Analysis and visualization of multidimensional cancer genomics data

Michael Philipp Schroeder

TESI DOCTORAL UPF / 2014

DIRECTORS DE LA TESI

Dra. Núria López Bigas & Dr. Abel González-Pérez

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES

Universitat Pompeu Fabra Barcelona

Acknowledgments

I'd like to thank Núria and Abel for their encouraging guidance throughout my thesis: It has always been a pleasure to work with you and under your supervision, which I deeply admire. Núria, thank you for giving me the chance to make my PhD in your group and for always seeing things positive and open-minded. It is an inspiration to learn from you.

It has been and is a pleasure to work with all the people that have been a part of our research group, Biomedical Genomics, where everybody contributed to the openness that has always been a ground value. I have to give special thanks to Jordi and Abel who were always available to share their expertise and allowed me to borrow their knowledge so many times so I could improve my understanding of cancer, computational biology and gain advanced programming skills.

Apart from my colleagues, I'd like to thank Ana infinitely and once more: my dedication during busy times would not have been possible without your love and support. Equally, my family back in Switzerland has always been supportive, thank you!

And last but not least I'd like to give thanks to all the people with whom I shared office space, football pitches, bar tables, improvised sofas, hikes, hugs and good times during these last four years. You know who you are and I love you!

Abstract

Cancer is a complex disease caused by somatic alterations of the genome and epigenome in tumor cells. Increased investments and cheaper access to various technologies have built momentum for the generation of cancer genomics data. The availability of such large datasets offers many new possibilities to gain insight into cancer molecular properties. Within this scope I present two methods that exploit the broad availability of cancer genomic data: OncodriveROLE, an approach to classify mutational cancer driver genes into activating and loss of function mode of actions and MutEx, a statistical measure to assess the trend of the somatic alterations in a set of genes to be mutually exclusive across tumor samples. Nevertheless, the unprecedented dimension of the available data raises new complications for its accessibility and exploration which we try to solve with new visualization solutions: i) Gitools interactive heatmaps with prepared large scale cancer genomics datasets ready to be explored, ii) jHeatmap, an interactive heatmap browser for the web capable of displaying multidimensional cancer genomics data and designed for its inclusion into web portals, and iii) SVGMap, a web server to project data onto customized SVG figures useful for mapping experimental measurements onto the model.

Resum

El cancer és una malaltia complexa causada per alteracions somàtiques del genoma i epigenoma de les cèl·lules tumorals. Un augment d'inversions i l'accés a tecnologies de baix cost ha provocat un increment important en la generació de dades genòmiques de càncer. La disponibilitat d'aquestes dades ofereix noves possibilitats per entendre millor les propietats moleculars del càncer. En aquest àmbit, presento dos mètodes que aprofiten aquesta gran disponibilitat de dades genòmiques de càncer: OncodriveROLE, un procediment per a classificar gens "drivers" del càncer segons si el seu mode d'acció ésl'activació o la pèrdua de funció del producte gènic; i MutEx, un estadístic per a mesurar la tendència de les mutacions

somàtiques a l'exclusió mútua. Tanmateix, la manca de precedents d'aquesta gran dimensió de dades fa sorgir nous problemes en quant a la seva accessibilitat i exploració, els quals intentem solventar amb noves eines de visualització: i) Heatmaps interactius de Gitools amb dades genòmiques de càncer a gran escala, a punt per ser explorades, ii) jHeatmap, un heatmap interactiu per la web capaç de mostrar dades genòmiques de cancer multidimensionals i dissenyat per la seva inclusió a portals web; i iii) SVGMap, un servidor web per traslladar dades en figures SVG customitzades, útil per a la transl·lació de mesures experimentals en un model visual.

Table of contents

Index of figures

1 INTRODUCTION

1.1 Oncogenomics

Big efforts from both academia and industry have been and are being put into the study of the group of diseases described as cancer. Many of the different cancer diseases have been described in detail from a morphological point of view. The *International Classification of Diseases for Oncology*[1] *(ICDO)* reflects the complexity of the cancer disease describing the numerous different cancer types which emerge from each of the many tissues of the human body. Oncogenomics is a field within cancer research that studies the genome, epigenome and transcriptome of cancerous tissues in search of the genomic variables and alterations that determine the cancer cell morphology and physiology. The knowledge gained from oncogenomic studies should then enable new strategies for the cancer disease treatment.

1 http://www.who.int/classifications/icd/en/

1.1.1 Somatic alterations

Changes in the genomic sequence of somatic cells, the cells of which all organs and tissues are composed of, are referred to as *somatic alterations*. Screening the genomic sequence of a cancer sample and comparing it to a healthy cell from the same patient always yields an array of somatic alterations; the detection of these *cancer somatic alterations* is the first step in cancer genomics studies.

Somatic alterations appear spontaneously when DNA of cells replicates upon cell division – despite the complex machinery that a cell disposes to control that the two emerging daughter cells are identical replications of the parent cell. The probability of occurrence of mistakes in DNA replication is influenced by many co-variates. Such co-variates may be environmental factors such as radiation, viral infections or certain chemical substances that get into our body and interact with the DNA or intrinsic to the genetic replication system where the double-strand break repair mechanism itself is known to introduce errors (Lieber 2010; Minamoto, Mai, and Ronai 1999). Other known intrinsic co-variates are replication timing, the compaction of the chromatin and the transcriptional status of the DNA segment in question (Supek et al. 2014; Lawrence et al. 2013). As diverse the influencing factors may be, they all translate into somatic alterations of different types, of which the most common are listed below.

- Mutations
- Copy number changes
- Translocations, Insertions and Inversions
- Chromothripsis

These alterations may occur anywhere within the DNA of total length of 3 gigabases that makes up the human genome. Nevertheless, most research has been done on the alterations within the protein coding genes or it's immediate vicinity which makes up less than two percent of the genome.

Mutations: SNVs & other small-scale mutations

Mutations in the DNA are changes of one or very few nucleotides. Point mutations, or SNVs (Single nucleotide variants) are minimal changes in the genetic code where one nucleotide is substituted and the length of the DNA sequence in question is not altered. Although large-scale alterations of the DNA are also often referred to as mutations, I'll reserve the term for small-scale changes of the DNA.

The consequence of a mutation on the protein function depends on the nucleotide substitution, insertion or deletion. As shown in Figure 1a, some codons are redundant as they code for the same amino acid. Point mutation that transform a codon into another that codes for the same amino acid are referred to as *synonymous* mutations, as the protein sequence is not altered after all. Even so, recent research has suggested that synonymous mutations are not entirely silent as could be expected. Possible consequences may be the alteration of recognition patterns used for co-factors or splicing recognition or influencing the transcription rate determined by the availability of tRNA (Czech et al. 2010; Sauna and Kimchi-Sarfaty 2011; Supek et al. 2014). If the nucleotide substitution entails an amino acid substitution in the protein sequence, it is classified as a *missense* (or *non-synonymous*) mutation. Another possible consequence of point mutations is the conversion of an amino acid codon into a premature stop codon (*stop gained* consequence*)* which truncates the rest of the protein sequence. The inverse scenario, the *stop lost* consequence, occurs when the stop codon is mutated into a regular amino acid codon and may add a nonsense sequence to the mRNA and the protein.

Other small-scale mutations that alter the sequence length are classified either as insertions or deletions (also named indels), depending on whether nucleotides are added or removed. If the insertion or deletion is not a multiple of three nucleotides the reading frame of the coding sequence is shifted and the protein sequence will most likely be garbage downstream of the *frameshift*. But in cases where three or a multiple of three nucleotides are inserted or deleted (in-frame *insertions* and in-frame *deletions)* within the reading frame,

A) Amino acid codon table

nonpolar | polar | basic | acidic | (stop codon)

Standard genetic code

1st base	2nd base T		2nd base C		2nd base A		2nd base G		3rd base
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)	TGA	Stop (*Opal*)	A
	TTG		TCG		TAG	Stop (*Amber*)	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

B) Consequence types

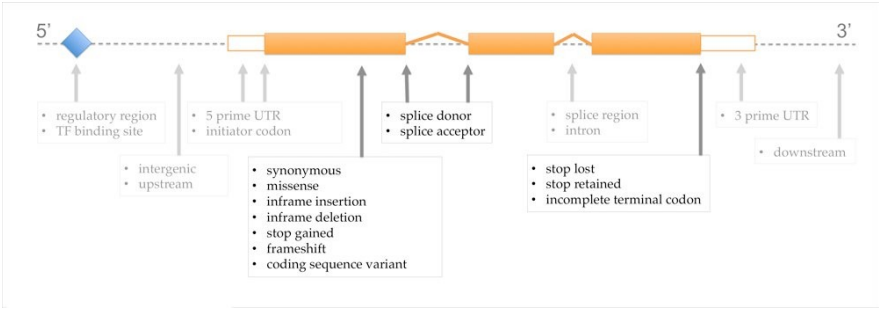

Figure 1: The amino acid codon table and consequence types of mutations
A) *The codons composed always of three nucleotides are grouped by the amino acids they code for. Leucine, for example, is encoded by six different codons. (Wikipedia 2014)* ***B)*** *A schema from the ensembl variant effect predictor website (Fiona 2014) which represents a gene annotated with mutational consequence types. The exonic consequence types are the ones highlighted in the figure and discussed in this chapter*

the protein product is not changed dramatically, disregarding possible influences on alternative splicing or the disappearance or displacement of key amino acid residues. Indel events are short-range events, depending on the study they have been described to be from 1-10'000 bp in length (Mullaney et al. 2010). Mutations at exon boundaries may prompt an alternative splicing of the mRNA and their consequences termed *splice donor* & *splice acceptor* mutations.

All the different consequences that point mutations, insertions and deletions may have on the gene are organized in the sequence ontology (Eilbeck et al. 2005) most of them are displayed in Figure 1b. Various tools exist to classify a mutation into their consequence types (Gonzalez-Perez, Mustonen, et al. 2013). A broad but important classification of mutations is the distinctions between protein truncating and protein altering mutations. The truncating mutations, such as frame-shifts and premature stop codons dramatically change the protein product. In most cases, it is safe to assume that the protein function is entirely lost. But mutations which alter the protein sequence without being truncating may abolish, change or add a certain function of the protein, while other functions are maintained.

Copy Number Alterations

Deletion and insertions as described in the preceding section are of short range. Even so, large genomic regions can be deleted or inserted as described hereafter.

Upon comparing the genomes of two individuals certain genomic regions may be found absent, duplicated or repeated multiple times in one of the individuals. This type of structural variation or chromosome abnormality is called copy number variation (CNV). The same phenomena can be observed when comparing the genomes of the tumor and the normal tissue. It is very common to observe genomic regions that are duplicated, repeated several times or entirely lost in one or both homologous chromosomes. These somatic alterations are termed copy number alterations (CNA). The extent of CNAs may vary from focal events including regions comprising some genes to loss or replications of whole chromosomal arms

(Zack et al. 2013).

The change of copy numbers may ultimately translate to a change in the number of copies of the gene transcripts. Genes that are subject to copy number gain – one chromosome carries multiple copies of them – are generally expected to be transcribed at higher levels, given that the necessary promoters, up- and downstream enhancers are replicated along with the gene.

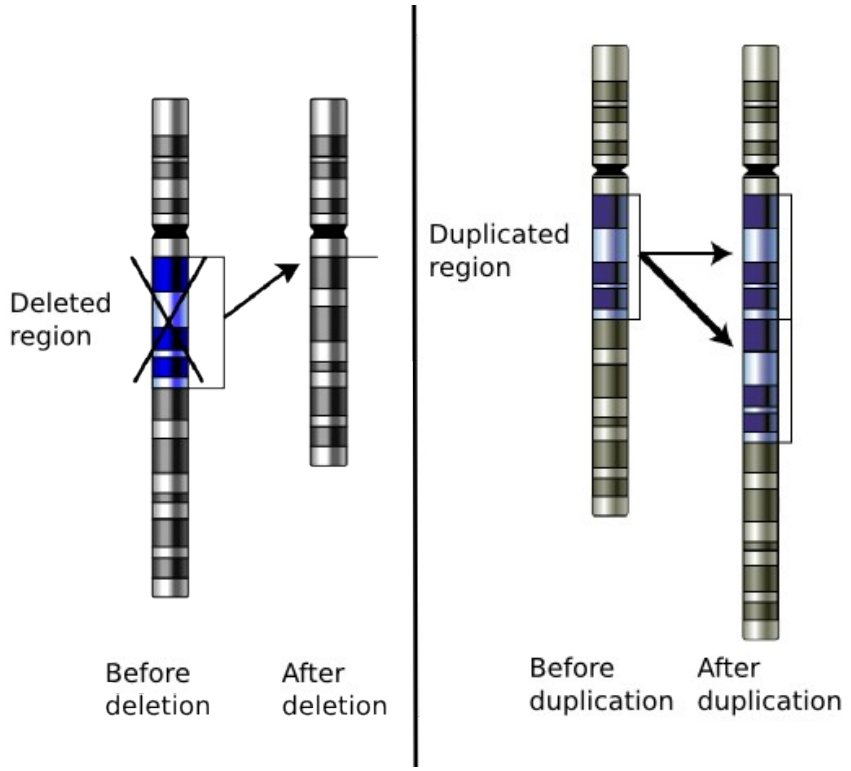

Figure 2: A cartoon of a deletion and a duplication event of a genomic region. *In the left part, a the chromosome looses a stretch of about 6 bands. The right part depicts the simplest case of copy number gain, a duplication. Figure adapted from (Wikimedia Commons 2013)*

Deletions of genomic regions may either erase one copy of the gene by a heterozygous deletion or both copies by a homozygous deletion event. A heterozygous deletion is sometimes also referred to by loss of heterozygosity (LOH) and a copy-neutral LOH occurs

when the lost DNA segment is replaced with a new copy of the other allele. Normal LOH can decrease the total amount of gene products present in a cell. Whether this is the case or not may depend on the regulation of the gene in question, as transcriptional compensation mechanism for a lost allele have been described in scientific literature (Guidi et al. 2004).

Considering the implications that copy number changes may have on the abundance and therefore function of the resulting protein, it is important to distinguish the deletions from translocations and insertions which normally don't affect the copy number.

Translocations, Insertions and Inversions

Other chromosomal abnormalities delete a genomic region from a chromosome and insert it into another and therefore does not alter the copy number but the locus of the genomic sequence. In contrast to copy number alterations, a translocation or insertion event may affect multiple chromosomes.

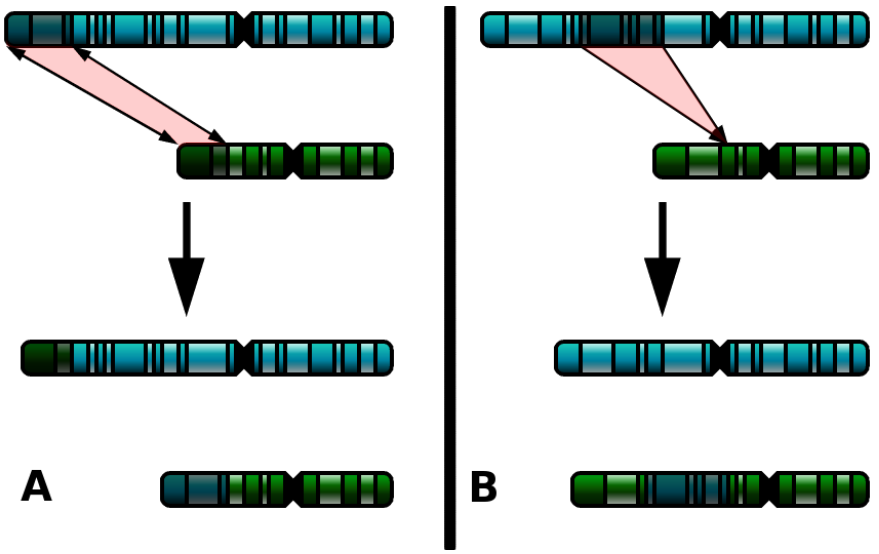

Figure 3: A cartoon of a translocation and an insertion event
A *A translocation event.* ***B*** *An insertion event. Adapted from. Image adapted from (Wheeler 2007).*

A reciprocal translocation interchanges two segments of DNA from two non-homologous chromosomes. The boundaries of the translocation event may be far from any gene and therefore have minimal effect on gene transcription. If the translocation occurs near or within gene boundaries the effect of translocations is difficult to predict. Nevertheless, for leukemia especially several translocations that have been described to produce fusion genes are causative of the disease. This incident happens when the translocation event causes a gene to be inserted directly up- or downstream an existing gene on the other chromosome. A fusion does not necessarily include both genes in their entirety as just parts may be fused. Many fusion genes are a product of the promoter of one gene that has been fused upstream of the second gene. This scenario changes the regulation of the second gene but not necessarily its protein function. Thus, depending on which parts have been fused together, fusion genes can yield a gene product with a new, altered or differently regulated function. A famous translocation between the band q11 of chromosome 22 and band q34 of chromosome 9, yielding the Philadelphia chromosome. This translocation results in different fusion products between the genes Bcr and Abl to form the Bcr-Abl fusion genes which are associated to different types of leukemia reviewed in (Advani and Pendergast 2002). Insertions at a chromosomal level are events where a genomic sequence is cut out from one chromosome and inserted into another. It therefore resembles the translocation with the distinction that no genetic material is introduced where the genetic sequence has been cut out. Note that fusion genes may also be a product of insertions, deletions or inversions.

Chromosomal inversions occur when a stretch of genomic sequence is cut out and inserted in the opposite sense. As with the other genomic abnormalities these do not necessarily cause any malfunctions, as the genes that are affected by the inversions can be read in the opposite sense in the other strand. If the inversion occurs at or within the coding sequence fusion genes may be produced. Various cases of inversions associated with different cancers have been described (Grimwade et al. 2010; Zech et al. 1984; Speleman et al. 2005).

In extreme cases, cancer samples have been described that contain

tens to hundreds of genomic rearrangements that are acquired in a single event, termed chromothripsis. Chromothripsis has been reported to occur in 2%-3% of all cancer samples and in about 25% of bone cancers (Stephens et al. 2011).

Epigenetic alterations

Epigenetics is the study of inheritable traits within the genome that are not caused by alterations in the nucleotide sequence of the DNA. Chemical modifications of nucleotides have direct influence on how the DNA is packed and organized and which parts of the DNA are accessible to transcription factors. In particular modifications on histones and cystosine methylation changes are two types of epigenetic alterations that are passed on to the daughter cells upon cell division, providing mechanisms to inherit the state of gene activity and expression (Richards 2006).

