Human genomics

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

May 16, 2022

Contents

Part I

Notes

Chapter 1

Introduction

1.1 Definitions

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 moth 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 Differences in genetic make-up

The genetic make-up of individual is different between individuals and it is responsible for human diversity. These variants can be inherited or acquired.

1.2.1 Inherited variants

Between inherited variants single nucleotide polymorphisms and copy number variants can be identified.

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.

1.2.1.3 Characteristics of inherited variants

Inherited variants can be characterized by penetrance and allele frequency.

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

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

1.2.2 Acquired DNA aberrations

DNA aberrations happen in diseased or aged cells and are the key to cancer genomics. They are also called somatic variants: they are not inherited from parents and are not transmitted to offspring. They can be single nucleotide variants or SNV or point mutation, indels or deletions, rearrangements and somatic copy number aberrations or SCNA.

1.2.2.1 Types of acquired DNA aberrations

1.2.2.1.1 Translocation Translocation happens when a sequence is moved from one genetic locus to another. It can be an insertion or unbalanced, where only one sequence move or balanced, when two sequences exchange locus.

1.2.2.1.2 Inversion Inversion happens when a sequence inverts its orientation.

1.2.2.1.3 Duplication In duplication a sequence doubles its copy number.

1.2.2.1.4 Deletion In deletion a sequence is lost.

1.2.2.1.5 Chromoplexy Chromoplexy is 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 in a single cell cycle.

1.2.2.1.6 Chromothripsis Chromotripsis is a clustered chromosomal rearrangement in confined genomic regions that result from a single catastrophic event, usually limited to one chromosome.

1.2.2.1.7 Kataegis Kataegis is a phenomenon that is characterized by large clusters of mutations in the genome of cancer cells. An APOBEC family enzyme might be responsible for the kataegis process.

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.