Human genomics

Giacomo Fantoni
telegram: @GiacomoFantoni
Elisa Pettinà
telegram: @elisapettina

Github: https://github.com/giacThePhantom/human-genomics

Notes taken from lectures' recordings and:

Github: https://github.com/Maurizio319/HumanGenomics

July 1, 2022

Contents

Part I

Notes

Chapter 1

Introduction

1.1 Genetics and genomics

Genetics is the study related to specific genes and their variants with a known effect on the phenotype. Genomics is focused on the function and structure of the human genome, evolution and anything relating to the whole genome:

- Coding regions.
- Non-coding regions.
- Linear and non-linear structures.
- Cell physiology and pathology.

1.1.1 Genetics

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

1.1.2 Genomics

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

1.1.3 Differences between genetics and genomics

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

1.1.4 The role of computational biology

Computational Biology encompasses a wide range of numerical methods to analyse and integrate large scale data towards the understanding of molecular, cellular and structural biology. Possible studies of computational biology are:

- Semi-quantitative simulations of metabolic pathways.
- 3D protein-protein interaction.
- Characterization of 3D chromatic structure,
- Discovery and characterization of disease related variants.

The main subjects involved are:

- Biology.
- Genetics.
- Statistics.
- Calculus.
- Computer science.
- Bioinformatics.

The focus of this work is on hot to mine raw data, mainly from sequencing, how to exploit it for quality control (QC) and hot to interpret the obtained results in the context of human diseases, especially in cancer.

1.2 Basis of human genomics

The genetic make-up is different in all human it is responsible for our diversity. SNPs (single nucleotide polymorphisms) and CNVs (copy number variants) contribute to make us different. The majority of the external phenotypes come from genetic variance that are inherited or, in minor measure, acquired.

1.2.1 Single nucleotide polymorphisms

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

1.2.2 Copy number variants

Copy number variants or CNVs are the differences in number of alleles for a gene present in one individual. They contribute much more than SNPs in the difference between unrelated individuals, but they're less known as inherited variants, as they are harder to quantify and detect. CNV are distinguished as gain-CNV (where the number of alleles is greater than two) and loss-CNV (where the number of alleles is 2, 1, or 0). It is of note that if both parents have only one copy of a gene their child can have none. CNV span $\gg 1\%$ difference between two unrelated individual genomes.

1.2.3 Inherited variants

Inherited variants can be characterized by penetrance and allele frequency.

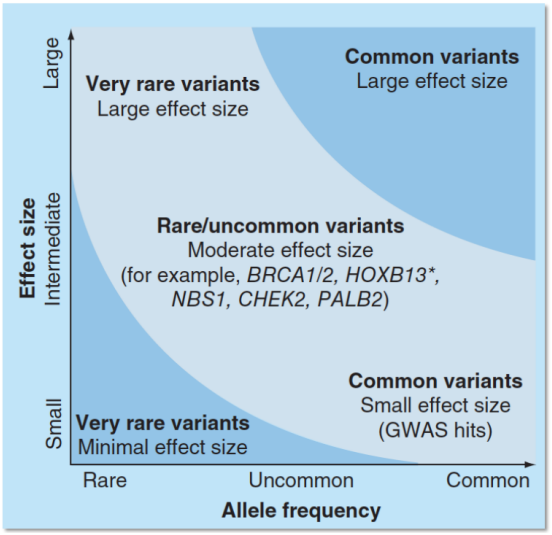


Figure 1.1: R. Eeles, Future Sci. OA (2016) 2(1), FSO87. Review on prostate cancer.
Common SNPs: $\frac{1}{4}$ of individual is homozygous for A1, $\frac{1}{4}$ is homozygous for A2, $\frac{1}{2}$ is heterozygous. The minor allele frequency is around 30-50% with low penetrance.
Rare SNPs: typically have very large size effect: if they are related to very specific traits they will have high penetrance. They constitute deleterious variants.

1.2.3.1 Penetrance

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

1.2.3.2 Allele frequency

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

$$AF = \frac{\#allele\ of\ interest}{\#copies\ of\ all\ thealleles\ at\ the\ genetic\ locus}$$

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

1.2.3.3 Differences in Genetic Make-Up, an example

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

able to change the ability of the organism to react to certain compounds causing pharmacokinetic variability and influencing the patient's treatment response. Both common and rare variants are involved. Figure 1.2 represent a subset of them.

Table 1. Comparison between pharmacogenomics approaches.

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

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

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

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

1.2.4 Somatic Variants

Somatic variants are **not** inherited from parents and are not transmitted to offspring. Most of them, but not all, are harmless. They can also be present in only a subset of the cells in an individual.

1.2.4.1 Classification

Somatic variants can be classified as:

- Single Nucleotide Variants (SNV). SNVs are a single point mutation restricted to a certain population of cells in an individual.
- Indels few nucleotides deletions.
- Rearrangements, like gene translocation, chromosome breakage and chromothripsis (which falls in the subcategory of chromosomal rearrangements).
- Somatic Copy Number Aberrations (SCNA).

1.2.4.2 Types of acquired DNA aberrations

1.2.4.2.1 Translocation Translocation happens when a sequence is moved from one genetic locus to another. It can be:

- Balanced: two sequences exchange locus and the overall quantity of DNA is maintained.
- Unbalanced: only one sequence move, generating an insertion.

1.2.4.2.2 Inversion Inversion happens when a sequence inverts its orientation. This aberration involves only one chromosome. The sequence of the inversion doesn't change, and the event will only be detected at its head and tail.

1.2.4.2.3 Copy number changes Copy number changes are events in which the quantity of DNA changes. They could involve one or more chromosomes. They can be:

- Duplication: a sequence doubles its copy number.
- Deletion: a sequence is lost.

1.2.4.2.4 Chromoplexy Chromoplexy derives from the Greek *pleko*, meaning to weave, or to braid. It describes 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 Chromothripsis derives from the Greek *thripsis*, meaning shattering into pieces. It describes a clustered chromosomal rearrangement in confined genomic regions that results from a single catastrophic event, usually limited to one chromosome.

Kataegis Kataegis derives from the Greek kataigis, meaning thunder. It describes a phenomenon that is characterized by large clusters of mutations (hypermutation) in the genome of cancer cells. An APOBEC family enzyme might be responsible for the kataegis process.

1.3 Experimental techniques to detect variants/aberrations

1.3.1 Karyotyping

Karyotyping is the process of pairing and ordering all the chromosome of an organism, providing a genome-wide snapshot of an individual's chromosomes. This experiment was used to try and detect genomic aberration, but it proved inadequate because its resolution wasn't enough. In particular it missed all sequence specific variants, breakpoints that could not be detected until the development of NGS.

1.3.2 Sequence capture for cancer genomics

In the paper summarized in 19 it is described a typical sequence capture for cancer genomics.

1.3.2.1 Reference

After sequencing there is a need to align the reads to a reference genome. This is especially needed when studying somatic changes. The best reference for this king of studies is the individual's own genome, usually retrieved from white blood samples. Both cancer and normal DNA can be aligned to detect if an aberration is cancer specific or is present in both normal and cancer DNA. The **match normal** tool is used to distinguish SNV from rare SNPs, somatic and germline indels, but also to make sure that copy number variations are somatic. Baits are nowadays used in the sequencing step, in order to sequence only the exome or specific part of the genome, as to make the sequencing process more cheap.

1.3.2.2 Deepness

Another fundamental parameter in sequencing is the deepness. A more deep sequencing is needed to find subclonal event that could increase cancers' fitness, like the ability to escape the immune system. It is necessary because not all cells present all the mutations characterizing the subclonal event and because the purity of a tumour sample is not always optimal.

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

1.3.2.3.1 Single end sequencing Single-read sequencing involves sequencing DNA from only one end. This solution delivers large volumes of high-quality data, rapidly and economically.

1.3.2.3.2 Paired end sequencing Paired-end sequencing allow to sequence both ends of a fragment and generate high-quality alignable data. It facilitates the detection of genomic rearrangements and repetitive sequence elements, as well as gene fusions and novel transcripts while providing double the coverage as a single-end protocol. It gives important information of the relative position of a molecule with respect to the reference, and is a necessary choice when structural aberration need to be assessed. This protocol is however twice as expensive as the single-end one.

1.3.2.3.3 Ability of paired end sequencing to detect genomic aberrations Figure 1.3 gives a nice graphical overview of genomic aberrations detectable by NGS, especially using PE sequencing. In particular the traslocation breakpoint in figure 1.3 would not be detectable without paired end sequencing. The most important parameter when studying deletions and insertion is to have enough coverage to perform significant downstream analysis, besides structural informations.

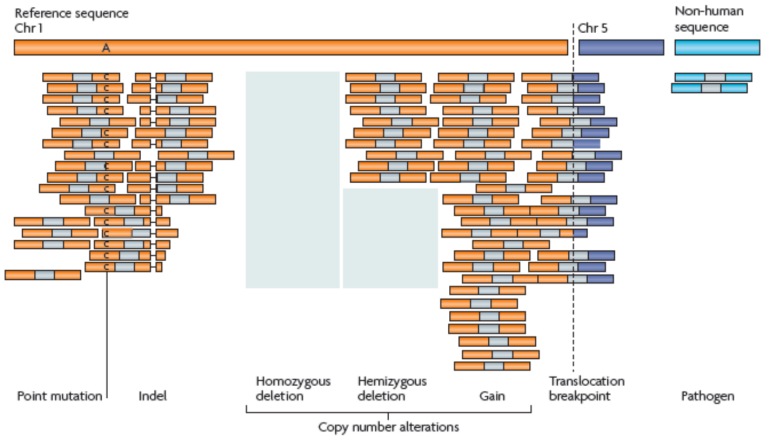


Figure 1.3: *Advances in understanding cancer genomes through second-generation sequencing*, Meyerson et al., Nature Reviews Genetics 2010

Chapter 2

Coverage

2.1 Computing coverage

The three key concepts needed when performing genomics analysis are the local and global coverage and allelic fraction.

2.1.1 Local coverage

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

$$cov = \#r_i : r_i \in p_i$$

2.1.2 Allelic fraction

The allelic fraction AF at position i is the proportion of reads that supports the reference (or alternative) base in p_i over the total number of reads that span p_i.

$$AF = \frac{\#r_{alternative_i\ (reference_i)}}{\#r_i} : r_i \in p_i$$

2.1.3 NGS global coverage

The Lander-Waterman equation is used to compute NGS global coverage. The equation is:

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

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

2.2 Mapping in NGS

When mapping NGS reads to a reference genome the number of correctly mapped reads is always lower than expected. This is due to sequencing or aligning errors or due to major translocations that will impair a good mapping of the reads. When considering which part of the reference genome is

covered by the mapped reads, a distinction between physical and sequence coverage has to be made. A schematic representation of the problem is displayed in figure 2.1.

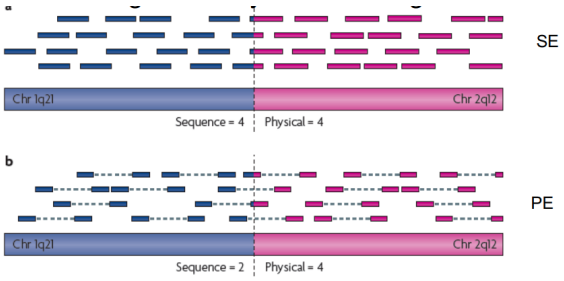


Figure 2.1: Schematic difference between sequence (on the left) and physical coverage (on the right).

2.2.1 Sequence coverage

Sequence coverage is the amount of oversampling (how many times a base is sequenced). It is needed to detect nucleotide alterations with high sensitivity. To do so for the 3 billion bases of the human genome a 30-fold coverage on average ($30X$) is usually required. This means that 90 million of bases need to be mapped for each sample.

2.2.2 Physical coverage

The expected distance between the paired reads is used to uniquely place the reads on the genome. Physical coverage refers to the average number of times a base is read or spanned by paired end reads. Unexpected read pairs are used to detect structural anomalies. So, while sequence coverage refers to a single base resolution, physical coverage refers to structural coverage of the genome.

2.3 Tuning coverage

In some experiments setups there's the need to carefully control the amount of intended coverage.

2.3.1 SNP detection

SNPs are by definition present in all of the cells of an individual, so only enough redundancy or local coverage to distinguish the reference from the alternative base is needed. Ideally for heterozygous SNPs half of the reads will support the reference base while the other half the alternative one. Typically to perform this kind of detections only 10 to $15X$ coverage is needed.

2.3.2 Subclonal events

Subclonal events from a tumour or hematopoietic sample are not harboured by all the cells in an individual, so the coverage must be increased with respect to the one needed for SNPs detection. The same approach must be taken to detect momozygous mutations and low abundant events, like low-expression transcripts and to reduce the effect of weak binding in chIP-seq.

2.3.3 Amplicon-based approaches

Amplicon sequencing is a highly targeted approach that enables researchers to analyse genetic variation in specific genomic regions. The ultra-deep sequencing of PCR products or amplicons allows for efficient variant identification and characterization. This methods uses oligonucleotide probes designed to target and capture regions of interest, which are amplified by PCR and sequenced by NGS.

2.3.3.1 Heterozygous deletion of PTEN

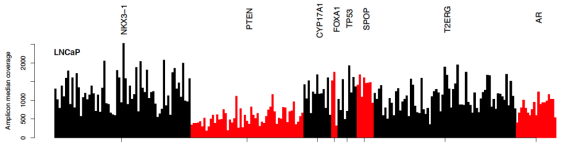


Figure 2.2: Example of local coverage of 10 genes for cell line LNCaP. On the x axis genomic location, while on the y axis the amount of amplicons sequenced.

In figure 2.2 a barplot representing the local coverage (y axis) in the gene locations (x axis) can be observed. The coverage is on average about $600x$ and it's not evenly distributed, as it typically is for this kind of experiment. Averaging the coverage for each gene it is clear how for PTEN it is significantly less. Since some coverage is still present for that gene the most probable event to have caused this type of event is an heterozygous deletion.

2.3.3.2 Homozygous deletion of PTEN

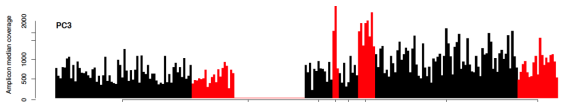


Figure 2.3: Example of local coverage of 10 genes for cell line PC3. On the x axis genomic location, while on the y axis the amount of amplicons sequenced.

In the plot 2.3 from a different cell line a clear monoallelic deletion and a partial biallelic deletion of PTEN can be observed.

2.3.3.3 Amplification of AR

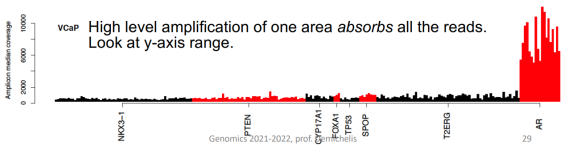


Figure 2.4: Example of local coverage of 10 genes for cell line VCaP. On the x axis genomic location, while on the y axis the amount of amplicons sequenced.

In figure 2.4 a massive amplification of the antigen receptor typical of prostate cancer can be seen. This massive amplification is probably due to a mistake of the assay. Amplification on the AR does not allow for the analysis and discovery of other copy number variations because all of the reads will be mapped to the AR site.

2.3.4 NGS-based approaches

NGS-based approach are based on the sequencing of all of the sample genome, skipping the probing and amplification passages.

2.3.4.1 Advantages over amplicon-based approaches

With this type of approaches it is easy to increase the experimental coverage at later point in time. It is enough to perform another run of sequencing on the same sample or library and then combine the output from different runs to increase coverage. This isn't possible with array-based technologies.

2.3.4.2 Problems of NGS-based approaches

There are limiting factors to be considered when performing NGS DNA sequencing experiments. In particular:

- Repeated regions are difficult to sequence and map correctly.
- Structural information are difficult to obtain.

This problems could be mitigated by increasing the read length, but this would make the sequencing process more error prone. There is a need to make a trade-off between read length and single-base accuracy.

2.4 Databases for NGS analysis

Two very known databases for NGS analysis are:

- Genome Reference Consortium.
- USCS Genome Browser.

2.4.1 Genome reference consortium

The genome reference Consortium assembles a reference genome reflecting the most common sequences in population at each position while tracking information on polymorphisms.

2.4.2 USCS genome browser

The USCS Genome Browser allows to select a reference genome and query all of its known features.

Chapter 3

Genetic Fingerprinting

3.1 Introduction

Genetic fingerprinting is a technique used to investigate some characteristics of a genome, typically a pattern of variable elements, like SNPs or minisatellites, in order to uniquely characterize a genome. It 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.

3.1.1 Fields of interest

DNA fingerprinting is used in different fields, like:

- In forensics for identification purposes.
- In lineage related tests, for cells or humans like paternity or 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 genetic drifts. It is necessary when using certain cell lines in an experiment for publishing purposes.
- To identify and remove samples that would skew the data. For example members of the same family when performing a GWAS study on a certain geographic area.

3.1.2 Variants used for genetic testing

Different variants can be used for genetic fingerprinting, such as Single Nucleotide Polymorphisms (SNPs) or inherited Copy Number Variations (CNVs). In the past, before sequencing and SNPs array, short tandem repeats were commonly used for genetic fingerprint.

3.1.2.1 CNVs

CNVs are a phenomenon in which sections of the genome are repeated or deleted, changing the number of times those regions appear in the genome. The most amenable type of inherited CNVs for genetic fingerprinting are the loss-CNVs. For a loss CNV in the population the copy number can vary between 2 and 0. In particular if both parents are heterozygous for a particular CNV their offspring could have an homozygous deletion. If instead both parents have 2 copies at a site that is polymorphic in the population, all of their offspring will have a copy number of 2. Gain-CNVs are

difficult to analyse for this tests because when combining multiple copies the origin of a single copy number cannot be traced to a parent and so they cannot be used for genetic fingerprinting.

3.1.2.2 SNPs

SNPs are substitutions of a single nucleotide at a specific position in the genome. They are the most amenable type of variation as they are simple, abundant in the genome and easy to detect in sequencing data even with low coverage. For these reasons the focus of this work will be on SNP-based genetic tests.

3.2 SNPs features

3.2.1 Hardy-Weinberg equilibrium

One property of SNPs which has to be taken into account when using them 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.

In the simplest case of a single locus with two alleles denoted A and a with frequencies $f(A) = p$ and $f(a) = q$ in a population the expected genotype frequencies under random mating are:

- $f(AA) = p^2$ for AA homozygotes.
- $f(aa) = q^2$ for aa homozygotes.
- $f(Aa) = 2pq$ for Aa heterozygotes.

In the absence of selection, allele frequencies p and q are constant between generations, reaching an equilibrium. The general equation that the allele frequencies need to fit in to be considered in equilibrium is:

$$(P + Q)^2 = 1 \qquad \wedge \qquad P^2 + 2PQ + Q^2 = 1$$

SNPs that respect this equilibrium are the most studied and thus the most informative.

3.2.2 Minor allele frequency

Minor allele frequency is the frequency at which the second most common allele occurs in a given population. 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 highest possible. Highest variability in the population allows to distinguish better more individuals.

3.2.2.1 Optimal MAF values for genetic fingerprinting

Number-wise, a frequency of $\frac{1}{3}$ for each SNP would maximize the variability in a population, but those SNPs wouldn't be in Hardy-Weinberg equilibrium generating possible missed calls. Therefore, the optimal SNPs to detect individuals' differences and similarities are those with genotype frequencies:

- $P_{AA} = 0.25$.
- $P_{BB} = 0.25$.
- $P_{AB} = 0.5$.

So the best SNPs will be those with $MAF = 0.5$.

3.2.3 Haplotype blocks

Another important feature to consider for SNPs selection are Haplotype blocks. They are blocks along the genome that tend to be inherited together as segments. In these regions there is little evidence for historical recombination and only a few common haplotypes are observed. Because of this to perform genetic fingerprinting it is enough to consider only a SNPs per haplotype block because the other won't bring additional information.

3.2.3.1 Linkage disequilibrium

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

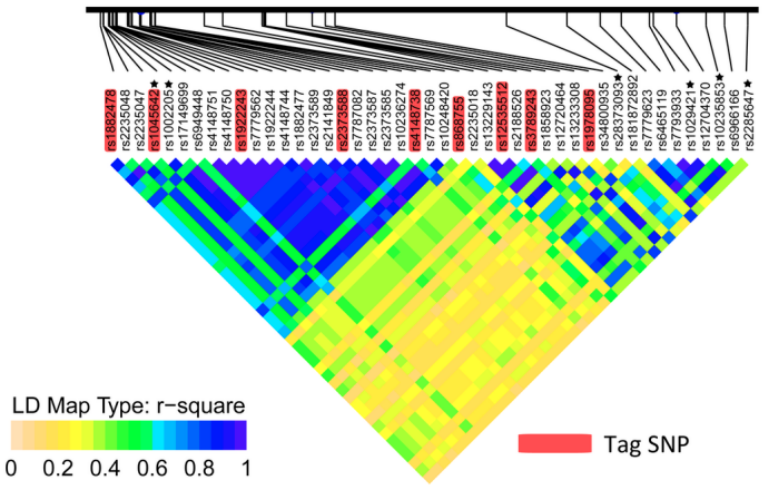


Figure 3.1: LD plot of SNPs with top-ranked bayes factors in CHB (Han Chinese in Beijing) of 1000 Genome Phase I. The colors indicate the strength of pairwise LD according to $r2$ metrics. The SNPs marked with asterisks represent independent strong associations. Tag SNPs are here shadowed in pink.

The colors indicate the strength of pairwise linkage disequilibrium (LD) according to the $r2$ metrics, the proportion of the variation in the dependent variable that is predictable from the independent ones. In fact, not all of the SNPs are informative to distinguish between individuals.

3.2.3.2 Tag SNPs

In figure 3.1 tag SNPs are shadowed in pink. A tag SNP is representative of a region with high linkage disequilibrium and represents a group of SNPs or haplotype. This tag SNPs are the ones

that will be included in a genetic fingerprinting test.

3.2.4 Other SNP features

Other SNP features to take into consideration when performing a genetic fingerprinting test are:

- Exclude chromosomal locations which undergo frequent somatic aberrations. For example areas commonly deleted in tumour will produce LOH but probably also no calls, since there is no DNA to be sequenced.
- Choose SNPs equally spread all around the genome so to represent it all.
- Select only autosomal SNPs.
- Select SNPs in exons, so to have signal even from a non-DNA assay.
- Exclude or include disease or drug response associated loci.
- Include or exclude loci with significantly different MAF in different ethnicity. Including them allow to perform a lineage test in the same assay.

3.2.5 Number of SNPs to select when performing a genetic test

The number of SNPs needed to run a genetic fingerprinting test must be assessed. This number should be allow the measure of the test to differentiate unrelated individuals. When choosing this number experimental and biological mismatches must be taken into account, weighted based on their likelihood. When choosing the number of SNP all the possible mismatches must be taken into account, increasing it to allow to identify individuals.

3.2.5.1 Experimental mismatches - Genotype call error rate

During sequencing errors can occur, resulting in no data available for some loci, that if they include SNPs of interest will cause a loss of a call for that SNP. These experimental mismatches are related to the error rate of the technology used and because of this they are platform dependent.

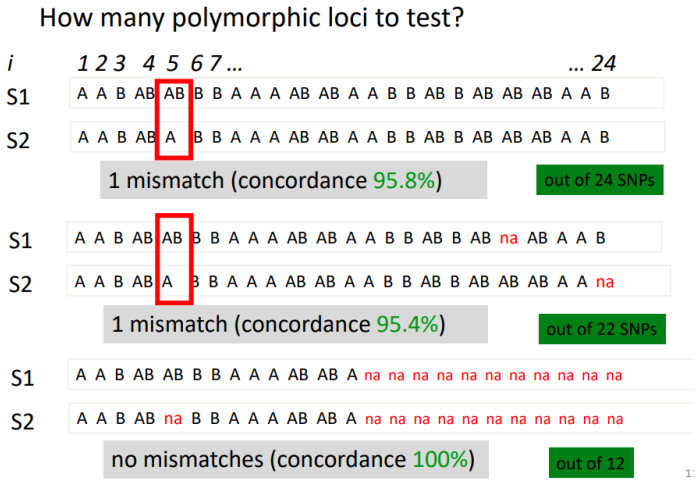


Figure 3.2: Effect of the error rate on concordance

3.2.5.1.1 An example on the effect of experimental mismatches In each example in figure 3.2 two sample for which 24 SNPs are called are analysed. To determine the difference of the two samples genotype for each position is considered and mismatches counted. In the figure:

- 'A' stands for homozygous for the reference.
- 'B' stands for homozygous for the alternative.
- 'AB' stands for heterozygous.

Analysing now SNPs calls:

- In the first test over the 24 loci, there is only one mismatch. The level of concordance is 95.8%. So the individual are highly related or the DNA come from the same sample.
- In the second test there is only one mismatch but there are position without a call. Therefore the concordance is measured out of 22 SNPs and is equal 95.4%.
- In the third test there are a lot of missing calls, so only 12 SNPs are available and concordance is 100%.

This different level of concordance is due to missing calls in the second and third test. The first is the most accurate as it considers the greater number of SNPs between the three, providing the most reliable information.

3.2.5.2 Biological mismatches

In the context of disease samples and tumours, many somatic events can happen, like deletions, gains of copies or homozygous deletions. This events can change the genetic make-up of a person and have to be taken into account when performing a genetic test.

