Human genomics

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

June 1, 2022

Contents

I Notes **4**

1 Introduction **5**
1.1 Genetics vs Genomics . . . 5
1.1.1 Genetics . . . 5
1.1.2 Genomics . . . 5
1.1.3 Differences . . . 5
1.1.4 Role of computational biology . . . 5
1.2 Human Genomics - the Basis . . . 6
1.2.1 Genetic Make-Up . . . 6
1.2.2 Inherited variants' relevance . . . 6
1.2.3 Differences in Genetic Make-Up, an example . . . 7
1.2.4 Acquired DNA aberrations . . . 7
1.3 Experimental techniques to detect variants/aberrations . . . 8
1.3.1 Cariotyping . . . 8
1.4 Sequence capture for cancer genomics . . . 8
1.4.1 Single End (SE) and Paired End (PE) reads . . . 8

II Papers **10**

2 Role of non-coding sequence variants in cancer **11**
2.1 Abstract . . . 11
2.1.1 Introduction . . . 11
2.2 Genomic sequence variants . . . 11
2.3 Non-coding element annotation . . . 12
2.3.1 Cis regulatory regions . . . 12
2.3.2 Distal regulatory elements . . . 12
2.3.3 RNA-seq . . . 12
2.3.4 Transcribed pseudogenes . . . 12
2.3.5 Evolutionary conservation . . . 13
2.4 Roles for somatic variants in cancer . . . 13
2.4.1 Gain of TF-binding sites . . . 13
2.4.2 Fusion events due to genomic rearrangements . . . 13
2.4.3 ncRNAs and their binding sites . . . 13
2.4.4 Role of pseudogenes in modulating the expression of a parental gene . . . 14
2.5 Roles for germline variants in cancer . . . 14

2.5.1 Promoter mutations . . . 14
2.5.2 SNPs in enhancers . . . 14
2.5.3 Variants in introns . . . 14
2.5.4 SNPs in ncRNA and their binding sites . . . 14
2.5.5 Others . . . 14
2.6 Interplay between germline and somatic variants . . . 14
2.7 Computational methods for identifying variants . . . 15
2.8 Experimental approaches for functional validation . . . 15

3 Advances in understanding cancer genomics through second-generation sequencing 17
3.1 Abstract . . . 17
3.1.1 Introduction . . . 17
3.2 Cancer-specific consideration . . . 17
3.2.1 Characteristics of cancer samples for genomic analysis . . . 18
3.2.2 Structural variability of cancer genomes . . . 18
3.3 Experimental approaches . . . 18
3.3.1 Whole genome sequencing . . . 18
3.3.2 Exome sequencing . . . 18
3.3.3 Transcriptome sequencing . . . 19
3.4 Detecting classes of genome alterations . . . 19
3.4.1 Somatic nucleotide substitutions and small insertion and deletion mutations . 19
3.4.2 Copy number . . . 19
3.4.3 Chromosomal rearrangements . . . 19
3.4.4 Microbe-discovery methods . . . 20
3.5 Computational issues . . . 20
3.5.1 Alignment and assembly . . . 20
3.5.2 mutations detection . . . 20
3.5.3 Validation of mutation and rearrangement calls . . . 20

4 Integrative genomics viewer 22
4.1 Introduction . . . 22

5 Tumour heterogeneity and resistance to cancer therapies 23
5.1 Abstract . . . 23
5.1.1 Introduction . . . 23
5.2 Causes of intratumoral heterogeneity . . . 23
5.2.1 Genomic instability . . . 23
5.2.2 The clonal evolution and selection hypothesis . . . 24
5.3 The spectrum of tumour heterogeneity . . . 24
5.3.1 Spatial heterogeneity . . . 24
5.3.2 Temporal heterogeneity . . . 25
5.4 Noninvasive monitoring of heterogeneity . . . 26
5.4.1 Analysis of ctDNA . . . 26
5.5 Overcoming heterogeneity . . . 26