DNA methylation is a regulation mechanism of gene transcription. It consists in adding methyl groups to cytosines typically within CpG islands, regions with high recurrence of CG dinucleotides. Genes whose promoters contain CpG islands can be silenced by strong methylation of the CpG sites. Methyl groups are believed to abolish the ability of the transcription factors to bind to the promoter. Moreover, methylated DNA can indirectly recruit proteins such as histone deacetylases and other factors that modify histones and therefore alter the organization of DNA packing. Tightly packed DNA, called heterochromatin, hinders the transcription of genes contained in tightly packed regions.

Studies of DNA methylation patterns within cancer samples have revealed many cases of DNA hyper- or hypomethylation upon comparing the methylation patterns of the cancer tissue with a normal tissue (Jones and Laird 1999). Particularly, tumor suppressor genes activities have been observed to be lost via hypermethylation of the promoter CpG islands (Baylin and Jones 2011). Possible use of hypermethylated genes as biomarkers for drug response has been reviewed in (Heyn and Esteller 2012) as hypermethylation events are normally observed locally at site-specific loci whereas generally the cancer cells show genome-wide hypomethylation (Rodríguez-Pare-

des and Esteller 2011).

1.1.2 Driving tumorigenesis

The genome of cancer cells bears a plethora of somatic alterations. A principal aim of oncogenomics is to find out how these changes in the genome and epigenome work mechanistically to give rise to the cancer disease. In fact, many of the alterations are not the cause of the cancer disease, but a consequence. So how does tumorigenesis work?

Clonal evolution: accumulating alterations

We know from evolutionary and molecular biology that a species accumulates germ line mutations or alterations which are then passed along with each new generation. The effect a new mutation on an offspring may vary from none at all to a visible phenotypic or measurable physiologic trait. In extreme cases such a trait may confer the offspring an advantage or disadvantage for survival amongst pears. This mechanism is the very basis for species evolution.

As described in the preceding chapter, mutations not only appear in germ line, but also in somatic tissues. In few cases, a mutation goes undetected by repair mechanisms of the cell and, if it is not lethal, is established in the tissue. Given that one mutation is present in a cell which gives rise the several generation of daughter cells, over time a new clone of cells forms in the tissue with a genotype differing by one mutation to the bulk of cells in the tissue. Analogously to the evolution theory for species, a mutation in a tissue-context may confer the cell an advantage which further helps the stabilization of the clone in the tissue or a disadvantage with the opposite effect. Advantages of various forms occur: accelerated and/or indefinite cell division or increased capability of energy uptake and/or metabolic efficiency to name a few (Douglas Hanahan and Weinberg 2011). If the new cell or clone of cells performs all the necessary functions for the tissue where it is situated, it poses no problem for the organism as such. In any case, while the described process repeats itself further mutations are acquired in a sub-clones

of the tissue.

An acquired alteration that specifically interferes with repair mechanisms of the cell confer a certain instability to the cell replication process which increases the probability of the introduction of new alterations. As the alteration level is increased in time more mutations potentially conferring strong advantages appear in the genome of the clone, up to the point where the cells cease to perform the tissue function they were programmed to and becomes somewhat autonomous – a neoplasm or tumor is emerging. The often cited definition of neoplasms is originating the British oncologist Willis: "A neoplasm is an abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of the normal tissues and persists in the same excessive manner after cessation of the stimuli which evoked the change" (Baserga 1985). In order to become a malignant tumor the neoplasm has to acquire a series of capabilities which often are referred to as the hallmarks of cancer, as discussed in the next section.

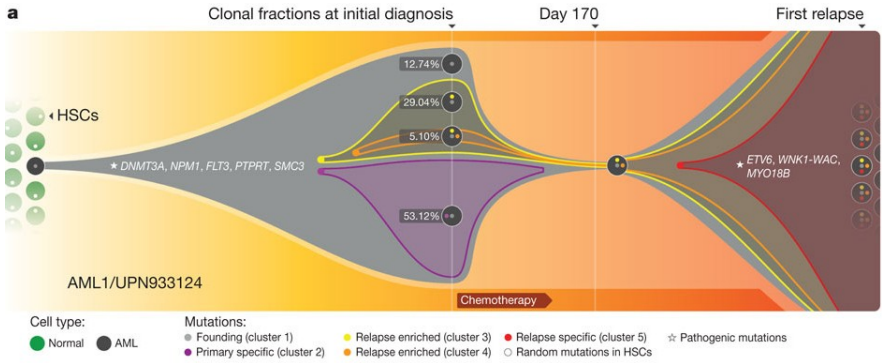

Figure 4: An outline of a clonal evolution found within a patient.
The founder clone boar five mutations in the genes DNMT3A, NPM1, FLT3, PTPRT, SMC3 as those mutations are present in all the detected sub-clones at first diagnosis. Chemotherapy diminished the present clones, but also exerted strong pressure for new clones with resistance alterations tor rise: ETV6, WNK-1-WAC, MYO18B. Image adapted from (Greaves and Maley 2012)

This process of accumulation of alterations is called clonal evolution. It has normally been going on already for many generation of cells when a malignant tumor is detected in a patient and a sample of it is obtained. Henceforth the tumoral tissue often contains a series of sub-clones, distinct in their somatic alteration patterns as

shown in Figure 4. However, only recently the concept has received more attention, as current high-throughput sequencing technologies have yielded high resolution which allow the researchers to determine the clonal architecture of a tumor tissue sample (Nielsen et al. 2011). Via the fraction of sequence reads that bear a certain somatic mutations it can be estimated which alterations were present only in the founder clone, represented in gray in Figure 4. New sub-clones can emerge spontaneously and outcompete other tumoral clones which is represented by the yellow, orange and purple areas in Figure 4. The setting for the clonal evolution is defined by natural or physiological restraints, environmentally derived genotoxicity and cancer therapy, jointly termed selective pressure (Greaves and Maley 2012).

Which type of cells undergo the clonal evolution of tumorigenesis is not clear as of yet. If the tumor type is known to follow the cancer stem cells (CSC) model the explanation would be that it in the CSCs is where the tumorigenesis happens. The CSC model is an alternative to the clonal evolution paradigm which is based on the discovery of cells with stem cell-like properties within the tumor tissue that have a hierarchical relationship with other cells in the tumor tissue. It is thought that the CSCs give rise to the rest of the tumor cells, which are differentiated, in contrast to the tumorigenic CSCs (Shackleton et al. 2009). The fraction of CSCs within a cancer can vary: depending on the tumor type, the tissue consists of almost only tumorigenic cells or contain only a low fraction of tumorigenic CSCs. Some CSCs have been described as specific markers, as revised in (Yu et al. 2012).

Drivers and passengers

The selective pressure that governs the local tissue environment defines which mutations are advantageous, deleterious or neutral for the cells and therefore drives the clonal selection. Those alterations which are advantageous for the tumorigenesis are called *driver alterations* and the genes in which they fall are referred to as *driver genes*. Some of the mutations in the founder clone are initial drivers of the tumorigenesis. Along the way other drivers may be acquired

in sub-clones that outcompete the original clone. As clones evolve and explore the mutational space many neutral mutations are picked up whereas cells that acquire disadvantageous alterations disappear quickly or only form a small sub-clone. The non-lethal neutral alterations which get fixed in the tumor tissue are called *passenger alterations* and the carrying genes *passenger genes* as those alterations do not support the expansion of the cancer cell clones. (Haber and Settleman 2007)

Tumor suppressor genes and oncogenes

Driver alterations may have several consequences for the affected genes. Those consequences are broadly distinguished upon the effect they have on the protein function. As *loss of function (LoF)* we consider those alterations that abolish a certain or all functions of a protein and as *activating* alterations those that potentiate a protein function or cause the appearance of a new function. Hence, cancer drivers are generally classified into two classes which represent two different mode of actions of the driver alteration or gene on tumor formation: tumor suppressors and oncogenes.

Tumor suppressors are genes that counteract tumorigenic behavior. Especially the p53 protein has been dubbed, the “guardian of the genome” due to it's preservative function to maintain genome stability (Lane 1992). As a general rule, a tumor suppressor gene is defined by the fact that it is beneficial for the tumorigenic process and the cancerous state of the cell, if its protein product, or at least its function, is *not available* to the cell. Therefore tumor suppressor genes are often lost in deletion events or truncated by mutations.

Oncogenes are the counterpart of tumor suppressors. Their high activity is beneficial for the tumorigenesis or cancerous state of the cell. Therefore the cancer cell often transcribes oncogenes in high amounts as their activity is needed. Insertions, translocations and mutation events may give rise to oncogenic protein products which exert new functions that are not originally available in the cell. A pathway may induce or maintain tumorigenesis by receiving an oncogenic activation where a key component is always *activated* due to over-expression or an activating mutation.

Meanwhile, the type of alteration event a driver genes suffers can be taken as indicator of its mode of action, but not all cases are clear-cut. Lots of experimental studies have been performed in order to study the oncogenic or tumor suppressor activities of all kinds of proteins, which has often lead to contradicting statements about the mode of action of a particular gene or protein. The difficulty lies in reproducing an *in vitro* system whose regulatory programs resembles those *in vivo*, but also that depending on the origin tissue and therefore on the activated genes and pathways, the tumorigenic role of the same protein may be opposite in different settings (Licciulli et al. 2013; Liu, Zhang, and Ji 2013). Knowing in which category a cancer driver falls has important implications on the interpretation of effects on the pathway and possible treatment possibilities. Generally, oncogenes are easier to target as the goal would be to abolish their function. Developing a treatment that compensates the loss function of a tumor suppressor gene is a much more complex task, although it has been achieved (Lambert et al. 2009).

Hallmarks of cancer

Detecting which alterations are driver alterations is complicated, because cancer is, considered from a genomic point of view, a very heterogeneous disease. This is not only so when comparing cancer types from distinct tissues, but also different tumor samples of the same cancer type and tissue of disease. This may be explained by the assumption that there is not one specific way for a neoplasm to arise and turn malignant, although some common routes have been identified. What capabilities the driver mutations should confer to the cell in order to turn cancerous, has been discussed by Hanahan & Weinberg in two crucial papers defining the "hallmarks of cancer" (D. Hanahan and Weinberg 2000; Douglas Hanahan and Weinberg 2011).

As insinuated earlier in the text, one important acquisition is the capability of introducing genome instability and circumventing cell cycle and DNA damage checkpoints and other control mechanisms that are at the disposal of the cell. Cells which do not behave as ex-

pected are confronted with possible immune destruction, senescence and/or apoptosis (D. Hanahan and Weinberg 2000). Evading these are some of the hallmarks of cancer.

Recent research suggests that immune surveillance control for the presence of possibly cancerous cells: tumor-infiltrating lymphocytes (TIL) infiltrate and initiate eradication of tumor cells (Kim, Emi, and Tanabe 2007). Cancerous cells which are weakly immunogenic may escape this "immunoediting" process and colonize a tissue. The immunoediting both protects the host from cancerous cells and sculpts an emerging tumor by exerting selective pressure in function of the immunocompetence of the host (Dunn et al. 2002).

Cellular or replicative senescence is a phenomenon in which cells cease to divide due to "aging". Senescence may play an important role in cancer as some research suggests that strong oncogene signals within the cell may initiate senescence (Braig et al. 2005; D. Hanahan and Weinberg 2000; Collado and Serrano 2010) or apoptosis in response to signal imbalances. Apoptosis, also known as programmed cell death, is a mechanism of multicellular organisms that targets possibly disturbing cells whose elimination is an advantage for the organism. Thus, malignant cancer cells often develop strategies to avoid or resist apoptosis (Lowe, Cepero, and Evan 2004). As a result of this and the avoidance of senescence, tumor cells gain unlimited replicative potential and start to proliferate quickly. Extreme proliferation can itself lead to senescence and also stands in conflict with various programs of the cell that negatively regulate its proliferation. Nevertheless, senescence can be circumvented by means of enlarging or maintaining the DNA telomere length (D. Hanahan and Weinberg 2000) which are normally shortened upon cell division and mark the age of the cell.

The typical tumor suppressor activities of RB1 and TP53, just to name two examples, consist in controlling the decision making with regard to cell proliferation. Signals of growth suppression are integrated by tumor suppressors proteins and therefore mutations in those may render the cell deaf to such stimuli.

Apart from avoiding and resisting the multiple programs that control the healthy state of the cell, cancerous cells must ensure that

enough nutrients and oxygen are delivered to them in order to sustain the rapid growth of the tumor tissue. One mechanism to achieve

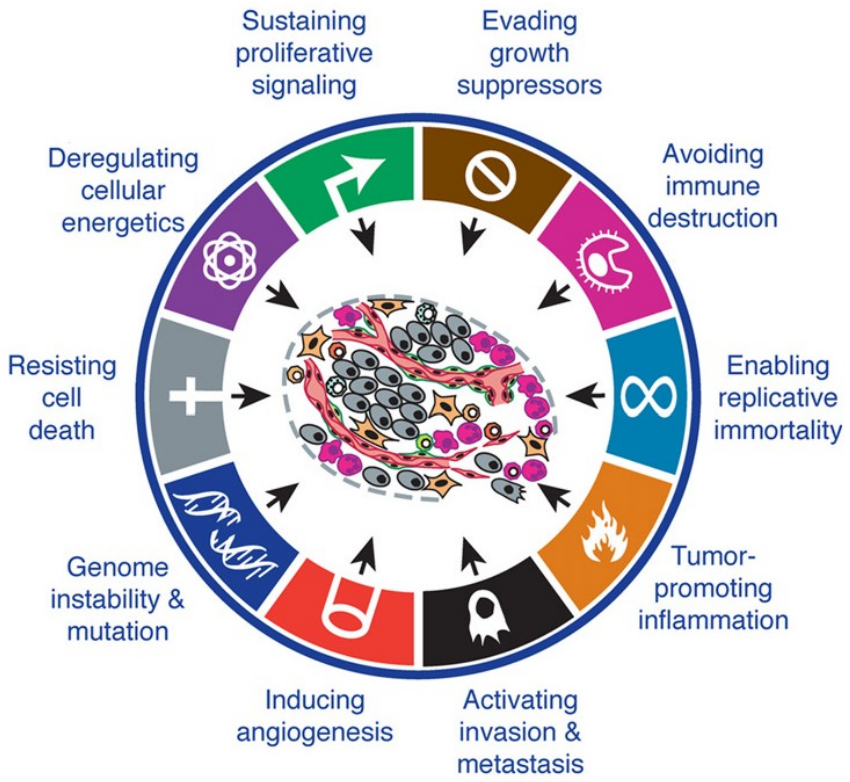

Figure 5: The hallmarks of cancer
The six classical hallmarks of cancer Sustaining proliverative signaling, evading growth suppressors, activating invasion & metastasis, Enabling replicative immortality, Inducing angiogenesis and resisting cell death from (D. Hanahan and Weinberg 2000) completed with four energy metabolism reprogramming, avoiding immune destruction, tumor-promoting inflammation and genome instability & mutation (Douglas Hanahan and Weinberg 2011; also source of image).

this consists in switching on angiogenesis with the purpose of building new vessels that deliver increased amounts of blood to the cells. Inflammation may similarly boost cancer growth because it can promote several of the aforementioned hallmarks by providing growth, survival and pro-angiogenic factors (D. Hanahan and Weinberg

2000; Douglas Hanahan and Weinberg 2011).

In summary, the hallmarks of cancer are a set of cell functions – or pathways – that tumor become frequently altered in the tumorigenic process. Some of the alterations may be more important than others and the order in which they occur is not clear. These factors may vary in function of the tissue where a tumor occurs. A certain alteration may be prevalent within a tumor type of a specific tissue, but at the same time this alteration may never be found in the samples of another cancer type. This is exemplified by TP53, RB1 or PTEN, all known tumor suppressors and accepted cancer drivers that are often targeted by gene alterations. They are often not altered in cancer samples obtained from patients as they are not the only elements of the pathways constituting the hallmarks. Tumorigenesis as a process depends on the alteration of the hallmark pathways irrespective of the precise gene which carries it.

Some proteins play very central roles in the pathway signaling cascade. Altering a hub in a protein-protein interaction network will have a major effect which, if beneficial for tumorigenesis, is a good target for alteration. These good targets are therefore often observed when screening a cohort of cancer samples and are easily identified. Regardless, unlikely alterations occur also and can result in the same functional pathway signaling aberration or gain that is needed by the tumorigenesis as one of the highly recurrent alterations may. Consequently, driver mutations that are less likely to occur are automatically less likely to be identified, as discussed in the next section.

1.2 Identification of cancer drivers

Given the hallmarks of cancer, the interpretation of the alterations that are observed in cancer cells is somewhat aided. Any alteration observed in a gene whose protein product is known to be implicated in one of the hallmarks becomes a suspect of being implicated in tumorigenesis. The main hurdle, as discussed in the preceding section, is that not all driver mutations are evenly likely to occur. Another hurdle is that many proteins are not thoroughly studied, and even well studied proteins may exert yet unknown functions and prove to be involved in tumorigenesis. Thus, how are cancer drivers being identified? The first step is to identify somatic alterations within the cancer samples by means of available technologies and creating a data cohort. This cohort can then be studied in a second step with bioinformatic tools in order to find putative cancer drivers.

1.2.1 Technologies

The scientific questions that can be investigated depend largely on the technology and its accessibility that is available to obtain data. A bit more than twenty years ago, microarrays have been introduced, revolutionizing genetics with new possibilities and arguably enabling *transcriptomics.* During the last decade there has been a shift from array-based data generation to the use of sequencing technologies, paired with big investments in cancer research by governments world-wide. This has led to a massive surge in high-quality cancer genomics data available to the scientific community around the globe.

Polymerase chain reaction (PCR)

Polymerase chain reaction (PCR) is a technology employed to clone fragments of DNA in order for signal quantification. By means of a heat-stable DNA polymerase DNA fragments are treated with successive cycles of heating and cooling in order to induce DNA melting, dissociate the two DNA strands and then facilitate DNA replication. With the help of short DNA fragments, called primers, complementary to the extremes of the DNA sequence being amplified, specific regions of the genome can be targeted for the PCR amplification. With each cycle, the targeted DNA stretches are replicated and doubled in amount, and thus amplified exponentially.

The reverse-transcription PCR (rtPCR) allows to retro-transcribe DNA from RNA. With this variation of PCR, it is possible to amplify the signal of mRNA and deduce the quantity of gene expression activity. The quantification has seen it's revolution with the introduction of DNA microarrays.