3.2.5.2.1 Loss of Heterozygosity (LOH) LOH is an event that results in the loss of one parental copy of a region which results in the genome having just one copy of that region. If that region contained a heterozygous locus, there will be loss of heterozygosity. Its probability of arising is:

$$P(AB, A) = P(AB) \cdot P(A|AB)$$

3.2.5.2.2 Gain Of Heterozygosity (GOH) GOH is due to a mutation in a site. It is polymorphic through inheritance. These types of events are pretty rare and their probability of arising is:

$$P(A, AB) = P(A) \cdot P(AB|A)$$

3.2.5.2.3 Double Mutation (DM) Double mutations are event when a mutation occurs on an already mutated event. They are very rare and their probability of arising is:

$$P(A, B) = P(A) \cdot P(B|A)$$

3.2.5.2.4 Modelling biological mismatches Biological mismatches can be modelled in an assay in a data driven way, assessing the error rate for genotyping for some specific SNPs or running specific tests. The resolution for the errors of these mismatches can be patient-wide or tissue-specific.

3.2.6 Project regarding SNPs

Some useful projects developed over the years are:

- dbSNPs.
- HapMap3.

3.2.6.1 dbSNPs

dbSNPs is a database of small scale nucleotide variants. It includes both common and rare single base nucleotide variation (SNV), short ($\leq$ 50bp) deletion/insertion polymorphisms, and other classes of small genetic variations. It can be found on `https://www.ncbi.nlm.nih.gov/snp/`.

3.2.6.2 HapMap3

HapMap3 is the 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. It can be found on `https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html`

3.3 Genetic distance

Having defined the number of SNPs to use, with maximum MAF and other amenable characteristics, the genetic test should provide a measure of the genetic distance between individuals associated with a probability of the measure to be correct.

3.3.1 Measuring distance

As a simple measure, the number of loci where two samples show different genotype can be counted and normalized on the total number of queried loci, defining a certain level of discordance. The output value will be the genetic distance between the two samples given the selected loci, which will be proportional to the number of discordant calls. In figure 3.3 a typical graph used to measure the genetic distance using SNP-based genetic testing is depicted. 4 samples with a set of 5 SNPs for each one are analysed. The distance is measured among all possible pairs, whose indexes are reported on the x-axis.

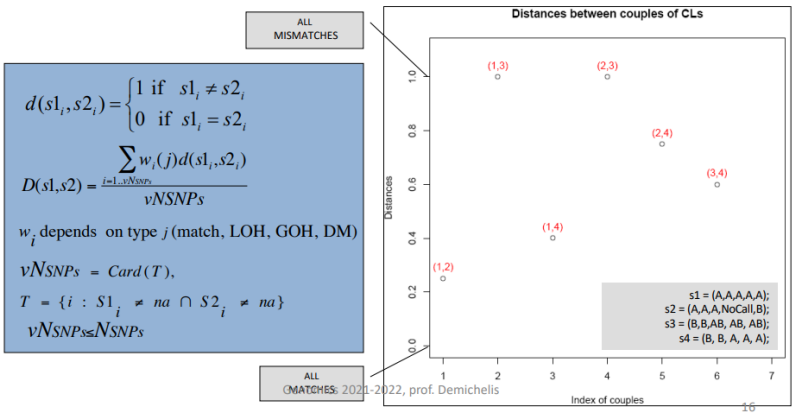


Figure 3.3: Genetic Distance graph with 4 samples. On the y axis the genotype distance is represented, varying from 0, total concordance to 1 total discordance. On the x-axis all possible pairs in the dataset are included.

In particular in figure 3.3:

- $s1$ and $s2$ have 3 equal calls, one locus without call and a mismatch. The genetic distance is 0.25.
- Samples $s1$ and $s3$ have 5 mismatches out of 5 SNPs, so the genetic distance is 1.

The equation in 3.3 is the one used to compute genetic distance. The weight w_i can be used to associate different importance to different mismatches. The distance is normalized by the total number of SNPs for which there are available calls $vNSNPs$, such that $vNSPNs \leq NSNPs$, the total number of SNPs considered in the test.

3.3.1.1 Expected distance

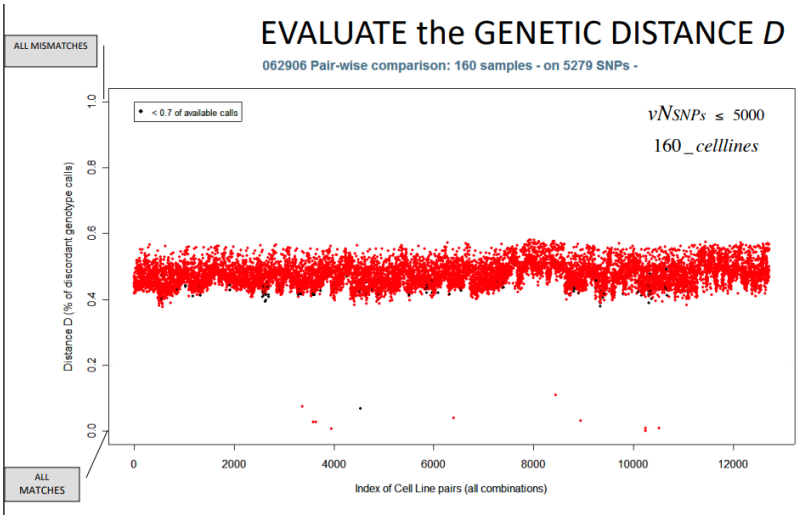


Figure 3.4: Genetic Distance graph with 160 samples

Figure 3.4 shows the distance, measured by genetic fingerprinting, of a collection of 160 samples of cell lines. There are $\frac{160 \cdot 159}{2}$ possible pairs. Applying the distance measure to a larger collection of samples an average distance among all possible pairs is expected. This, for different samples will be close to 1, never reaching it due to genetic variance and errors. In this case the average distance is around 0.5, since by chances some samples share some genotypes. Some pairs have a low distance: this is due to a mislabelling to the cell lines used: some believed to be two different cell lines were in fact the same one.

3.3.1.2 Identifying a sample origin from RAP samples

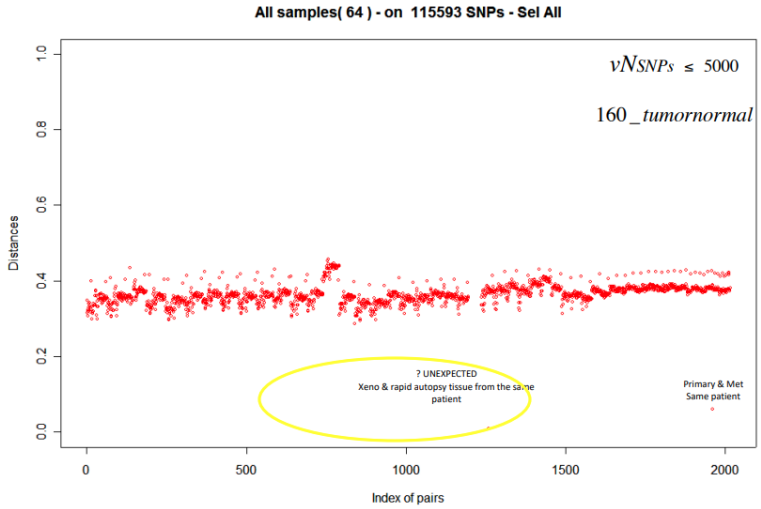


Figure 3.5: Genetic Distance graph with tumour samples

Figure 3.5 depicts a genetic fingerprinting experiment performed on a collection of 160 tumour samples, with a SNP array composed of more thant 100000 SNPs. Two samples with very low distance can be observed. One of the is from a Rapid Autopsy program and the other from a xenograft model. Rapid autopsy programs or RAP are programs for which patients at the end of their life agree to donate their tumour tissues for research purposes. The material must be taken within two hours after death. Those sample are usually highly characterized but the patient's identity is lost. So the metastasis model used for the xenograft was the one obtained from the RAP.

3.3.2 Distance changes with different numbers of SNPs

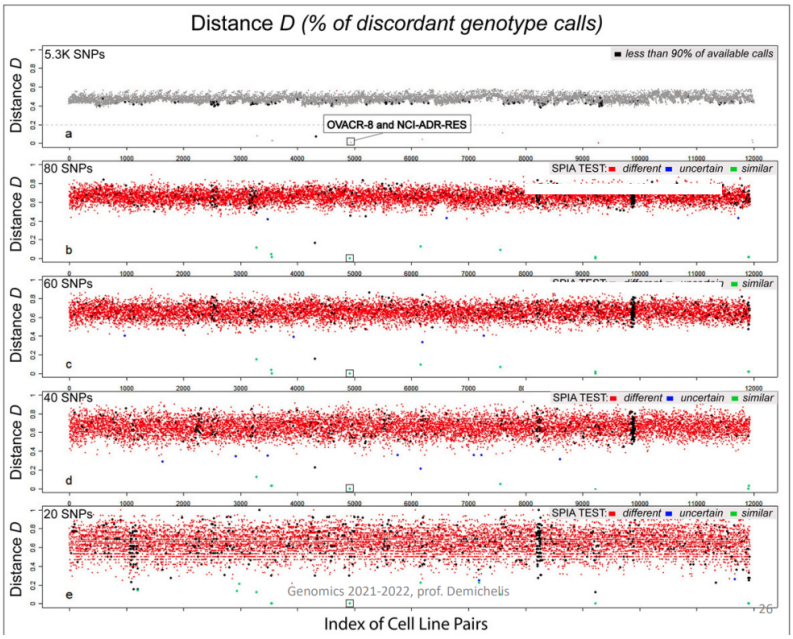


Figure 3.6: Genetic Distance graph at decreasing number of selected SNPs

Figure 3.6 depicts an experiment in which 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. It is noticeable that, in the second plot where 80 SNPs matching the required characteristics were selected, the average distance and the standard deviation across all pairs is higher than in the previous example, in which all available SNPs were used. Decreasing the number of SNPs to 60, then to 40 and 20 leads to have the same average distance between pairs, which settles around 0.66, but higher standard deviation. So enough SNPs in order to prevent unexpected issues and to trust the measure are needed. Maximizing the likelihood that SNPs have a different genotype, the average distance of unrelated individuals will increase.

3.4 Building a SNP-based genetic test

3.4.1 Pipeline

Building an identity test base on SNPs is a multi-step process, consisting in:

1. Definition of a genetic distance to compare samples.
2. Definition of SNPs requirements, 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;

- MAF and Hardy-Weinberg equilibrium from HapMap data.

4. Implementation of a probabilistic test of classification.
5. In silico validation on independent datasets.
6. Validation on cell lines genotyped on independent platform.

3.4.2 Implementation of a probabilistic test to identify samples

When classifying samples there is a need to assess the threshold on the genotype distance to assign samples to a class, the confidence of the classification and the minimum number of loci needed for a robust test. A probabilistic test could be used to define the threshold and the confidence. The gold standard to do so is to compare observation with expectations. Assuming that SNP calls at different loci are independent, this process can be modelled as a binomial distribution: each SNP call is a trial, with n the number of SNPs in the assay, k the number of matches, p the probability of a match and $1-p$ the probability of a mismatch. The probability of having k matches out of n SNPs follows a binomial distribution. Moreover with n, np and $np(1-p)$ large enough, the Gaussian approximation of the binomial distribution can be used. Doing so the probability of having k matches out of n SNPs can be modelled as a Gaussian distribution with:

$$\mu = np \qquad \wedge \qquad \sigma = np(1-p)$$

The area of confidence o the assay can be defined as $1m \pm m \cdot \sigma$, where m is the number of standard deviations used to define the threshold.

3.4.2.1 Classification

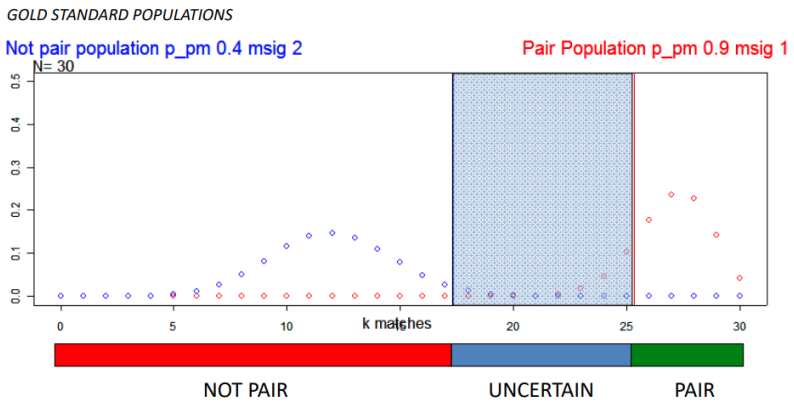


Figure 3.7

Having modelled the match distribution, the probability mass function for unrelated individual (blue dotted line in figure 3.7) can be plotted The same can be done with the probability mass function for related individual (red dotted line in figure 3.7). These functions can be used to set thresholds to define 3 regions dependent on the number of matche:

- Not pair: samples in this region are different.
- Pair: samples in this regions are similar.
- Uncertain: no certain result.

The grey area changes with m and defines the level of confidence. Decreasing the number of SNPs the grey zone will decrease, making the results difficult to interpret.

3.4.3 Some examples of genetic tests

3.4.3.1 Investigating cell line passages

A massive use of these genetic tests is done to assess genetic changes during in-vitro cultivation and in studies for tumor evolution, lineage plasticity and heterogeneity across metastasis across individuals or in 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.

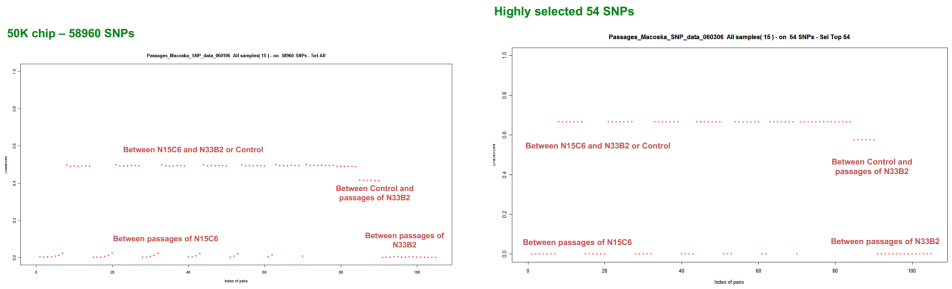


Figure 3.8: Genetic distance between cell line passages

In the example 3.8, two types of prostate cell lines which underwent multiple passages were used:

- $N15C6$ from passages from 48 to 63.
- $N33B2$ from passages from 21 to 39.

The cell lines were profiled with a SNPs array and a genetic fingerprinting test was performed. All passages of each cell lines were compared with all other passages. All passages from the same cell line should be classified as similar in the test. However the results obtained using the full array of SNPs (50k), showed that some pairs which should be exactly identical are different. They can be seen on the bottom-left of 3.8. Using only 54 SNPs this distance is not detectable. This increased distance is due to a major difference for certain passages with respect to the initial one on chromosome 11 of $N15C6$. This is due to the insertion in chromosome 11 of a viral sequence to immortalize the cell line.

3.4.3.2 Investigating individual relatedness

The HapMap consortium sequenced hundreds of individuals and trios for different ethnicities and also used trios. Trio sequencing is a technique which involves the sequencing of the genome of mother, father and their child. Trios provide major information of haplotype blocks and for identifying regions related to inheritance.

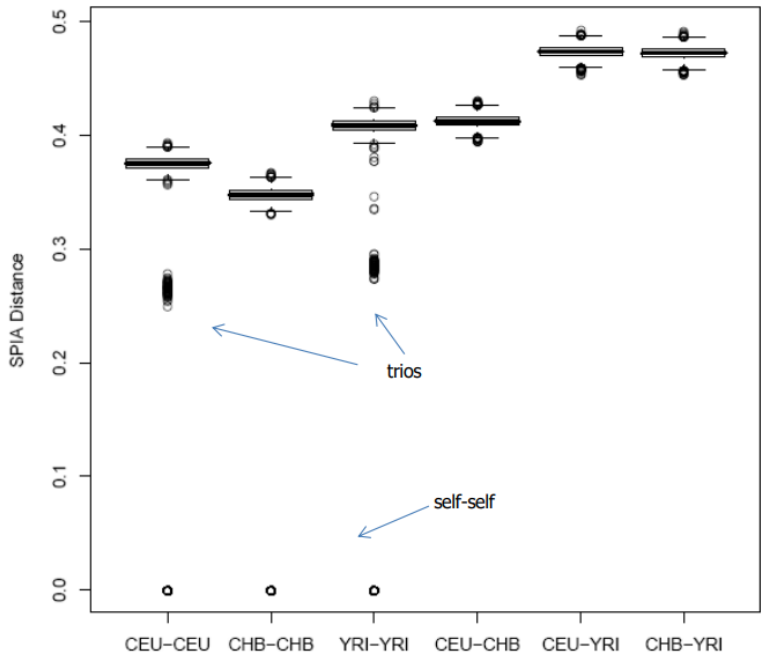


Figure 3.9: Trio genotype distance

Figure 3.9 depicts the genetic distance of individuals computed by the SPIA Assay. Self-self pairs have a genetic distance of 0. Each ethnicity has an average genetic distance different from the global one and trios have a lower distance from individual from the same ethnicity. Different in MAF in population could explain differences in distance among mixed samples. This shows how this type of test can be used for paternity tests of for forensic science.

3.4.4 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. Understanding the genetic structure of human populations is of fundamental interest to medial, forensic and anthropological sciences. Knowing the genetic substructure of data is important because:

- 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 disease 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 GWAS study using a specific ethnicity a worldwide marker for susceptibility for a molecule or disease cannot be trusted.

In medicine and in the study of human evolution 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 and those could be used to infer the individual genetic background in terms of 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 different ethnical groups can be distinguished.

3.4.4.1 Genes mirror geography in Europe

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

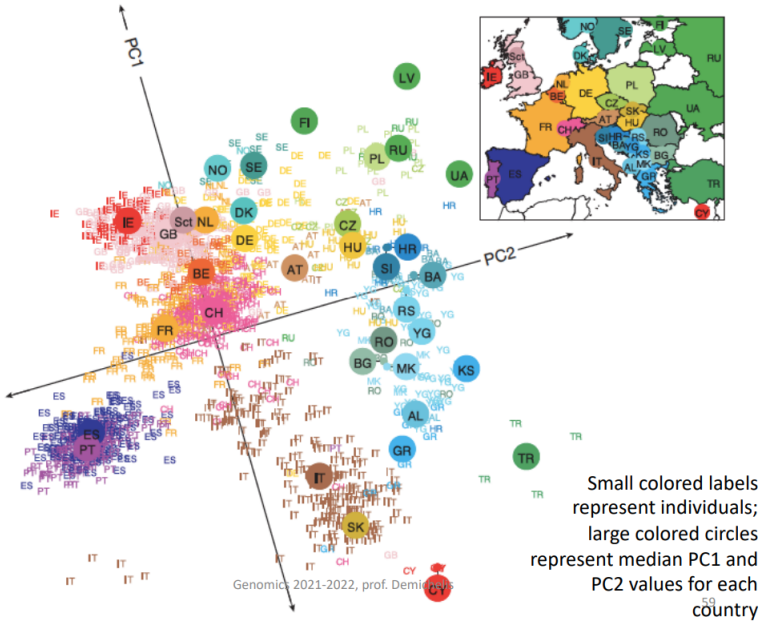


Figure 3.10

Properly selecting variants it should be possible to distinguish individuals coming from different

countries. SNPs were selected following the genetic fingerprinting approach, adding for SNPs with different MAF in different population. More dense and distant clusters could be due to the fact that SNPs selected are typical for that area and are able to maximize the difference with respect to it.

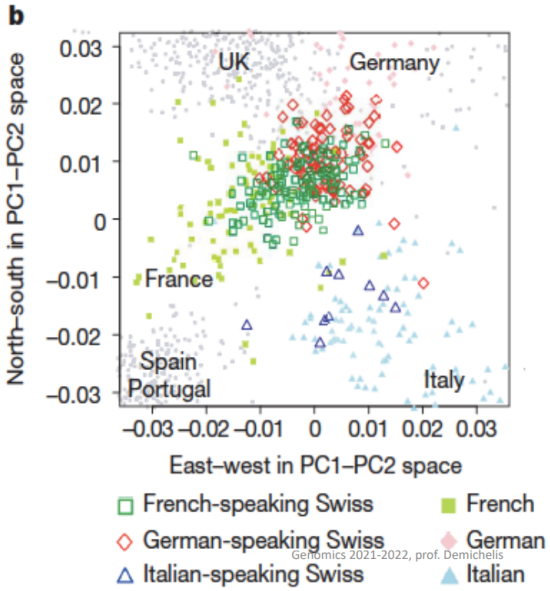


Figure 3.11

Focusing on Switzerland, they could even make inference on the linguistic canton 3.11. It is possibly true that in country where some regions have very different habits might lead to have similar genetic fingerprint. Moreover low-frequency alleles tend to be the result of a recent mutation and are expected to geographically cluster around the location at which the mutation first arose. Hence, they can be highly informative about the fine-scale population structure. Despite low average levels of genetic differentiation among Europeans, close correspondence between genetic and geographic distances was found. When mapping the genetic basis of a disease phenotype, spurious associations can arise if genetic structure is not properly accounted for.

Chapter 4

IGV (Integrative Genomics Viewer)

4.1 IGV characteristics

4.1.1 Introduction

The human genome nowadays is being explored extensively thanks to:

- Exons and whole-genome sequencing.
- Epigenetic surveys.
- Expression profiling of coding and noncoding RNAs.
- SNPs and copy number profiling.
- Functional assays.

IGV allow to visually explore the data generated from this kind of study, which is mostly used for the development of precision medicine, an approach for disease treatment and prevention that takes into account individual variability in:

- Genes.
- Living environment.
- Lifestyle.

The objective of precision medicine is to define which drug, at what time and at what dose to administer to an ill individual to have optimal response. The IGV software is a 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. In IGV, each vertical bar corresponds to a read. A coloured sign in a read indicates the presence of a polymorphism. The browser also gives info about the quality of the read and the bases from the loaded BAM file. Other information regarding IGV are present in the Supplementaty information - Integrative Genomics Viewer pdf file.

4.1.2 Main uses of IGV

Some utilization of IGV are:

- NGS alignment.
- Epigenomics studies.
- Copy number evaluations.
- RNA-sequencing.
- Identification of variants and genotypes.

Some of the main utilization are represented in figure 4.1.

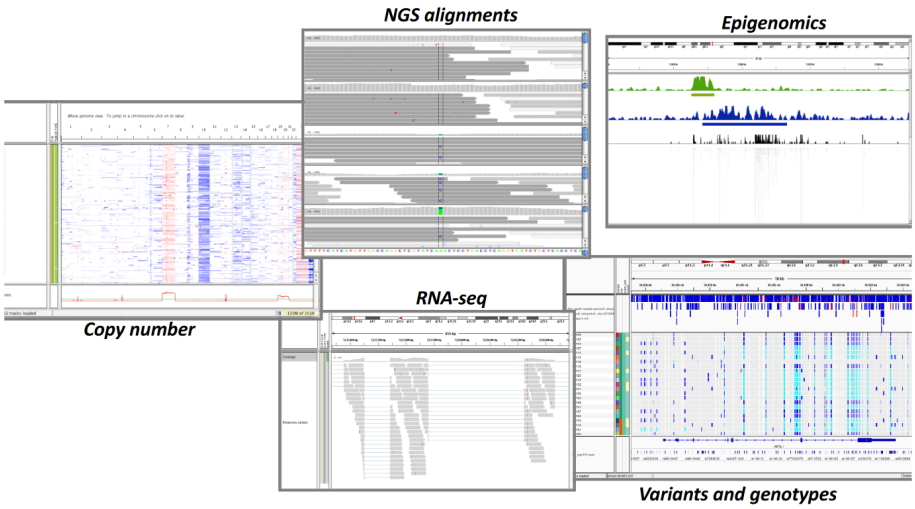


Figure 4.1: Main type of data exploration done with IGV.

4.1.2.1 RNA sequence analysis

IGV can be used to explore RNA sequence analysis. For example figure 4.2 represents reads that span exon junctions, with their heights depending on the number of reads that connect them.

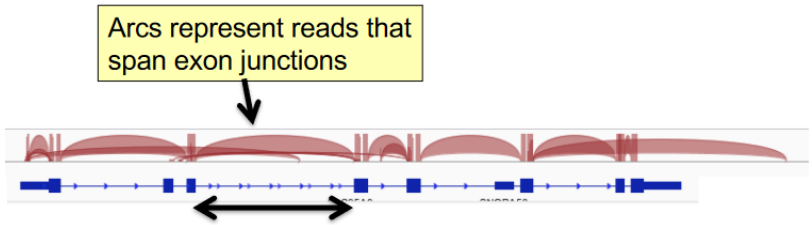


Figure 4.2: View of RNA sequence data for a gene.

Another thing that can be done are Sashami plots. In these plots the number of reads connecting exosomes are represented on curved lines. The peaks represents coverage within exons. An example of a Sashami plot is depicted in figure 4.3.