6 Unravelling the clonal hierarchy of somatic genomic aberrations **28**
6.1 Introduction . . . 28
6.1.1 Abstract . . . 28
6.1.2 Background . . . 28
6.2 Results . . . 29
6.2.1 Clonality assessment of aberrations from sequencing reads . . . 29
6.2.2 Inferring the order of mutations in a tumour sample . . . 29
6.2.3 In silico and in situ experimental validation . . . 30
6.2.4 Comparative analysis reveals different mechanisms of tumour deregulation . . . 30
6.2.5 Clonal hierarchy of genomic aberrations . . . 31
6.3 Materials and methods . . . 31
6.3.1 CLONET pipeline . . . 31
6.3.2 CLONET on exome and targeted sequencing data . . . 31
6.3.3 Expected distribution of the allelic fraction of a genomic segment . . . 31
6.3.4 Estimated proportion of neutral reads for a genomic segment . . . 32
6.3.5 From neutral to non-aberrant reads . . . 32
6.3.6 From aberrant reads to aberrant cells . . . 32
6.3.7 Uncertainty assessment and its propagation to clonality estimates . . . 33
6.3.8 Clonality of bi-allelic deletion . . . 33

7 TPES: timor purity estimation from SNVs **34**
7.1 Abstract . . . 34
7.1.1 Introduction . . . 34
7.2 Materials and methods . . . 34

8 SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines **36**
8.1 Abstract . . . 36
8.1.1 Introduction . . . 36
8.2 Material and methods . . . 36
8.2.1 Genotype distance . . . 36
8.2.2 SNP panel selection procedure . . . 37
8.2.3 SPIA probabilistic test on cell line genotype distance . . . 37

Part I

Notes

Chapter 1

Introduction

1.1 Genetics vs Genomics

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 math techniques known as bioinformatics, genomics researchers analyse enormous amounts of DNA-sequence data to find variations that affect health, disease or drug response. In human that means searching through about 3 billion units of DNA across 23000 genes.

1.1.3 Differences

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

1.1.4 Role of computational biology

Computational biology offer a wide range of numerical methods to analyse and integrate large scale data towards the understanding of molecular, cellular and structural biology. The focus of this course is on human genomics and how to mine raw data, how to exploit it for quality control and how to interpret the results in the context of human disease, especially cancer.

1.2 Human Genomics - the Basis

1.2.1 Genetic Make-Up

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

1.2.1.1 Single nucleotide polymorphisms

Single nucleotide polymorphisms or SNPs are changes of one nucleotide in the sequence of a gene. They constitute 1% of the difference between two unrelated individuals' genomes.

1.2.1.2 Copy number variants

Copy number variants or CNVs are difference of the number of allele for a gene present in one individual. They contribute much more than SNPs in the difference between unrelated individuals. If both parents are monozygous in one gene the child could have zero copy of the gene.

1.2.2 Inherited variants' relevance

Inherited variants can be characterized by penetrance and allele frequency.

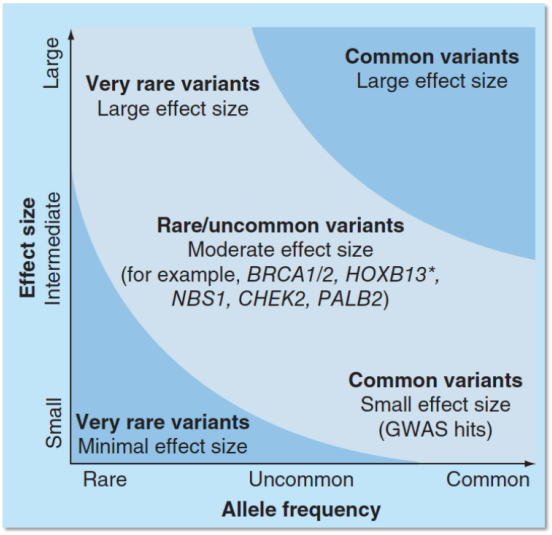

Figure 1.1: R. Eeles, Future Sci. OA (2016) 2(1), FSO87. Review on prostate cancer

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

1.2.2.0.2 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.

