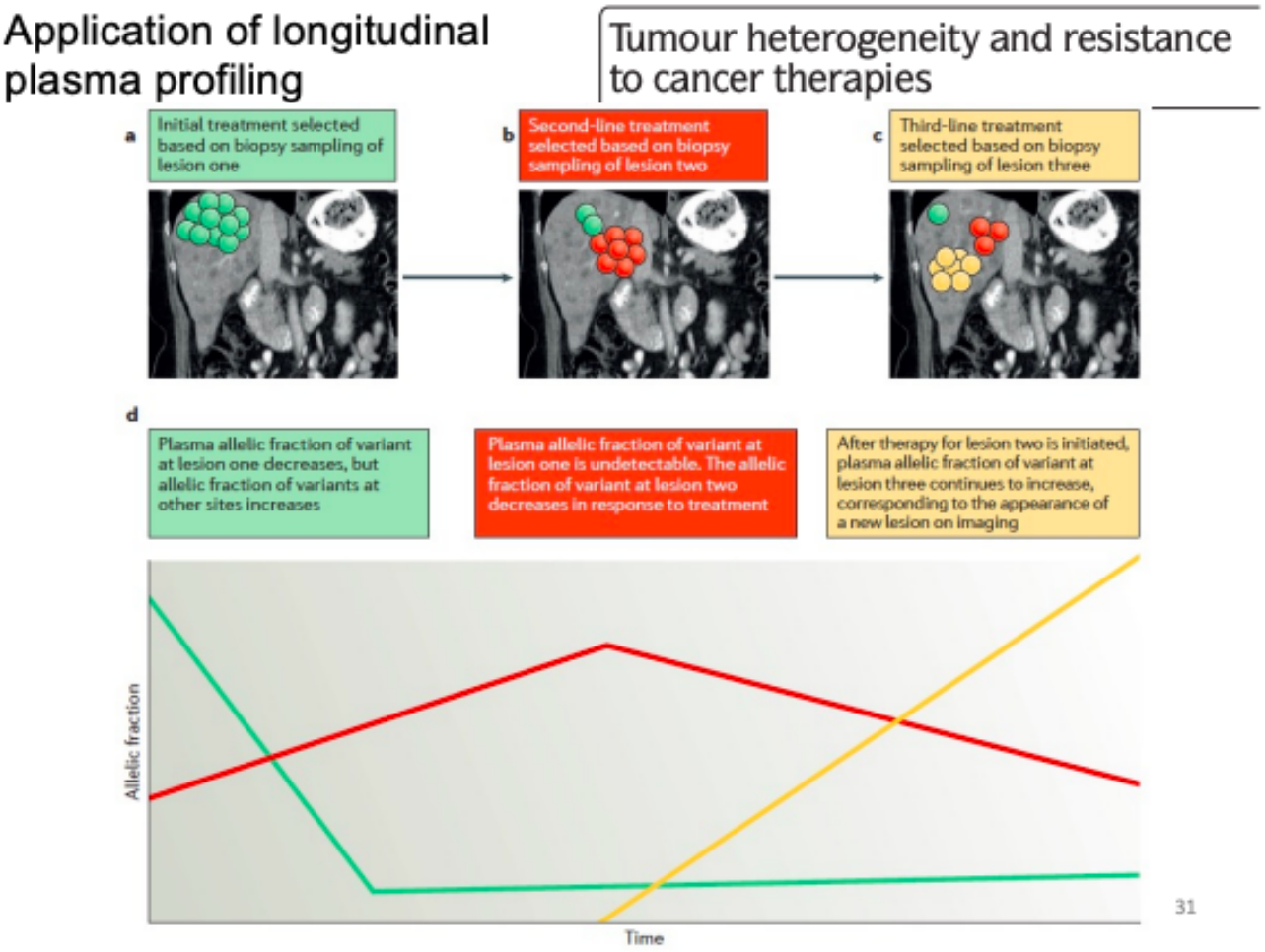
We have a local metastasis and two distant mets.

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.