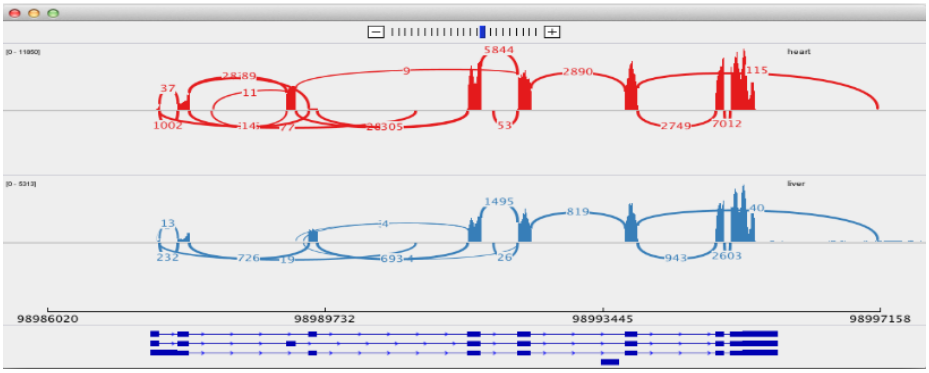
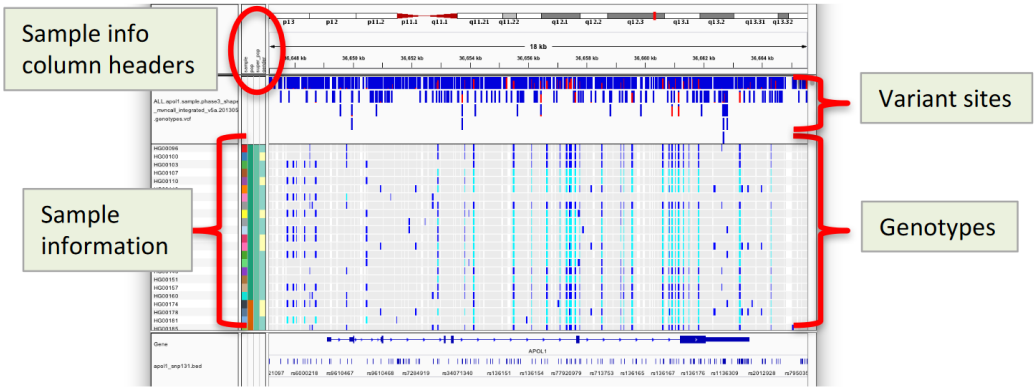


Figure 4.3: Sashami plots

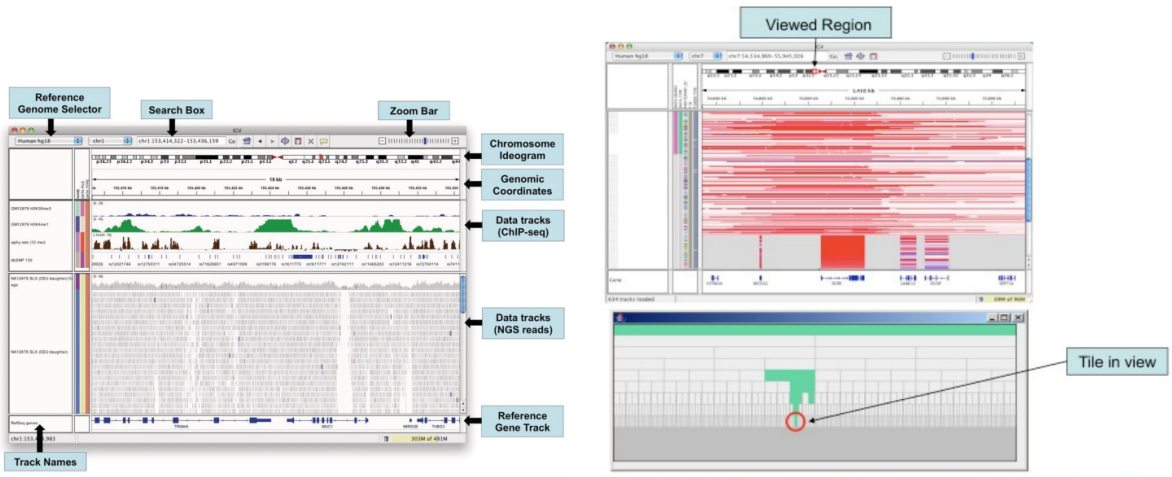
4.1.2.2 Variant discovery

IGV allows to study variants from different samples, like it can be seen in figure **??**



4.1.3 Operations

IGV allows for the simultaneous visualization of an arbitrary number of BAM files. The user can move, zoom in and out quickly over different genomic scales, panel *a* of figure 4.4), and also to jump in precise positions of the sequence. It is possible to search for genomic coordinates or gene names. For each resolution scale called zoom levels, the aggregated data is divided into tiles, panel *b* of figure 4.4) that correspond to a region viewable on a typical user display. Each tile is subdivided into bins, with the width of a bin chosen to correspond to the width represented by a pixel at that resolution scale. It is also possible to sort the samples in different ways and to group them considering different characteristics.



(a) IGV interface main features

(b) Tiles view in IGV

Figure 4.4

Navigation through a data set is similar to that of *Google Maps*, allowing the user to zoom and pan seamlessly across the genome at any level of detail from whole genome to base pair. Annotations for specific genomes could be found consulting the UCSC Table Browser (UCSC table).

4.1.4 Tiles

The corresponding data tiles for each zoom level are stored in the binary Tiled Data Format, or TDF, which has been optimized for fast tile retrieval. A tiled data file (.tdf) is a binary file that contains data that has been preprocessed for faster display in IGV. TDF files are generated by using the *igvtools* package (*toTDF* command). Tile sizes for each zoom level are constant and small. A single tile at the lowest resolution (spanning the entire genome) has the same memory footprint as a tile at the very high zoom levels (might span only a few kilobases). Tiles no longer in view are discarded as needed to free memory.

4.1.4.1 Pixel resolution error

Pixel resolution errors occur when data density exceeds the constraint given by the number of pixels available for display. This can be solved through data aggregation. As the user zooms below the $\sim 50kb$ range, individual aligned reads become visible, like in figure 4.5. It is possible then to zoom further, and see the bases at each position.

4.1.4.2 Data displayed at different resolutions

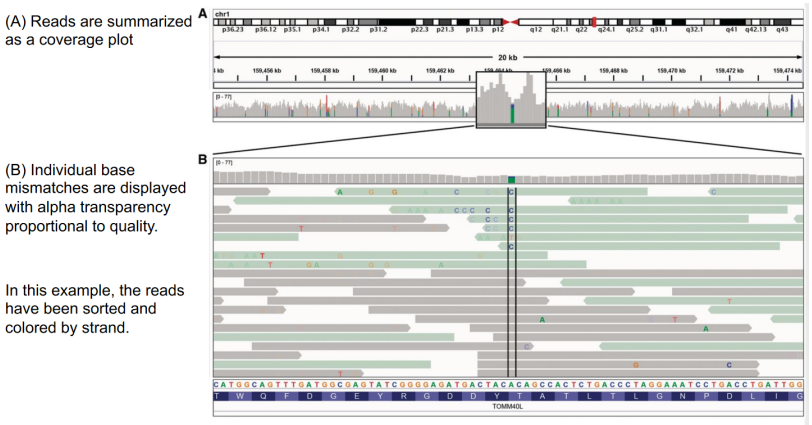


Figure 4.5: Different resolution of IGV

In panel a of figure 4.5 reads are summarized as a coverage plot. In panel b individual base mismatches are displayed with alpha transparency proportional to quality. In this example, the reads have been sorted and coloured by strand and span a $20kb$ genomic region for 1 individual. The coverage is between 0 and 77 at certain positions there is more than one base supported by the reads. This is visualised by the green and blue column: some reads support C instead of A. The coloured area is proportional to the allelic fraction for each base.

4.1.5 IGVTools

Igvtools comprises a set of utilities to prepare large files for efficient display. Some commands are reported in figure 4.6. In particular the count function takes as input a BAM file to generate coverage data. The obtained file could be then loaded with the "Load pre-computed coverage data" command for faster loading times.

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

Figure 4.6: igvtools possible operations.

4.1.6 Session files

Sessions allow users to share their data and views with other users simply and accurately. Session files describe the session state in XML.

4.2 Interpreting pair orientations

IGV allows to discover genetic aberrations through the interpretation of pair orientation. A paired end protocol can uncover:

- Inversions.
- Duplications.
- Translocations.

To detect these kind of events reads that span the breakpoint are needed. This type of reads can be obtained through long reads or paired-end sequencing.

4.2.1 Element to consider when interpreting pair orientations

Different parameters are to be considered when interpreting pair orientations:

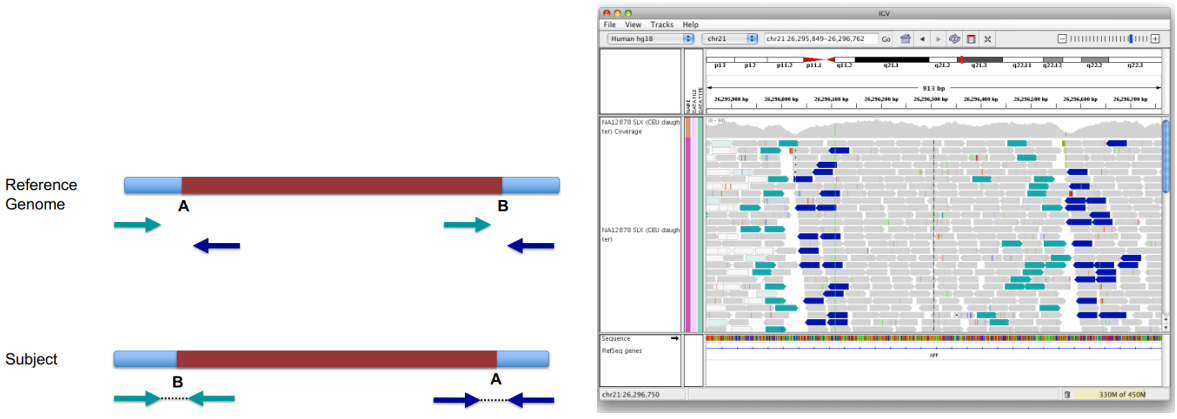
- Pair ends relative orientation.
- Insert size length.
- Coverage within the aberrant region.
- Coverage outside of the aberrant region (flanking genomic segments).
- Coverage at the breakpoints.

In particular when considering pair ends relative orientation some considerations can be done:

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

4.2.2 Inversion

Figure 4.7 depicts an inversion. A local drop in coverage and reads direction *LL* or *RR* can be observed. Moreover a drop on the breakpoint can be observed.

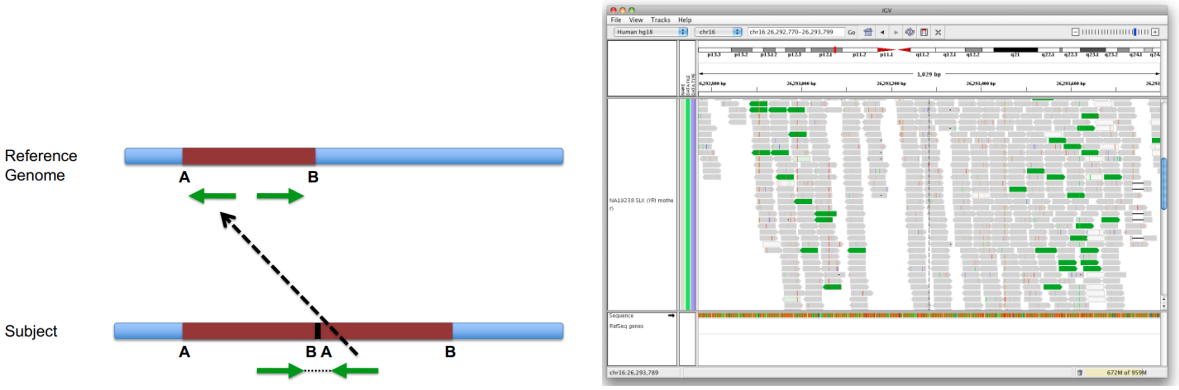


(a) Schematic (b) IGV view

Figure 4.7: Inversion discovering exploiting PE reads.

4.2.3 Tandem duplication

Figure 4.8 depicts a tandem duplication. All the reads that do not cover the junction point align perfectly to the reference. Moreover it can be observed a 50% increase of coverage proportional to the extra copy. Junctions A and B are not modified. A read mapping BA would be partially aligned at B on the reference. There is no drop in coverage.

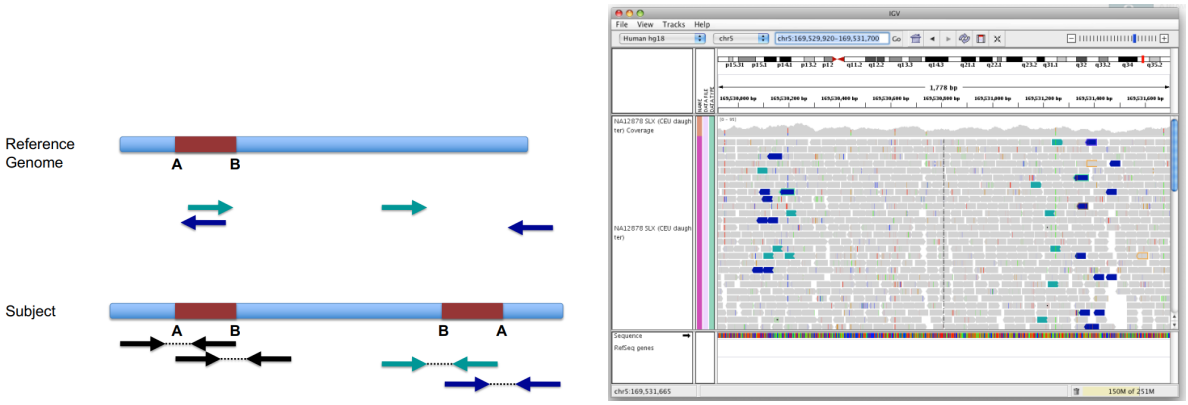


(a) Schematic (b) IGV view

Figure 4.8: Tandem duplication discovering exploiting PE reads.

4.2.4 Inverted duplication

Figure 4.9 depicts an inverted duplication. Coverage increase in the duplicated site in the reference gene is expected. Both A and B on the first segment are LR, while the second will have reads oriented LL or RR. The insert size will be significantly longer. An overlapping of left and right reads can be found on the reference.



(a) Schematic (b) IGV view

Figure 4.9: Inverted duplication discovering exploiting PE reads.

4.2.5 Deletion

Deletion can be discovered by observing drop in coverage or the observed distance between reads, which gives a clean indication of the size of the deletion. For small deletions like indels the sequence within the reads has to be observed.

4.3 Uncovering genetic aberrations - some examples

4.3.1 First example

Figure 4.10 depicts the genomic region chr1:11,050,009-11,055,137. The event visualized could be a tandem duplication on one of the two alleles and a deletion on the other allele. This is due to the fact that despite of the duplication the coverage on remains constant on the region.

Figure 4.10: A tandem duplication followed by a deletion

4.3.2 Second example

Figure 4.11 depicts the genomic region chr5:9,410,315-9,413,699. It is clear due to the absence of coverage how both allele of that region were deleted.

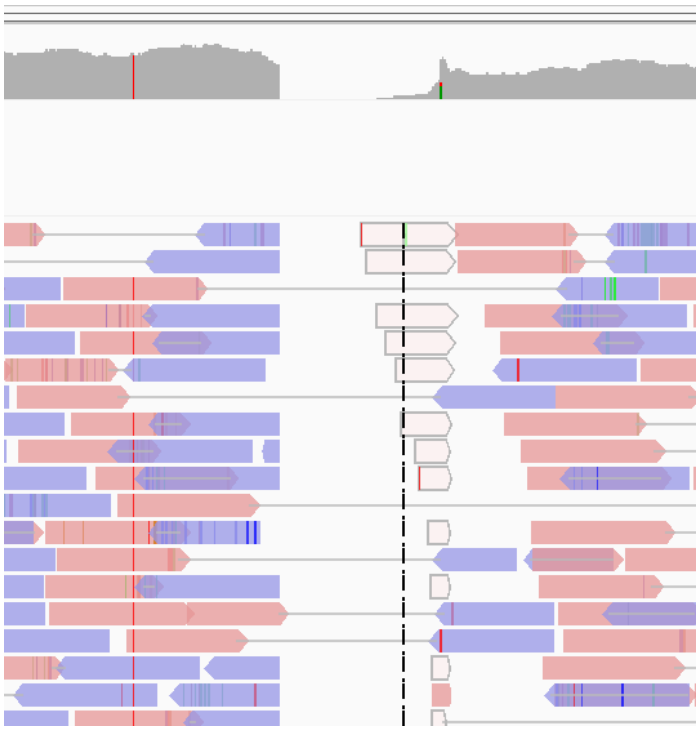
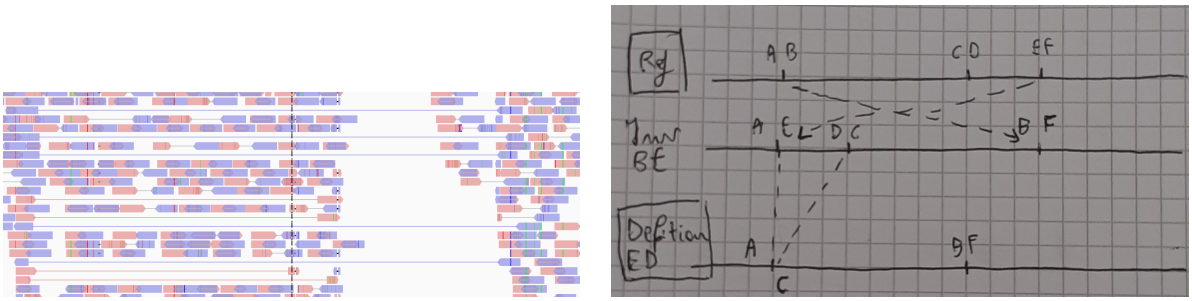


Figure 4.11: An homozygous deletion

4.3.3 Third example

Figure 4.12 and 4.13 depicts an inverted duplication followed by a deletion. This prediction is justified because of the presence of overlapping LR reads and LL and RR reads with a correspondent increase of coverage, followed in the following region by a considerable drop in coverage, indicating an heterozygous deletion.

Figure 4.12: Coverage for an inverted duplication followed by a deletion



(a) IGV view (b) written passages

Figure 4.13: Coverage for an inverted duplication followed by a deletion

Chapter 5

Tumor Evolution Studies via NGS data

5.1 Tumour evolution

5.1.1 Introduction

To be fully able to treat cancer it is important to understand what are the somatic events that occur during tumor genesis and evolution and when they arise. Cancer cells accumulate mutations due to both cell division and toxic agents like radiations or UV light. These mutations are maintained by the cell and lead to clonal expansion. Cancer could then originate or evolve through an arbitrary number of driver mutations that could, for example, activate an oncogene or disrupt some molecular pathways.

5.1.1.1 Typical traits of cancer

Typical traits of cancer are:

- Cancer is a dynamic disease: tracking its evolution if fundamental.
- During the course of disease, cancers generally become more heterogeneous, which is often related to treatment resistance.
- The bulk tumour includes a diverse collection of cells harbouring distinct molecular signatures with differential levels of sensitivity to treatment.
- This heterogeneity might result in a non-uniform distribution of genetically distinct tumour-cell sub-populations across and within disease sites (spatial heterogeneity) or temporal variations in the molecular make-up of cancer cells (temporal heterogeneity).

5.1.1.2 Tumour boards

Tumour boards are teams of specialists that follow a cancer patient through treatment. They collect data and suggest course of actions during the illnesses and can teach new doctors how to manage difficult cases.

5.1.2 Heterogeneity

Every site of the genome with somatic mutations is going to be differentially represented in tumour samples. Heterogeneity drives cancer resistance and is a great obstacle in treatment. Therefore, an accurate assessment of tumour heterogeneity is essential for the development of effective therapies. Emerging techniques to study with considerable potential to dissect the complex clonal architecture of cancers heterogeneity are:

- Multi-region sequencing.
- Single cell sequencing.
- Analysis of autopsy samples.
- Longitudinal analysis of liquid biopsy samples.

However, techniques to study tumour heterogeneity are hindered by intra-patient heterogeneity, which can be spatial or temporal. Figures 5.1 describes tumour heterogeneity.

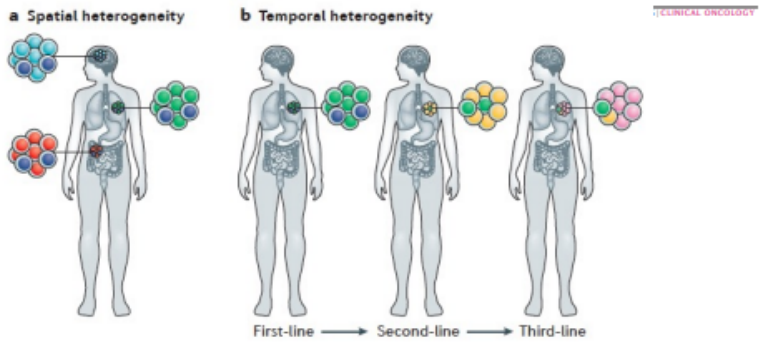


Figure 5.1: a) Spatial heterogeneity denotes an uneven distribution of cancer subclones across different regions of the primary tumour and/or metastatic sites. b) Temporal heterogeneity refers to variations in the molecular make-up of a single lesion over time, either as a result of natural progression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

5.1.2.1 Temporal heterogeneity

Temporal heterogeneity refers to the change in time of a single tumour mass. This can happen naturally or under particular selective pressures.

5.1.2.2 Spatial heterogeneity

Spatial heterogeneity describes different independent tumour masses that can be found in patients which share certain cells while having a unique genetic make-up.

5.1.3 Type of evolutions

It could be possible that some cells positively respond to treatment and others not, creating a heterogeneous population in the mass. The features of this set of cells changes over time. This evolution happens either because the new population replaces the older, or there's a branching and

the tumour mass becomes heterogeneous. Moreover a metastatic mass could have a monoclonal (from one mass) or polyclonal (from different masses) origin.

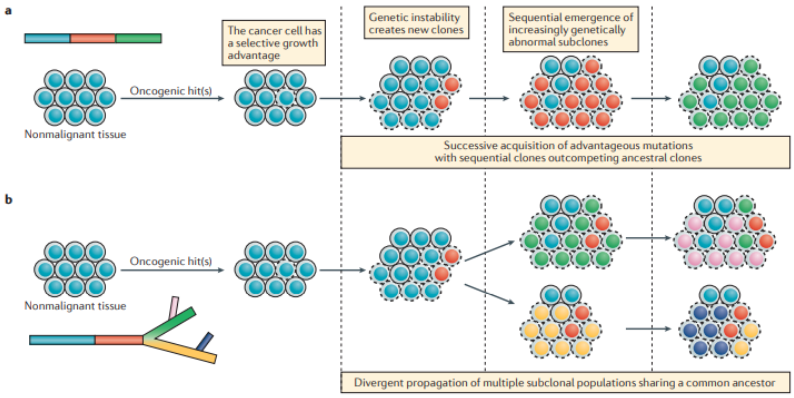


Figure 5.2: a) everything branches out from the monoclonal origin, but b) polyclonal origin, independent metastatic processes. Cells from independent lesions meet and form a highly diverse metastatic tumour.

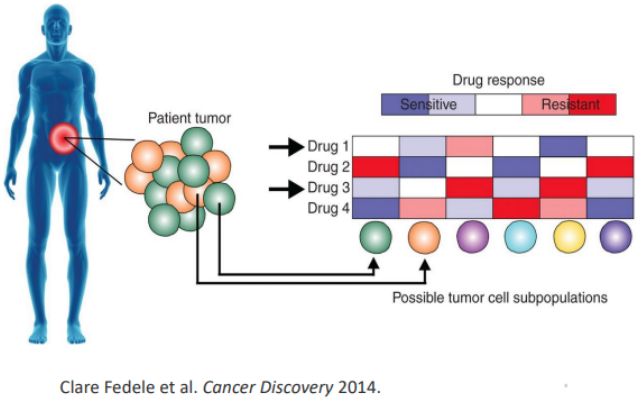
5.1.3.1 Linear evolution

Sequential genetic alterations confer a fitness advantage such that successive generations are able to out compete the preceding clones, which lack this fitness advantage. Surviving dominant clones harbour the ancestral mutation.

5.1.3.2 Branched evolution

Multiple genetically distinct populations can emerge from a common ancestral clone, with certain subclonal populations diverging from the common ancestor before others.

5.1.4 Treatment resistance



Clare Fedele et al. *Cancer Discovery* 2014.

Figure 5.3: Certain cells of the tumour mass respond to treatment and some don't. Red cells are resistant to the drug, while blue cells are sensitive to it.

Tumour resistance to treatment can be encoded in the original cells or can be driven by the treatment. Figure 5.4 depicts the processes that drive resistance that originates from treatment. This can be due to the selection of clones that provide resistance or the transformation of clones under treatment pressure.

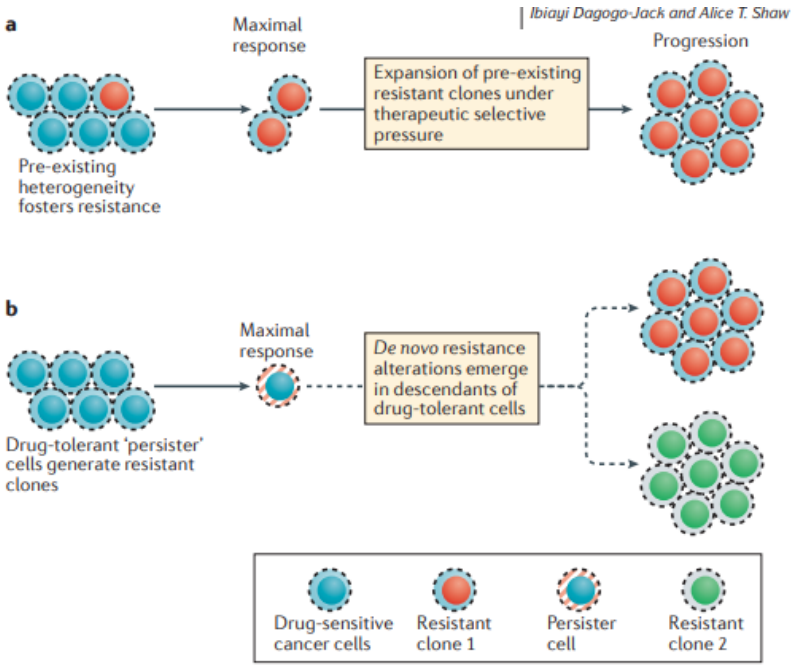


Figure 5.4: Tumor cells evolution driven by treatment.