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

1.2.3 Differences in Genetic Make-Up, an example

One example of how the genetic make-up plays a role in deases is the ADME genes. ADME stands for *Absorption, distribution, metabolism and elimination.* It is a set of genetic variants that are able to change ability of the organism to react to certain compounds (pharmacokinetic variability),influencing the patients' treatment response. Both common and rare variants are involved. A therapeutic approach that considers these variations could be very useful in precision medicine.

Somatic variance: not acquired from parents. SNV and SNP are basically the same thing, but SNV are restricted to a certain population of cells, while SNP are genetically encoded in all cells of the organism. Other types of somatic variance are rearrangemets (gene translocation, chromosome breakage, chromotripsy), somatic copy number changes are equivalt to copy number variance but somatic (SCNA).

1.2.4 Acquired DNA aberrations

Variants that are not inherited from parents and are not transmitted to offspring are called **somatic variants**. They are usually caused by DNA aberrations. DNA aberrations happen in diseased or aged cells and are the key to cancer genomics. They can be

- Single nucleotide variants (or SNV or point mutation). SNV and SNP are basically the same thing, but SNV are restricted to a certain population of cells, while SNP are genetically encoded in all cells of the organism.
- Indels or deletions
- Rearrangements
- Somatic copy number aberrations or SCNA

1.2.4.1 Types of acquired DNA aberrations

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

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

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

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

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

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

1.3 Experimental techniques to detect variants/aberrations

1.3.1 Cariotyping

Basically all the aberrations described in 1.2.4.1 were discovered in the last 10-15 years because there's the need of NGS. Cariotyping indeed is not enough! Sequence specific variants, breakpoints, etc. could not be detected until NGS.

1.4 Sequence capture for cancer genomics

In the paper [2] it is described a typical sequence capture for cancer genomics. One of the main realizations is that the test reference genome is the normal (non-cancer) DNA. Intuitively, we align both cancer and normal DNA so that we can detect if an aberration is cancer specific or it is present also in the normal DNA. The **match normal** is used to distinguish SNV from rare SNPs, but also copy number variation. Baits are nowadays used in the sequencing step, in order to sequence only the exome (usually, money issue). Another concept is the need to sequence *deeply*, to find subclonal events that give the cancer maybe some fitness (escaping immune system), but also because if the sample comes from the tissue, there are both cancer and some healthy cells and we need to able to distinguish them. A more in depth discussion is provided in *paper*.

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

Paired end useful especially for detecting structural variance. Massive information of relative position of a molecule wrt the reference gneome.

Figure 1.2 gives a nice graphical overview of genomic aberrations detectable by NGS, especially using PE sequencing. Important to notice is that performing PE sequencing is like having double the coverage. Both for homozygous and hemizygous deletions and insertions the single most important

[1] See Khurana E et al, NATURE REVIEWS | GENETICS, 2016
[2] Meyerson et al., Nature Reviews Genetics 2010

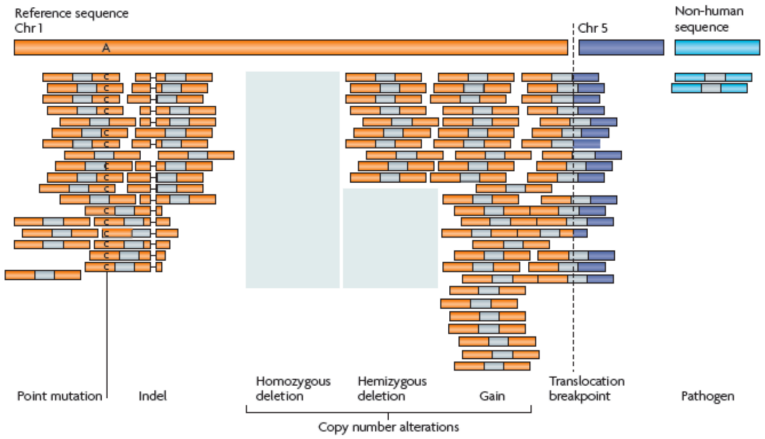

Figure 1.2: *Advances in understanding cancer genomes through second-generation sequencing*, Meyerson et al., Nature Reviews Genetics 2010

thing we need to care about is to have enough coverage on the whole experiment to perform significative downstream analysis. An important thing to notice in 1.2 is the translocation breakpoint: without PE we would not be able to detect the translocation event.

