

Human genomics

Giacomo Fantoni

telegram: @GiacomoFantoni

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

May 19, 2022

Contents

I	Notes	3
1	Introduction	4
1.1	Genetics vs Genomics	4
1.1.1	Genetics	4
1.1.2	Genomics	4
1.1.3	Differences	4
1.1.4	Role of computational biology	4
1.2	Human Genomics - the Basis	5
1.2.1	Genetic Make-Up	5
1.2.2	Inherited variants' relevance	5
1.2.3	Differences in Genetic Make-Up, an example	5
1.2.4	Acquired DNA aberrations	7
II	Papers	8
2	Role of non-coding sequence variants in cancer	9
2.1	Abstract	9
2.1.1	Introduction	9
2.2	Genomic sequence variants	9
2.3	Non-coding element annotation	10
2.3.1	Cis regulatory regions	10
2.3.2	Distal regulatory elements	10
2.3.3	RNA-seq	10
2.3.4	Transcribed pseudogenes	10
2.3.5	Evolutionary conservation	11
2.4	Roles for somatic variants in cancer	11
2.4.1	Gain of TF-binding sites	11
2.4.2	Fusion events due to genomic rearrangements	11
2.4.3	ncRNAs and their binding sites	11
2.4.4	Role of pseudogenes in modulating the expression of a parental gene	12
2.5	Roles for germline variants in cancer	12
2.5.1	Promoter mutations	12
2.5.2	SNPs in enhancers	12
2.5.3	Variants in introns	12
2.5.4	SNPs in ncRNA and their binding sites	12

2.5.5	Others	12
2.6	Interplay between germline and somatic variants	12
2.7	Computational methods for identifying variants	13
2.8	Experimental approaches for functional validation	13
3	Advances in understanding cancer genomics through second-generation sequencing	15
3.1	Abstract	15
3.1.1	Introduction	15
3.2	Cancer-specific consideration	15
3.2.1	Characteristics of cancer samples for genomic analysis	16
3.2.2	Structural variability of cancer genomes	16
3.3	Experimental approaches	16
3.3.1	Whole genome sequencing	16
3.3.2	Exome sequencing	16
3.3.3	Transcriptome sequencing	17
3.4	Detecting classes of genome alterations	17
3.4.1	Somatic nucleotide substitutions and small insertion and deletion mutations	17
3.4.2	Copy number	17
3.4.3	Chromosomal rearrangements	17
3.4.4	Microbe-discovery methods	18
3.5	Computational issues	18
3.5.1	Alignment and assembly	18
3.5.2	mutations detection	18
3.5.3	Validation of mutation and rearrangement calls	18
4	Integrative genomics viewer	20
4.1	Introduction	20
5	Tumour heterogeneity and resistance to cancer therapies	21
5.1	Abstract	21
5.1.1	Introduction	21
5.2	Causes of intratumoral heterogeneity	21
5.2.1	Genomic instability	21
5.2.2	The clonal evolution and selection hypothesis	22
5.3	The spectrum of tumour heterogeneity	22
5.3.1	Spatial heterogeneity	22
5.3.2	Temporal heterogeneity	23

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.

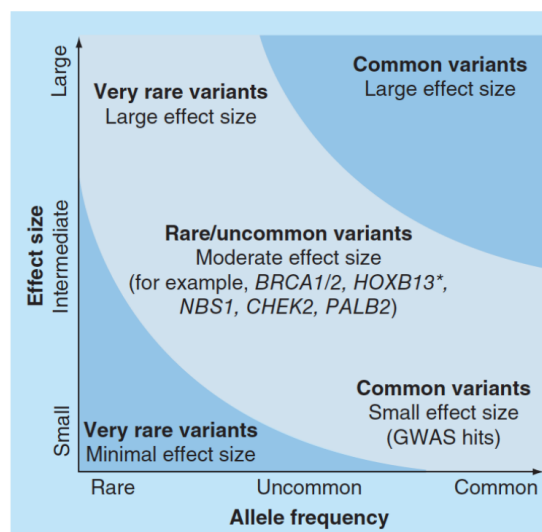


Figure 1.1: R. Eeles, Future Sci. OA (2016) 2(1), FSO87. Review on prostate cancer

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

1.2.3 Differences in Genetic Make-Up, an example

One example of how the genetic make-up plays a role in diseases 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 rearrangements (gene translocation, chromosome breakage, chromotripsy), somatic copy number changes are equivalent 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 moves (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 deletion. Could involve one or more chromosomes

1.3. EXPERIMENTAL TECHNIQUES TO DETECT VARIANTS/ABERRATIONS

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 ?? 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 be 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 genome.

Figure ?? 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

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

²Meyerson et al., Nature Reviews Genetics 2010

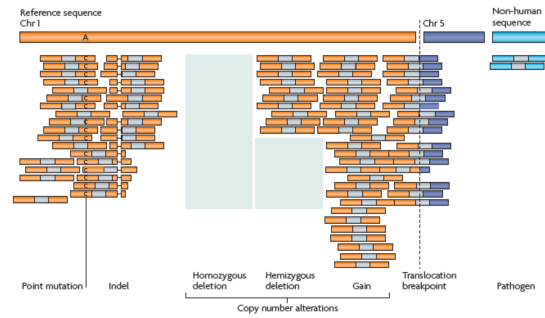


Figure 1.2: *Advances in understanding cancer genomes through second-generation sequencing*, Meyerson et al., Nature Reviews Genetics 2010

thing we need to care about is to have enough coverage on the whole experiment to perform significant downstream analysis. An important thing to notice in ?? 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 germline 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 chromatin 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 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 marks 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 of 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 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 variations 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 expression 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 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.

2.3.4 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 structure and contain a 3'-poly(A) tail. These pseudogenes can be transcribed

and regulated the expression of their parent genes, generating endo-siRNA and participating 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

because 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.

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 disrupt 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 happen 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 covariates 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 mutations 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 exist to annotate and prioritize potentially functional non-coding variants with high impact. These tools can interpret 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 introduce sequence variants involves the use of CRISPR-Cas9 systems. Then the effect is 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 barcode 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 is independent of their location, so they can be incorporated into high-throughput 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 assess the resulting expression level of the reporter driven by each variant element. Visible reporter assays using synthetic transcription reporter constructs that contain the regulatory sequence of the reporter gene enable direct validation. Other approaches are needed to validate variants in ncRNA, UTR and introns. 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 helps 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.

2.8. EXPERIMENTAL APPROACHES FOR FUNCTIONAL VALIDATION

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 clinical 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 create artefactual alteration. Nucleic acids from cancer are of lower quality due to formalin fixation and paraffin embedding necessary for microscopic 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 limitations: 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

3.4. DETECTING CLASSES OF GENOME ALTERATIONS

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-targeted hybridization can occur. The higher coverage makes 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 hampered 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 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 is 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 passenger 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 allows to estimate the tumour to normal copy number ratio 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 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.

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 issues 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. 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.

3.5.2 mutations detection

As somatic genome alterations 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.

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 fractions. Current efforts are focused on applying deep targeted second generation sequencing to

3.5. COMPUTATIONAL ISSUES

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 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 result 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 might 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. Intratumoral 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-scale 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 advantage. 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 of 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

5.3. THE SPECTRUM OF TUMOUR HETEROGENEITY

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 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 exert selective pressures on oncogene-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 tricar 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.

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