5.1.4.1 Primary resistance

Pre-existing heterogeneity fosters resistance. Only susceptible cells die and resistant cells continue dividing, allowing the tumour to regrow. An high tumour mass will be visible.

5.1.4.2 Acquired resistance

Drug-tolerant, persistent cells generate resistant clones. At the time of diagnosis there are no markers of the resistance alteration because it is developed afterwards. The resistant cells will form over time new tumours not responding to the drug.

5.2 Using NGS data to uncover tumour evolution

5.2.1 Introduction

A tumour is a collection of multiple independent lesions. Sequencing can be used to obtain a representation of tumour burden and of the features of all of the different areas. In order to study tumor evolution common and private lesions across multiple samples from the same individual can be used to reconstruct the evolutionary path. Data from multiple individuals can be used, selecting the most clonal lesion, to build a common clonal evolution map. From this type of data it has

been found, for example that in prostate cancer if CHD1 is mutated, then a subsequent PTNEN mutation is found. The interest is on finding which lesions occurred first and which is the model that fits better the data. In particular when dealing with tumour data there is a need to take into account:

- Intra tumor heterogeneity.
- Inter tumor/intra patient heterogeneity.
- Inter-patient heterogeneity.
- Clinical/treatment relevance.
- Time dependency.
- Admixture DNA (tumor purity).

This characteristics, if properly investigated, can provide insightful hints during the analysis.

5.2.2 Admixture

A sample coming from a patient's tissue contains multiple cell types. So a sample will never be composed only of cancer cells. DNA admixture refers to the percentage of cells that are not tumoral. Purity, instead, is the percentage of cancer cells in a sample. It is computed as:

$$Pur = 1 - Adm$$

A particular lesion is clonal if all tumour cells harbour it. If only a portion of cells harbour it it is subclonal. Purity and admixture can be used to distinguish between clonal and subclonal lesions.

5.2.3 Informative SNPs

SNPs can be exploited to characterize tumour evolution. Informative SNPs are heterozygous SNPs: the allelic fraction can be counted and the proportion of reads supporting the alternative base can be assessed. The allelic fraction will change for somatic events that involve the genomic locus of a particular SNP, allowing to detect them. This is represented in 5.5. For example in a deletion the allelic fraction of a heterozygous SNP will change from 0.5 to either 0 or 1. In particular when selecting informative SNPs to design an assay is important to consider their MAF together with data from databases like dbSNP to select the one that are most probable to are in an heterozygous genotype.

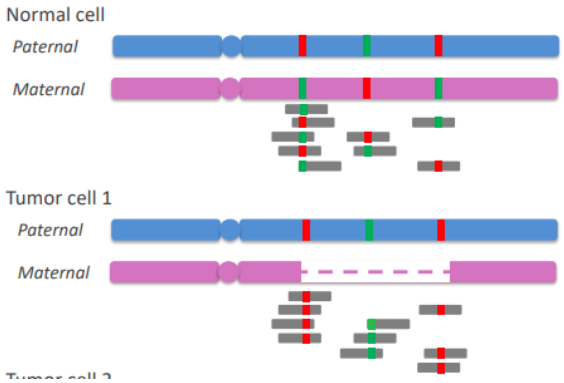


Figure 5.5: Thanks to the presence of informative SNPs is easy to detected the loss of an allele in the tumour cell.

5.2.4 Beta value

When dealing with non pure samples the β value needs to be introduced to uncover events concerning informative SNPs. This is because the allelic fraction signal will change based on the purity: in fact when only some cells in the sample are from a tumour the allelic fraction distribution will have two peaks. β is the percentage of neutral reads, or the number of reads that can be coupled, one with the reference base and one with the alternative, over the total number of reads at that SNP. In particular:

- $\beta = 1$ both alleles equally represented.
- $\beta = 0$ only one allele represented.

The more β is far from 1, the more the sample is admixed or the more the lesion is clonal. Moreover $Nref$, the percentage of the reference base in the non-deleted allele can be introduced. Figure 5.6 links together the concepts of β, sample purity and allelic fraction.

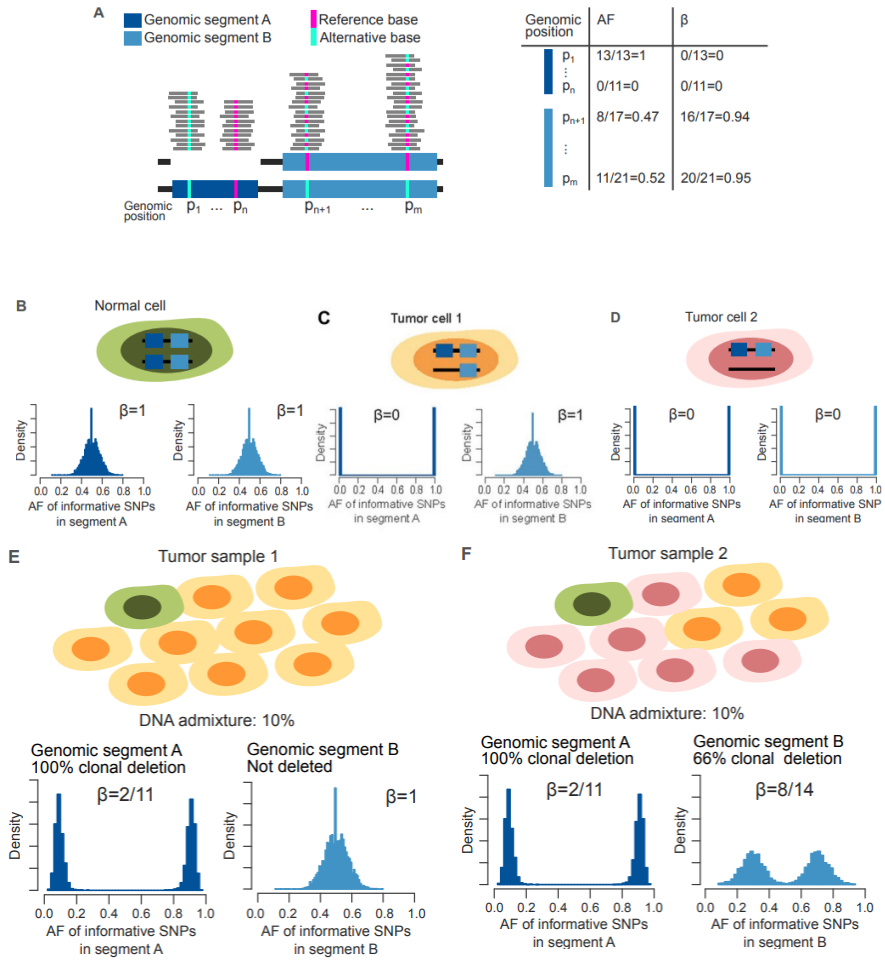


Figure 5.6: A)Example of the allelic fraction (AF) and beta (β) computed in five genomic positions (p_1 to p_m). Positions p_1 to p_n are within a hemizygous deleted genomic segment A, while genomic positions p_{n+1} to p_m lie within a wild type genomic segment B.
B-D) Examples of a normal cell and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

5.2.4.1 Computing beta

β can be computed for each genomic segment S:

1. Compute the observed distribution of the AF of informative SNPs in the genomic segment S.
2. Find the values of β and $Nref$ such that the expected distribution of the AF matches the observed AF.
3. Compute uncertainty around β as a function of:
   (a) The mean coverage of S.
   (b) The number of informative SNPs in S.

5.2.4.2 Effect of coverage on beta

The mean coverage of an experiment impact the accuracy of β. The more deep a sequencing is more two close peaks of AF can be distinguished. This is especially important when β is close to 0. An example of the effect of coverage on β is shown in figure **??**.

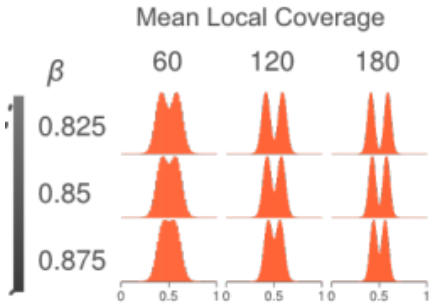


Figure 5.7: 2X experiment: on average there are 2 reads per gene. It is easy to not recognize the correct distribution.
10X experiment: it is easy to recognize the correct distribution. The process becomes more difficult for a tumour sample. This image shows how the higher the sequencing depth, the more the two peaks of the distribution are distinguishable.

5.2.5 Estimates of global and local admixture

To understand how to determine the admixture two samples will be considered.

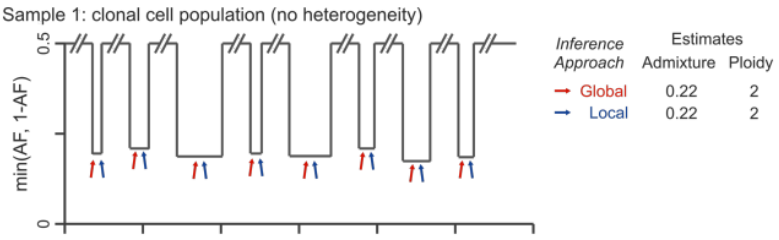


Figure 5.8: In the clonal cell population how much the distribution deviates from AF is the same globally and locally.

In sample one (figure 5.8) a clonal cell population is represented. So there is no heterogeneity. On the x-axis there are the genomic coordinates indexed by informative SNP, on the y-axis the

AF. Drops in the allelic fraction represent lesions, and it can be seen how all drops are of the same length on the y axis. This means that the amount of DNA loss or gain is identical in all of this drops. This means that the level of admixture remains constant over all the sample and is the same both globally and locally.

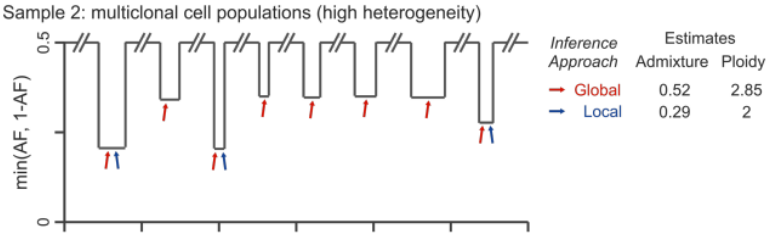


Figure 5.9: This sample harbours heterogeneity: the global and local values of admixture are different. A global value is relative to tumor purity, while local to clonality.

In sample two (figure 5.9) a multiclonal cell population is represented. Multiclonality is characterized by different depth of lesions. In fact the depth of the drop is proportional to the number of cells that carry that lesion. So, in this case the global admixture refers to tumor purity, while the local one to the clonality of the lesion in the diseased cell population.

5.2.5.1 Estimates of DNA admixture

The estimates of global and local admixture can be combined together with the β-value and the log2 ratio, as is depicted in figure 5.10. The log2 ratio is computed as:

$$\log_2 \frac{tumour_{samples}}{normal_{samples}}$$

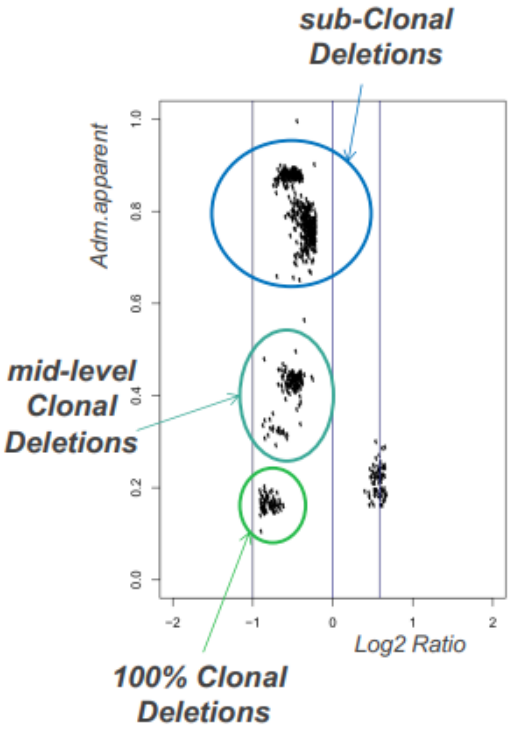


Figure 5.10: Each dot is a genomic segment and different clusters are visible. Lower clusters are used to determine the admixture while the other represent subclonal cells. Close points probably represents lesions that happened close in time.

On the x axis the log2 ratio is measured while on the y the apparent admixture, which is proportional to β. The log2 ratio allow to interpret information about every segment in the genome when coupled with β. Apparent admixture is computed as:

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

And associates an apparent DNA admixture to each monoallelic deletion. The clonality is computed from the apparent admixture and the global one:

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

5.2.6 PR-2741 - an example

Figure depicts data of a real case of prostate cancer in which a clear drop in coverage in region 2 of the $5th$ chromosome, while region 1 and 3 have equal coverage can be seen. The DNA present in region 2 could come either from admixing cells or from cells that do not have the deletion. Looking at the allelic fractions of region 1, 2 and 3, both from the tumour and the match normal normal sample more or less the same two modes of distribution are observed. In the tumour sample in fact the expected peaks at 0 and 1 are not observed. This could be due to signal coming from intervening

normal cells, that bring the modes to the center, or due to subclonality events. It is clear how a subclonality event could be uncovered through a lesion with admixture.

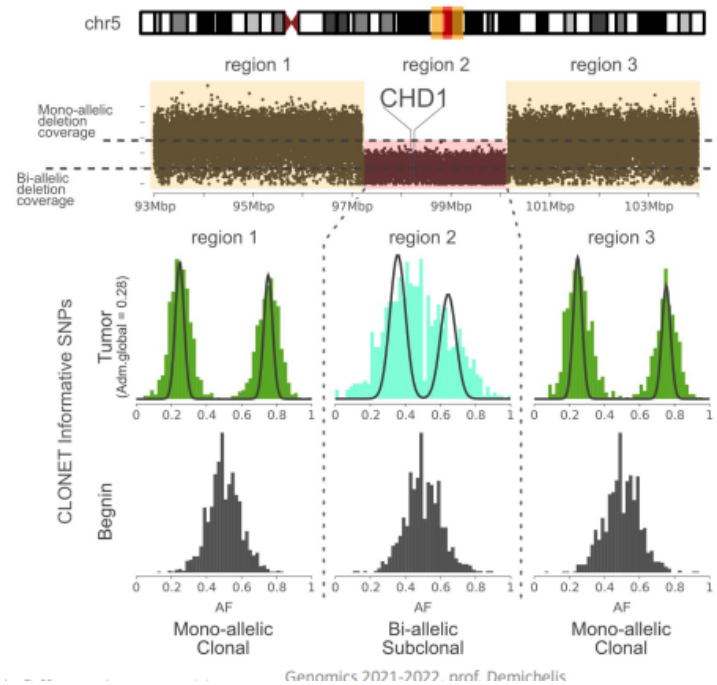


Figure 5.11: The distribution of heterozygous SNPs in benign cells is peaked at 0.5, while two modes in tumour cells are observed. The distance between the two modes is equal in 1 and 3. In the middle the two modes are moving towards the center, suggesting that the deletion is not likely 100% clonal or it is clonal and shifts is due to lack of deletion in 1 and 3

Chapter 6

Tumor evolution studies: copy number based methods

6.1 Analysis of clonality

6.1.1 Informative SNPs

Informative SNPs are at the basis of tumour evolution studies. An informative SNP is a SNP for which a specific individual has an heterozygous call. The set of informative SNP is unique for each individual. The somatic loss of an allele of a genomic region that contains an informative SNP will change its allelic fraction. In this way the allelic fraction of that informative SNP will be informative of the lesion and of its depth. In this way clonality of the tumour will be determined by the fraction of tumour cells that harbour the lesion:

- A set of lesions in a sample is said to be clonal when all tumour cells harbour it.
- A set of lesions in a sample is said to be subclonal when only a subpopulation of tumour cells harbour it.

6.1.2 Log2 ratio

The Log2 ratio is computed as:

$$\log_2 \frac{\#tumour\ content}{\#normalcontent}$$

This measure can be obtained through the intensity of signals in array data or through the local coverage over a tumour BAM file over a normal one. The log2 ratio help to uncover the ploidy and copy number changes of the tumour population in a sample. In particular:

- $\log_2 R < 0$ loss of copy number in tumour.
- $\log_2 R = 0$ the tumour is copy number neutral.
- $\log_2 R > 0$ gain of copy number in tumour.

6.1.3 Beta value

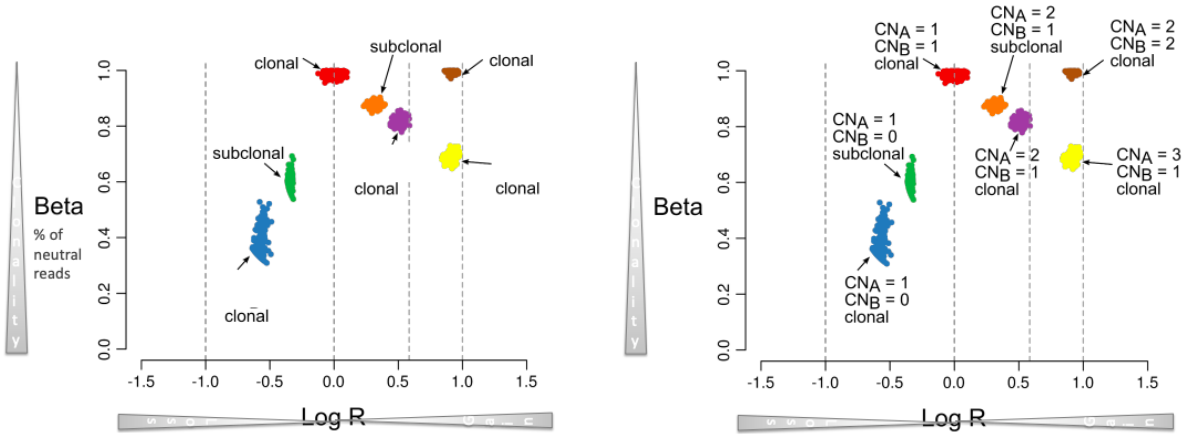
The β value is the percentage of neutral reads, or the number of reads that can be coupled, one with the reference base and one with the alternative, over the total number of reads at a SNP. β is used to evaluate the purity of a sample. In particular:

- When $\beta = 1$ the purity of the sample is 0, so there is no tumour content with copy-number variations and so no tumour signal.
- If $\beta = 0$ the purity of the sample is 1, so there is only tumour content with copy-number variations.

It can be seen how, the more β tends to 0, the higher the tumour content and the clonality of the sample.

6.1.4 Cluster analysis in the beta-log2 ratio space

Figure 6.1 is an example of a sample in which cells cluster in population in the $\log_2 R \times \beta$ space.



(a) Clonality (b) CNV

Figure 6.1: $\log_2 R \times \beta$ space of a tumour sample.

Considering the β value it can be said that lower clusters are more clonal, while considering the $\log_2 R$ value moving to the left side the more the cells have lost DNA, while moving to the right side the more the cells have gained DNA. Panel A allow the visualization of the cluster, while panel B adds information about the copy number changes of two specific alleles. The dotted lines are built in a data-driven manner and they represent events of deletion of gains for an allele. The informations that can be extrapolated from these graphs are:

- The blue cluster with deletions is the most clonal one as it is the lower one.
- Both blue and green clusters have deletions, since they have a negative $\log_2 R$, but the green one is less clonal than the blue one as it is higher. It can be seen hot both the clusters have the same copy number for the two alleles, and ho they have lost a copy for one and both for the other.
- The red cluster is found in $\log_2 R = 0$

and $\beta = 1$, so the cells in it are wild-type in terms of copy-number changes the total number of alleles is the same to healthy cells.

- All the other clusters with a positive $\log_2 R$ had a gain of DNA.

So, the $\log_2 R \times \beta$ space allow to map the status of clonality and the number of copies for a specific segment in the genome.

6.2 Evolution maps

6.2.1 Introduction

The information obtained on clonality can be used to build evolution maps. These maps are useful to track the evolution of a tumour over time with specific conditions like treatments.

6.2.2 Building an evolution map

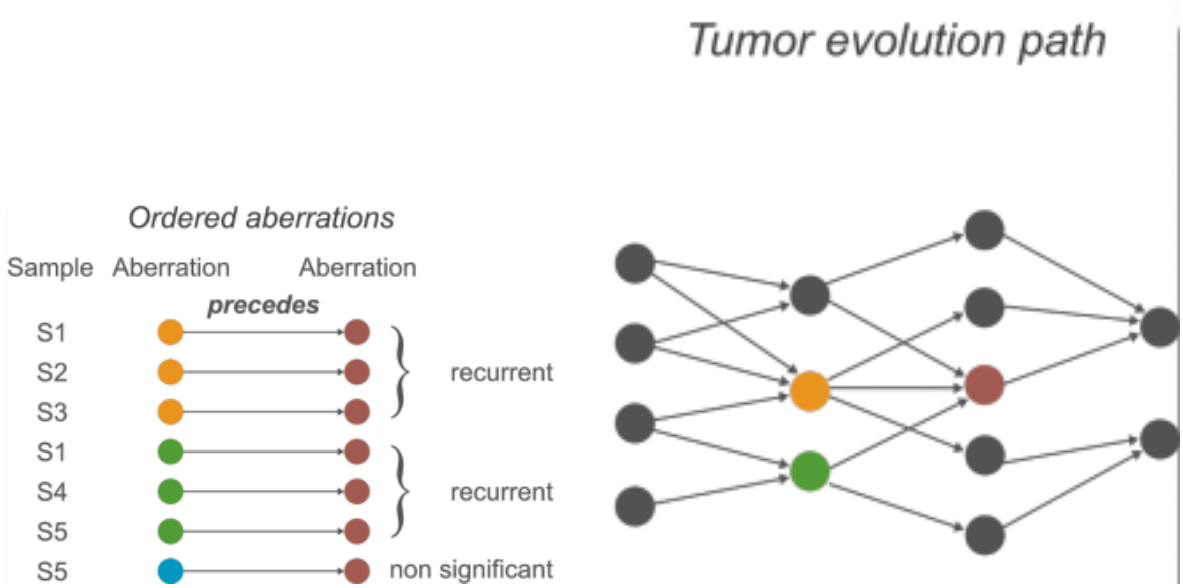
An evolution map is represented by a graph where each node represents a lesion and each arc represents a timing relationship between lesions. To build one a number of samples from different individual are collected and all concomitant lesion within each sample are considered. Then:

- For each sample an edge is drawn from a more clonal lesion to a more subclonal one, compiling a list of lesions.
- Then the number of times this relationship is found in different samples is counted.
- If the number of times an edge is observed deems it significant it is added to the graph, with a weight proportional to the significance of the relationship. The significance of the relationship depends on the number of observations and the total number of co-occurrences.

So, it is assumed that the more clonal a lesion is, the earlier it appeared during tumour evolution. This is assumed only if this relationship is found in enough samples. In this way clonality is linked with time and an evolution graph can be built.

6.2.3 A toy example

The first thing to consider when building an evolution map is to look within each individual at concomitant deletion where one is subclonal to the other one. Figure 6.2 depicts a number of concurrent lesions in different samples.



(a) Ordered aberration (b) Tumour evolution path

Figure 6.2: From ordered aberrations to a tumour evolution path.

Panel A of the figure represents 5 samples. The direction of the arrow indicates the subclonality of the lesions. From this panel it can be inferred that:

- In samples $S1, S2$ and $S3$ the brown lesion is subclonal to the yellow one.
- In samples $S2, S4$ and $S5$ the brown lesion is subclonal to the green one.
- In sample $S5$ the brown lesion is subclonal to the blue one.

So, from these it can be inferred that the yellow and green lesions preceded in the tumour evolution the brown one. The same cannot be said for the blue one as only one sample provides that information, making the statistical analysis non significant. Panel B represent instead a graph of the tumour evolution path:

- The yellow and the green lesions are on the same depth and they are not connected.
- The brown lesion is more deep than the yellow and green ones and an edge from them comes into it.

From this it can be inferred than the green and yellow lesions are at the same time point and this subpopulation both were subject, later in time, to the brown lesion.

6.2.4 A real world data example

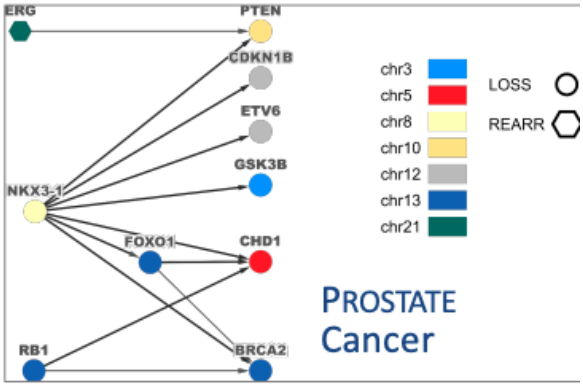


Figure 6.3: A tumour evolution graph built from real data

Performing this analysis all the dependencies supported by more than one individual can be considered. For example it has been found that in prostate cancer a loss in $NKX3\text{-}1$ precedes the deletion of $PTEN$.

6.2.4.1 Limitations

This type of studies are pretty limited: even with hundreds of whole exon sequencing data from large collections, the deepest graph had three layers. One reason for this is the harsh statistical requirements: two lesions must be co-occurrent and one must be subclonal to the other in a significant number of individual compared to the total number of individuals in which co-occurrence is found. The limiting factor is co-occurrence of the lesions.