Part II

Papers

Chapter 2

Role of non-coding sequence variants in cancer

2.1 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.

2.1.1 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.

2.2 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.

2.3 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.

2.3.1 Cis regulatory regions

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

2.3.2 Distal regulatory elements

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

2.3.3 RNA-seq

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

2.3.4 Transcribed pseudogenes

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

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

2.3.5 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.

2.4 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.

2.4.1 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.

2.4.2 Fusion events due to genomic rearrangements

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

2.4.3 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.

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

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

2.5 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.

2.5.1 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.

2.5.2 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.

2.5.3 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.

2.5.4 SNPs in ncRNA and their binding sites

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

2.5.5 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.

2.6 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.

2.7 Computational methods for identifying variants

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

2.8 Experimental approaches for functional validation

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

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

Chapter 3

Advances in understanding cancer genomics through second-generation sequencing

3.1 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.

3.1.1 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.

3.2 Cancer-specific consideration

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

3.2.1 Characteristics of cancer samples for genomic analysis

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

3.2.2 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.

3.3 Experimental approaches

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

3.3.1 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.

3.3.2 Exome sequencing

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

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

3.3.3 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.

3.4 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.

3.4.1 Somatic nucleotide substitutions and small insertion and deletion mutations

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

3.4.2 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.

3.4.3 Chromosomal rearrangements

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

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

3.4.4 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.

3.5 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.

3.5.1 Alignment and assembly

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

3.5.2 mutations detection

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

3.5.3 Validation of mutation and rearrangement calls

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

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

Chapter 4

Integrative genomics viewer

4.1 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 5

Tumour heterogeneity and resistance to cancer therapies

5.1 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.

5.1.1 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.

5.2 Causes of intratumoral heterogeneity

5.2.1 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.

5.2.2 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.

5.3 The spectrum of tumour heterogeneity

5.3.1 Spatial heterogeneity

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

5.3.1.1 Heterogeneity at a single disease site

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

5.3.1.2 Comparison of spatially distinct disease sites

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

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

5.3.2 Temporal heterogeneity

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

5.3.2.1 Genomic complexity might increase with exposure to targeted therapies

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

5.3.2.2 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.

5.3.2.3 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.

5.4 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.

5.4.1 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.

5.5 Overcoming heterogeneity

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

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

Chapter 6

Unravelling the clonal hierarchy of somatic genomic aberrations

6.1 Introduction

6.1.1 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.

6.1.2 Background

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

6.2 Results

6.2.1 Clonality assessment of aberrations from sequencing reads

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

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

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

6.2.2 Inferring the order of mutations in a tumour sample

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

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

6.2.3 In silico and in situ experimental validation

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

6.2.4 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.

6.2.5 Clonal hierarchy of genomic aberrations

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

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

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

6.3 Materials and methods

6.3.1 CLONET pipeline

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

6.3.2 CLONET on exome and targeted sequencing data

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

6.3.3 Expected distribution of the allelic fraction of a genomic segment

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

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

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

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

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

6.3.4 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.

6.3.5 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.

6.3.6 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:

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

6.3.7 Uncertainty assessment and its propagation to clonality estimates

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

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

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

6.3.8 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 7

TPES: timor purity estimation from SNVs

7.1 Abstract

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

7.1.1 Introduction

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

7.2 Materials and methods

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

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

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

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

Chapter 8

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

8.1 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.

8.1.1 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.

8.2 Material and methods

8.2.1 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.

8.2.2 SNP panel selection procedure

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

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

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

8.2.3 SPIA probabilistic test on cell line genotype distance

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

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

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

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