DNA microarrays

DNA microarrays, short *arrays,* are chips with many tiny “wells” called spots that contain little quantities of a specific DNA fragment, called probes. The probes are designed complementary to cer-

tain sites of a gene or other complementary DNA (cDNA) elements in order to specifically hybridize that target sequence.

Before hybridizing the probes, the cDNA fragments are labeled with dyes that turn fluorescent upon hybridizing the chip probes. The emitted light is used as signal. Therefore the signal strength depends on the amount of cDNA that binds to the spots, which is a relative proxy of how much of the original sequence, e.g. an mRNA of a certain gene, was in the sample. The cDNA fragments can be amplified via the PCR or rtPCR reaction in order to ensure that enough signal is available for measurement.

Different microarray technologies exist for various applications. The most common ones are chips for gene expression profiling, single nucleotide polymorphism (SNP) detection per SNP array and chromosomal abnormality detection via comparative genomic hybridization.

The different arrays have served well the scientific community, as discussed with the example of expression arrays in cancer research (see page 38). Microarrays have introduced a new era in biology of measuring *en masse*. Never before the age of arrays, molecular biologists had to cope with that big amounts of measurable features at once. The GeneChip Human Genome array from the company Affymetrix, a DNA array of the latest generation, claims to report on the abundance of 18'400 transcripts and variants including around 14'500 human genes (http://www.affymetrix.com/). The analysis and interpretation of this data has enabled a new generation of bioinformatics and biostatistics oriented scientists.

Besides the hurdles imposed for the biological interpretation of several thousands of gene measurement for multiple samples, the microarray poses challenges caused by technological artifacts.

A batch effect problem arises when comparing multiple samples that have been processed over different time points and possibly even with different microarray platforms (Leek et al. 2010). This is especially a problem when collecting data that has been posted by different researchers to public array data bases such as GEO (Edgar, Domrachev, and Lash 2002) to be jointly analyzed. The batch effect

appears because the specific conditions and factors, such as reagents, that have been used to process the microarray influence the measurements, the used platform or the personnel. Therefore the intensities of two different samples of the same experimental group, sample A and sample B, that are analyzed on the same machine, on the same chip and under the same conditions may correlate better than a replica of sample A on a second machine, another chip or simply processed another day.

It is important to consider that the measurement of fluorescent intensities is a relative quantification: The measurement is only informative in the context of other samples and relative differences between them and thus possible batch effects complicate the interpretation as the quantities that are measured are influenced by non-biological factors that vary in between different processing batches. Various normalization solutions have been proposed. The virtualArray package combines many proposed solutions from mean centering to the empirical Bayes method (Heider and Alt 2013). Generally, a good experimental design of the batches may diminish their influence on later erroneous correlation (Leek et al. 2010).

As for expression arrays, another problem is that the lowly expressed genes may be indistinguishable from random noise. After the binding of the cDNA segments to the probes on the chip, a washing is applied to eliminate all the sporadic association between cDNA fragments and probes. In any case, some non-specifically bound cDNA fragments will remain on the chip and cause noise signal. Therefore, microarrays are not able to report a non-zero (unexpressed) status, as for the spots where no complementary cDNA is available and would bind strongly to the probe, unspecific binding is more likely to happen and falsify the signal.

DNA Sequencing

Sequencing is generally defined as the process of determining the primary structure of biopolymers. In other words, sequencing is used to determine the exact order of amino acid in a polypeptide chain or nucleotides in polynucleotide chains as DNA or RNA.

In the 1970s the first DNA sequences were determined with the help of sequence-specific primers, as employed in the PCR method (Jay et al. 1974) known as Sanger sequencing. This type of sequencing dominated until the early 2000s. Thereafter many techniques and protocols have been described to improve DNA sequencing in quality and speed and have enabled the creation several commercially available high-throughput sequencing machines, allowing to perform whole-exome and whole-genome sequencing in a realatively short time. Paradoxically, they are still referred to as "next-generation sequencing". A time line featuring the different sequencing techniques and major sequencing achievements is displayed in Figure 6.

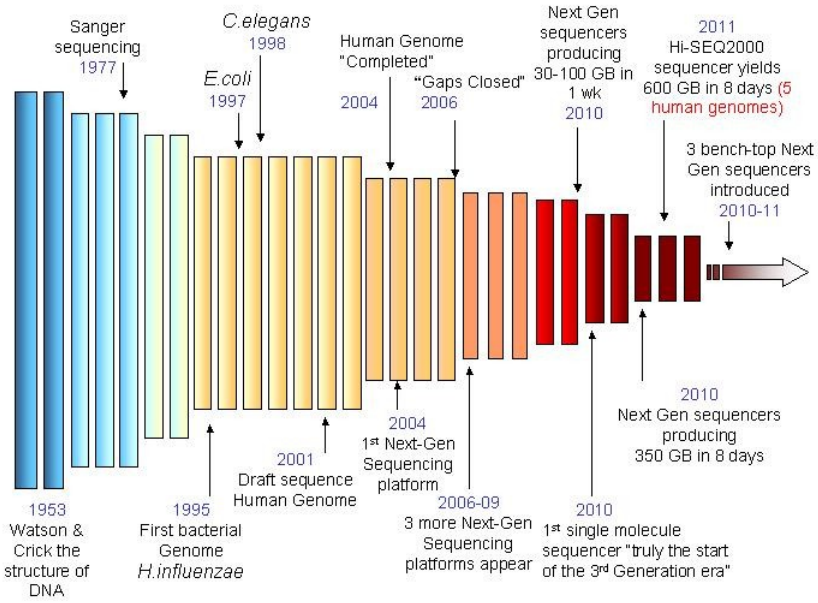

Figure 6: Timeline of sequencing techniques and achievements
Image from (Laing 2011).

Further generations of improved sequencing platforms are to be expected, because available machines are manufactured by different companies and employ different techniques which still are error-prone. Additionally, the scientific community and in particular the

health sector exert high pressure to lower sequencing costs.

The read length of DNA fragments that are sequenced are rather short, between 50 and 1000 bp, depending on the technology. This entails various problems for the assembly of the read sequences: in general they are hard to align to the reference genome, in particular in repetitive regions belong to and each platform has its bias to certain errors.

Generally, the higher the coverage the less errors are introduced as non-recurrent bases can be identified as false positives. High coverage also makes the alignment process somewhat easier as the higher the coverage the more overlapping parts are read and less gaps have to be aligned. In cancer studies two samples are required for sequencing: one sequence from healthy tissue and another from the cancer tissue such that somatic alterations can be determined from the comparison of the aligned reads of both. The amount of data stored is rather large which imposes a big challenge, especially for large projects. It is important to store the original read data as the aligner software and the human reference genome are both evolving and thus the resulting alignment may result differently if repeated after some time. The choice of an adequate sequence aligner is somewhat of a challenge now, as different tools yield different results. Li and Homer reported in 2010 that already 20 short-read alignment software packages have been published. Popular choices are Bowtie or BWA due to their speed (H. Li and Homer 2010).

Overall the reads are more error-prone towards their ends. The bases in each end are often discarded if high accuracy is needed as is the case when calling genotypic variants for an individual. The problem of the introduced sequencing errors is, that they may be interpreted as alterations by the different bioinformatic approaches that have been developed to make the variant calling. Additionally, each method has its advantages and pitfalls when it comes to detecting lowly frequent variants or the before-mentioned false variants (Wang et al. 2013; Kim, Emi, and Tanabe 2007)

RNA sequencing

The sequencing of RNA, also called RNA-seq, is superseding the use of micro-array platforms for the quantification of the transcribed DNA in a sample. The biggest advantage of RNA-seq over micro-arrays is that theoretically all RNA present in the cell can be detected and no prior knowledge is needed as is the case for designing the microarray probes, except for a reference genome which is already available. Furthermore, RNA-seq has more applications than micro-arrays as not only abundance, but also the exact sequence is being detected. Besides gene expression quantification by mRNA sequencing, RNA-seq can be used for detection of single nucleotide variants (SNVs) or even post-transcriptional SNVs, intron-exon boundaries, fusion genes and as well for the study of isoform balance and other RNA populations than mRNA. The manifold opportunities of interpretation of RNA-seq goes hand in hand with the space the resulting data takes up on disk. Similar to DNA-sequencing, as RNA-seq is employed for big projects, it's data management is also becoming a challenge which has to be managed well.

A particular problem for the study of RNA transcription in cancer samples is that the matching of normal and cancer samples is more difficult as not all tissues allow for taking normal samples (e.g. brain) and the transcription of many genes is tissue-specific. Therefore often times, when no paired normal sample is available, blood samples or samples from healthy donors are used as backup. Although these may be viable solutions, they do not reflect accurately the transcription abundance in the healthy tissue of the cancer-site of the same patient.

1.2.2 Large scale cancer genomic studies

The heterogeneous nature of the cancer disease makes it difficult to reproduce singular findings such as de-regulation of a gene (group) or a specific mutation in limited cohorts. On top of this, few groups or institutes have the power to generate large enough cohorts to address this statistical problem and identify the cancer alterations

in depth. But in recent years, orchestrated efforts have been going on to generate high-quality cancer data cohorts available for research around the globe.

TCGA: The Cancer Genome Atlas

TCGA is a North American effort to characterize genomic aberrations in cancer patients. In a pilot study, published in 2008 (McLendon et al. 2008) showing that by coordination of multiple centers it is possible to gather large high-quality genomic cohorts that give new molecular insights of the cancer disease, in this case of Glioblastoma multiforme samples. For the ongoing phase, TCGA has set itself the goal of gathering multidimensional genomic data, including exon or whole-genome sequences, expression profiles, copy number status and others, of minimum 500 cancer samples from up to 25 different cancer types.

In 2013 researchers associated to TCGA released a series of studies under the title TCGA pan-cancer, in which a cancer cohort encompassing samples from twelve tumor types were analyzed together in order to gain new insight of differences and similarities between them (The Cancer Genome Atlas Research Network et al. 2013). The TCGA pan-cancer datasets guarantee certain processing standards of the data and therefore allow to pull together different datasets. The TCGA pan-cancer mutation data for example, has been produced with the same aligner and mutation callers.

On the data portal launch site, TCGA reported to have 10'206 analyzed cancer samples for which data is available (https://tcga-data.nci.nih.gov/tcga/, accessed July 24th 2014).

ICGC: International Cancer Genome Consortium

With similar goals in mind as TCGA to create large cancer genomics resources, an international community of scientists started to think about an international effort for collecting and providing cancer genomics data. Altough the ICGC does not provide direct funding for sequencing projects, it creates a powerful platform where researchers can discuss, define common goals and also lobby

the respective agencies to generate funding. In 2010, the resulting ICGC presented their intentions of obtaining large-scale characterizations of 25'000 cancer samples from 50 cancer types in a white paper (Hudson et al. 2010). To date, the ICGC cites on their website 73 projects across the globe with funding commitment from their respective countries. As the ICGC is a multinational effort, TCGA provides an important part of the data coming from North America. Spain is contributing to the ICGC with its CLL project. Following TCGA pan-cancer, the ICGG has issued a call for whole-genome pan-cancer studies and is providing collaboration coordination for all participants.

1.2.3 Computational analysis of cancer genomic data

The advanced knowledge of cancer genomics combined with the large cancer genomics cohorts that are being produced creates a setting where cancer genomics properties can be studied with novel computational approaches. A simple question such as which genes and alterations are driving different cancer types can be answered with many different approaches, as discussed further on.

Mutational patterns: identifying mutational drivers

As cohorts of cancer samples are being collected, and all the mutations in the genome are being registered, it is possible to survey for mutational patterns of each gene. For example, the mutational recurrence in a cancer sample cohort is a straight-forward indicator that the gene is a driver candidate. A beautiful example of this is the APC gene which has been reported mutated in about 80% of the colorectal cancer samples (Stratton, Futreal, and Wooster 2004). As easy as it seems, this interpretation harbors problems because genes can also be recurrently mutated due to reasons other than the tumorigenesis. The reasons are several: the longer the gene is the more likely it is that a mutation will fall within it's coding sequence, such as in the case of TTN (Tamborero et al. 2013). Furthermore not all genes are critical for the cancer cell to function and survive

as they are not transcribed in a tissue or do not interfere with maintenance of the cancerous state of the cell. Such passenger mutations may be carried along without problem. Conversely, if a mutation falls in a gene which is critical to yield protein products for the cancer cells, this mutation is under negative selection and is therefore less likely to pass to the next generation of cancer cells. Other genes are more likely to be mutated as they lie in late-replication genomic regions which are more error-prone (Koren et al. 2012). Therefore the probability of receiving a mutation is not equal for all the genes. In order to correctly model an enrichment in mutations across a sample cohort, one would need to take into account the background mutation rate (BMR) which reflects the likelihood that a gene is mutated in a cell of a given tumor type. This approach is employed by recurrence-based methods such as MuSic and MutSig (Lawrence et al. 2013; Dees et al. 2012). Modeling the BMR is complicated as, one cannot directly assume that the same BMR applies to all the tissues and we do not control all the co-variates of it.

In any event, as there are mutations that are under negative selection during tumorigenesis, the contrary is also the case. We know that mutations in genes that confer the cancer cell an advantage are selected for. The selection process also leaves behind traces other than recurrence, as listed in Figure 7. The functional impact (FM) bias explores the selection for mutations that have a high impact on protein function. The rationale is that sporadic passenger mutations occur across the entire spectrum of functional impact. Driver mutations on the other hand are not sporadic and are expected to either abolish or alter the function of the protein. OncodriveFM takes up this idea by exploring the functional impact scores of all mutations in a gene and assessing if the gene is particularly targeted by mutations with great impact on protein function more ofthen than passenger ones (Gonzalez-Perez and Lopez-Bigas 2012). The functional impact scoring depends on the mutation location and inferred consequence. Truncating mutations have the major impact imaginable as they directly cancel any protein function. Missense mutations may have very different consequences depending on where they fall within the gene. The scores of functional impact for missense mutations are assessed by approaches such as SIFT,

Polyphen2 or MutationAssessor (P. Kumar, Henikoff, and Ng 2009; Adzhubei et al. 2010; Reva 2013) that largely depend on protein alignment between different species of the gene in question in order to assess the conservation of the nucleotide where the mutation falls. A very conserved residue is thought to be critical for protein function while a highly variable residue is not.

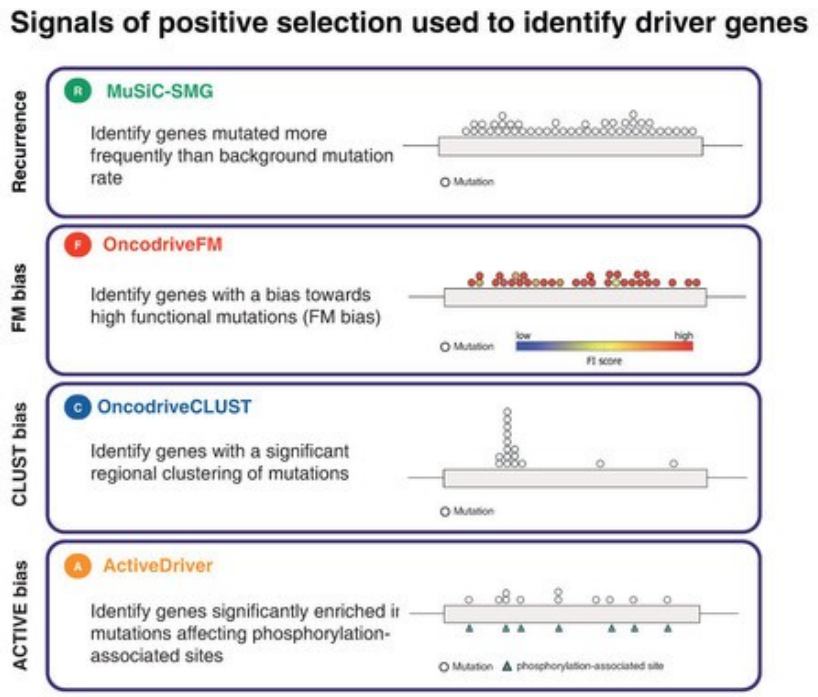

Figure 7: Mutational patterns of cancer driver genes
The positive selection for tumorigenic mutations leave behind mutational patterns within a cancer sample cohort which can be used to identify mutational cancer drivers. Figure adapted from (Tamborero et al. 2013).

Thus, a gene that shows no bias towards the accumulation of mutations with high functional impact is arguably a gene bearing passenger mutations and may therefore be discarded as a driver even if it has a highly recurrent mutational pattern. A gene with very few mutations, but all with great impact, signals a good candidate for cancer driver. This allows to identify lowly recurrent driver genes, and the bias measure of this approach allows to disregard correction for the background mutation rate. An example is found in IntOGen-

TCGA (Tamborero et al. 2013), which reports the DICER1 gene, a cancer gene annotated as such in the cancer gene census (Futreal et al, 2004), to have a functional impact bias within the TCGA uterus datasets where 9 mutations have been registered amongst 230 samples where only 3 out of the 9 mutations are annotated as lowly impacting.

In any case, not all cancer drivers bear mutations that are truncating or fall into very conserved sites. Very subtle changes with low impact bias in the genomic sequence may assure that the resulting protein is always in an activated or inactivated state. Only one or very few mutational options may be available to achieve a very specific behavior or function of the protein. OncodriveCLUST is a tool that tests if the mutations are accumulating at the same site or in a cluster across a tumor samples cohort (Tamborero, Gonzalez-Perez, and Lopez-Bigas 2013). The BRAF V600E is a classic example of an oncogenic mutation that is found in most melanoma tumor samples. Even within the TCGA Glioblastoma dataset, IntOGen reports 5 out 7 samples with mutations within 290 tumor samples to have the same mutation. Two further mutations affect amino acids that are 2 and 4 positions apart only, very proximate to the well-known mutational hot spot.

ActiveDriver is another tool designed to detect mutational patterns associated to phosphorylation sites. Similarly to the rationale of clustering, this approach tests if the registered mutations are affecting the capability of signal transduction. The authors of ActiveDriver have reported a new phosphorylation site in the well known oncogene EGFR and tissue-specific phosphorylation site affecting mutations within the EGFR signaling module (Reimand, Wagih, and Bader 2013).

In summary, many genes can be identified as cancer driver genes by means of their mutational pattern. Two recent efforts, focusing on somatic mutation data from cancer sample cohorts of 12 or more tumor types have suggested around 250-290 driver genes (Tamborero, Gonzalez-Perez, and Lopez-Bigas 2013; Lawrence et al. 2014).