6.2.5 Pathway-based evolutionary maps

To increase the ability to built this evolutionary maps single gene function information is aggregated into gene families or pathways. An example of this for prostate cancer is depicted in 6.4.

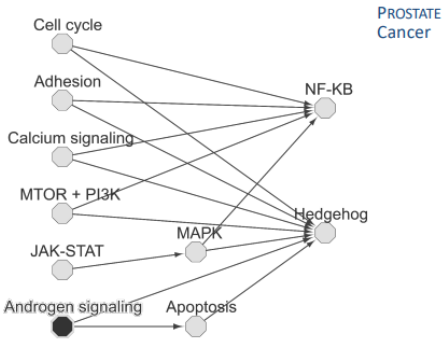


Figure 6.4: Pathway evolution study example for prostate cancer

In this types of reconstructions, instead of considering lesions for a single genes, co-occurrences are counted including all the gene lesions with the same function in a pathway compared with lesion with the same function in another pathway. This method allow to have more data available to consider major changes during the evolution of the tumour pathway. So, a relationships between gene B and gene A, and another between gene C and gene A can be assumed with the same effect and aggregated on the landing gene if B and C are from the same pathway. For example, considering PTEN, a tumour-suppressive gene relevant in pathway PF3K, all information that alter that pathway can be aggregated into one. In conclusion, in pathway-based evolutionary maps a set of pathways that are more or less altered at some level in earlier stages of the disease and that then trigger changes in other pathways are exploited. Doing so allow to learn more of the biology of the evolution of the disease.

6.2.5.1 Timing

There are also more complicated ways to make inference of tumor evolution. Some of this try to avoid the hypothesis that the more clonal a lesion is the more likely it is to happen early. This is because this is not always true: for example treatment could impair this process.

6.2.5.2 Treatment

In a treatment regiment, because of drug pressure selection, specific resistant clones harbouring a specific lesion can take over due to their higher rate of proliferation. In this case, a lesion that appears to be more clonal, could have happened later in time. It appears more clonal because of the higher proliferation rate due to an evolutionary advantage in resisting the treatment.

6.3 Ploidy and purity corrections

6.3.1 Introduction

To compare two different samples for which completely different levels of tumour content are quantified without having to convert every time and for every lesion the depth of the lesion based on tumour content there is a need to adjust the signal for tumour purity and ploidy. This is necessary also because the accuracy in calls for pure or admixed tissue with the same coverage is different and more false positive calls are likely in a more admixed sample. The coverage makes data coming from different samples comparable: everything is normalized to the total coverage, but this is not enough for diseased cells because of admixture. This necessity adds another step in data pre-processing and to do it it is necessary to know both tumour purity and ploidy.

6.3.2 A melanoma example

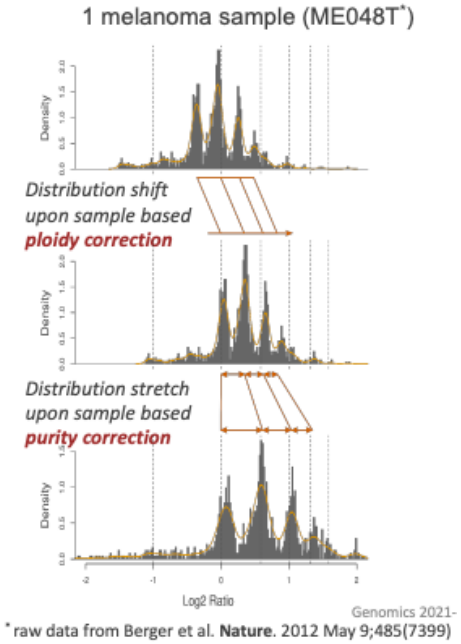


Figure 6.5: Highly aberrant melanoma sample: genome duplication and intervening lesions.

In figure 6.5 data from whole genome sequencing of one melanoma sample is depiced. Multiple peaks in the density of $\log_2 R$ correspond to different copy number states. Most of the time the best way to detect copy-number variants is to compare the tumour signals with the normal ones. Matching tumour data with no somatic changes to a normal genome the two samples will be similar and the histogram of the $\log_2 R$ will look like a normal distribution centred around 0. A tumour sample with many heterozygous or homozygous deletions or copy-number gains will heave, other than the peak at 0, two smaller peaks toward the left, with the first representing an heterozygous deletion and the second an homozygous one. Copy-number gains will be represented by peaks toward the right. Noise effect from admixture will make this secondary peaks difficult to distinguish, bringing them closer to 0.

6.3.2.1 Ploidy correction

The ploidy is computationally assessed through the copy number space. In the example of 6.5 the first plot represents non-corrected data. Due to the genome duplication the main peak will not be in 0, but more to the right. So, assessing the ploidy and looking at a backbone state of three copies for the genome, the main peak should be shifted to the right, as can be seen in the second graph. If instead, the tumour genome underwent more deletions, the signal will be moved to the left. In conclusion the distribution will be shifted toward zero.

6.3.2.2 Purity correction

Tumour admixture is computed and used to correct the graph. As tumour admixture dilutes the signal coming from the tumour, correcting for purity will stretch the peaks, making them more distant and easily obtainable. This can be seen in the last plot of 6.5.

6.3.3 A melanoma example considering more samples

Global copy number changes will shift the peak of the distribution away from zero, while local ones will create different peaks in the distribution. Admixture will compress the distribution around its centre. In figure 6.6 an example is shown for 25 samples of an highly aberrant melanoma tumour.

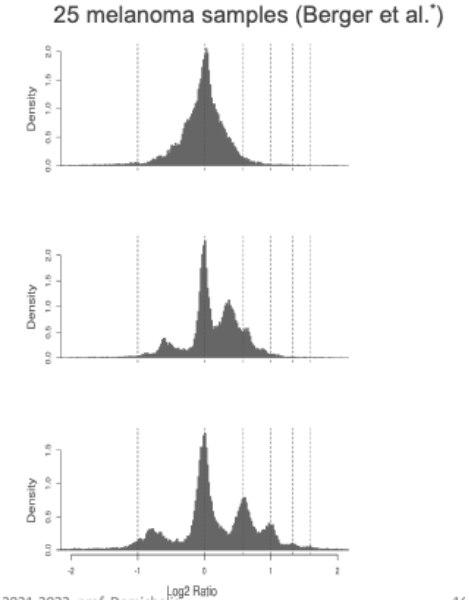


Figure 6.6: 25 melanoma samples. With ploidy correction the signals start being interpretable, and with purity correction, while not perfect they become clear.

In the first graph in the figure it is represented the distribution of the $\log_2 R$ data of uncorrected signal. Every melanoma sample is highly aberrant with ploidy and purity different between individuals. The data is corrected to get rid of the noise. In the second graph data is corrected for ploidy, while in the third for both ploidy and purity. From the corrected data it is inferable that:

- A lot of tumours have a backbone ploidy of 2.
- There are some homozygous deletion not perfectly centred in 1 but closer to that value after the correction.
- Some signal is compatible with an homozygous deletion.
- A reasonable amount of signal for three copies is present: some tumours could have a threeploid status.

6.3.4 An example with TGCA data

In this study data from TGCA was analysed to see how suboptimal tumour purity affect proper copy number data analysis and how common it is that purify is not 100% and ploidy is not 2 in the primary disease.

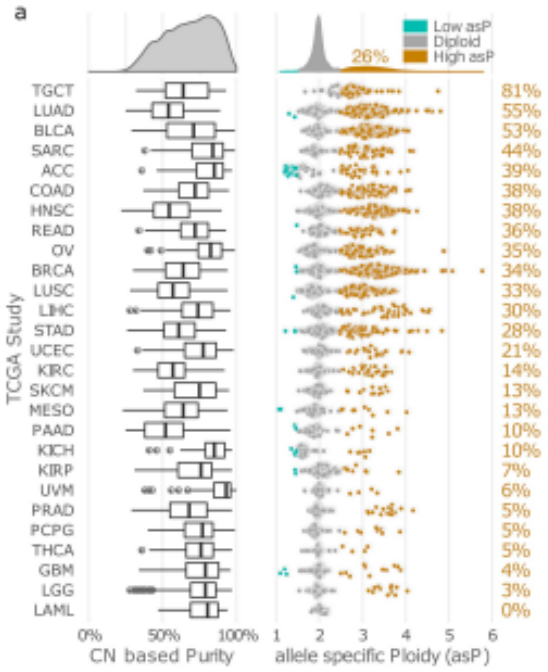


Figure 6.7: Tumour types and purity,

Figure 6.7 depicts a list of tumour types, where every line is a tumour type. On the x-axis tumour purity is depicted and for each type of tumour its distribution id depicted. Every tumour type has a different number of sample profiles. The middle vertical line represent the median signal of the distribution, the black horizontal line represent the interquartile range and some outlier are shown. Altogether, 8183 primary cancer samples matched to 27 tumour types profiled with WES from TCGA. $4,950$ cases with overall high tumour cellularity were identified. The majority had 69% tumor purity. There are some outliers: for example ovarian cancer has high purity.

6.3.4.1 Ploidy correction

Considering ploidy there is a need to investigate what is the fraction within each tumour type with a ploidy significantly above 2. In the plot the tumour types are sorted by decreasing percentage of tumours with a ploidy higher than 2. For example for the first two types more than 50% of the primary tumours have a ploidy status over 2. This mean that either they underwent whole genome duplication or they have a triploidy status. Some tumours have very low ploidy: at least one of the copy of the genome is lost and there is low allele specific ploidy assessment. Figure 6.8 shows the change of data distribution when the samples are corrected for ploidy and purity.

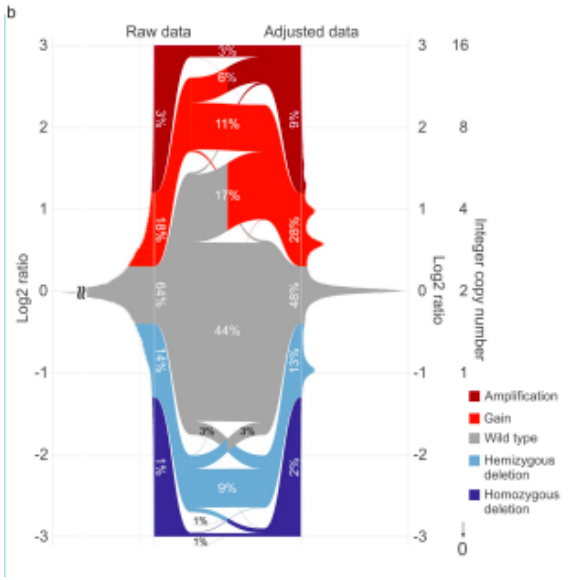


Figure 6.8: The y axis is the $\log_2 R$. On the left raw data are represented, while on the right the adjusted ones.

Focusing on the first half of the graph it can be seen how there is noise similar for the melanoma uncorrected data. The correction of the data resulted in the reclassification of 30% of the totality of the segments. The correction led to the doubling of the observed homozygous deletions, meaning that the total number of unavailable protein products weren't being produced.

6.4 Allele-specific analysis

6.4.1 Introduction

Once tumour data has been corrected for purity and polity there is still a need to analyse allele specific information. Some allele could produce a non functional protein, or there could be copy-number neutral events. These aberration wouldn't be uncovered by the analysis done until now, so there is a need to add a step to evaluate them and their functional effects.

6.4.2 An example

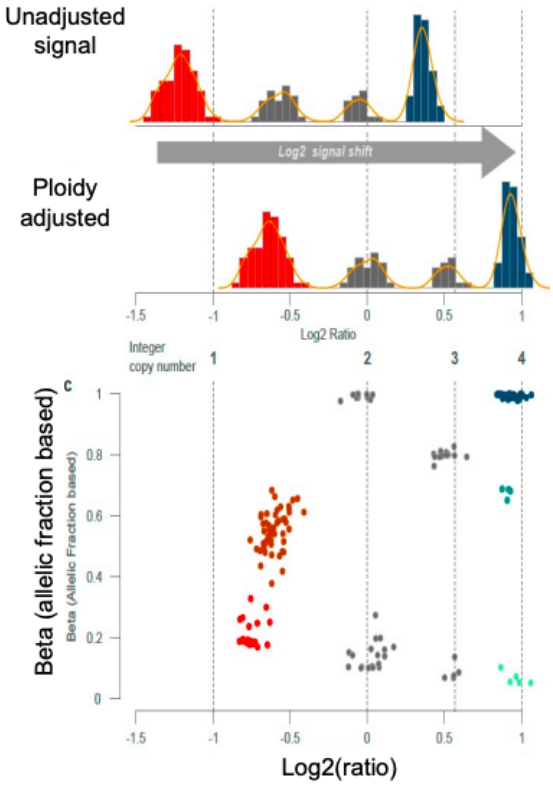


Figure 6.9: Top panel: loss of an allele on A so we'll have 2-1-2 copies. Bottom panel: same situation on allele A but allele B is doubled so we'll have 3-2-3 copies. So, in this situation, the gene x will have two copies but both of them coming from the same allele (B).
By computing the log2 ratio in 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).

Figure 6.9 represents adjusted data for some samples. Then those samples are represented in the $\log_2 R \times \beta$ space. It can be seen how data underneath the peaks belong to specific clusters. This suggests how looking only at the $\log_2 R$ the presence of clusters with different clonalities is undetectable. The lower cluster with $\log_2 R = 0$ only one allele is visible, so there is a copy neutral loss of heterozygosity or CN-LOH: there are two copies of one allele and zero of the other. This becomes visible from the $\log_2 R \times \beta$ space: some equations allow to extrapolate the number of copies for each allele, as depicted in 6.10.

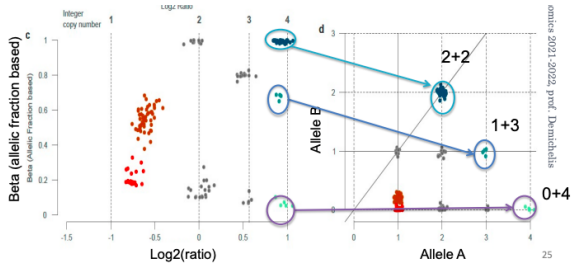


Figure 6.10: From $\log_2 R \times \beta$ space to allele-specific analysis

It can be seen how for 2 different allele and a copy number of 4 there are three different combinations:

- 2 copies of A and 2 copies of B.
- 3 copies of A and 1 copy of B.
- 4 copies of A and 0 copies of B.

6.4.2.1 Changes of copy number for specific alleles

So it can be seen how once the data has been corrected for ploidy and purity the analysis can be shifted to the level of the number of copies of each allele for each gene. This is important because some alleles don't produce a functional protein, so in case when only them are present the function will still be lost, even with an allele present. Distinguishing the alleles is fundamental in distinguishing a possible loss of function.

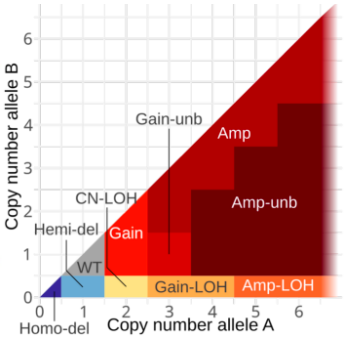


Figure 6.11: Effect of changes of copy number for specific alleles.

6.4.2.2 Copy number neutral events

This analysis allow to reclassify copy number status in the $\log_2 R \times \beta$ space and assign an allele-specific copy number to every segment of the genome, or to every gene. In this way a significant fraction of the genome that would appear wild-type for the copy number underwent loss of one allele and gain of the other. This event is called copy-neutral loss of heterozygosity or CN-LOH. A relevant fraction of high copy number levels from the TGCA data seen before came from the same allele.

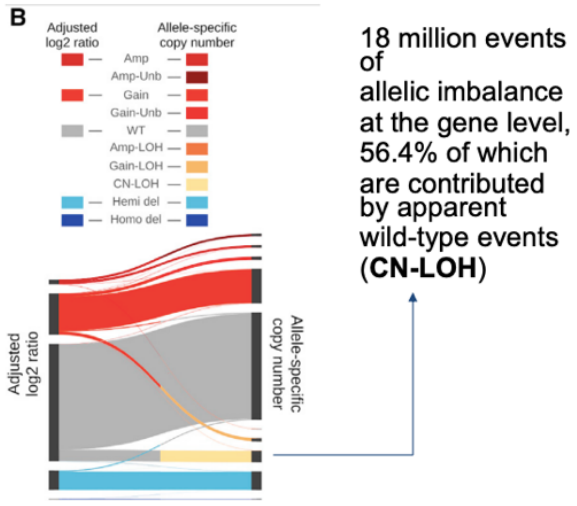


Figure 6.12: Ciani et al, Cell Syst 2022

So, allele analysis is fundamental to get all the information necessary for a functional analysis of the genome. This information is relevant in precision medicine because there are ways to target genes exploiting loss of heterozygosity and this pipeline allow to discover this event even in the case of a wild type segment in term of copy number.

6.4.3 A case study

In the example depicted in 6.13 data from a patient's primary and metastasis sample regarding the copy number for the alleles of sequenced genes are visible.

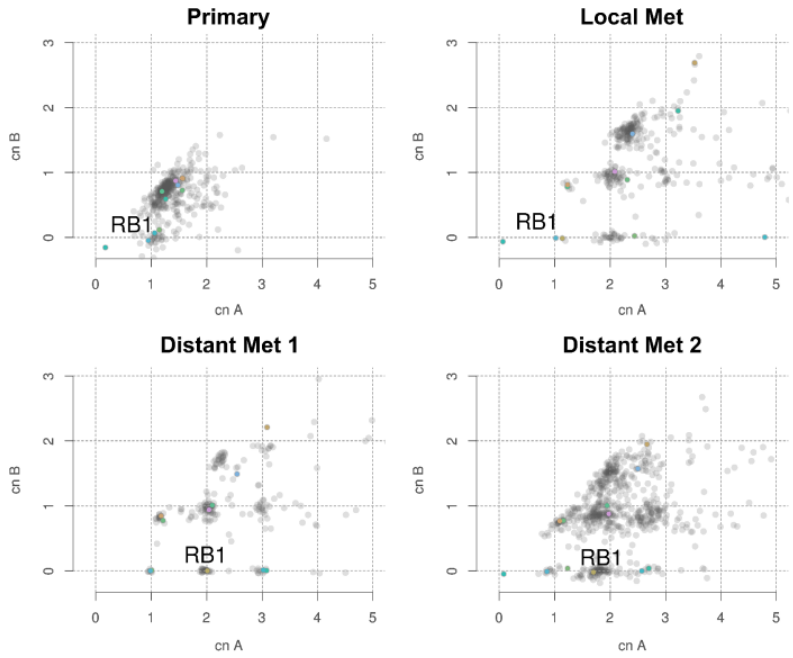


Figure 6.13: Copy number of CN_A and CN_B from multi-sample data from the same patient.

6.4.3.1 Primary site

In the primary site is visible that:

- A cloud of dots has a total copy number of two $(1,1),(2,0),(0,2)$.
- A cluster underwent hemizygous deletion and only one copy is visible $(1,0),(0,1)$.
- A gene that underwent an homozygous deletion is visible $(0,0)$.

6.4.3.2 Metastasis

Other than the primary sites a local and two distance metastasis has been sequenced. From the data of those it can be seen how:

- In distant met 1 there are no homozygous deletion, so probably this metastasis happened before the deletion in the primary site.
- In both the distant metastasis $RB1$ gained an extra copy of allele A.
- In all the metastasis there are extra gains of copies of all the genes, so maybe a whole genome duplication happened.
- In distant met 1 it can be seen how the data point in yellow are subclonal: they are close enough to $(1,1)$ that purity shifted them from $(1,0)$.
- Extra copies of the whole genome are likely in the local metastasis after the loss of the second copy of $RB1$.
- There is a copy number neutral loss of heterozygosity for many genes, including $RB1$.
- The level of subclonality is overall not high.

6.5 Longitudinal plasma profiling

Another way to track tumour evolution is to have data of the genomic landscape from different time points.

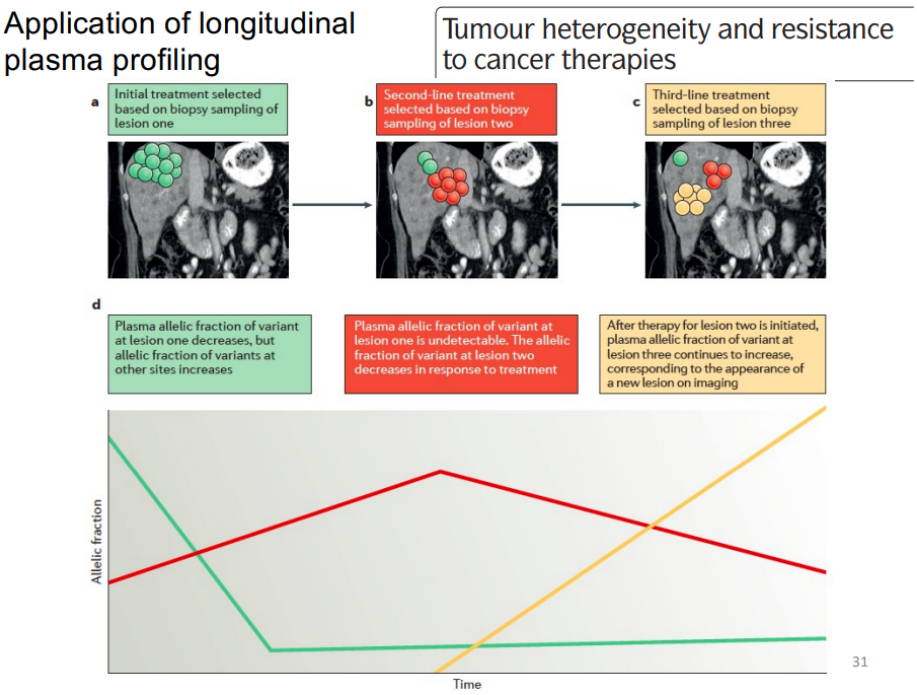


Figure 6.14: Application of longitudinal plasma profiling.

Longitudinal monitoring of alterations in circulating tumour DNA has the potential to enable molecular relapses to be detected before the emergence of disease relapse on imaging.

6.5.1 An example

In the hypothetical example depicted in figure 6.14. A biopsy sample from lesion 1 in green leads to the use of a targeted agent directed at the alterations in lesion one (a). A failure to also lesion 2 in red might then lead to outgrowth of clones harbouring alternative molecular alterations, prompting the use of a combination of targeted agents or use of a single targeted therapy capable Of overcoming both molecular alterations (b). The emergence Of lesion three in yellow might then be missed by biopsy sampling until this lesion becomes detectable on imaging (c). Longitudinal analysis of liquid biopsy Samples would enable the detection and determination of the allelic fractions of the variants at all three lesions before their detection on imaging (d). The figure illustrates the ability of the molecular analysis of plasma to convey the full spectrum of resistance alterations and shows the dynamic nature of resistance.

6.5.2 Tracking evolution

Dealing with biopsies over time the evolution of the disease can be tracked using the allelic fraction of a lesion. Naturally, these allelic fractions at any time point needs to be corrected for tumour content, otherwise different time points would be incomparable.

Chapter 7

Tumor evolution studies: SNVs-based methods

7.1 Introduction

7.1.1 Copy-number neutral tumours

There is a large number of tumours where copy-number aberrations are minimal. These tumours are said to have quiet genomes and make about 1 to 3% of all primary tumours, like the one in figure 7.1. Consequently, it is difficult to use copy number based approaches for these kind of tumours as they display very few copy number changes. These tumours are typically correlated to a better prognosis both in overall survival and in progression-free interval, but relapses are still present so the assessment of these tumours is important.

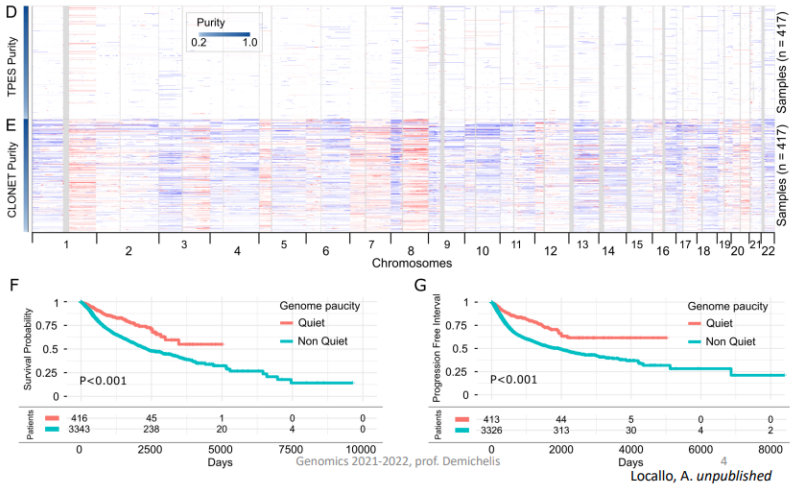


Figure 7.1: Data on quiet genomes.

7.1.2 Different methods for tumour assessment

Multiple Tumor Purity Assessment assays have been proposed, based on:

- RNA.
- DNA methylation.
- SNVs.