Copy number drivers

Not all driver genes act through mutational alterations. Several cancer drivers are known to be subject to chromosome abnormalities such as copy number alterations. CDKN2A is a tumor suppressor gene, upstream of TP53 signaling, that falls in a chromosomal region where large deletions have been reported with high recurrence. Figure 8 shows that CDKN2A is homozygously deleted in almost 50% of the tumor samples from a cohort of Glioblastoma patients. EGFR, an oncogene is subject to recurrent copy number gain in the same dataset. 45% of the samples are reported to have more than two copies of EGFR. A quarter of the samples with EGFR gains even have mutations in the EGFR gene, indicating that different alteration types may have complementary effects on tumorigenesis.

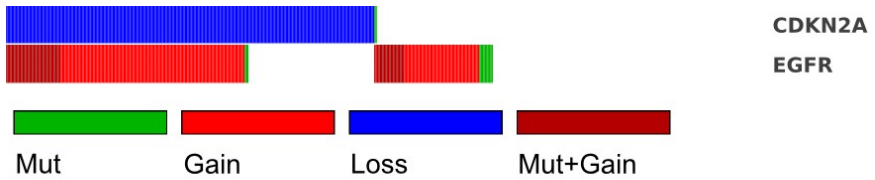

Figure 8: Copy number drivers of the TCGA Glioblastoma datasets
The above figure shows in alterations across the 561 Glioblastoma cancer samples for CDKN2A and EGFR. CDKN2A is lost due to large range deletions in 270 (48%) of the samples. EGFR, a oncogene, has been detected to have multiple copies in 252 (45%) of the Glioblastoma samples. A subset of 61 samples have EGFR gained and mutated.

The recurrence shown in Figure 8 can be interpreted as a clear sign of positive selection for the CDKN2A and EGFR, especially so, as both of the mentioned genes are well-known cancer drivers. But copy number gains and losses are chromosomal abnormalities that can affect many dozens of genes, which makes interpretation very difficult for a single cancer sample. Thus, cancer sample cohorts can help the detection of cancer genes by aligning the chromosome abnormalities from multiple samples. The overlapping regions between the samples will put the focus on the genes that are recurrently targeted by the copy number alteration events. GISTIC is a

computational method that assesses the distribution of the copy number events, identifies the peak regions and reports the genes that fall within the peak regions as putative copy number drivers (Mermel et al. 2011). Another important criteria to take into account when identifying copy number drivers is the zygosity of the event. A copy number loss of a tumor suppressor may only be effective if both copies of the gene are lost, although some work suggests that dosage may play a role in copy number drivers also (Davoli et al. 2013). Similarly, the increase in transcription of an oncogene may increase with each copy gained – the more copies the more mRNA can possibly be transcribed simultaneously. Gistic2 classifies the copy number alterations into weak and strong events for each sample. The recurrence of strong events (homozygous loss and multi-copy gain) are more reliable indicators for copy number drivers.

Detection of mode of action

Other approaches to identify cancer drivers have been described which directly part from the assumption that the cancer driver genes are either proto-onocgenes or tumor suppressor genes. It has been proposed to classify the recorded mutations into truncating and repetitive missense which can then be used as proxy for tumor suppressors and oncogenes. A critical underlying assumption for this approach is that truncating mutations are not observed in oncogenes (Vogelstein et al. 2013).

A somewhat more complex approach makes use of a set of classifiers that distinguish the mutational and copy number patterns between oncogenes, tumor suppressor and neutral genes as proposed by (Davoli et al. 2013).

Both approaches are interesting, as the first is relatively simple, and the latter includes multiple genomic evidences. Both methods try to solve two problems in one step: identifying cancer drivers and classifying them into oncogenes and tumor suppressors. It is less clear if those approaches are equally apt to discard passenger alterations and genes as the methods described in the preceding chapters, most of which have been developed specifically exactly to distinguish driver and passenger genes.

Mutual exclusive alteration patterns within driver gene sets

As mentioned earlier, driver alterations are expected to fall into hallmark pathways to confer the cell with tumorigenic capabilities. Those pathways are therefore under a strong selective pressure to be altered. Once altered this selective pressure is relaxed. As mentioned before, some of the proteins within a pathway may be more easily targeted by alterations than others, but each component of the signal-cascade is a possible target.

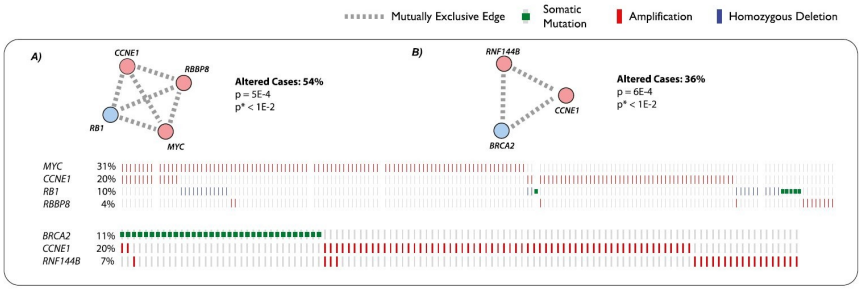

Figure 9: Mutual exclusive interaction modules detected by MeMo
The two cases show interacting proteins which are altered in the genomic sequence or an altered copy number within an ovarian cancer sample cohort. Almost all alterations are mutually exclusive which is a hint at a positive selection for alterations within the module. Figure taken from (Ciriello et al. 2011)

Hence, the multiple targets and the selective pressure contribute to the idea that one alteration per hallmark pathway and sample may be enough to give rise to tumorigenesis. This hypothesis would be reflected by a mutual exclusive alteration pattern within driver pathways in tumor samples. Several methods have been developed in order to detect gene groups with mutually exclusive alteration patterns. Approaches for *de novo* identification of possible gene-gene interactions from mutually exclusive patterns across cancer samples have been proposed (Ciriello et al. 2011; Ciriello, Cerami, et al. 2013; Vandin, Upfal, and Raphael 2011). A hurdle for the approach to detect *de novo* cancer driver gens is the size of the combinatorial possibilities to form gene modules. If we'd like to asses possible combination of 22'000 genes in groups of three, we'd have to test more than one thousand billion (1.3×10^{15}) combinations. In MeMo and Dendrix (Ciriello, Cerami, et al. 2013; Vandin, Upfal, and

Raphael 2011) this problem is somewhat alleviated by using prior knowledge: network-based combinations. The opposite approach relies on the hypothesis driven rationale of testing well-known cancer pathways and modules.

Expression patterns

Besides alterations in the coding sequence of the genes, regulatory mechanisms can lead to the same tumorigenic effect by suppression of the transcription or translation of a tumor suppressor protein or the increased transcription, also called *over-expression*, of an oncogene. Additionally, the possibility to interrogate the transcriptome provides researchers with a global view of the transcriptional status of almost any gene within a tumor sample. Insight can be gained of what genes and pathways are up-, down- and co-regulated across tumor and normal samples (Alon et al. 1999).

Employing gene expression profiling, cancer subtypes have been identified and gene expression signatures of carefully selected genes have been shown to predict the clinical outcome of cancer treatment (Sørlie et al. 2001; van ’t Veer et al. 2002).

1.3 Visual data exploration & cancer genomics data analysis

The large amounts of data that are being generated have to be exploited and analyzed. Researchers and medical staff are and will continue to need to access the data that is released to public domain by TCGA, the ICGC and independent research groups as multidimensional large-scale oncogenomic data sets. Raw data coming out of the various platforms for a single sample has to be processed and analyzed before it is interpretable. This can constitute a barrier for the researcher as hurdles and questions come up: How to process the raw data, what analyses to use? How to merge various datasets of possibly different platforms and cancer types? How to store the processed data in order to guarantee efficient exploration? How to visually explore the multidimensional oncogenomics data?

1.3.1 Interpretation and availability of multidimensional cancer genomics data

The pan-cancer datasets that are now available – containing somatic mutations, CNA alteration status, methylation levels and many more genomic data points of thousands of cancer samples - allow complex analysis. The plethora of data also bears dangers of prompting wrong conclusions and correlations which are not causative. Thus, the context of every data point may prove crucial for interpretation. For example clinical annotations about the cancer patients and samples or molecular details about genes support the interpretation of the results. For Glioblastoma multiforme and breast cancer samples, TCGA data already provides cancer subtype information. When using this information as criteria to group cancer samples for the visualization of the expression data, group-specific expression patterns become visible. The distinct expression patterns may be a reflection of the tumorigenic process that affects genes or pathways in distinct manners in various cancer types and subtypes.

This heterogeneity is the very reason, hypotheses may prove right in certain sub-cohorts but not in others. However, for cancer treatments to become ever more adequate it is important to identify and extract cancer sub-cohorts that have specific molecular patterns. How such a specific cohort is identified depends on how the data is available. Data preparation and normalization is an expensive step in the analysis pipeline and may be unnecessarily repeated by independent groups that download raw cancer genomics data. Resources that provide pre-analyzed and prepared data sets therefore eliminate a big hurdle for many researchers. A good example is the cBio Cancer Genomics Portal (Cerami et al. 2012) which allows the user to select a TCGA cancer cohort and see some preliminary analysis and alteration data. More advanced analyses may help to prefilter a TCGA cancer sample cohort. IntOGen mutations (Gonzalez-Perez, Perez-Llamas, et al. 2013) and IntOGen arrays (Gundem and Lopez-Bigas 2012) similarly let the user browse and filter combined cancer data sets and download gene-based results related to cancer

and tumorigenesis. Both web resources combine a database for analyzed cancer data with visualization approaches. This kind of resources help speed up cancer research in general as quick consultation of preliminary hypotheses is possible as data in analyzed form may be downloaded. Thus, the more effort is put into the combination of intuitive data browsing and easy data accessibility the more the research community will profit. Analysis tools that are used widely in the research community should therefore have easy access to widely used and prepared data sets for them to become more valuable to the field.

1.3.2 Visualizing multidimensional cancer genomics data

In 2013 we published a review on tools and approaches that specifically aid the visualization and exploration of multidimensional cancer genomics data sets. As in this work we made a literature review and with the purpose of avoiding self-plagiarism and repetition, I include the review as a chapter in the introduction.

Schroeder, M.P., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). Visualizing multidimensional cancer genomics data. Genome Medicine 5, 9.

Schroeder MP, Gonzalez-Perez A, Lopez-Bigas N. Visualizing multidimensional cancer genomics data. Genome Medicine. 2013; 5: 9. DOI 10.1186/gm413

Schroeder *et al. Genome Medicine* 2013, **5**:9
http://genomemedicine.com/content/5/1/9

Genome Medicine

REVIEW

Visualizing multidimensional cancer genomics data

Michael P Schroeder[1], Abel Gonzalez-Perez[1] and Nuria Lopez-Bigas*[1,2]

Abstract

Cancer genomics projects employ high-throughput technologies to identify the complete catalog of somatic alterations that characterize the genome, transcriptome and epigenome of cohorts of tumor samples. Examples include projects carried out by the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). A crucial step in the extraction of knowledge from the data is the exploration by experts of the different alterations, as well as the multiple relationships between them. To that end, the use of intuitive visualization tools that can integrate different types of alterations with clinical data is essential to the field of cancer genomics. Here, we review effective and common visualization techniques for exploring oncogenomics data and discuss a selection of tools that allow researchers to effectively visualize multidimensional oncogenomics datasets. The review covers visualization methods employed by tools such as Circos, Gitools, the Integrative Genomics Viewer, Cytoscape, Savant Genome Browser, StratomeX and platforms such as cBio Cancer Genomics Portal, IntOGen, the UCSC Cancer Genomics Browser, the Regulome Explorer and the Cancer Genome Workbench.

Oncogenomics data and their dimensions

Cancer genomics benefits from high-throughput technologies that allow the comparison of the genomic sequences, epigenomic profiles, and transcriptomes of tumor cells with those of normal cells. These technologies often characterize different types of somatic alterations (or variations) in a tumor cell population that are absent from normal cells - including copy number alterations (CNAs), mutations, gene expression changes and methylation changes [1-4]. Together, these somatic alterations constitute multidimensional oncogenomics datasets that describe the variations that coexist in common elements (for example, the genes) of the genome (or transcriptome) of a particular cohort of tumor cells. Such data are currently being used to identify cancer-driver genes and pathways, to discover molecular targets for new therapies, and to define molecular profiles that characterize clinically meaningful patient categories. An array of analytical methods are currently used to exploit the information contained within this multidimensional layout [5-12].

Along with computational and statistical methodologies, effective visual exploration by experts is crucial to successful extraction of knowledge from oncogenomics data. For example, this step might be key to unraveling rare genomic events, verifying data quality at maximum resolution or identifying key players in cancer development. Thus, researchers need intuitive tools that allow the visual integration and simultaneous exploration of both different types of alterations and clinical information. Many data visualization tools have been developed in recent years to support genomic studies. In this review, we revisit the most common ways in which these data are visualized, and present selected tools that allow researchers to visualize multidimensional oncogenomics datasets effectively (Table 1).

To aid our review of the tools, we describe four case studies that illustrate their use: the visual exploration of 1) alterations in cancer-driver genes per tumor through a representation based on OncoPrint (described below); 2) cause-effect relationships between different alteration types in tumor samples, through the use of Gitools and the Network viewer from the cBio Cancer Genomics Portal; 3) the stratification of tumor samples based on clinical annotations, using CircleMap, the Integrative Genomics Viewer (IGV) and Gitools; and 4) dramatic structural alterations that encompass the rearrangement of large chromosomal regions, employing the Circos tool and data obtained from the Catalogue of Somatic Mutations in Cancer (Cosmic).

Types of genomic data visualization

Numerous methods have been developed to automate the analysis of genomic data [13-15]. Nonetheless, the visual exploration of alterations in cancer genomes, epigenomes and transcriptomes in multidimensional datasets, and of

*Correspondence: nuria.lopez@upf.edu
[1]Research Program on Biomedical Informatics - GRIB, Universitat Pompeu Fabra (UPF), Parc de Recerca Biomèdica de Barcelona (PRBB), Dr. Aiguader 88, E-08003 Barcelona, Spain
Full list of author information is available at the end of the article

Table 1. Tools and resources for visualizing multidimensional cancer genomics data

Name	Description	Visualization type	Tool type	Data that can be visualized
cBio Cancer Genomics Portal [32] http://www.cbioportal.org	Resource for visualizing TCGA and other data sets with many features, of which the network viewer and OncoPrint are of special interest. In the network viewer, the portal overlays multidimensional genomics data onto all nodes that are representing genes. This provides the frequency of mutations and copy number alterations (and optionally, mRNA up-/downregulation). OncoPrint shows the same alteration data in a matrix heatmap	Networks Matrix Heatmaps	Web tool	Pre-calculated TCGA and other data sets
CircleMap [8] http://sysbio.soe.ucsc.edu/nets	Tool that produces heatmaps with a circular layout. Different data sets coming from the same samples can be plotted as different layered circles that form a node. The data layers are plotted maintaining the sample order, which can be adjusted by the user	Circular heatmaps	Command line application web tool	Any user-prepared data
Circos [24] http://circos.ca/	Tool for visualizing data and information in a circular layout. It allows intuitive exploration of the relationships between genomic positions, which are depicted as ribbons. Different genomic data types can be represented in different layers of the circle. To a great extent, the color code and plot style for each layer (or data set) can be adjusted by the user	Circular genomic coordinates	Command line application	Any user-prepared data
Caleydo StratomeX [34] http://stratomex.caleydo.org	Tool prepared for the visualization of interdependencies between multiple datasets. It allows exploration of relationships between multiple groupings and different datasets. It can cluster genomics data of different alterations and represents them as matrix heatmaps. The different groupings are connected by ribbons whose width corresponds to the number of samples shared by the connected clusters. Clinical data and pathway maps can be integrated to characterize the clusters	Matrix heatmap with option to visualize pathway maps	Desktop application (Java)	Any user-prepared data (matrices, clusterings). Prepared TCGA data available at http://compbio.med.harvard.edu/tcga/stratomex
Cytoscape [36] http://www.cytoscape.org	Software for visualizing complex networks and integrating these with any type of attribute data such as genomics data and clinical patient information. An extensive library of community-developed plugins is available, some of which (for example, Reactome FIs) focus on cancer data analysis [38]	Networks	Desktop application (Java)	The stand-alone application supports any user-prepared network or attribute data. Additional data are available via various plugins (for example, GeneMANIA [72] for networks)
Genomica [73] http://genomica.weizmann.ac.il	Tool that can be used to analyze and visualize genomic data. Data can be visualized as heatmaps or along genomic coordinates. Module maps and module networks can be created from expression data and can integrate gene expression data, DNA sequence data, and gene and experiment annotations	Matrix heatmap Genomic coordinates	Desktop application (Java)	User-prepared data
Gitools [31] http://www.gitools.org	Tool for analysis and visualization of genomic data using interactive heatmaps. It allows loading of multidimensional matrices (with several values per cell), and thus is very well suited for the visualization and exploration of multidimensional cancer genomics data. It contains several analyses and options that are specifically designed for the exploration of cancer genomics data	Matrix heatmap with interactive features	Desktop application (Java)	Any user-prepared data and data imported from IntOGen [33] database, as well as any Biomart [69,73] database
Integrative Genomics Viewer (IGV) [20] http://www.broadinstitute.org/igv	Visualization tool for interactive exploration of integrated genomics datasets, with a focus on good performance when working with large data sets. All tracks can be annotated with color-coded sample and clinical information; genomic regions can be annotated with text labels. All of the common genomic file formats are supported, including array-based data, next-generation sequence data formats and genomic annotations	Genomic coordinates	Desktop application (Java)	User-prepared data and data from the IGV server, including some TCGA data. In addition, IGV can be accessed from external tools such as GenePattern [68]

Continued overleaf

Table 1. Continued

Name	Description	Visualization type	Tool type	Data that can be visualized
IntOGen [33] http://beta.intogen.org	Resource that is used to analyze and visualize cancer genomics data, including expression, copy number variation and somatic mutation data from cancer genomic projects. Various visualization options are offered, of which web-interactive heatmaps (using jheatmap [74]) are of special interest. These are used to display alterations per gene in a cohort of tumor samples or in a set of tumor types	Matrix heatmaps with interactive features	Web tool	Pre-calculated data from more than 300 cancer genomic experiments and user-prepared data for somatic mutations in tumors
NAViGaTOR [75] http://ophid.utoronto.ca/navigator	Tool for visualizing and analyzing protein-protein interaction networks (Network Analysis, Visualization and Graphing TORonto). The network visualization options can be customized to represent genomic data properties by automatically mapping attribute values to visual properties	Networks	Desktop application (Java)	User-prepared data. Data can also be loaded via plugins from multiple portals (such as Reactome [76] or KEGG [77])
Regulome Explorer [70] http://explorer.cancerregulome.org	Tool for the integrative exploration of associations between clinical and molecular features of data from the TCGA project. The visualization is interactive and the displayed data can be filtered according to different criteria. Visualization options include circular and linear genomic coordinates and networks	Circular and linear genomic coordinates Networks	Web tool	Pre-calculated TCGA data
Savant Genome Browser [22] http://genomesavant.com/savant	Desktop visualization and analysis browser for genomics data. This tool was primarily developed for the effective visualization of large sets of high-throughput sequencing data, similar to IGV. Multiple visualization modes enable the exploration of genome-based sequence, points, intervals, or continuous datasets. Plugins are available, amongst which is the WikiPathways [78] plugin, which aids the navigation of the data by the integration of pathways	Genomic coordinates	Desktop application (Java)	User-prepared data or data that can be downloaded through plugins such as the USCS Explorer plugin
The Cancer Genome Workbench (CGWB) [79] https://cgwb.nci.nih.gov/	Host for mutation, copy number, expression, and methylation data from a number of projects. It has tools for visualizing sample-level genomic and transcription alterations in various cancers. The main viewers in CGWB are Integrated tracks view, Heatmap view and Bambino, an alignment viewer. The interface of CGWB is based on the UCSC Genome Browser [80]	Genomic coordinates Heatmap	Web tool	Pre-calculated data from various resources (such as Cosmic, NCI-60 and TCGA [40,65,81]) The user can also add custom data tracks for visualization
UCSC Cancer Genomics Browser [21] https://genome-cancer.ucsc.edu	Tool for hosting, visualizing, and analyzing cancer genomics datasets. The browser can display genome-wide experimental measurements for multiple samples, which can originate from multiple data sets alongside their associated color-coded clinical information. The browser provides interactive views of data from genomic regions to annotated biological pathways and user-contributed collections of genes. Integrated statistical tools provide quantitative analysis within all available datasets	Genomic coordinates Heatmap	Web tool	TCGA data and data from independent publications available from the UCSC server. In addition to open access to public datasets, the browser provides controlled access to private project data

the relationships between these alterations, presents specific challenges. This review focuses on the visualization principles, methods and tools employed to analyze these multidimensional oncogenomics datasets. (For general reviews on omics data visualization, see [16-19].)