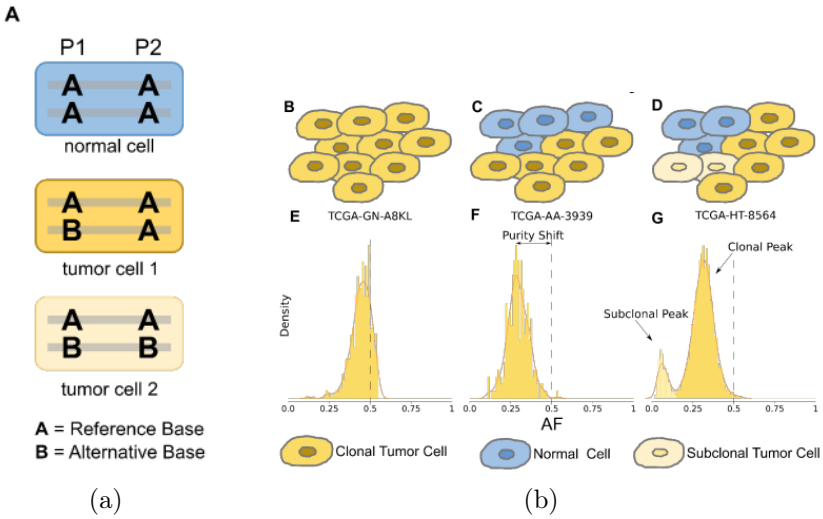
This chapter focuses on methods exploiting SNVs.

7.1.3 Rationale of SNV based methods

Figure 7.2 depicts the call for 2 SNV respectively in 2 genomic loci P_1 and P_2. Let A be the reference allele and B the alternative one associated with a somatic point mutation. Three populations of cells are considered:

- Normal cells show a genotype of AA for P_1 and AA for P_2.
- Clonal cells show a genotype of AB for P_1 and AA for P_2.
- Subclonal cells show a genotype of AB for P_1 and AB for P_2.

It can be seen how the distribution of allelic fraction of the clonal population (B-E) is symmetric, with the main peak around 0.5. A mixed population of clonal tumour and normal healthy cells show a peak shifted toward 0 (C-F), as healthy cells don't carry the mutation. The distance from 0.5 is proportional to the fraction of normal cells. A subclonal population (D-F) is identified thanks to a second peak of allelic fraction closer to 0.



(a) (b)

Figure 7.2: Allelic fraction change for SNVs in a clonal, subclonal and admixed tumour cell population.

7.1.4 Advantages and limitations of SNVs based methods

Advantages and limitations of SNVs based methods are described in table 7.1.

Advantages	Limitations
Best suited for copy number neutral tumour genomes.	Needs a reasonable number of putative clonal somatic heterozygous SNVs per sample.
Applicable to a range of NGS techniques.	
Fast and low demanding of computational resources.	Sensible to subclonal cell populations which could influence clonal peak detection.
TPES is available as an R package on CRAN.	

Table 7.1: Advantages and limitations of SNVs based methods

7.1.5 Reference mapping bias

When performing an SNV based itest it is important to consider the Reference Mapping Bias. A polymorphic locus carrying a non-reference base is less likely to be mapped during the alignment process. With a perfect SNV that is clonal, monoallelic and in highly pure tumour, the allelic fraction will not be at 0.5 because the aligner considers the variation as an error and sometimes discards the read containing it loosing some signal with a bias for the alternative base.

7.2 TPES - tumour purity estimation from SNVs

7.2.1 Introduction

TPES (tumour purity estimation from SNVs) is a tool able to assess the purity of a tumour sample basing its process on the detection of SNVs. The process that the tool employs is depicted in figure 7.3

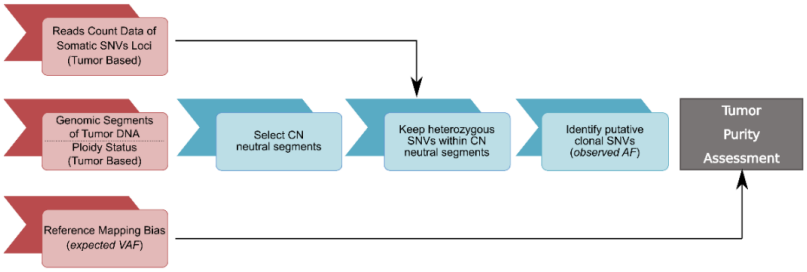


Figure 7.3: Workflow of the TPES tool

7.2.2 SNV identification

TPES pays great attention in which SNV chooses to perform purity assessment. The choice is done in different steps considering different property of each SNV. In particular it selects SNVs according to:

- SNVs in copy number neutral segments.
- Allelic fraction threshold.
- Identification of putative clonal SNVs.

7.2.2.1 Selection of copy-number neutral segments

TPES selects for point mutations that are flat in terms of copy-number. These are perfect for flat genomes and easier to deal with. This is the first filter implemented by this tool: a threshold is set on the $\log_2 R$ to filter the segments in which SNVs are searched.

7.2.2.2 Allelic fraction

Considering all of the somatic point mutations of whole genome, a major peak is expected around 0.5. This is called expected VAF. Other peaks can be originated from things that escaped the previous filter or from monoallelic mutations with copy-neutral loss of heterozygosity. In this case the allelic fraction results doubled. TPES poses another threshold on the allelic fraction to filter out all the SNVs deriving from these segment: the $\max AF$ allowed is of 0.55.

7.2.2.3 Identification of putative clonal SNVs

TPES ranks all the remaining SNV according to their allelic fraction: the peak closer to 0.5 is the most useful to determine tumour purity. This is because the others are related to subclonal events.

7.2.3 Purity estimation

With enough point mutations and after peaks identification the purity is computed as:

$$1 - purity = admixture = 1 - \frac{observed\ VAF}{expected\ VAF}$$

7.2.4 Minimal number of SNV for tumour identification

The number of SNVs changes for each tumour type, so not all tumour types guarantee enough SNVs. The minimum number of SNVs needed to obtain reliable results can be assessed with a comparative analysis. Spearman's correlations between the results of two different purity caller algorithms (TPES and CLONET) using decreasing number of SNVs are computed. SNVs are subsampled as many times as possible to increase the confidence on the results. At each iteration as many samples as possible are used, but the number of samples available decreases as the number of SNVs increases. These statistical test determined 10 as the minimum number of SNVs needed to infer tumour purity, as reported in figure 10.3. With this number, tumour content was detected in 80% of the samples by combining TPES and CLONET, which is copy-number based. The 20% could be tumour-free or samples non detected. Since both SNVs and CN based methods failed, the tumour content of the remaining 20% could be possibly detected with methylation analysis.

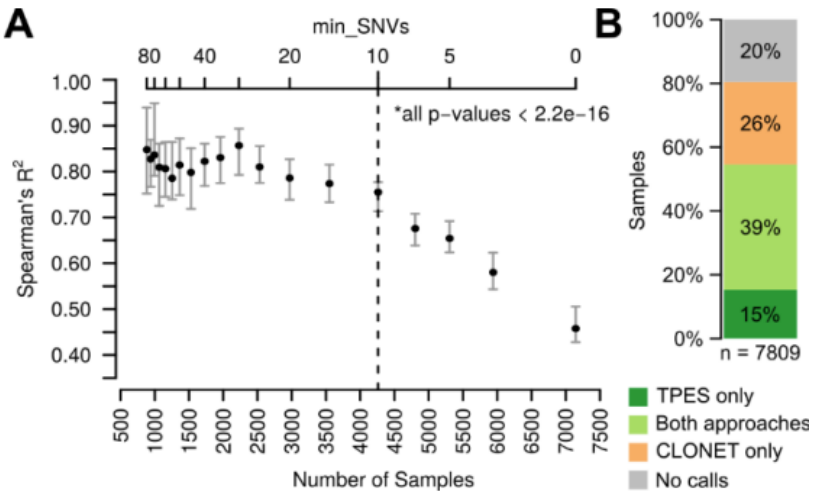


Figure 7.4: A) Correlation between TPES and CLONET with decreasing number of SNVs. **B)** Percentages of samples where tumour purity was assessed considering 10 SNVs.

7.2.5 Comparison with other tumour callers

TPES was compared to other tumour purity caller with a range of different methodologies: good correlation between the results was found, in particular with copy number based algorithms. This shows that genomics is more reproducible in general to assess purity, while methods relying for example on image analysis give different results. The best solution to assess tumour purity is to couple a copy number-based and a SNV-based approach: some samples are only detected by one of the two so a combination gives the best results globally.

Chapter 8

Liquid biopsies in oncology

8.1 Introduction

8.1.1 Tracking tumour progression

It is more feasible to track the tumour progression stage for a patient from liquid biopsies rather than from tissue biopsies. Liquid biopsies give a panoramic overview at a particular time point of patient's state, while tissue biopsies an highly accurate snapshot in a specific site at a particular time point. This is because liquid biopsies are minimally invasive, allowing to collect more time points. Furthermore it is easier for early diagnostics or screening and to quantify the presence of minimal residues of diseases. However the accuracy of a liquid biopsy depends on many factors. In particular, an homogeneous sample is able to achieve similar quality to the one from a tissue biopsy.

8.1.2 Differences between tissue and liquid biopsies

The main difference between tissue and liquid biopsies are listed in table 8.1.

Tissue Biopsy	Liquid Biopsy
Accurate and detailed view of one tissue only.	Landscape overview, with resolution depending on tumour burden, releasing rates, metastases and tumour heterogeneity. It is possible to get an aggregated signal of different tumour cell populations.
Single tumour.	Possibility of getting signal from multiple tumour masses.
Signal relative to a specific point in time.	Signal relative to different specifics point in time obtained through longitudinal sampling.
Invasive and painful for the patient, not feasible for all the tissues or in presence of metastatic sites.	Minimally invasive, it can be coupled with a routine blood draw, allowing to collect samples multiple times.

Table 8.1: Main differences between tissue and liquid biopsy.

8.1.2.1 Liquid biopsies allow to perform a number of analysis

Moreover liquid biopsies allow to perform data analysis otherwise impossible with tissue biopsies:

- Specific assays can be designed to detect minimal quantities of tumour cells. This is useful to detect minimal residual disease (MRD) that can remain after surgery and avoid tumour recurrence.
- The collection of serial samples allows to:
  - Track clonal evolution of the tumour over time.
  - Catch treatment resistances early on.
  - Monitor the patient's response to the treatment.
- It can be used for early detection of cancer.

8.1.2.2 Material availability

Samples obtained through tissue biopsy come from:

- Needle biopsies.
- Biopsies.
- Surgical resection (if some material is left after the clinical protocol and the patient agrees to a research protocol).

Samples obtained through liquid biopsies come from:

- Circulating tumour cells.
- Extracellular vescicles.
- Cell-free DNA.

In particular considering cell-free DNA a difference in its concentration in a sample is visible from a healthy and tumour sample. In particular for healthy donor it has a concentration of $\sim 4\frac{ng}{ml}$ and belo $10\frac{ng}{ml}$. In tumour patients the range of this concentration varies more and can reach $\sim 100\frac{ng}{ml}$. This concentration tends to increase as the tumour becomes metastatic. Another factor that influences cel-free DNA concentration is treatment: tumour patients under treatment tend to have a concentration of cfDNA similar to healthy individuals. cfDNA concentration is influenced by a lot of factors, making it impossible to use it as a good prognostic feature by itself.

8.1.2.3 Tumour content

Tumour content in tissue biopsies can be assessed through a visualization protocol: the proportion of tumour cells compared to healthy one is measured counting the cell in the sample. Tumour and healthy cells are distinguished through staining or through their differences in morphology. Concentration is computed considering the magnification of the image. Subtyping is performed through a staining for biomarkers. Also some computational methods are available. Instead, for liquid biopsies the fraction of circulating tumour DNA or ctDNA is inferred with methods based on genomics or on methylomics.

8.1.2.4 Assessment of tumour ploidy

For tissue biopsies tumour ploidy is assessed through:

- Cytogenetics.
- FISH.
- NGS data.

The average ploidy of samples from liquid biopsies is inferred computationally based on a genomic process, but it is difficult to reach significative results.

8.1.3 Application-dependent requirements

8.1.3.1 Early tumour, MRD and recurrence detection

When performing early tumour detection minimal residual detection and recurrence detection tumour quantity is low. So a low signal is expected and an higher quantity of starting higher material is required. The quantity to be used need to be finely tuned to balance false positives, which arise from too much material and false negatives, which arise from too little material.

8.1.3.2 Analysis of tumour dynamics, treatment response and analysis of the mechanisms of resistance

When performing analysis of tumour dynamics, treatment response and analysis of the mechanisms of resistance the assay should be designed in order to be able to distinguish between different clones. So a subclonality analysis should be possible.

8.1.3.3 Single biomarker assessment

When performing biomarker assessment the only important thing to take into consideration is to detect whether one point mutation is present or not. In this case tumour content is not important. A target assay is used and specific locations associated with the SNV are sequenced as deep as possible to detect the mutation.

8.1.4 Whole-genome and targeted sequencing

Whole genome sequencing has higher computational cost, while targeted assays have higher sample preparation time. The sequencing cost is higher for whole genomes but it does not decrease linearly with the length of the genome that needs to be sequenced in target sequencing. Figure 8.1 depicts a comparison between whole genome and targeted sequencing.

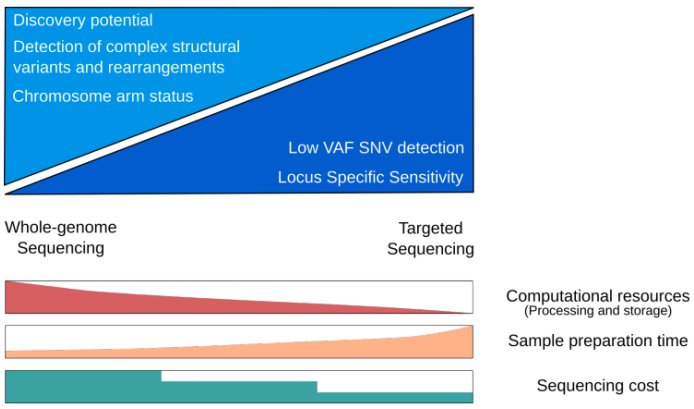


Figure 8.1: Whole genome vs targeted sequencing

8.1.5 Challenges in tracking tumour evolution with liquid biopsies

When tracking tumour evolution with liquid biopsies different parameters need to be considered:

- ctDNA content: fraction of tumour content in circulation/
- Polyploidy: allelic imbalance events.
- The ability to detect signal is gene region dependent and individual dependent.
- Different metastasis have different DNA release rates.

8.1.5.1 An example of similarity between two plasma samples between two different tumours

Figure 8.2 depicts metastatic biopsy time points during clinical progression from CRPC-Adeno to CRPC-NE. Plasma sample at time of CRPCA-Adeno with lymph node and bone metastases displayed a genomic ctDNA profile similar to the one of CRPC-NE liver metastasis observed on imaging and biopsied 3 months later at time of progression on abiraterone.

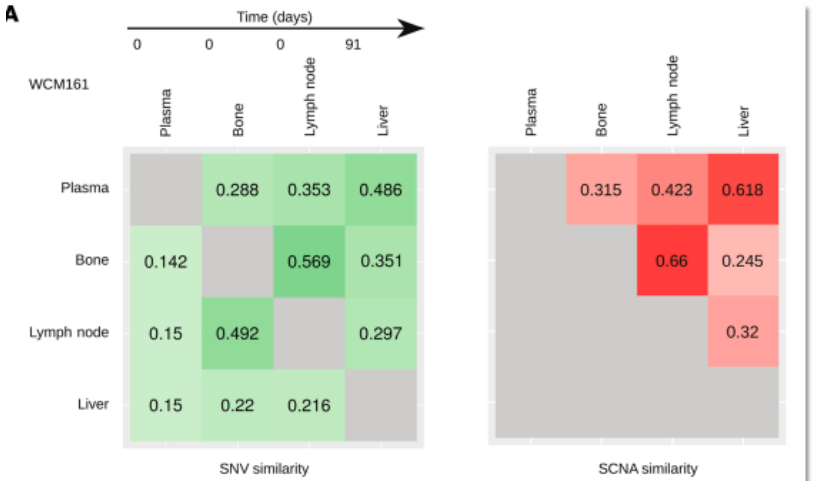


Figure 8.2: Metastatic biopsy time points during clinical progression.

8.2 Interpretation of cell free DNA data

8.2.1 Introduction

Interpretation of cell free DNA data from liquid biopsies poses a number of challenges to detect SNVs of the tumour fraction.

8.2.2 Normalization

When interpreting data from liquid biopsies, it is fundamental to contextualize a mutation after observing it. In order to associate a particular mutation to a particular diagnosis, the signal has to be normalized based on tumour content. Without normalization, tumour content is the most influential variable on the patient's prognosis and can mislead the analysis, making them useless.

For example in figure 8.3 depicts one mutation that is detected only in a specific type of tumour when it is actually present in other types too. It is not detected in other tumours because of the low tumour content of some samples. For this reason not all the literature available about liquid biopsies is reliable: lack of normalization leads to completely wrong conclusions. All assays that study cell-free DNA, from microarrays to extracellular vesicles need to be normalized.

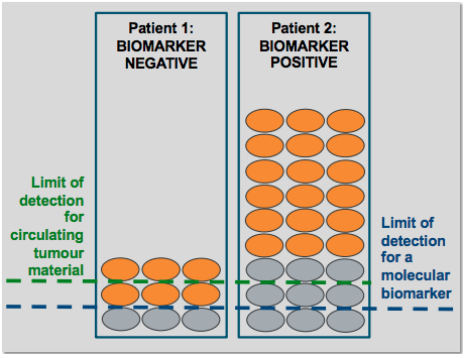


Figure 8.3: Two samples with the same percentage of tumour cells: the first one results negative for the marker because of its low tumour content.

8.2.3 Quantity of input material

Another source of errors in the interpretation of cfDNA data is the amount of input material: if the patient's tumour content is high, the results could be obtained with a limited amount of extracted nucleic acid, but if the tumour content is low, too little material can lower the chances of detecting tumour cells. This is visualized in figure 8.4, in which the same analyses are repeated with different quantities of starting material. It is visible that for a patient with high tumour content the results do not change even with as little as $5ng$ of cfDNA.

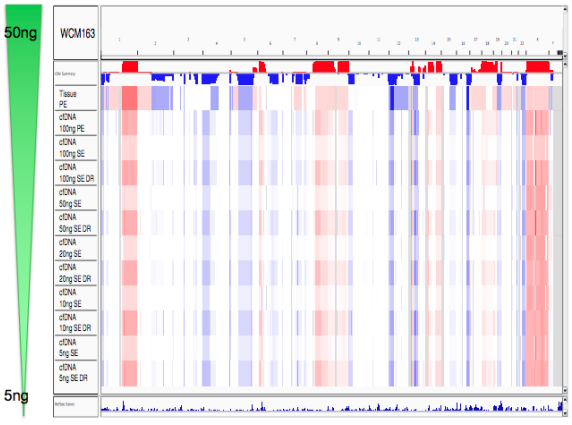


Figure 8.4: Change of resolution of analysis with different quantities of starting material.

In most cases the tumour content is unknown before the analysis. This has to be taken into account when designing an experiment. Usually the standard procedure is to begin with $2ml$ of

plasma, so the DNA obtained should be from $50ng$ to $5ng$. For a pure sample $10ng$ of DNA should correspond to $\sim$ 1500 diploid tumour genomes. In the case that no tumour is detected, the assay should be repeated with more material to be sure that the tumour is not present and not just undetected. Information about the patient state can guide the selection of the quantity of input material: for example, a patient in remission will require more material. Figure 8.5 depicts how copy number signal correlates well between different initial amounts of cfDNA for a sample with high tumour content.

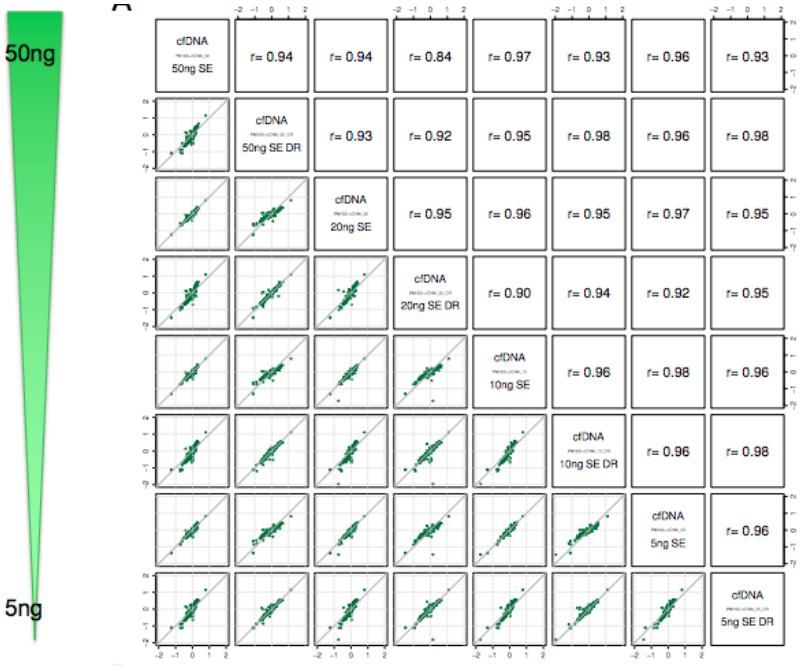


Figure 8.5: Correlation between copy number signal and initial amount of cfDNA for high tumour content.

8.2.4 SNVs detection

Multiple tools are available to detect SNVs from liquid biopsies data. Each tool will probably give different results or partially concordant ones. Each tool can be tuned to favour some types of calls, so the tuning parameters should be carefully selected. SNVs detection in liquid biopsies faces different problems, both of technical and biological nature.

8.2.4.1 Technical problems

Technical problems that need to be addressed when detecting SNVs though liquid biopsies are:

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

8.2.4.2 Biological problems

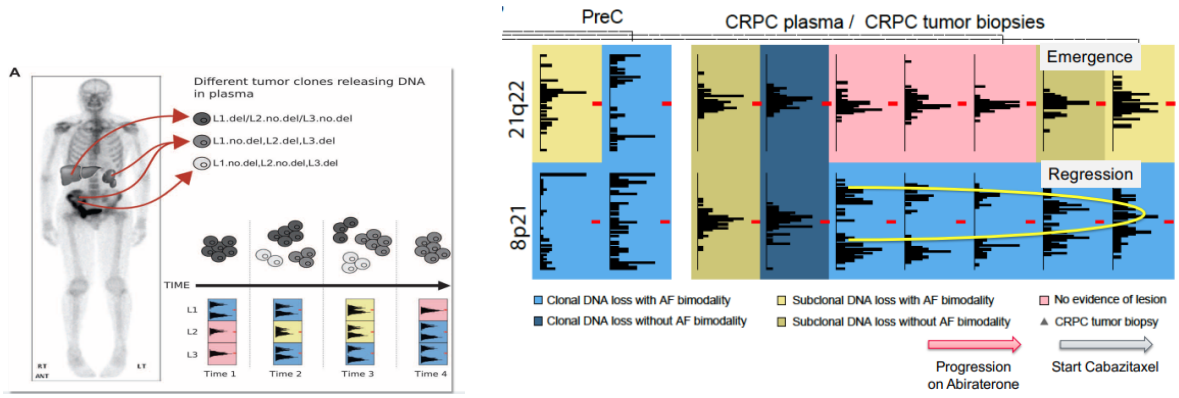
Biological problems that need to be addressed when detecting SNVs though liquid biopsies are:

- Low tumour content: ctDNA to cfDNA ratio.
- Clonal hematopoiesis: an hematopoietic stem cell starts making cells with the same genetic mutation. To distinguish the signal coming from clonal hematopoiesis, it is compared with what has been sequenced before from solid tumours. It is rare to observe something in liquid biopsy that has never been noticed in solid ones.
- Copy number variations and ploidy: with a whole genome duplication and a SNV only present on one allele, the signal corresponding to the mutation is only 25% of what it should be and has to be correctly interpreted.
- Intra-patient tumour heterogeneity: very low allelic fractions for SNVs that are not clonal can be difficult to observe.

8.2.5 Two case studies

8.2.5.1 Signal distribution in prostate cancer

In figure 8.6 an assay to study specific signal distribution for prostate cancer is depicted. By analyzing this as a dynamic process, the overall distribution of clonal DNA can be derived. One of the lesion tracked was 21q22: the distribution on the top and the bottom are different at each time point with different dynamics. When the patient regressed 8p21, 21q22 emerged.



(a) (b)

Figure 8.6: a) Longitudinal sampling process. **b)** Emergence and progression of lesions over time.

8.2.5.2 Non-AR driven castration resistance prostate cancer

Non-AR driven castration resistant prostate cancer is really rare as a de novo disease, but it has a high rising incidence, especially after potent AR-pathway modifications. It is hard to treat and to diagnose. Both tissue and liquid biopsy analysis are performed to find:

- Potential biomarkers for liquid biopsy.
- Distinguish the transition to severe stage.

The first analysis from tissue biopsies sample allowed to investigate tumour heterogeneity, and determined the similarity of the metastasis. Homogeneity is higher for NE, the most aggressive phenotype. The same was observed in liquid biopsies, confirming the potential of the use of NE biomarkers. A liquid biopsy reported equal result to a lymph node metastasis, with a clear signal. In other patients, certain genes switch from one cluster to another from tissue to plasma, suggesting that the metastatic representation was partially heterogeneous. It is clear how liquid biopsies provide a landscape overview.

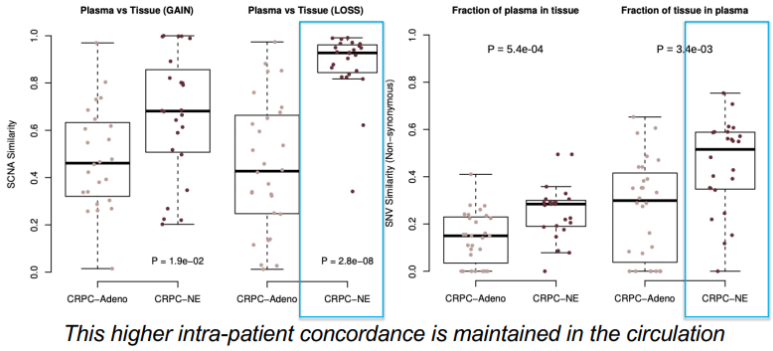


Figure 8.7: Concordance between two patients

Figure 8.7 depicts SCNA and SNV similarity for a man initially diagnosed with adenocarcinoma. The diagnosis switched to NE after clinical assessment. Plasma sample come before NE diagnosis and multiple tissue biopsies were performed. It was noted that the liver metastasis NE had lesion represented in the plasma sample, suggesting that a clone that was transforming was already present in the past. This makes clear how in certain cases, a liquid biopsy can be informative of something that would only emerge later clinically. Comparing the genomic content of each metastasis with the data coming from a liquid biopsy, a measure of the modification contributing more to the disease can be obtained.

Chapter 9

Extracellular vesicles

9.1 Introduction

9.1.1 Definition

Extracellular vesicles are membrane-enclosed nanoscale particles released from essentially all prokaryotic and eukaryotic cells that carry proteins, lipids, RNA and DNA.

9.1.2 Compartments of extracellular vesicles

Different compartments of extracellular vesicles can be identified. Figure 9.1 represents all the main components of a vesicle.

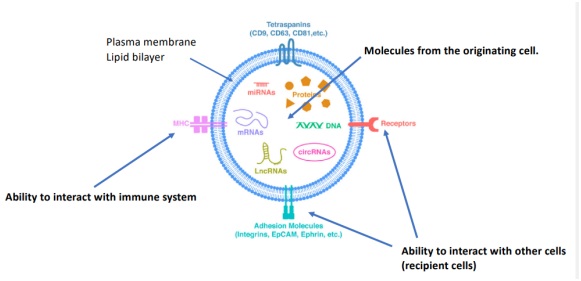


Figure 9.1: Sketch representing the main components of an EV.

9.1.2.1 Outside layer

The outside layer of extracellular vesicles is made of the lipidic membrane of the cells from which the vesicles were originated.

9.1.2.2 Content

The content of extracellular vesicles can vary. The can contain:

- RNAs of different length.
- DNA.
- Proteins.

This molecules are derived from the cell that originated the vesicle.

9.1.2.3 Membrane proteins

On the membrane of extracellular vesicles there are proteins that are able to to interact with:

- The immune system.
- Receptors.
- Adhesion molecules.
- Tetraspanins.

Tetraspanins are proteins that span the membrane four times and act as markers or recognition proteins of the vesicle.

9.1.3 Characterization of extracellular vesicles

Extracellular vesicles are very different from each other. Especially in older studies, each research group used to study extracellular vesicles from a certain site and gave them a specific name, for example:

- EVs from the prostate were called *prostatosomes*.
- EVs from a tumor sample were called *oncosomes*.

A consortium was created to order the nomenclature and to determine which are the parameters to characterize and study extracellular vesicles.

9.1.3.1 Size

Extracellular vesicles can be characterized by their size:

- 100-1000nm: microvesicles.
- 100-5000nm: apoptotic extracellular vesicles and apoptotic bodies.
- 50-150nm: exosomes.
- 30-50nm and no lipid bilayer: exomeres.

The categories overlap and there is no clear cut.

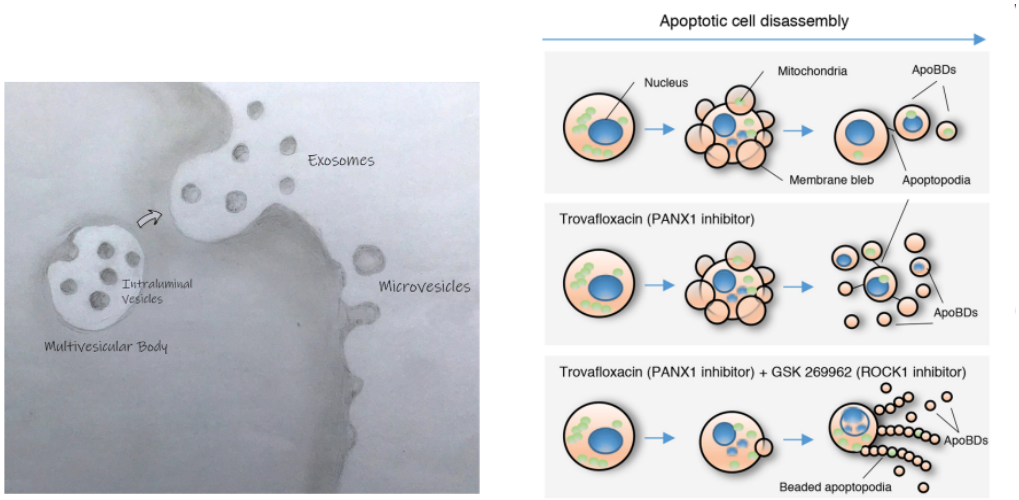
9.1.3.2 Origin

Extracellular vesicles can be further categorized according to their origin:

- Exosomes originate from the endocytic pathway.
- Microvesicles and apoptotic bodies originate from the plasma membrane.

In particular microvesicles originate directly from the membrane of the cell, while the exosomes come from the multivesicular body, which contains the intraluminal vesicles, as shown in panel a) of figure 9.2. Apoptotic bodies instead are the result of cell death. In the experiment performed in panel b) of figure 9.2, the researchers induced apoptosis using different methods and the vesicles had different sizes.

9.1.3.2.1 Process of origin of exosomes Exosomes are the most interesting extracellular vesicles and originate from a multi-step process:



(a) Origin of exosomes and microvesicles

(b) Origin of apoptotic body

Figure 9.2: Origin of EVs

1. Endocytosis: the cell either capture everything in the ECM, or the substrate is selected by receptors.
2. Formation of early and late endosomes: lysosomes are organelles which go through a process of maturation, in the late endosomes enzymes complete the packaging of the substrate.
3. Formation of multivesicular bodies: multivesicular bodies contain intralumenal vesicles, the precursors of the exosomes.

9.1.3.3 Content

9.1.3.3.1 Exosomes and microvesicles Exosomes and microvesicles mainly carry:

- Proteins.
- Nucleic acids:
  - mRNA.
  - miRNA.
  - other non-coding RNAs.

Vesicles are really smalls and cannot contain big fragments of DNA. Further studies however proved the presence of longer, protein coding transcripts. A lot of RNA transcripts have important regulatory functions like miRNA or circRNA.

9.1.3.3.2 Apoptotic bodies Apoptotic bodies instead are the entire representation of the cell's cytoplasm. The contain an equal representation of the cell content.

9.1.3.3.3 Exosomes Exosomes select their cargo through an active process in which:

1. Abundant RNA is selected.
2. Fragmentation of the RNA occurs.
3. RNA is selected through:
   - Specific sequence motifs.
   - Unique secondary structures.
   - RNA modification like mRNA uridylation.

9.1.4 Functions of extracellular vesicles

Each extracellular vesicle carries information about its cell of origin and its putative function. Each extracellular vesicle can have a different function, based on the presence on other recipient cells.

9.1.4.1 Functions of extracellular vesicles in cancer

In cancer, EVs have important functions. In prostate cancer it has been shown that exosomes are able to modulate the immune system, by changing the preferential maturation of the cells of the immune systems. They aid in the proliferation of endothelial cells, in stromal fibroblast differentiation, creating population that are pro/anti tumorigenic. They remodel the ECM, which is extremely important for metastasis, as EVs provide for a way for cells to bind to a different substrate and create metastatic sites (especially for bone cancer). An example of how exosomes can boost metastasis is reported in figure 9.3, in which prostate cancer extracellular vesicles mediate intercellular communication with bone marrow cells and promote metastasis in a cholesterol-dependent manner. In this process exosomes from prostate cancer travel in the body, arrive to the bone and boost metastasis.

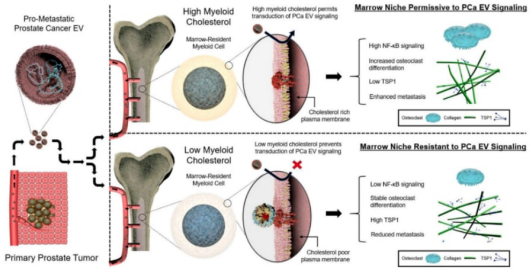


Figure 9.3: Exosomes-mediated metastasis in prostate cancer.

9.2 Tumour studies through extracellular vesicles analysis

9.2.1 Introduction

Liquid biopsies can ve used as a novel tool for cancer detection and monitoring. By just drawing some blood in a serial matter a lot of extracellular vesicles coming from all body's tissues, including cancer cells, can be retrieved. This is of importance because in the same tumour there could be different populations, each harbouring different mutations and having different proliferation rates. This cause them to respond in different ways to therapies. Liquid biopsies allow to analyse data coming from:

- Tumour.
- Cell free DNA.
- Extracellular vescicles.
- Ribolipoproteins.

Analysis of liquid biopsies can be considered with a multi-analyte approach using different molecular cues, coming for example, from DNA and RNA, to detect tumour related signal.

9.2.2 Breast cancer - an example

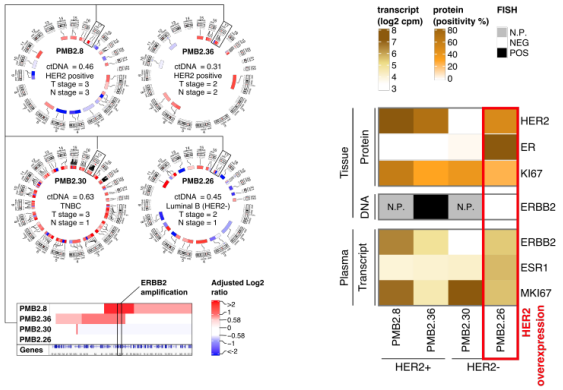


Figure 9.4: 4 Breast cancer patients. On the left: Whole Exome Sequencing of cfDNA from plasma.

Breast cancer usually stratifies in different subtype and one of the main molecular feature to distinguish them is the presence of hormone receptors, in particular the one expressed by HER2. The experiment depicted in figure 9.4 collected data from two patients that were HER2 positive and two patients HER2 negative. The two positive patients were confirmed by high protein level, but one of the negative patients has a signal for the protein, but no amplification, as the data was collected through FISH. The clinicians classified this patient as ambiguous. Integrating the data from RNA coming from extracellular vesicles the HER2 positive patient expressed high level of ERBB2, concordant with the previous results, but this data was aslo found for the ambiguous HER2 negative. This is concordant with he immuno istochemistry but not by FISH. This is because the regulation of HER2 is not only regulated by the amplification, but also by over-expression. From this experiment is clear that performing only geneic analysis some information are missed, because with EVs we were able to identify over-expression even in absence of amplification.

9.2.3 Tracking tumour signal in serial samples

To track tumour signal evolution a serial approach is taken. The signal of different biomarkers is tracked in time. This allow to track the response to a drug treatment. In this process sequence data from digital PCR of the sample from the blood is performed and the biomarkers are followed in time to discover how well the patient is responding to the treatment.

9.2.4 Extracellular vesicles isolation methods

In literature different ways to isolate EVs are reported:

- Nickel-base isolation (NBI): exploits the charge of the EVs (negative) using metallic beads.
- Size exclusion chromatography (SEC): filter the sample based on the size of the molecule or of the extracellular vesicle.
- Ultracentrifugation (UCFG): heaviest molecules are compressed in a pallet, which is discarded, and only the surnatant composed of EVs is retained.

Usually these three methods are used together because they are not perfect: the extracellular vesicles are heterogeneous and each method is able to isolate different subpopulation. This is particularly important in liquid biopsies: collecting only some subpopulation would introduce batch effects in the down-stream analysis.

9.2.5 Challenges in studying tumour evolution through extracellular vesicles

Some difficulties do consider when studying tumour evolution through extracellular vesicles are:

- Evidence of high variability of EVs populations in terms of EVs isolation method.
- No standard de facto for the analysis of the EVs transcriptome.
- Dealing with plasma samples, multiple populations of EVs are present, some of which are relevant to cancer, while other are not.
- The RNA signal deriving from multiple populations is difficult to interpret.

9.2.6 Deconvolution

Deconvolution is the process through which, from mixed signal after EVs isolation and RNA sequencing, the different population of vesicles and molecules can be characterized and differentiated.

9.2.6.1 Supervised deconvolution

Supervised deconvolution is the standard to perform deconvolution. The process is shown in 9.5. This process has several limitations, mainly due to the fact that it exploits known signature matrices, making it impossible to detect new extracellular vesicles species.

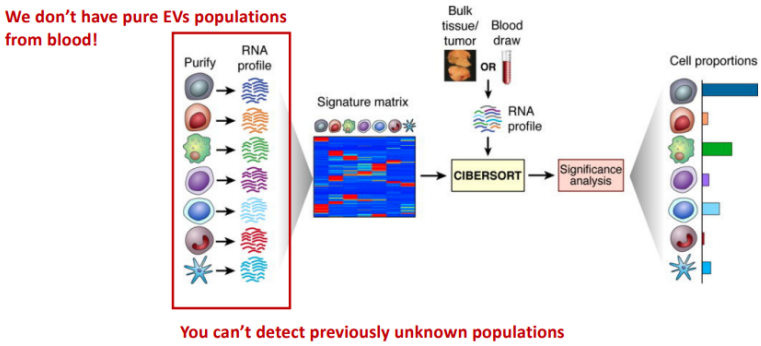


Figure 9.5: Supervised deconvolution.

9.2.6.2 Unsupervised deconvolution

Another possible approach in current development is to implement an unsupervised approach, as the one depicted in figure 9.6.

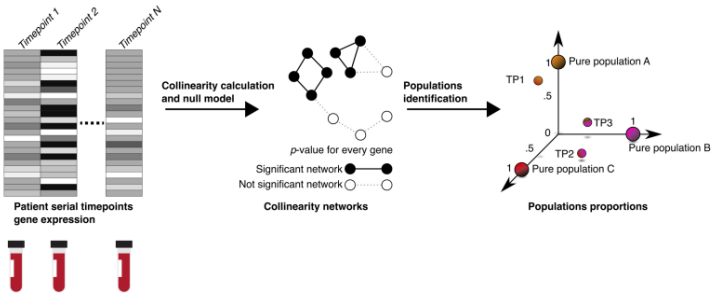


Figure 9.6: Unsupervised deconvolution.

9.2.7 Conclusion

In conclusion it can be said that:

- Extracellular vesicles can be isolated and analysed from biofluids.
- Extracellular vesicles are involved in tumour related processes.
- Extracellular vesicles carry information about their cell of origin and function.
- Extracellular vesicles populations in blood are heterogeneous. Making their analysis difficult.
- Deconvolution approaches are possible but not yet well established.
- The analysis of extracellular vesicles together with other molecules in liquid biopsies like cfDNA can be more informative in respect to the analysis of a single analyte alone. A multi-analyte approach is convenient and increase the predictive power of the analysis.

Chapter 10

Epigenetic profiling of cell-free DNA

10.1 Introduction

10.1.1 Epigenetic

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

10.1.1.1 Epigenetic modifications

The main epigenetic modifications are:

- DNA methylation: in humans they are mainly found on CpG islands, genomic regions with high CG content.
- Histone post translational modifications like PTMs.
- chromatine architecture changes.

10.1.1.2 Epigenetic plasticity

The epigenetic landscape is very different from the genetic one. DNA mutations are directional: they cannot be reverted so they accumulate with subsequent cells generations. The epigenome is plastic, so it can be reverted, physiologically or through therapies. Moreover, the human epigenome is tissue and cell specific while the genome is unique.

10.1.1.3 Epigenetic deregulation

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

10.1.2 DNA methylation

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

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

10.1.2.1 CpG islands

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

10.1.2.2 Cancer tissues

In cancer tissue, hypomethylated and hypermethylated regions are often observed, leading to an abnormal regulation of gene expression. In addition to that, hypermethylation of pericentrometric heterochromatin in cancer can lead to mitotic recombination and thus genomic instability. Figure 10.1 depicts methylation patterns that are altered in cancer.

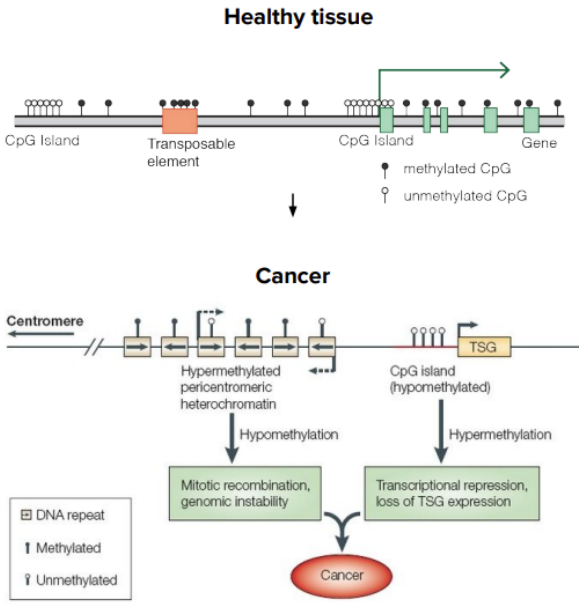


Figure 10.1: Methylation patterns altered in cancer

10.1.2.3 Landscape of regulation

This methylation landscape is highly regulated and tissue-specific. This landscape of regulation is very complex: DNA methylation can regulate gene expression but it is not the only regulating factor, some histone modifications also contribute for example. DNA methylations are not inherited across generations, so there is no accumulation of methylation variants, as it happens with regular DNA mutations. Each individual is born with a brand new methylation landscape that is then disrupted during life, not only due to cancer or disease. Interestingly, it could be possible to exploit variations in the DNA methylome to measure age by computing how many cell divisions led to that specific methylation state.

10.1.2.4 DNA methylation markers

DNA methylation markers can be:

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

10.1.3 Measuring DNA methylation

Figure 10.2 depicts the main methods for DNA methylation measurements. Immunoprecipitation-based methods are also available, but the most frequently used methods are based on whole genome profiling.

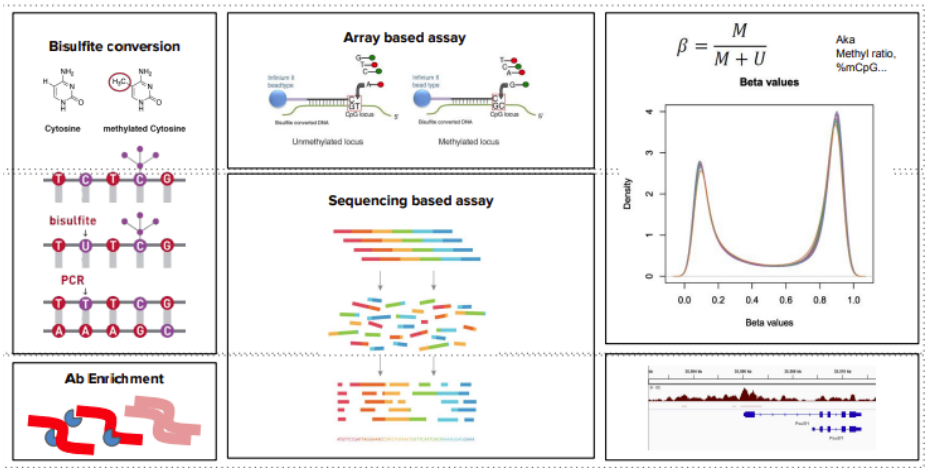


Figure 10.2: Main methods for DNA methylation measurement

10.1.3.1 Bisulphite conversion

The fist step is to perform bisulfite conversion on DNA: thanks to bisulfate ions, unmethylated Cs are converted into Us.

10.1.3.2 Sequencing and alignment

The DNA so modified is then sequenced and aligned. The alignment is done through a particular alignment algorithm aware of the bisulfite conversion. In this way errors and methylations can be detected. Both array-based and shotgun-sequencing-based assays are used to this aim.

10.1.3.3 Computing beta values

The result of this assay is a series of **beta values**. These represent the fraction of reads corresponding to 1 methylated genome site.

10.1.3.4 Single read analysis

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

10.2 DNA methylation based liquid biopsy

10.2.1 Introduction

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

10.2.2 Comparison with genomics-based liquid biopsies

For this purpose, when compared to genomic DNA, the analysis of the methylation landscape has positive and negative aspects. For genomic DNA, the percentage of actually informative signal on the whole information that is obtained can be small and difficult to observe, on the other hand, for DNA methylations it is difficult to discriminate between what is aberrant and what is not because the modifications are tissue-specific and it is difficult to obtain clear background references to make a comparison. A comparison of the advantages and limitation of both approaches is depicted in figure 10.3.

	Genomic DNA	DNA methylation
Molecular signal	Signal is limited to genomic alterations, and thus might be low for SNVs or quiet tumors	Extended and multi-facet signal, amenable to genome wide detection
Background/reference	A single well known background: the normal human genome, as profiled by the control germline sample	Multiple cell populations with distinct profiles, each contributing to the DNA methylation signal
Variability	Low rate of biological variability, discrete signal and overall acceptable technical errors	Discrete degree of biological variability, continuous signal with variable confidence (ᵛ coverage, platform, experimental approach...)
Information content	Limited to genomic information (SNV, SCNA...) but possible fragmentomics applications	Could potentially capture transcriptional state of cancer cells, offering a snapshot of processes such as lineage switching
State of the art	Highly characterized and interpretable, extended literature and high quality samples are available to aid interpretation	Fewer datasets available, but promising results in the past few years. Currently a mostly uncharted territory

Figure 10.3: Comparison of genomics and methylation analysis for cancer detection

10.2.3 Workflow

A common analysis workflow for DNA methylation based liquid biopsies is depicted in figure 10.4.

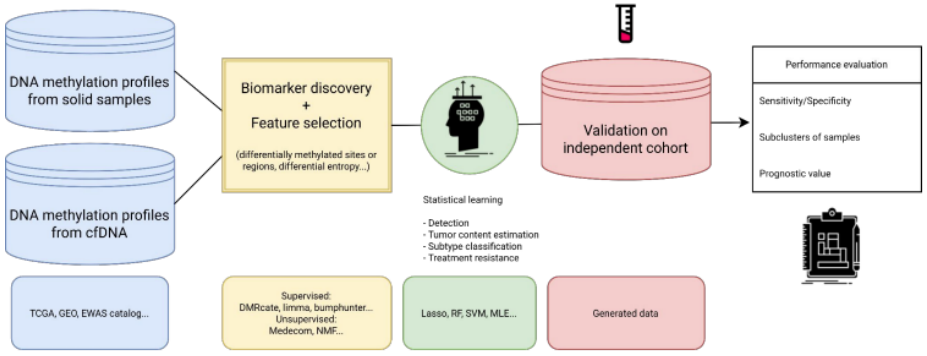


Figure 10.4: Common DNA methylation based liquid biopsies analysis workflow

Different steps in the workflow can be identified:

1. The data is sequenced from solid and liquid samples because the methylation profiles from solid samples are needed as reference. The reference profiles for liquid biopsies analysis are derived from white blood cells and from the cancer type of interest. White blood cells are the background reference for cfDNA, since the most frequent genomic material in circulation originates from this type of cells.
2. If a methylation pattern different from the one of blood cells is found in cfDNA it means that cells of some other tissue are dying and their material is going into circulation. This is a possible signal of disease.
3. Once having obtained these reference pattern, biomarkers are searched and feature selection is performed.
4. A model is fitted and optimized to perform predictions on new data.
5. Performance evaluation of the model is done.

10.2.4 CCGA study

The Circulating Cell-free Genome Atlas (CCGA) is a study conducted by Grail designed to characterize the landscape of genomic cancer signals in the blood of people with and without cancer. The study enrolled approximately 15000 participants. Their goal is early and simple detection of cancer from analysis of methylations on cell-free DNA.

10.2.4.1 Obtaining methylation profiles

They performed whole genome methylation profiles after choosing between three different independent methods:

- Whole genome sequencing.
- Targeted sequencing.
- Whole genome sequencing for CNVs.

In the second phase an assay for a targeted methylation study has been developed: the best features to discriminate between the two classes (cancer vs non-cancer) were selected in order to sequence the areas with these modifications without whole genome analysis. A model for this classification was developed, trained and validated. The last step is a large-scale clinical validation with a 5 years follow-up that is still in progress.

10.2.4.2 Results

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

10.2.5 Deconvolution approach

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

10.2.5.1 Considering liquid biopsies

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

10.2.5.2 High-quality reference atlases

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

10.2.6 Targeted panel approaches for tumour content estimation

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

Part II

Laboratory

Chapter 11

Relevant file's formats

11.1 FASTA format

FASTA files contain information about a sequence and its quality score.

11.1.1 Components

FASTA files are composed of:

1. A line beginning with > followed by a sequence ID and a sequence description.
2. The sequence.
3. Quality scores for each base.

A multi-fasta file is obtained as a simple concatenation of individual FASTA, with > as the separator.

11.1.2 Alphabet

The alphabet of FASTA is built such that for DNA and RNA:

- $ATCG$ for the normal bases.
- N for an unknown base.
- R for either A or G.
- Y for either C or T.
- For RNA T is replaced by the U.

For Proteins:

- Standard letter for standard amino acids.
- X for unknown amino acids.
- $OBZJ$ for protein amino acid modifications.

11.1.3 DNA sequence quality

DNA sequences have a quality value associated with each nucleotide. This score is a measure of the reliability for each base, as it is derived from the physical process of sequencing. The quality score

is formalized by the Phred software for the human genome project: let P be the probability of a base call being incorrect, then the quality score Q is computed as:

$$Q = -10 \log P \Leftrightarrow P = 10^{-\frac{Q}{10}}$$

11.2 FASTQ format

FASTQ files are similar as FASTA files, their differences being that the start symbol is an @ instead of an >, and the quality score is separated from the sequence with a + followed by a blank line. Moreover the quality encoding uses letters or symbols to represent numbers.