We distinguish between three main approaches commonly used to represent multidimensional oncogenomics data: genomic coordinates, heatmaps and networks (Figure 1). These three approaches complement each other, and each is best suited to answer different specific questions.

Genomic coordinates

A common way to visualize oncogenomics data is to show alterations tied to their genomic loci. This approach is well suited to provide answers to questions about the genomic topography of alterations or to inspect particular genome loci. We distinguish between two main visualization approaches that use genomic coordinates: Genome Browsers and Circular Plots. Three of the most popular genome browsers employed to visualize cancer alterations are the Integrative Genomics Viewer (IGV) [20], the UCSC Cancer Genomics Browser [21], and the Savant Genome Browser [22]. All three support multiple data formats that are used to represent various types of alterations. They display the alterations in each tumor sample as genomic tracks, which can be loaded onto the browser and navigated by zooming and by scrolling to particular genomic regions.

The IGV and Savant genome browsers work as desktop applications and are particularly suited to the display of aligned sequencing data. IGV has a special focus on visualizing integrated datasets that include both array-based and sequencing-based data as well as clinical information about tumor samples and donors. The clinical information displayed in vertical lines in conjunction with the data tracks can be used to sort and group the tracks, thus simplifying the stratification of samples (Figures 2 and 3e). A further advantage of IGV is the split screen view, which allows multiple loci to be displayed next to each other. On the other hand, Savant offers an application programming interface (API) that allows third-party developers to extend and add visual, analytic, navigational, and data loading functions to the genome browser. Available plugins include edgeR [23], aimed at detecting differentially expressed genes or regions. Other plugins are described in the Savant Genome Browser manuscript [22]. Another strength of the Savant genome browser is the visualization of paired-end reads [19].

The web-based UCSC Cancer Genomics Browser offers an easy-to-use interface that can be used to browse cancer genomics datasets, such as those of The Cancer Genome Atlas (TCGA), which have been pre-analyzed with various tools and include clinical information. The user can choose between different plotting types: heatmaps, box plots and proportions. The features are shown in either the classic representation bound to genomic locations or in a gene-set visualization, analogous to the IGV split-screen view, resulting in a browser-like heatmap (Figure 2). Unlike IGV and Savant, the UCSC browser does not allow users to upload data.

Circos [24] is a flexible and popular tool that can be used in many different research fields to plot circular ideograms. In the case of multidimensional oncogenomics data, the genomic coordinates of all chromosomes are represented in a circular layout (Figure 3f). This tool aptly illustrates relationships between distinct alterations, represented as data tracks outside the ideogram, that take place at different locations within the genome. These relationships between regions are normally depicted as ribbons. Intra- and inter-chromosomal translocations are particularly well represented in Circos.

Genome browser tools in general have limited capacity to display relationships between genomic features that are independent of location, such as the coordinated expression of genes. The IGV and UCSC Cancer Genomics Browser attempt to tackle this problem using the split-screen and heatmap approaches, respectively. Another issue with visualization-based genomic reference is that it falls short in visualizing extensive genomic rearrangements. The circular layout of Circos can compensate for this deficit, or it can be resolved by the use of specific tools such as Gremlin [25]. Many other tools also perform specific tasks, exploiting the genomic coordinates representation scheme. For example, putative translocation events can be verified by the command-line tool Pairoscope [26], which generates relational diagrams of paired-end sequencing reads to aid in the discovery of translocation events. To view and analyze single nucleotide polymorphism (SNP) and comparative genomic hybridization (CGH) array alteration data tools and methods such as VAMP [27] and waviCGH [28] are options based on web technologies, whereas Genome Alteration Print [29] is a desktop application. Furthermore, it has been proposed that there should be a move towards visualizing genomic rearrangements, such as gene fusions, graphically to emphasize the order of the rearranged segments rather than the genomic distance between the breakpoints [30].

Heatmaps

Heatmaps are graphical representations that are frequently used to describe transcriptomics and genomics data stored in the form of matrices. In oncogenomics datasets, the columns in a heatmap usually correspond to tumor samples, whereas the rows are genes, transcripts, microarray probes, or other genomic elements (Figure 1). The color of each cell represents a value indicating a measurement of, let's say, for simplicity, the gene in the

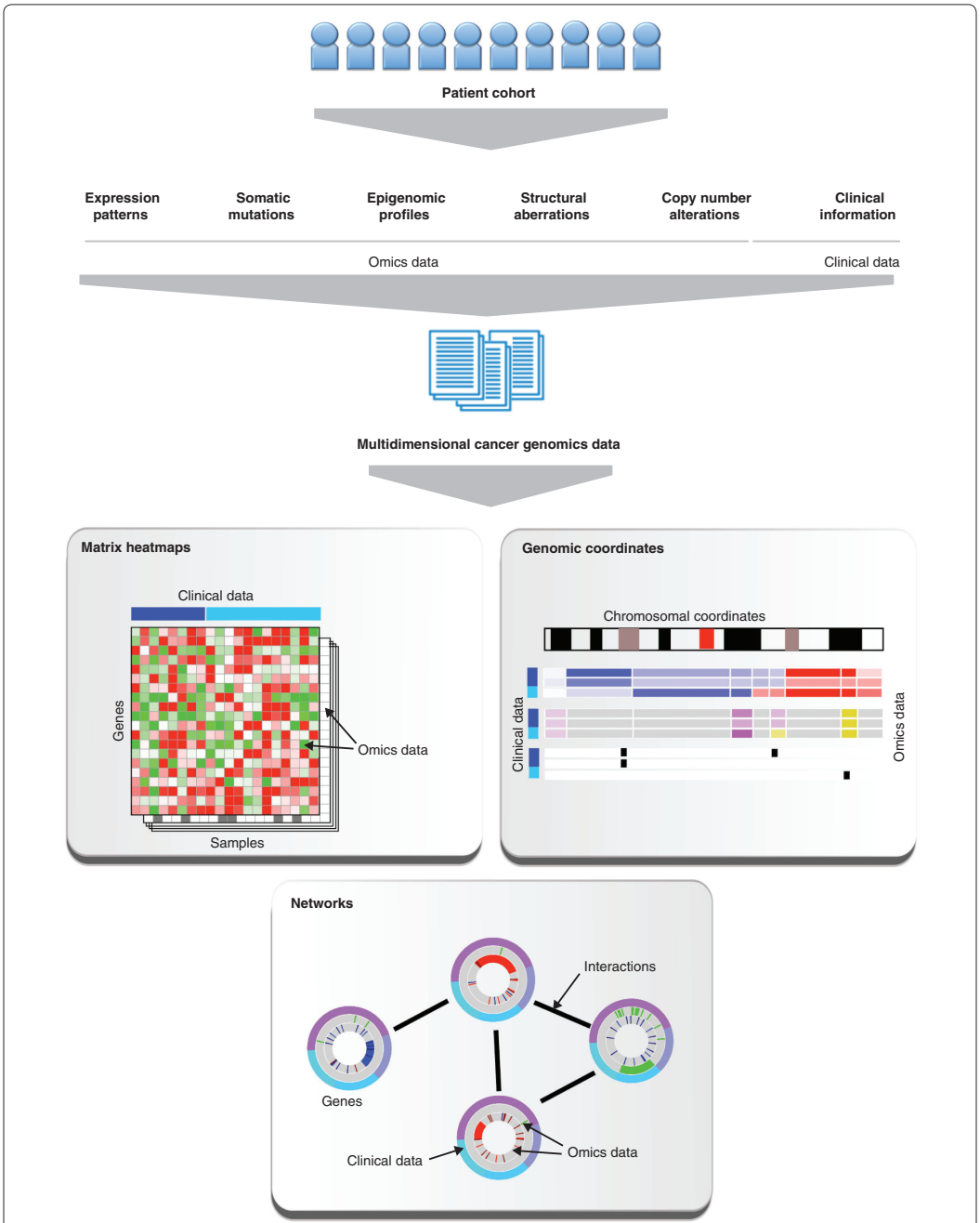

Figure 1. Cancer genomics projects generate multidimensional data for a cohort of patients. Different technological platforms will screen for different genomic and epigenomic changes in each patient, generating multidimensional data sets. The data are usually represented by clinical data along with one or more of the three main types of visualization tools: genomic coordinates, matrix heatmaps and networks.

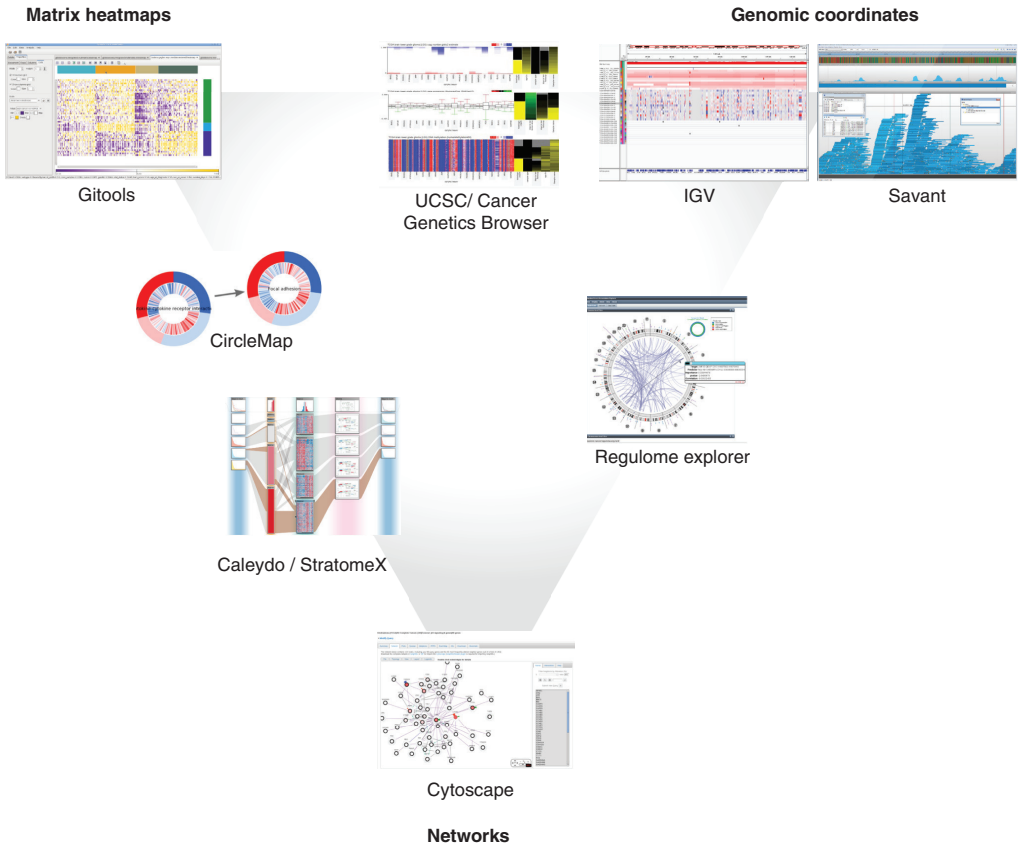

Figure 2. Screenshots of tools that are frequently used in cancer genomics research distributed according to their visualization principles. Each of the three visualization methods - matrix heatmaps, genomic coordinates and networks - are associated with a point of the triangle. Tools that are placed close to one of these points mainly use the visualization method associated with that point; those placed in between use a mixed-model visualization method.

tumor, such as its expression level or mutational status. As matrices, heatmaps impose no restriction on the order of the data. This allows data from distant genome loci to be grouped and visualized together for comparison. For example, genes in the same pathway or genes that are associated with certain tumor types might be grouped together. In other words, rows or columns can be clustered according to molecular or clinical features. It is precisely this flexibility to explore visually patterns within the alterations that are correlated to external characteristics, such as the function of genes or the features of the tumor samples, that make heatmaps so popular as a way of representing multidimensional oncogenomics data.

Many tools and programs generate heatmaps from numerical or categorical matrices. We focus here on tools that have features that are particularly well suited to the visual exploration of multidimensional oncogenomics data.

Gitools [31] is an open-source java application for the analysis and visualization of matrices using interactive heatmaps. The heatmaps in Gitools can contain multiple dimensions, that is, multiple values in each cell, which makes it especially well suited to the exploration of multidimensional cancer genomics data. Its interactive capabilities allow the user to filter, sort, move, and hide rows and columns in the heatmap and to launch several common exploratory analyses (such as correlation, clustering, enrichment and differential expression analyses). Multi-value data matrices, which can contain all types of alterations detected across a cohort of cancer samples, can be explored visually in Gitools, either focusing on a single

dimension (that is, one type of alteration) or fixing one dimension to explore its influence on others. Gitools also allows the integration of these data with clinical information.

The cBio Cancer Genomics Portal [32] is a web resource for visualization of oncogenomics datasets that uses heatmap representation, among other options. The OncoPrint heatmaps display alterations in arrays of genes across tumor samples. Individual genes are represented as rows, and individual cases or patients as columns. Different colors and shapes are used to show different alteration types, so that multiple alterations in a patient's gene can be distinguished easily.

IntOGen is a resource that can be used to analyze and visualize oncogenomics data [33]. It presents different values, estimating the accumulation of somatic mutations, CNA or transcriptional alterations in genes and pathways across tumor samples. Pre-computed data for more than 300 cancer genome experiments are currently available. Web-interactive heatmaps are used to explore gene and pathway alterations across samples and tumor types.

Caleydo StratomeX [34] is a visualization tool built upon the Caleydo framework [35], with a focus on exploring interdependencies between different stratifications of cancer samples within a given study. Genomics data on different alterations can be clustered and visualized as matrix heatmaps. The clusters of different alterations are connected by ribbons whose widths correspond to the number of samples shared by the connected clusters. Clusters can also be visualized as pathway diagrams, allowing the researcher to observe the impact of alterations on pathway function (Figure 2).

Heatmaps can also be represented not as rectangles but as circles, as with CircleMap [8] (Figure 2). With this command-line tool, dimensions can be aligned in a circular plot accompanying a gene, which is represented as a circle that can be attached to other genes in a network layout (Figure 3d).

A general limitation of the heatmap visualization is that structural relationships between genes are difficult to grasp. For instance, it is very hard to discern whether the coincidence of CNA in several genes reflects a possible synergy or is simply a result of their location within a recurrently amplified or deleted chromosomic fragment. Gitools tries to solve this problem by offering the possibility of adding genomic annotations to the rows that can encode functional or structural information. Caleydo StratomeX solves this problem by incorporating pathway diagrams displaying functional relationships between the genes, and CircleMap plots can also be used as nodes to construct a network diagram for this purpose.

Networks

Networks represent functional relationships between different entities, such as genes. This type of information is difficult to represent in heatmaps and non-circular visualizations of genomic coordinates. Genetic features can be coded in node attributes such as color, size, or shape. Different alterations can be displayed as additional halos around the node. The network arrangement allows the researcher to explore visually clusters of nodes representing highly interconnected altered genes that can constitute driver pathways or subnetworks.

Cytoscape [36], a collaborative open-source project, is a widely used and intuitive network visualization and analysis tool in genomics research. No special bioinformatics knowledge is needed to use Cytoscape. The properties of the nodes and the edges and the network layout are customizable, and the comprehensive array of plugins constitutes an added value for researchers. This tool has proven useful for integrating expression data into a gene network [37], as well as for mapping genes with cancer somatic alterations directly to a functional interactions (FI) network [38] that identifies subnetworks of altered genes in order to find cancer drivers. A web version, Cytoscape-web [39], is compatible with common internet browsers and facilitates interaction with the networks displayed. The cBio Cancer Genomics Portal [32] implements an adaption of this tool optimized for visually exploring multidimensional oncogenomics data from TCGA [40]. Node colors and halos encode the alteration status of cancer genes.

Representation of the genomic alterations present in individual tumor samples in network viewers presents a challenge. As a consequence, many details about the individual tumor samples are normally left out of network figures. In the case of the cBio Cancer Genomics Portal network viewer, this problem is alleviated by the inclusion of plots that show the proportions of samples with different genomic alterations. Similar effects can be achieved with plugins for Cytoscape that transform nodes into pie charts (such as GoogleChartFunctions [41] and nodeCharts [42]).

Case studies

The case studies presented here elaborate on four different oncogenomic research questions that can be answered visually with the available tools and resources. The description of the case studies focuses on their biological interpretation. Supporting documentation on how to generate images corresponding to those in Figure 3 is included in the 'Additional file 1 and 2'. Learning to use most of these tools requires a certain investment of time, and tutorials provided by the developers are highly recommended as a starting point.

Visual exploration of cancer drivers

Distinguishing the alterations that give cancer cells a selective advantage (drivers) from those that are merely

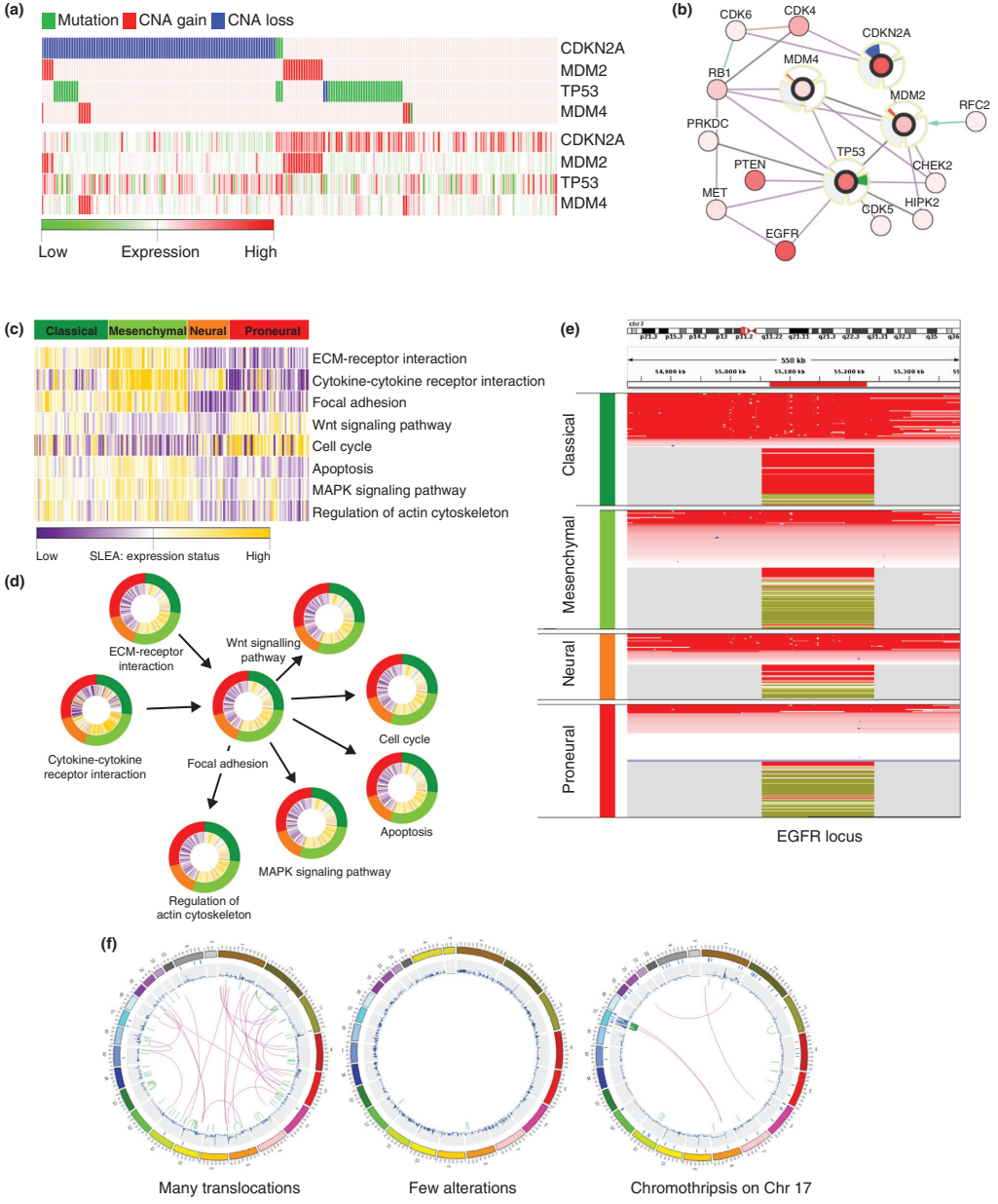

Figure 3. See next page for legend.

Figure 3. Four case studies are represented using one or several of the major visualization methods applied in oncogenomics. **(a)** Heatmap of oncogenomic alterations ordered by mutual exclusivity plotted with Gitools. In the upper half of the image, colors indicate the type of alteration: mutations (green), CNA gain (red) and CNA loss (blue). The heatmap below shows expression data (high expression in red and low expression in green) for the same samples and genes, allowing the visual observation that genomics regions whose copy number is amplified tend to have higher expression values. **(b)** The same data as in (a), with the same color code for alterations, represented as a network of functional interactions between the genes, extracted from the cBio Cancer Genomics Portal. The halo around the four selected nodes is divided into three sectors. Changes in the proportion of samples with altered copy number are indicated in red (gain) or blue (loss) in the top sector, whereas changes in the proportion of samples with mutations are indicated in green in the lower-right sector. Expression changes are shown in light red (increase) and light blue (decrease) in the lower-left sector of the halo. Panels (c-e) include clinical information. Each tumor sample is assigned to one of four subtypes of glioblastoma, color-coded as dark green (classical), light green (mesenchymal), orange (neural) and red (proneural). **(c)** Heatmap of pathway expression levels plotted with Gitools. Each column is a tumor sample. The subtype is represented in colors in the top row and each row represents a biological pathway. The color of each cell indicates the Zscore of the sample level enrichment analysis (SLEA) of the pathway in the sample. Clear differences in the expression values in different pathways can be observed for different cell subtypes. **(d)** Same data as in (c) represented in the form of a network, drawn using CircleMap. Each node is a pathway and its edges indicate functional interactions between pathways as extracted from KEGG. The two halos around each node indicate the Zscore of the pathway in each sample and the clinical subtype. **(e)** CNA and expression data for the EGFR gene region of glioblastoma samples as shown by IGV. The top part of the plot indicates the genomic position we are observing. Each sample is shown as a horizontal track, ordered by clinical subtype. Within each clinical subtype, the tracks in the upper half illustrate CNA whereas those below show expression. This visualization reveals clear differences in the CNA and expression of the EGFR locus in different clinical subtypes. **(f)** Adaptations of Circos plots of three breast tumors with three very different alteration landscapes. The four circles in each plot, from outermost inwards, represent the human chromosomes, mutations, copy number alterations, and structural rearrangement.

side effects (passengers) of the destabilization of the cancer genome is a major problem in oncogenomics research. Several new methodologies [5-8,11,38,43-46] address this problem by exploiting the properties of driver genes. For example, the mutually exclusive alteration of genes in a pathway is a characteristic of cancer drivers [5,6,47]. One plausible explanation of this behavior is that an alteration that targets an affected pathway does not confer further selective advantage to the cancer cell. A built-in Gitools option sorts genes and samples within a heatmap to present the pattern of mutually exclusive alterations, which is one approach to visual exploration of driver genes that are involved in the same pathway (Figure 3a) [48]. Oncoprint (cBio Cancer Genomics Portal) uses the same principle to display the alterations across TCGA datasets of a gene set provided by the user.

An alternative approach to identify cancer drivers involves mapping altered genes to a FI network (Figure 3b) [7,38,46]. The Reactome FI Cytoscape plugin offers this functionality. After a gene list is submitted, a FI network is constructed using so-called linker genes: genes that are not in the user-submitted list but that can connect two of the submitted genes. Usually, this approach identifies network regions in which recurrently altered genes, which are thought to point to driver genes and subnetworks, are enriched. The visualization of genes and their alterations in the form of FI networks is thus very useful (see Figure 3b for an example).

Visualizing cause-effect relationships between different types of alterations

The effect of genomic alterations can be manifested at the genome, transcriptome or proteome level. Single nucleotide variants (SNVs) might not directly influence transcription of the mutated gene but usually affect protein functionality. On the other hand, CNA and changes in methylation status frequently perturb the expression levels of the altered genes or other genes under their control. Determining the cause-effect relationships of such alterations is important to our understanding of cancer mechanisms. One approach is to plot one type of alteration (for example, CNAs) in a heatmap, sorting the tumor samples to separate diploid genes from altered genes. Changes in gene expression values, presented in another heatmap, can then be readily compared between these two groups (Figure 3a), allowing the detection of any significant differences.

Gitools can load a multidimensional data matrix containing different alterations for each sample, and a simple switch between the values shown in the heatmap cells easily changes the display from one heatmap to the other [49].

Networks offer another way of visualizing cause-effect relationships. The interactions between genes in a network can represent their functional relationships, for example, one gene might regulate the expression of another. Overlaying the alterations within a cohort of tumors on top of each node of the network might illustrate the effect of a gene alteration on the expression of other genes in the network (analogous to Figure 3b).

The network viewer of the cBio Cancer Genomics Portal supports the visualization of expression data, if available. Similar visual effects could be achieved in Cytoscape by mapping data onto node properties.

Visualizing cancer patient stratifications

Cancer is a complex disease. Tumors that seem very similar when examined through conventional diagnostic methods might look markedly different from the

molecular viewpoint, which can lead to different outcomes or treatment responses. Therefore, the molecular features of tumors can be used to stratify patients to support more accurate clinical and therapeutic decisions. Over the past decade, molecular stratification of tumors using expression microarrays has been an important area of cancer research [50-53]. The visualization of molecular alteration patterns in a heatmap is often used to explore subgroups of tumors and to associate them with particular clinical features. These heatmaps usually portray the expression patterns of genes or transcripts across samples, but the benefit of data analysis at the level of gene groups, for example pathways [54-56], is increasingly evident. Stratification and visualization can also be done at the level of pathways or other gene modules (Figure 3c), for example using sample level enrichment analysis (SLEA) [57,58], which analyzes the transcriptional status of pathways (or other gene sets) in each tumor sample.

In the case of multidimensional oncogenomics data, various clinical features and alterations such as CNA or changes in mRNA or microRNA expression can be used to cluster or stratify tumors, leading to different groupings of samples. In Figure 3c,d, we show ways of representing the results of applying SLEA to the TCGA glioblastoma dataset, with the samples grouped by the corresponding glioblastoma subtype. The alterations are visualized using both Gitools and CircleMap. Please see the 'Additional file 1 and 2' for a more detailed description of this process.

Stratifications can also be meaningful when exploring a single locus. Figure 3e illustrates the same grouping of samples by glioblastoma subtype, employing copy number and expression data from the TCGA glioblastoma study using IGV (Figure 3e).

Caleydo StratomeX is especially well suited to exploring relationships between groups of samples (Figure 2). These relationships are visualized as ribbons of varying width drawn between neighboring columns. Wide ribbons encode a high co-occurrence of samples in different groupings, whereas their absence indicates mutual exclusion. This coding provides a straightforward and scalable overview of the consistency of group memberships of tumor samples across different data types.

Visualizing global alteration profile patterns

Various alteration phenotypes have been observed in cancer cells. One of the most conspicuous of these is the mutator phenotype [59]: tumor cells typically have an abnormally high mutational burden. Tumor samples with chromothripsis [60,61] or many chromosomal translocations are also common. Categorization of the alteration events in a cancer cell population could influence the therapeutic decision, and requires a simultaneous exploratory view of all the alteration events.

One approach to exploring visually all the alterations of a sample is the circular genome mapping proposed by Krzywinski *et al.* using their tool Circos [24]. Several cancer studies [59,62-64] have used Circos to show the landscape of alterations. This tool is highly configurable, which is evident from the figures in the cited publications. One compact figure can represent all somatic alteration events in a given tumor sample. Data from different alteration types can be organized in layered circles while rearrangement events occupy the innermost space. Figure 3f is composed of three Circos plots of breast cancer samples [59] as they are represented on the Cosmic website [65]. The outer-most circle of each diagram represents the human chromosomes, followed by a plot of ticks showing point mutations. The next layer plots CNA along all the chromosomes; the links in the middle visualize the structural rearrangements.

The recently developed ggbio package [66] for the R programming environment allows, among other things, the creation of circular genome plots, and supports a variety of data formats for sequencing data.

Interfacing of tools

Researchers often need to use several of the complementary tools described here to explore their datasets. Nevertheless, the landscape of visualization tools for multidimensional oncogenomics data seems rather fragmented. This is the result of different groups focusing on the development of tools optimized to solve one particular visualization issue, which is probably a more efficient way of investing resources instead of engineering one single monolithic tool that has all possible visualization capabilities. Unfortunately, this fragmentation makes the use of different tools problematic: they accept very different data formats, they look different to users and so on. Thus, users need to spend time learning how to use each tool and reformatting their data to each tool's requirements. This extra effort could be alleviated if developers were to facilitate the combined use of tools.

One of the major efforts to develop a universal interface that will bridge the gap between different bioinformatic tools is the GenomeSpace project [67]. GenomeSpace allows the user to store data in a common repository and the same web interface guides users to execute the integrated tools, load data, and store results. Conveniently, it contains several built-in converters for some often-used data formats. Several tools listed in Table 1 (IGV, Genomica, Cytoscape and Gitools) are included in this pilot project. This platform interface approach is promising and possibly the most user-friendly option for users who lack a background in bioinformatics.

Another approach to facilitate the use of several tools is the creation of direct tool-to-tool interfaces. These are

possible when a tool offers an API that defines the form of communication between the tool and the rest of the world. There are different kinds of APIs, which allows the implementation of different approaches. If the API offers external control, it can send the tool a command and indicate whether the execution of this command has been successful or not. This is the case, for example, with IGV and Gitools: both offer a set of commands that the other application can use. Gitools has a built-in link that sends a 'find locus' command to IGV, whereas IGV exports data into a matrix format and commands Gitools to load it. In practical terms, this means that the user can explore the same data with two complementary visualization tools that can communicate with each other.

Another kind of API can be used for plugin development. This is a general way of creating new capabilities for established tools. As mentioned above, Cytoscape and Savant support plugability, meaning that they possess internal commands that can be used by an application to extend the functions of the tool.

Unidirectional APIs are typically employed by databases and allow easy data transfer between the data source and tools. For example, IGV's external control of the software allows the cBio Cancer Genomics Portal and GenePattern [68] to load data directly into IGV, and Gitools accepts imported data for all BioMart [69] databases.

Conclusions and future directions

The cancer genomics research field is rapidly evolving in parallel with advances in high-throughput genomics technologies. This evolution of the field requires continuous advancement in visualization techniques and tools. For instance, the amount of data it is possible to generate for an oncogenomics project continues to increase, requiring visualization tools that very efficiently load and process large amounts of data.

As this rapid scientific evolution continues, cancer researchers are highly dependent on computational scientists and bioinformatics professionals to help them manage, analyze and visualize data. To speed up research advances, the barrier between the large amount of data generated in oncogenomics projects and the effective exploration of these data by cancer researchers must be minimized. Visualization and exploration tools should be intuitive and easy to use, not requiring computational or bioinformatics expertise. Not all tools currently meet these standards, as some programming or even technological knowledge is required of the user. In recent years, however, there has been an important effort to facilitate access by 'non-bioinformaticians' to visualization tools for the analysis of oncogenomics data [20,31,32]. Continued work to improve the usability of visualization software is highly important, but requires great effort from developers for low scientific reward when compared to the development of new methods or visualization techniques. Funding agencies must understand that increased investment in personnel dedicated to the development and maintenance of new tools, as well as user training and support, is crucial to the achievement of improvements in the field.

The complexity of oncogenomics data and the multitude of questions to be addressed ensure that a static plot is often insufficient for data visualization. The user needs to explore the data interactively in order to address a wide range of questions. Several tools listed in Table 1 (including IGV, Gitools and Caleydo) make use of interactive visualization techniques to make this possible. Other web frameworks with various visualization and some optional analysis possibilities are being developed, including the cBio Cancer Genomics Portal [32], IntOGen [33] and Regulome Explorer [70]. Open source and plug-in architecture facilitates quick adoption of these new platforms.

Although not discussed at length in this review, the use of cancer genomics data visualization in the clinical setting is likely to become a key topic in the near future, as the results of cancer genome projects begin to be translated into personalized cancer medicine. Clinicians will be the main users of this information as they make decisions regarding patient treatment. In this regard, simple, efficient tools that support the visual stratification of tumor genomic profiles and that highlight their relationships to known drugs or treatments will be more useful than the existing research-oriented tools. As a result, it will probably be necessary to develop specialized clinical tools or to adapt existing ones to the clinical setting. This has been achieved in the case of the MedSavant Browser [71], a clinical adaption of the Savant Genome Browser.

In summary, visualization of multidimensional oncogenomics data is essential for the extraction of useful knowledge from the vast amount of data generated by high-throughput technologies. Important efforts have been made in recent years to create visualization tools that can explore these datasets. Further efforts are needed to develop those resources and to create new tools to meet the changing needs of the field. Long-term investment and funding are needed to guarantee the maintenance, improvement, and evolution of visualization tools beyond their first publication.

Additional files

Additional file 1. The following additional data are available with the online version of this paper. Additional file 1 provides information on how to generate visualization images for the case studies covered.

Additional file 2. Instructions on using Additional file 1.

Abbreviations
API, application programming interface; CNA, copy-number alteration; Cosmic, Catalogue of Somatic Mutations in Cancer; FI, functional interactions; IGV, Integrative Genomics Viewer; TCGA, The Cancer Genome Atlas.

Competing interests
The authors declare no competing financial interests and declare the authorship of Gitools and IntOGen.

Acknowledgements
We acknowledge funding from the Spanish Ministry of Science and Technology (grant number SAF009-06954 and FPI fellowship assigned to MPS), and from the Spanish National Institute of Bioinformatics (INB). We are thankful to Elaine M Lilly for assistance with language editing.