11.2.1 Data compression

FASTQ files are very big, as they are typically more than $10GB$. Therefore they will often be compressed with *gzip*, in order to get to $\leq 20\%$ of their original size.

11.3 SAM and BAM formats

11.3.1 SAM files

SAM files contain information about on the alignment of each read, optimized for readability and sequential access.

11.3.1.1 COmposition

They are composed by an header containing information about:

- Version.
- Reference sequences.
- Read groups with platform information.
- Processing history.

Following the header alignment records are found, containing:

- Query name.
- Flag.
- Mapping quality.
- CIGAR.
- Sequence.
- Quality.

11.3.2 BAM files

BAM files are binary SAM files, compressed and optimized for size. They may be sorted and indexed at the location query. A sorted and indexed BAM is the default for an analysis pipeline and it is converted into a SAM file only to allow visualization.

11.3.3 Operation with SAM and BAM files

Samtools by default expects a BAM file as input and will produce a SAM file as output. Alignment results are typically stored as a sorted and indexed BAM file. Aligners produce SAM files.

11.3.3.1 Converting from SAM to BAM

1. samtools view -Sbh Normal.sam > Normal.bam
2. samtools sort Normal.bam > Normal.sorted.bam
3. samtools index Normal.sorted.bam

The file is sorted so that read pairs are next to one another, typically with the same order as the FASTQ file. Sorting will depend on the next analysis method to be used.

11.3.3.2 Filtering

Before performing downstream analysis, the BAM file must be cleaned and processed to eliminate biases:

1. Count reads: samtools view -c Normal.sorted.bam
2. Reads mapping to reverse strand f: samtools view -c -f 16 Normal.sorted.bam
3. Reads mapping to forward strand F: samtools view -c -F 16 Normal.sorted.bam
4. Mapping quality > 30: samtools view -c -q 30 Normal.sorted.bam

11.3.3.3 Explore statistics

Different operations are allowed to explore statistics of a bam file:

1. General statistics: samtools stats Normal.sorted.bam > Stats.txt
2. Single base sum coverage per region: samtools bedcov CG100.bed Normal.sorted.bam > BEDCov.txt
3. Single base depth: samtools depth -b CG100.bed Normal.sorted.bam > BEDDepth.txt

11.3.3.4 Mpileup

Mpileup allow to compute the pileup, the number of reads aligned to a reference sequence in a region:

1. samtools mpileup -r 1:3410684-3410690 Normal.sorted.bam
2. samtools mpileup -r 1:3410684-3410690 -q 60 -Q 60 Normal.sorted.bam

The different read support for a specific gene could be checked through:

samtools bedcov CG100.bed Normal.sorted.bam | grep "TP53"

The numbers obtained from this commands should always be normalized with the total number of reads.

Chapter 12

Data pre-processing

12.1 Realignment

12.1.1 Introduction

The identification of indels is not easy to be done by mappers. In particular indels across the ends or in complex regions of reads are of difficult detection. This generates some artefact mismatches, which if not corrected can introduce biases that propagates across the downstream analysis. Realignment is an operation performed on BAM file in order to improve accuracy of other pre-processing steps.

12.1.2 An example

A typical situation in which realignment helps is one in which a small homopolymeric region is in the middle of two regions with consecutive SNPS. Given the complexity the aligner is not able to identify an indel in that particolar region. Because of this some artefacts will be found at the end of the reads. The scoring function allow to accept an alignment even in the presence of gaps and mismatches. In this case a misalignment could introduce mismatches, reducing the score of the alignment.

12.1.3 Objective of realignment

The objective of realignment is to identify regions hiding indels. This process can be done in two steps:

1. Known sites collected in databases like 1000 genomes and dbSNP are checked.

2. The CIGAR line gives information about the goodness of an alignment. All reads are explored and regions where indels are present should be identified. If this is not feasible the density of variation of quality at difference reads should be explored. For example an high density of mismatches with respect to expectancy suggests the presence of an indel.

12.1.4 GATK

GATK implements realignment considering the default CIGAR line and the density of variation to provide known indes. To do so GATK considers each alignment and:

1. Inserts an indel and finds a better alignment with an alternative consensus sequence.
2. The score for that alternative consensus is computed as the total sum of the quality scores of mismatching bases.
3. If the score of the best alternative consensus is deemed significantly better by a LOD score the proposed alignment of the reads is accepted.

12.1.5 Protocol

- Apply RealignedTargetCreator to the BAM file to identify which regions need to be realigned. By default the tool uses the density criteria and the CIGAR one, but it is also possible to provide a file with a list of known sites.
- IndelRealigner performs the actual realignment at the RTC target intervals using the same input files.

12.1.6 Realignment results

When the new BAM file is created realigned reads change their CIGAR but maintain the original one with an OC tag. Because of this it is easy to check how the realignment was performed. Realignment is a useful step as position artefacts might lead to an incorrect correlation between a patient and pathogenic SNPs.

12.2 Recalibration

12.2.1 Introduction

Base quality score recalibration involves assigning accurate confidence scores to each sequence. Quality scores are critical for all the downstream analysis and systematic biases are a major contributor to bad calls.

12.2.1.1 Dealing with systematic errors

Systematic errors correlate with base call feature like:

- Reported quality score.
- Position within the read due to the machine cycle.
- Sequence context due to the chemistry of sequencing.

The error distribution and how error varies with base call features can be computed empirically. This process is made possible by looking at works per read groups, or entire lane of data.

12.2.2 Computing empirical qualities

Except for known variants any sequence mismatch represents an error. The number of observation and the number of errors are taken into account as a function of the various covariates that give rise to systematic errors.

12.2.2.1 Base quality score recalibration

A PHRED scaled quality score is computed as:

$$\frac{\#of\ reference\ mismatches + 1}{\#of\ observed\ bases + 2}$$

Base quality score recalibration or BQRS is a method that adjust the PHRED quality scores to be more accurate by looking at every base in a BAM file. To run BQSR an additional file is necessary to make the process aware of all known single nucleic polimoprhisms SNPs. Each read base is grouped into separate group bins, as different sequencing machines may be calibrated differently. Then the data is split in a read group by the quality scores. Then BQSR compare the scores reported by the sequencing machines with the empirical scores derived from the empirical error counts. An empirical PHRED score is computed as:

$$Q_{actual} = Q_{global} + \Delta all + \Delta readgroup + \Delta quality + \sum_{i}^{n} \Delta covariate_i$$

Post-recalibration quality scores should fit the empirically-driven quality scores well, without obvious systematic biases.

12.2.3 Protocol

A typical process for base quality score recalibration is composed of 4 steps:

1. BaseRecalibrator: model the error modes and recalibrate qualities. Its inputs are a BAM file and the known sites.
2. PrintReads: write recalibrated data to a BAM file thanks to the recalibration table produced in the previous step. Original qualities are retained with the OC flag.
3. The process is repeated to build the after model to evaluate remaining error.
4. AnalyzeCovariates: before and after plots are made based on recalibration tables.

12.3 Marking duplicates

12.3.1 Introduction

Duplicates are non-independent measurements of a sequence, as they are sampled from the exact same template of DNA, violating the assumptions of variant calling. Errors in sample or library preparation will be propagated to all the duplicates. Therefore, the best copy among all the duplicates should be taken to mitigate the effects of errors.

12.3.2 Identification of duplicates

Duplicates come from the same input DNA template, so they should have the same start position on reference. This is even more true for paired end, in which both the reverse and the forward read should have the same starting position. Duplicate sets are first identified, then the representative read based on base quality scores an other criteria is chosen for each set.

12.3.2.1 Consideration on Borrow-Wheeler aligner

Borrow-Wheeler aligner sometimes clip bases from the ends of the alignment, so fragments mapped to the reverse strand are specified by their $3'$ position instead of the $5'$. SAM flags and CIGAR strings are needed to determine the unclipped $5'$ ends.

12.3.3 Protocol

Marking the duplicates can be done in two different ways:

- MarkDUplicates from Picard, the golden standard.
- markdup from samtool. It requires the addition of mate tags to the BAM file through the fixmate command.

The samtools command is faster than the Picard one as it exploits the Cigar of the mate read to correct with a simple iteration. However the Picard command retains more reads because samtools' command removes all the reads that have a mae mapped to a different chromosome, removing in this way structural variants.

Chapter 13

Variant calling

13.1 Introduction

13.1.1 Objective of variant calling

The objective of variant calling is to compute allelic fractions across all positions. Thresholds or basic statistics can be used to distinguish between genotypes. The goal is to identify germline variants. Quality thresholds can be added.

13.1.2 Bayes' rule for variant calling

Refined methods can find the genotype of each sample by calculating via Bayes' rule the probability of each possible genotype/ Let $P(G)$ be the prior of the genotype, $P(D|G)$ the likelihood of the genotype and $P(D|H)$ the haploid likelihood function. Then, given $G = H_1H_2$ the diploid assumption, the Bayesian model is computed as:

$$\begin{cases} P(G|D) = \frac{P(G)P(D|G)}{\sum\limits_i P(G_i)P(D|G_i)} \\ P(D|G) = \prod\limits_j \left(\frac{P(D_j|H_1)}{2} + \frac{P(D_j|H_2)}{2}\right) \end{cases}$$

Where:

$$P(D_j|H) = P(D_j|b)$$

Is a single base pileup and:

$$P(D_j|b) = \begin{cases} 1 - \epsilon_j & D_j = b \\ \epsilon_j & D_j \neq b \end{cases}$$

13.2 Likelihood estimation for variant calling

The inference used as a gold standard relies on a likelihood function to estimate the probability of sample data given the proposed haplotype. The probability is computed by calculating the support of the alternative base based on quality. All diploid genotypes are considered at each base. The likelihood of genotype is computed using only pileup of bases and associated quality scores at a given

locus. Only bases satisfying minimum base quality, mapping read quality, pair mapping quality or other quality parameter are considered. This approach can be implemented as a multi-sample calling, where it is possible to obtain joint estimates across samples.

13.2.1 Available tools

Available tools to perform this task are:

- Bcftools: bcftools mpileup -Ou -a DP -f human_g1k_v37.fasta Sample.sorted.bam | bcftools call -Ov -c -v > Sample.BCF.vcf
- GATK: java -jar GATK/GenomeAnalysisTK.jar -T UnifiedGenotyper -R human_g1k_v37.fasta -I Sample.sorted.bam -o Sample.GATK.vcf -L chr20.bed

13.3 VCF files

VCF is a text file format, most likely stored in a compressed matter, that contains meta-information lines, a header line and data lines, each containing information about a position in the genome.

13.3.1 Composition

The fileformat field is required and it should detail the VCF format version number. Descriptions of field INFO, FILTER and FORMAT are optional but strongly encouraged. The header line names 8 fixed, mandatory columns:

- CHROM: chromosome.
- POS: position.
- ID: semi-colon separated list of unique identifiers.
- REF: reference bases.
- ALT: comma separated list of alternative non-reference alleles called on at least one of the samples.
- QUAL: prhed-scaled quality score for the assertion made in ALT.
- FILTER: pASS if the position has passed all filters. "." when filters are not applied.
- INFO: additional information.

Genotype data are followed by a FORMAT column header and an arbitrary number of sample IDs. The header line is tab-delimited.

13.3.2 Vcftools

The package vcftools can be used to perform a number of operations on VCF files:

- Filter out specific variants for quality, mean depth or allelic fraction.
- Compare files.
- Summarize variants.
- Convert to different file types.
- Validate and merge files.
- Create intersections and subsets of variants.

13.4 A variant calling protocol

A variant calling protocol consists of two steps. In the first bcftools pileup provides the supported bases and their quality. The highest support is checked through a Bayesian model. A position with an alternative allele is reported in the sample VCF file. Then in the second step vcftools or GATK are used to produce a new VCF with the corresponding informations. GATK keeps more variants with respect to vcftools, but there is no method clearly better than the other. Usually the consensus is used to compare the output of the two tools and an intersection is built. Moreover the two tools show a lot of similarities in their output. Vcftools are more confident in making calls while GATK can introduce a portion of false positives. This can be reduced by increasing the filter on minimal coverage and reaching a set of more trustable parameters.

Chapter 14

Variant annotation

14.1 Introduction

Variant annotation is a crucial step in linking sequence variants with changes in phenotype. Annotation results can have a strong influence on the ultimate conclusions of disease studies. Incorrect or incomplete annotations can cause researchers both to overlook potentially disease-relevant DNA variants and to dilute interesting variants in a pool of false positives. The annotation capabilities depend on the pipeline, the variant caller and on the filtering threshold used.

14.1.1 Annotation databases

Different databases can be used for annotation.

14.1.1.1 Genomic data repositories

Genomic data repositories include:

- 1000 genomes.
- ExAC.
- gnomAD: exome and genome.
- dbSNP.

14.1.1.2 Variant-disease databases

Variant-disease databases include:

- ClinVar: variation and phenotype.
- Gencode.
- HGMD: mutations.
- COSMIC: human cancer.
- OMIM: mendelian.

14.1.1.3 Conservation and pathogenicity prediction databases

Conservation and pathogenicity prediction databases provide computational methods for prediction and include:

- SIFT: impact of missense variants.
- PolyPhen2: deleteriousness of change.
- GERP++: sequence conservation.

14.2 SnpEff

14.2.1 Introduction

SnpEff is a variant effect predictor program categorizing each variant based on its relationship to coding sequenced in the genome and how it may change the coding sequence and affect the gene product.

14.2.2 Set of transcripts

Variant annotation depends on the set of transcripts used as the bases for annotation. Widely used annotation databases such as ENSEMBL, RefSeq and UCSC contain sets of transcripts that can be used for variant annotation.

14.2.3 Populating the VCF

SnpEff take information from the provided annotation database and populate a VCF file by adding it itno the INFO field name ANN. Data fields are encoded separated by the pipe sign | and the order of fields is written in the VCF header. More than $4GB$ are necessary to run a full annotation. Depending on the database a more or less complete annotation will be obtained.

14.2.4 Common annotations

Some common annotations include:

- Putative_impact or impact: a simple estimation of putative impact or deleteriousness, can take values HIGH, MODERATE, LOW, MODIFIER.
- Gene name: common gene name HGNC. Optionally it can use the closest gene when the variant is intergenic.
- Feature type: with type of feature, for example, transcript, motif or miRNA. It is preferred to use Sequence Ontology SO terms, but custom ones are allowed.
- Feature ID: this may be the transcript ID, the motif ID, the miRNA, ChipSeq peak or histone mark for example. Some features may not have a unique ID.
- Byotipe: a description on whether the transcript is coding or non-coding.

14.2.5 SnpSift

Each gene can be present in many transcript evrsions, so the estimated impact is different. SnpSift helps filter large genomic datasets in order to find the most significant variants. Complex expression to filter and combine can be used while annotation. All fields in ANN can be managed.

Chapter 15

Ancestry

15.1 Introduction

15.1.1 Population stratification

The degree of admixture and population of origin identification is important for population stratification correction in:

- GWAS studies.
- Ancestry.
- Migration pattern studies.
- Precision medicine.

15.1.2 Methods

15.1.2.1 Principal Component analysis based methods

In principal component analysis methods the low-dimensional projection of the data allows to maximally retain variance-covariance structure among the genotypes. Fast algorithms are available to solve the problem, but interpretation might be non-trivial. Some examples are:

- EIGENSTRAT.
- SMARTPCA.
- LASER.
- EthSEQ.

15.1.2.2 Model based methods

The explicit generative model for the data is based on Hardy-Weinberg equilibrium and linkage equilibrium. Probabilistic methods estimate ancestry information from inference of the best model parameters that fit the data. Some examples are:

- STRUCTURE.
- FastSTRUCTURE.
- ADMIXTURE.
- FRAPPE.

15.2 SMARTPCA

15.2.1 Introduction

SMARTPCA is a PCA-based method specifically developed for SNPs data. The input is the genotype of N SNPs for M individual in a tool-specific format. It is also possible to apply a conversion from

the PED format. The genotype data is always encoded in a matrix, the difference between data formats involves the SNP information.

15.2.2 PED format

A PED file contains a:

- Genotype file: information about the individual in the first 6 lines, like ID, sex, affection and genotype.
- Map file: chromosome, marker ID, genetic distance and physical position.

15.2.3 Output

SMARTPCA converts the two files of the PED format in a proprietary one. Then monomophic and suboptimal SNPs are excluded from the PCA. The output is a table of an user defined number of eigenvector. The result will be a plot with clusters of individuals divided according to geographical origin. A more detailed analysis could be performed increasing the number of individuals.

15.3 fastSTRUCTURE

15.3.1 Input

The input for fastSTRUCTURE is the genotype of N SNPs for M Individuals in BED, BIM or FAM formats. BED is a binary format with genotype data, while BIM and FAM contains SNPs and samples information.

15.3.2 Output

Data is clustered and for each individual the amount of genetic background coming from a specific population is determined. The method is supervised, with a number of cluster given as input beforehand. The output is a matrix containing a probability value for each of the inferred cluster for each individual. The final representation allow to visualize patterns of ancestry.

15.4 EthSEQ

15.4.1 Introduction

Starting from the idea of PCA methods, model-based reasoning is applied. The starting point is a set of genotype data and a reference model with a set of individuals from known populations, creating an aggregated model. After the PCA it is easier to quantify the level of admixture and population of origin exploiting the reference data. The EthSEQ pipeline is described in figure 15.1.

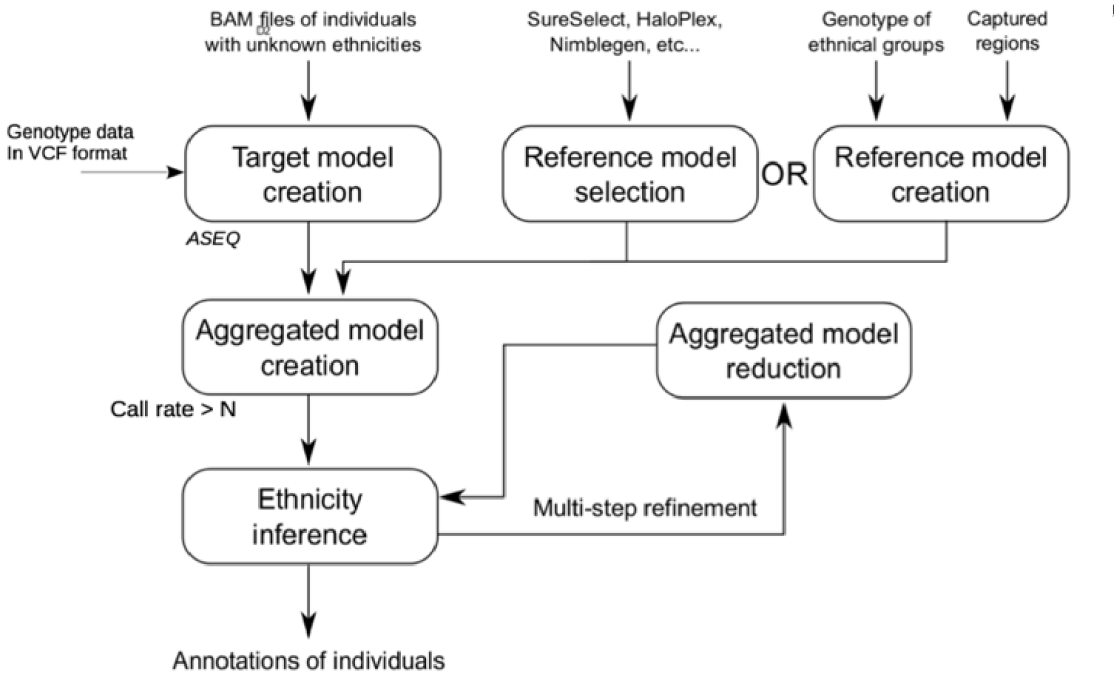


Figure 15.1: EthSEQ pipeline.

15.4.2 Analysing sequencing data

EthSEQ is design to analyse whole exome or targeted sequencing data from BAM or vcf file. The PCA in 3D shows that exploiting the aggregated model aids in the analysis, allowing to input just one target individual. The procedure relies on reference groups forming well-distinguished clusters. By considering the position in which the target sample falls in space, it can be assigned to a specific cluster.

15.4.3 Multi step refinement

The multi-step refinement process allows to repeat the analysis recursively on a smaller subset of close population to better distinguish between structures.

15.4.4 Dealing with ambiguous points

When a point lies between different clusters, a method computing the distance from the centroid is applied in order to infer the relative contribution of each population. This works well in the case of evident admixture.

Chapter 16

Somatic variant calling

Chapter 17

Somatic copy number calling

Part III

Papers

Chapter 18

Role of non-coding sequence variants in cancer

18.1 Abstract

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

18.1.1 Introduction

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

18.2 Genomic sequence variants

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

18.3 Non-coding element annotation

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

18.3.1 Cis regulatory regions

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

18.3.2 Distal regulatory elements

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

18.3.3 RNA-seq

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

18.3.4 Transcribed pseudogenes

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

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

18.3.5 Evolutionary conservation

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

18.4 Roles for somatic variants in cancer

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

18.4.1 Gain of TF-binding sites

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

18.4.2 Fusion events due to genomic rearrangements

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

18.4.3 ncRNAs and their binding sites

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

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

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

18.5 Roles for germline variants in cancer

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

18.5.1 Promoter mutations

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

18.5.2 SNPs in enhancers

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

18.5.3 Variants in introns

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

18.5.4 SNPs in ncRNA and their binding sites

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

18.5.5 Others

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

18.6 Interplay between germline and somatic variants

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

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

18.7 Computational methods for identifying variants

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

18.8 Experimental approaches for functional validation

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

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

Chapter 19

Advances in understanding cancer genomics through second-generation sequencing

19.1 Abstract

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

19.1.1 Introduction

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

19.2 Cancer-specific consideration

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

19.2.1 Characteristics of cancer samples for genomic analysis

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

19.2.2 Structural variability of cancer genomes

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

19.3 Experimental approaches

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

19.3.1 Whole genome sequencing

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

19.3.2 Exome sequencing

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

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

19.3.3 Transcriptome sequencing

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

19.4 Detecting classes of genome alterations

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

19.4.1 Somatic nucleotide substitutions and small insertion and deletion mutations

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

19.4.2 Copy number

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

19.4.3 Chromosomal rearrangements

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

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

19.4.4 Microbe-discovery methods

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

19.5 Computational issues

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

19.5.1 Alignment and assembly

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

19.5.2 mutations detection

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

19.5.3 Validation of mutation and rearrangement calls

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

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

Chapter 20

Integrative genomics viewer

20.1 Introduction

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

Chapter 21

Tumour heterogeneity and resistance to cancer therapies

21.1 Abstract

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

21.1.1 Introduction

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

21.2 Causes of intratumoral heterogeneity

21.2.1 Genomic instability

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

21.2.2 The clonal evolution and selection hypothesis

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

21.3 The spectrum of tumour heterogeneity

21.3.1 Spatial heterogeneity

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

21.3.1.1 Heterogeneity at a single disease site

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

21.3.1.2 Comparison of spatially distinct disease sites

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

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

21.3.2 Temporal heterogeneity

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

21.3.2.1 Genomic complexity might increase with exposure to targeted therapies

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

21.3.2.2 Longitudinal sampling provides insight into temporal heterogeneity

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

21.3.2.3 Residual drug-tolerant cells can foster temporal heterogeneity

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

21.4 Noninvasive monitoring of heterogeneity

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

21.4.1 Analysis of ctDNA

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

21.5 Overcoming heterogeneity

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

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

Chapter 22

Unravelling the clonal hierarchy of somatic genomic aberrations

22.1 Introduction

22.1.1 Abstract

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

22.1.2 Background

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

22.2 Results

22.2.1 Clonality assessment of aberrations from sequencing reads

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

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

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

22.2.2 Inferring the order of mutations in a tumour sample

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

22.2.3 In silico and in situ experimental validation

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

22.2.4 Comparative analysis reveals different mechanisms of tumour deregulation

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

pressure is acting later and may promote cancer progression at a later stage. PTEN is homogenously lost in metastatic melanoma. In lung cancer this loss is more rare. CLONET can identify tumour lineage specific subclonality.

22.2.5 Clonal hierarchy of genomic aberrations

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

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

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

22.3 Materials and methods

22.3.1 CLONET pipeline

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

22.3.2 CLONET on exome and targeted sequencing data

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

22.3.3 Expected distribution of the allelic fraction of a genomic segment

Consider a genomic segment that spans a set of informative SNPs for the individual of intersect. For any of them with coverage *cov* the total number of reads r supporting the reference base is the

sum of the neutral reads r_n and the active reads r_a supporting the reference base. β is the ratio between neutral reads and the total number of reads spanning the SNP of interest. The probability of having k reference reads is the convolution of the probability of observing $\beta \cdot k$ neutral reads and $(1-\beta) \cdot k$ active reads:

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

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

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

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

22.3.4 Estimated proportion of neutral reads for a genomic segment

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

22.3.5 From neutral to non-aberrant reads

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

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

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

22.3.6 From aberrant reads to aberrant cells

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

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

The proportion of non-aberrant reads covering M is:

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

22.3.7 Uncertainty assessment and its propagation to clonality estimates

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

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

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

22.3.8 Clonality of bi-allelic deletion

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

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

Chapter 23

TPES: timor purity estimation from SNVs

23.1 Abstract

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

23.1.1 Introduction

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

23.2 Materials and methods

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

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

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

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

Chapter 24

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

24.1 Abstract

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

24.1.1 Introduction

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

24.2 Material and methods

24.2.1 Genotype distance

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

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

Where:

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

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

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

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

24.2.2 SNP panel selection procedure

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

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

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

24.2.3 SPIA probabilistic test on cell line genotype distance

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

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

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