Author details
[1]Research Program on Biomedical Informatics - GRIB, Universitat Pompeu Fabra (UPF), Parc de Recerca Biomèdica de Barcelona (PRBB), Dr. Aiguader 88, E-08003 Barcelona, Spain. [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

Published: 31 January 2013

References

1. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I-M, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA, Hartigan J, Smith DR, Strausberg RL, Marie SKN, Shinjo SMO, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, *et al.*: **An integrated genomic analysis of human glioblastoma multiforme.** *Science* 2008, **321:**1807-1812.
2. The Cancer Genome Atlas Consortium: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455:**1061-1068.
3. The Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474:**609-615.
4. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, *et al.*: **International network of cancer genome projects.** *Nature* 2010, **464:**993-998.
5. Ciriello G, Cerami EG, Sander C, Schultz N: **Mutual exclusivity analysis identifies oncogenic network modules.** *Genome Res* 2012, **22:**398-406.
6. Vandin F, Upfal E, Raphael BJ: **De novo discovery of mutated driver pathways in cancer.** *Genome Res* 2012, **12:**375-385.
7. Vandin F, Upfal E, Raphael BJ: **Algorithms for detecting significantly mutated pathways in cancer.** *J Comput Biol* 2011, **18:**507-522.
8. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM: **Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.** *Bioinformatics* 2010, **26:**i237-i245.
9. Louhimo R, Lepikhova T, Monni O, Hautaniemi S: **Comparative analysis of algorithms for integration of copy number and expression data.** *Nat Methods* 2012, **9:**351-355.
10. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D: **An integrated approach to uncover drivers of cancer.** *Cell* 2010, **143:**1005-1017.
11. Tamborero D, Lopez-Bigas N, Gonzalez-Perez A: **Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression.** *PLoS ONE*, in press.
12. Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ: **Discovery of multi-dimensional modules by integrative analysis of cancer genomic data.** *Nucleic Acids Res* 2012, **40:**9379-9391.
13. Medvedev P, Stanciu M, Brudno M: **Computational methods for discovering structural variation with next-generation sequencing.** *Nat Methods* 2009, **6:**S13-S20.
14. Rueda OM, Diaz-Uriarte R: **Finding recurrent copy number alteration regions: a review of methods.** *Curr Bioinform* 2010, **5:**1-17.
15. Eifert C, Powers RS: **From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets.** *Nat Rev Cancer* 2012, **12:**572-578.
16. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin A-C: **Visualization of omics data for systems biology.** *Nat Methods* 2010, **7:**S56-S68.
17. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T: **Visualizing genomes: techniques and challenges.** *Nat Methods* 2010, **7:**S5-S15.
18. O'Donoghue SI, Gavin A-C, Gehlenborg N, Goodsell DS, Hériché J-K, Nielsen CB, North C, Olson AJ, Procter JB, Shattuck DW, Walter T, Wong B: **Visualizing biological data - now and in the future.** *Nat Methods* 2010, **7:**S2-S4.
19. Quinlan AR, Hall IM: **Characterizing complex structural variation in germline and somatic genomes.** *Trends Genet* 2012, **28:**43-53.
20. Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Brief Bioinform* 2012. doi: 10.1093/bib/bbs017.
21. Sanborn JZ, Benz SC, Craft B, Szeto C, Kober KM, Meyer L, Vaske CJ, Goldman M, Smith KE, Kuhn RM, Karolchik D, Kent WJ, Stuart JM, Haussler D, Zhu J: **The UCSC Cancer Genomics Browser: update 2011.** *Nucleic Acids Res* 2011, **39:**D951-959.
22. Fiume M, Smith EJM, Brook A, Strbenac D, Turner B, Mezlini AM, Robinson MD, Wodak SJ, Brudno M: **Savant Genome Browser 2: visualization and analysis for population-scale genomics.** *Nucleic Acids Res* 2012, **40 (Web Server issue):**W615-W621.
23. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
24. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19:**1639-1645.
25. O'Brien TM, Ritz AM, Raphael BJ, Laidlaw DH: **Gremlin: an interactive visualization model for analyzing genomic rearrangements.** *IEEE Trans Vis Comput Graph* 2010, **16:**918-926.
26. Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, *et al.*: **Genome remodelling in a basal-like breast cancer metastasis and xenograft.** *Nature* 2010, **464:**999-1005.
27. Rosa PL, Viara E, Hupé P, Pierron G, Liva S, Neuvial P, Brito I, Lair S, Servant N, Robine N, Manié E, Brennetot C, Janoueix-Lerosey I, Raynal V, Gruel N, Rouveirol C, Stransky N, Stern M-H, Delattre O, Aurias A, Radvanyi F, Barillot E: **VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles.** *Bioinformatics* 2006, **22:**2066-2073.
28. Carro A, Rico D, Rueda OM, Díaz-Uriarte R, Pisano DG: **waviCGH: a web application for the analysis and visualization of genomic copy number alterations.** *Nucleic Acids Res* 2010, **38:**W182-W187.
29. Popova T, Manié E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH: **Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays.** *Genome Biol* 2009, **10:**R128.
30. Nielsen C, Wong B: **Points of view: representing genomic structural variation.** *Nat Methods* 2012, **9:**631.
31. Perez-Llamas C, Lopez-Bigas N: **Gitools: analysis and visualisation of genomic data using interactive heat-maps.** *PLoS ONE* 2011, **6:**e19541.
32. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.** *Cancer Discov* 2012, **2:**401-404.
33. Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, Furney SJ, Lopez-Bigas N: **IntOGen: integration and data mining of multidimensional oncogenomic data.** *Nat Methods* 2010, **7:**92-93.
34. Lex A, Streit M, Schulz H-J, Partl C, Schmalstieg D, Park PJ, Gehlenborg N: **StratomeX: visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization.** *Comput Graph Forum* 2012, **31:**1175-1184.
35. Streit M, Lex A, Kalkusch M, Zatloukal K, Schmalstieg D: **Caleydo: connecting pathways and gene expression.** *Bioinformatics* 2009, **25:**2760-2761.
36. Shannon P: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
37. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang P-L, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, *et al.*: **Integration of biological networks and gene expression data using Cytoscape.** *Nat Protoc* 2007, **2:**2366-2382.
38. Wu G, Feng X, Stein L: **A human functional protein interaction network and its application to cancer data analysis.** *Genome Biol* 2010, **11:**R53.

39. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD: **Cytoscape Web: an interactive web-based network browser.** *Bioinformatics* 2010, **26:**2347-2348.
40. **The Cancer Genome Atlas** [http://cancergenome.nih.gov/]
41. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization.** *Bioinformatics* 2011, **27:**431-432.
42. **nodeCharts Cytoscape Plugin** [http://www.cgl.ucsf.edu/cytoscape/utilities/index.html#nodeCharts]
43. Gonzalez-Perez A, Lopez-Bigas N: **Functional impact bias reveals cancer drivers.** *Nucleic Acids Res* 2012, **40:**e169.
44. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.** *Genome Biol* 2011, **12:**R41.
45. Getz G, Höfling H, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES: **Comment on "The consensus coding sequences of human breast and colorectal cancers."** *Science* 2007, **317:**1500.
46. Cerami E, Demir E, Schultz N, Taylor BS, Sander C: **Automated network analysis identifies core pathways in glioblastoma.** *PLoS ONE* 2010, **5:**e8918.
47. Thomas RK, Baker AC, DeBiasi RM, Winckler W, LaFramboise T, Lin WM, Wang M, Feng W, Zander T, MacConaill LE, Lee JC, Nicoletti R, Hatton C, Goyette M, Girard L, Majmudar K, Ziaugra L, Wong K-K, Gabriel S, Beroukhim R, Peyton M, Barretina J, Dutt A, Emery C, Greulich H, Shah K, Sasaki H, Gazdar A, Minna J, Armstrong SA, *et al.*: **High-throughput oncogene mutation profiling in human cancer.** *Nat Genet* 2007, **39:**347-351.
48. **Visualizing mutually exclusive alteration patterns in cancer with Gitools** [http://bg.upf.edu/blog/2012/03/visualizing-mutually-exclusive-alteration-patterns-in-cancer-with-gitools/]
49. **Exploring the effect of cancer genomic alteration on expression with Gitools** [http://bg.upf.edu/blog/2012/03/exploring-the-effect-of-cancer-genomic-alteration-on-expression-with-gitools/]
50. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530-536.
51. van 't Veer LJ, Bernards R: **Enabling personalized cancer medicine through analysis of gene-expression patterns.** *Nature* 2008, **452:**564-570.
52. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J Clin Oncol* 2009, **27:**1160-1167.
53. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci U S A* 2001, **98:**10869-10874.
54. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.
55. Khatri P, Sirota M, Butte AJ: **Ten years of pathway analysis: current approaches and outstanding challenges.** *PLoS Comput Biol* 2012, **8:**e1002375.
56. Bild AH, Potti A, Nevins JR: **Linking oncogenic pathways with therapeutic opportunities.** *Nat Rev Cancer* 2006, **6:**735-741.
57. Gundem G, Lopez-Bigas N: **Sample level enrichment analysis (SLEA) unravels shared stress phenotypes among multiple cancer types.** *Genome Med* 2012, **4:**28.
58. **Sample Level Enrichment Analysis (SLEA) tutorial and Gitools 1.6.2** [http://bg.upf.edu/blog/2012/04/sample-level-enrichment-analysis-slea-tutorial-and-gitools-1-6-2/]
59. Stephens PJ, McBride DJ, Lin M-L, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, Greenman CD, Jia M, Latimer C, Teague JW, Lau KW, Burton J, Quail MA, Swerdlow H, Churcher C, Natrajan R, Sieuwerts AM, Martens JWM, Silver DP, Langerød A, Russnes HEG, Foekens JA, Reis-Filho JS, van 't Veer L, Richardson AL, Børresen-Dale A-L, *et al.*: **Complex landscapes of somatic rearrangement in human breast cancer genomes.** *Nature* 2009, **462:**1005-1010.
60. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin M-L, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, *et al.*: **Massive genomic rearrangement acquired in a single catastrophic event during cancer development.** *Cell* 2011, **144:**27-40.
61. Rausch T, Jones DTW, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, Jäger N, Remke M, Shih D, Northcott PA, Pfaff E, Tica J, Wang Q, Massimi L, Witt H, Bender S, Pleier S, Cin H, Hawkins C, Beck C, von Deimling A, Hans V, Brors B, Eils R, Scheurlen W, Blake J, Benes V, Kulozik AE, Witt O, Martin D, *et al.*: **Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations.** *Cell* 2012, **148:**59-71.
62. Kloosterman WP, Hoogstraat M, Paling O, Tavakoli-Yaraki M, Renkens I, Vermaat JS, van Roosmalen MJ, van Lieshout S, Nijman IJ, Roessingh W, van 't Slot R, van de Belt J, Guryev V, Koudijs M, Voest E, Cuppen E: **Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer.** *Genome Biol* 2011, **12:**R103.
63. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, Tine BAV, Hoog J, Goiffon RJ, Goldstein TC, Ng S, Lin L, Crowder R, Snider J, Ballman K, Weber J, Chen K, Koboldt DC, Kandoth C, Schierding WS, McMichael JF, Miller CA, Lu C, Harris CC, McLellan MD, Wendl MC, DeSchryver K, Allred DC, Esserman L, Unzeitig G, *et al.*: **Whole-genome analysis informs breast cancer response to aromatase inhibition.** *Nature* 2012, **486:**353-360.
64. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, *et al.*: **Mutational processes molding the genomes of 21 breast cancers.** *Cell* 2012, **149:**979-993.
65. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R: **The COSMIC (catalogue of somatic mutations in cancer) database and website.** *Br J Cancer* 2004, **91:**355-358.
66. Yin T, Cook D, Lawrence M: **ggbio: an R package for extending the grammar of graphics for genomic data.** *Genome Biol* 2012, **13:**R77.
67. **GenomeSpace** [http://www.genomespace.org/]
68. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: **GenePattern 2.0.** *Nat Genet* 2006, **38:**500-501.
69. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A: **BioMart Central Portal -unified access to biological data.** *Nucleic Acids Res* 2009, **37:**W23-W27.
70. Network TCGA: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487:**330-337.
71. **MedSavant** [http://genomesavant.com/medsavant/]
72. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q: **GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function.** *Genome Biol* 2008, **9:**S4.
73. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nat Genet* 2004, **36:**1090-1098.
74. **jHeatmap** [http://bg.upf.edu/jheatmap/]
75. Brown KR, Otasek D, Ali M, McGuffin MJ, Xie W, Devani B, van Toch IL, Jurisica I: **NAViGaTOR: Network Analysis, Visualization and Graphing Toronto.** *Bioinformatics* 2009, **25:**3327-3329.
76. Joshi-Tope G: **Reactome: a knowledgebase of biological pathways.** *Nucleic Acids Res* 2004, **33:**D428-D432.
77. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27:**29-34.
78. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR: **WikiPathways: building research communities on biological pathways.** *Nucleic Acids Res* 2012, **40(Database issue):**D1301-D1307.
79. Zhang J, Finney R, Edmonson M, Schaefer C, Rowe W, Yan C, Clifford R, Greenblum S, Wu G, Zhang H, Liu H, Nguyen C, Hu Y, Madhavan S, Ding L, Wheeler DA, Gerhard DS, Buetow KH: **The Cancer Genome Workbench: identifying and visualizing complex genetic alterations in tumors.** *NCI Nature Pathway Interaction Database* 2010, doi: 10.1038/pid.2010.1.
80. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The Human Genome Browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
81. Shoemaker RH: **The NCI60 human tumour cell line anticancer drug screen.** *Nat Rev Cancer* 2006, **6:**813-823.

doi:10.1186/gm413
Cite this article as: Schroeder MP, *et al.*: **Visualizing multidimensional cancer genomics data.** *Genome Medicine* 2013, **5:**9.

1.3.3 Modular data visualization: web data portals

In the preceding section, several tools which implement a specific visualization approach have been described. Dedicated tools are great for providing an in-depth experience of a specific approach but fail to offer different types of visualization of the same data which can be somewhat alleviated by inter-connectivity of different tools on the desktop.

But the latest developments in web technologies provide a new environment with a unified rule set for so called web-applications. That the browser and HTML documents can be used for research purposes has been shown already with the USCS Genome Browser, a web native application available since 2000. Today's HTML5 combined with JavaScript and potent web browsers allow for complex code to be executed within them – on desktops and mobile devices. Especially genome browsers are popular as stand-alone web-application (Meyer et al. 2012; Pak and Roth 2013; Westesson, Skinner, and Holmes 2012), but in the context of multidimensional cancer genomics data sets, the so-called portals are a very appealing concept: a portal is a data browser that combines different visualization approaches of data that may have been filtered according to choices from the user. The data can be precomputed or even on the fly computation may happen in the browser.

The development of approaches on data visualization web-applications is therefore much needed. As each developed component or web-application has to be embedded in a HTML5 document, a modular development approach makes sense as it serves the whole community as well. In the case of data visualization, each developed visualization component may be used alongside others developed by other groups in a data portal. BioJS (Gómez et al. 2013) exemplifies this by cataloging JavaScript applications that are related to biological data access and representation in the web browser. BioJS provides a loose framework for how each module has to be installed but still leaving great flexibility to the module creators.

The cBioPortal (Cerami et al. 2012) for cancer genomics is a good

example of how multidimensional datasets can be explored and some simple analyses can be performed within the portal. In the end, the user may download parts of the data for offline exploration and analysis by applying filters. As stated above, web portals can provide a wide range of views of data. Other dedicated web services such as IntOGen Mutations (Gonzalez-Perez, Perez-Llamas, et al. 2013) offer online analysis coupled with online results browsing for which independent visual components may be used.

One hurdle for complete web-application based data analysis and research is that most research groups that create independent services do not have resources (funding, time, know-how) to maintain web-servers that support computational intensive tasks for a large clientèle. Another hurdle is that in many cases researchers are not legally allowed or simply uncomfortable to submit genomic data of their patients samples to other services and thereafter offline or local solutions are still a requirement.

2 OBJECTIVES

In the light of the fields that have been introduced, I'd like to state the goals of my PhD thesis separated into two main objectives:

Cancer data analysis

- Develop a framework or method that can classify cancer driver genes into their respective roles of oncogenes and tumor suppressor genes.
- Develop a hypothesis-driven method to test for mutually exclusive alterations in cancer drivers

Biological data visualization

- Facilitate the accessibility, visualization and analysis of cancer genomics datasets with the help of interactive matrix heatmap visualization solutions for different use cases:
  - On the desktop, focusing on good performance of large datasets and easy data interpretation.
  - On the web, focusing on easy and interactive communication of complex datasets.
- Create an easy solution for mapping molecular biological data onto complex figures.

3 RESULTS

3.1 OncodriveROLE classifies cancer driver genes in Loss of Function and Activating mode of action

In this chapter I present the OncodriveROLE classifier, an approach to separate cancer driver genes into different mode of actions, namely *activating* and *loss of function*, with the premise of aiding on the identification of drug targets and provide valuable information for development of computer models of the cancer disease.

The classifier is based on mutational and copy number patterns within a cancer sample cohort. As we propose an alternative approach we compare ours to two preceding approaches in order to assess the capabilities of OncodriveROLE.

Schroeder, M.P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. OncodriveROLE classifies cancer driver genes in Loss of Function and Activating mode of action. Bioinformatics *30*.

Schroeder MP, Rubio-Perez C, Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. Bioinformatics. 2014 Sep 1; 30(17): i549-55. DOI: 10.1093/bioinformatics/btu467

BIOINFORMATICS

Vol. 30 ECCB 2014, pages i549–i555
doi:10.1093/bioinformatics/btu467

OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action

Michael P. Schroeder[1], Carlota Rubio-Perez[1], David Tamborero[1], Abel Gonzalez-Perez[1,*] and Nuria Lopez-Bigas[1,2,*]

[1]Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, E08003 Barcelona and [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), E08010 Barcelona, Spain

ABSTRACT

Motivation: Several computational methods have been developed to identify cancer drivers genes—genes responsible for cancer development upon specific alterations. These alterations can cause the loss of function (LoF) of the gene product, for instance, in tumor suppressors, or increase or change its activity or function, if it is an oncogene. Distinguishing between these two classes is important to understand tumorigenesis in patients and has implications for therapy decision making. Here, we assess the capacity of multiple gene features related to the pattern of genomic alterations across tumors to distinguish between activating and LoF cancer genes, and we present an automated approach to aid the classification of novel cancer drivers according to their role.

Result: OncodriveROLE is a machine learning-based approach that classifies driver genes according to their role, using several properties related to the pattern of alterations across tumors. The method shows an accuracy of 0.93 and Matthew's correlation coefficient of 0.84 classifying genes in the Cancer Gene Census. The OncodriveROLE classifier, its results when applied to two lists of predicted cancer drivers and TCGA-derived mutation and copy number features used by the classifier are available at http://bg.upf.edu/oncodrive-role.

Availability and implementation: The R implementation of the OncodriveROLE classifier is available at http://bg.upf.edu/oncodrive-role.

Contact: abel.gonzalez@upf.edu or nuria.lopez@upf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Research in cancer genomics has identified hundreds of genes involved in different stages of tumorigenesis due to specific somatic events. Single nucleotide variants, and large-scale amplifications and deletions of chromosomal regions have been identified as two of the main driver alterations in human tumors. The genes suffering these alterations are traditionally classified as oncogenes and tumor suppressors, depending on their role in cancer development. When the product of tumor suppressors lose their function, tumor cells tend to proliferate faster. Driver alterations in these genes frequently exhibit a recessive behavior. The loss of function (LoF) can be achieved through truncating or missense mutations, DNA deletions or hypermethylation of their promoters. Some known LoF genes, most notably BRCA1 and BRCA2, carry germline variants that increase the susceptibility to develop a tumor because only one hit is required to inactivate their function. Oncogenes, on the other hand, increase or change their function upon somatic variants in tumorigenesis. Therefore, their mode of action follow a dominant pattern, as one faulty copy of the gene is frequently enough to provide the required phenotype. A copy number gain may exponentiate the oncogenic function of the gene; a point mutation may achieve the same result by changing key amino acid residues, which results in constitutive activation of the protein, or produce a new biochemical function. These special cases are also regarded as activating driver mutations, as the new function is gained much like in the case of classic oncogenes. The Cancer Gene Census (CGC; Futreal *et al.*, 2004) is a regularly updated compilation of well-studied cancer genes, which classifies their mode of action as dominant or recessive, following the oncogene/tumor suppressor paradigm, *LoF* and *Act* (activated), hereafter. The CGC contains some 500 genes implicated in cancer (November 2013). This is a rather small fraction of the 20 000 genomes in the human genome (International Human Genome Sequencing Consortium, 2004), but recent large-scale re-sequencing projects of tumor genomes (Hudson *et al.*, 2010) suggest many additional genes may be involved in tumorigenesis. One important first step in the analysis of datasets of cancer genomics alterations is the identification of the genes that drive tumorigenesis. This is a non-trivial problem because tumor samples contain up to thousands of somatic alterations. The list of genes altered in tumors is heterogeneous, even within the same cancer type. Therefore, the difficult task is to distinguish between *driver* and *passenger* alterations.

The most intuitive way to identify driver genes is to detect signals of positive selection across tumor samples because cancer cell populations undergo a selection process during the progression of the disease. Different methods that aim to identify driver genes tackle different evidences to achieve their goal (Gonzalez-Perez *et al.*, 2013a). Two recent efforts to comprehensively identify driver genes across large cohorts carried out by Lawrence *et al.* (2014) and Tamborero *et al.* (2013b), combining several signals of positive selection (Dees *et al.*, 2012; Gonzalez-Perez and Lopez-Bigas, 2012; Lawrence *et al.*, 2013; Reimand *et al.*, 2013) detected, respectively, 291 and 260 likely driver genes.

Although years of experimental work have revealed the role of most well-known cancer genes, now our capability of detecting drivers has surpassed our capacity to probe their mode of action. Thus, revealing the mode of action of driver genes in tumorigenesis is becoming crucial to fully understand the mechanisms of tumorigenesis. This is essential for the development of new targeted cancer therapies because as a general rule only Act drivers are in principle susceptible to targeted drugs. Although exceptionally, some mutated tumor suppressors may be targeted (e.g.

*To whom correspondence should be addressed.

Lambert *et al*., 2009), other strategies, such as synthetic lethality, are needed to compensate for their LoF. This is the reason why we need to develop bioinformatics approaches to make this classification as accurately as possible. Vogelstein *et al*. recently described the so-called '20/20 rule' to detect tumor suppressor genes and oncogenes based on their mutational pattern across tumor samples (Vogelstein *et al*., 2013). It states that genes with ≥20% truncating mutations are tumor suppressors, whereas genes with >20% of missense mutations in recurrent positions are oncogenes. While it correctly detects and classifies most of the well-known cancer genes, the rule fails to identify drivers included in newer catalogs (Tamborero *et al*., 2013b), mostly the lowly recurrent ones.

Building upon the same idea, Davoli *et al*. developed a machine learning approach to directly identify tumor suppressor genes and oncogenes from the somatic alterations observed across cohorts of tumor samples through their mutational and copy number patterns. Many cancer drivers are recognized correctly by carefully selected features (Davoli *et al*., 2013).

We recently proposed a strategy to obtain a comprehensive list of drivers minimizing the probability of detecting false-positive findings by combining complementary methods that detected different signals of positive selection (Tamborero *et al*., 2013b).

Once a list of high-confidence drivers (HCDs) is obtained, it is important to classify those in their mode of action. To this aim, we first carefully assessed the capability of 30 features to differentiate between these two groups of cancer genes. Then, we combined different sets of features with various classification algorithms to create several automated classifiers. We trained these classifiers with CGC genes, and after careful check of their performance, we selected a random forest algorithm that achieves an accuracy (ACC) of 93%, which we call OncodriveROLE. It is the first freely available automatic classifier that undertakes the task of assessing the mode of action of driver genes. Used in this setting, it may shed light upon the mechanisms of tumorigenesis in major cancer types. We have used it to classify the two previously mentioned lists of mutational drivers that have been recently published, namely, HCDs (Tamborero *et al*., 2013b) and Cancer5000 (Lawrence *et al*., 2014), and describe the results of this analysis.

2 METHODS

2.1 Mutation data, copy number alteration data and cancer driver lists

We retrieved data for the 17 TCGA (The Cancer Gene Census) projects currently available without restriction: BLCA, BRCA, COAD/READ, GBM, HNSC, KIRC, LAML, LGG, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA and UCEC. We designed and computed several features that we hypothesized might be useful to classify driver genes according to the role using mutation and copy number data. These features are based on the patterns of mutations and copy number alterations (CNAs) across tumor samples. Tumors with at least one mutation in the TCGA pan-cancer 17 dataset available at Synapse (syn1729383.2) were retrieved after excluding those considered as hypermutators (Kandoth, 2014; Kandoth *et al*., 2013). Hypermutators of a tumor type contained more than ($Q3 + 4.5 \times IQR$) somatic mutations, where Q3 and IQR are the third quartile and the interquartile range of the distribution of mutations across all samples of the tumor type, respectively. After filtering, the pan-cancer 17 dataset was composed of 4327 samples. These mutations were mapped to protein positions, and their consequence types were assessed using the IntOGen-mutations pipeline (Gonzalez-Perez *et al*., 2013b), which makes use of the Ensembl Variant Effect Predictor (v70; Chen *et al*., 2010). The CNA status for all probed genes was downloaded from the January run of the TCGA FIREHOSE pipeline at the Broad Institute (http://gdac.broadinstitute.org/).

To apply the OncodriveROLE classifier, we gathered two lists of likely cancer drivers from the Supplementary Material of two independent papers (Lawrence *et al*., 2014; Tamborero *et al*., 2013b). From the Tamborero *et al*. (2013b), we selected the list of 291 genes annotated as HCDs, discarding one non-coding gene. From Lawrence *et al*. (2013), we obtained a list of 260 genes from the spreadsheet 'Individual q-values'.

For comparison purposes, we retrieved the classifications of genes carried out by the previous work by Davoli *et al*. from the Supplementary Material of their paper, applying the same cutoffs described in the manuscript (Davoli *et al*., 2013). We also obtained the classification carried out by applying the 20/20 rule (Vogelstein *et al*., 2013) to the mutational dataset of 17 tumors types.

Whenever possible, data were obtained associated to Ensembl gene identifiers (Flicek *et al*., 2013). Other identifiers have been mapped to Ensembl gene identifiers with a dataset obtained from Ensembl v70.

2.2 Classifiers

We chose six different classifiers to test: cforest.party (cforest method in R), conditionalTree (ctree), logisticRegression (glm), naiveBayes (train), simpleTree (rpart) and randomForest (Breiman, 2001; Hothorn *et al*., 2006; Kuhn, 2008; Olshen *et al*., 1984; R Core Team, 2013). Some classifiers either do not accept missing values or perform variable imputation for those. Therefore, we opted to remove genes if they had missing values in one or more of the features and leave them unclassified. From each classifier we obtained a score of the certainty that each gene belongs to the Act class.

2.3 Training set

To use cancer genes with well-established roles as training set, we downloaded the material available at the CGC in November 2013 (Futreal *et al*., 2004). See below details on the curation of this dataset for training the classifier.

The CGC contains extensive and manually annotated information on well-known cancer genes and classifies the cancer genes into dominant (Dom) and recessive (Rec) influence on tumorigenesis. We have used the CGC classification into Rec and Dom classes as proxy for LoF and Act genes. Genes with ambiguous annotation, such as 'Rec?' or 'Dom?' or not citing observed somatic mutations were discarded, leaving 381 entries (see Supplementary Table S7 for their classification). To only include CGC driver genes, which are likely to act across the TCGA pan-cancer 17 cohort, we used a *one-signal filter*: we discarded genes not detected as significant by MutSigCV (recurrence signal), OncodriveFM (mutations impact signal) or OncodriveCLUST (mutations clustering signal). We also rejected genes with <12 protein affecting mutations (PAMs; Gonzalez-Perez and Lopez-Bigas, 2012; Lawrence *et al*., 2013; Tamborero *et al*., 2013a). Only 115 CGC genes passed this filter. Equally, all CGC genes that were solely associated to translocation events—all labeled with Dom—were not allowed in the training set, finally leaving 76 entries in the training set.

2.4 Computing features

All features we computed are listed in Table 1 along with a brief explanation of their computation: some of them are similar to the ones used previously (Davoli *et al*., 2013; Vogelstein *et al*., 2013). Truncating mutations include mutations causing a frameshift, a gained or lost stop codon as well as mutations in splice donor or acceptor sites. PAMs include truncating mutations and missense mutations. Benign missense refers to missense mutations that

are categorized as low or unknown functional impact by TransFIC (Gonzalez-Perez *et al*., 2012). OncodriveFM *P*-values (Gonzalez-Perez and Lopez-Bigas, 2012) and the location of OncodriveCLUST clusters of mutations (Tamborero *et al*., 2013a) for all driver genes were obtained by running the IntOGen-mutations pipeline on the TCGA pan-cancer 17 dataset.

The R implementation of Wilcoxon's signed rank (R Core Team, 2013) was used to compare the distribution of each feature between the CGC Rec and CGC Dom genes. We also used the variable importance function from the party library (Hothorn *et al*., 2006; Strobl *et al*., 2008) to rank features for their selection to be taken into account by the classifiers.

2.5 Training and prediction

The selected CGC genes were therefore used as training set of the classifiers. With all different classification settings, we performed a leave-one-out cross-validation: each item in the training set is classified with a model built with the rest of the training set items. We found three genes whose initial classification extremely contradicted their CGC category: NOTCH1, NPM1 and CEBPA genes, which have evidence in the literature for a dual role (Halmos *et al*., 2002; Sportoletti *et al*., 2008; Vogelstein *et al*., 2013). Therefore, we decided to discard them from the training set. Thus, the final, trimmed CGC training set included 28 Dom and 45 Rec genes.

For the classification of HCD and Cancer5000 genes, we considered that values between 0.7 and 1 as Act and those with values between 0 and 0.3 as LoF. We computed the ACC and MCC (Matthew's correlation coefficient) of each classifier at the leave-one-out cross-validation of the training set. Furthermore, we calculated the coverage (COV) of the classifier, which reflects the percentage of the entire training set for which a prediction could be made.

3 RESULTS

3.1 Identifying features that differentiate Act from LoF driver genes

We tested 30 features that we initially hypothesized could be used to characterize and discriminate between LoF and Act drivers

Table 1. List of mutational and CNA features for cancer driver genes

Attribute name	Description
CNA_cbs_countGain	# samples in cohort with CBS value > 1.1
CNA_cbs_countLoss	# samples in cohort with CBS value < 1.1
CNA_cbs_logratio_GvL	Log10-ratio of countGain VS countLoss
CNA_gain_freq	# samples in cohort with CBS value > 1.1 / cohort size
CNA_loss_freq	# samples in cohort with CBS value < 1.1 / cohort size
MUTS_clusters_miss_VS_pam	Log10-ratio of missense VS PAM within OncodriveCLUST peaks
MUTS_freq_clustered	# of mutations in OncodriveCLUST peaks / # of samples with gene mutated
MUTS_freq_disruptive	# of samples with truncating mutations or high impact missense / # of samples having gene mutations
MUTS_freq_missH	# of high impact missense mutations not in OncodriveCLUST peaks / # samples with gene mutated
MUTS_freq_missHM	# of high and medium impact missense mutations not in OncodriveCLUST peaks / # samples with gene mutated
MUTS_freq_truncating	# of samples with truncating mutations / # of samples with at least one mutation
MUTS_missense_clustercov	# missense mutations in OncodriveCLUST peaks / # missense mutations / # amino acids covered by peaks
MUTS_missense_mutrec	# recurrent missense mutations / # high and medium impact missense mutations
MUTS_missense_rec_freq	# recurrent missense mutations / # mutations (as in Vogelstein *et al*.)
MUTS_missense_recHM	# samples with high and medium impact recurrent missense mutations / # samples with missense mutations
MUTS_OncoFM_pvalue	OncodriveFM *P*-value
MUTS_pams_count	# samples with PAM
MUTS_pams_freq	# samples with PAM / # samples with gene mutations
MUTS_pams_ratio	# samples with PAM VS # samples with no PAM
MUTS_pamsrec_freq	# samples with PAM VS # of samples with gene mutation
MUTS_trunc_count	# samples with truncating mutations
MUTS_trunc_freq_cohort	# of truncating mutations / # of samples with gene mutations
MUTS_trunc_mutfreq	# truncating mutations / # mutations (as in Vogelstein *et al*.)
MUTS_trunc_vs_missbenign_ratio	# samples with truncating mutations VS # samples with benign missense mutations
MUTS_trunc_vs_missense_ratio	# samples with truncating mutations VS # samples with missense mutations
MUTS_trunc_vs_notrunc_ratio	# samples with truncating mutations VS # samples without truncating mutations
MUTS_tuson_missHM_missbenign_ratio	# samples with high and medium impact mutations VS # samples with benign missense mutations (as described in Davoli *et al*.)
MUTS_tuson_splicing_missbenign_ratio	# samples splicing variants mutations VS # samples with benign missense mutations (as described in Davoli *et al*.)
MUTS_tuson_trunc_missbenign_ratio	# samples with truncating (excluding splicing variants) mutations VS # samples with benign missense mutations (as described in Davoli *et al*.)

Note: List of features initially created for characterizing LoF and Act genes. The description reflects the formula applied for the calculation of the features. All features elaborated describe either mutation or CNA characteristics. Abbreviations used in the descriptions are: **# (number sign)**: Count/number of, **/ (slash)**: divided by, **CBS** : circular binary segmentation, **truncating mutations**: frameshift, stop gained and lost, splice donor and acceptor, **missense**: all missense mutations and insertions and deletions not altering the reading frame, **high and medium impact mutations**: all missense mutations with and TransFIC impact of 1 and 2 , **benign missense**: all missense with low or unknown TransFIC impact, **PAM** : protein affecting: frameshift, stop gained and lost, splice donor and acceptor, missense, **(gene) mutations**: all mutations-affecting coding sequence, **VS** : versus—a ratio has been obtained.

(see Table 1 for detailed description of each). All features elaborate on somatic mutation and CNA patterns across data from the pan-cancer 17 cohort. We expected LoF genes to be affected more frequently by deleterious events such as CNA loss and truncating mutations. Act genes should be more frequently amplified and receive protein-affecting non-truncating mutations, which may increase and/or alter the protein function.

To select the most informative features for the task of distinguishing between Act and LoF genes, we compared the distribution of the features in both categories of CGC genes (Fig. 1). The features we considered can be divided into four broad categories (Fig. 1A): (i) features that measure the relative abundance of truncating mutations, (ii) features that reflect the CNA status of the gene across tumors, (iii) features that account for the relative abundance of PAMs and (iv) features that measure the degree of clustering of missense mutations along the protein sequence.

Features in Group iii show the poorest performance to discriminate between CGC Dom and CGC Rec genes (light blue in Fig. 1A). On the other hand, all the features in Group i (green in Fig. 1A) rank at the top of performance of all features analyzed. As expected, this reflects that Act genes (or proto-oncogenes) are intolerant to truncating mutations because an active protein

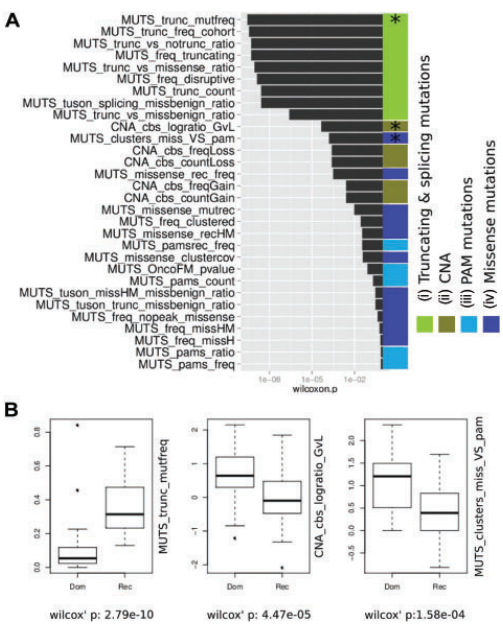

Fig. 1. **A**) The list of features ordered by Mann–Whitney–Wilcoxon rank sum test *P*-value significance. Features dependant on truncating mutations are the best discriminators for LoF and Act genes. Features described in (**B**) are marked with asterisk. A detailed explanation of each feature can be found in Table 1. (**B**) Box plots comparing the distribution of the three non-redundant top-ranking features that have been selected for the OncodriveROLE classifier in CGC genes annotated as Dom and Rec

product is required for tumorigenesis. In LoF (or tumor suppressor) genes the truncation of the protein product gene is positively selected, which facilitates the identification of LoF candidates. The best performing feature in this group was the ratio of truncating mutations to the total number of coding mutations in the protein (Fig. 1B).

The distribution of mutations within the gene (Group iv, dark blue in Fig. 1A) differs significantly between CGC Dom and CGC Rec genes. The CGC Dom genes have fewer mutational hotspots, detected as *clusters* by OncodriveCLUST, than CGC Rec genes, whose mutations tend to be more evenly distributed (Supplementary Fig. S1) along the protein sequence. This is probably because Act driver genes receive mutations that potentiate their function, e.g. by constitutively activating a regulatory site, or cause a switch of the protein function. To achieve such behavior through mutations, these must occur at specific places in the sequence, which results in fewer numbers of recurrent sites (clusters) than in CGC Rec genes (Supplementary Fig. S1). We elaborated a series of features based on impact, frequency and clustering of missense mutations. Many did not show any power of discrimination of CGC Rec and Dom. The features that perform reasonably well are based on the recurrence of missense mutations. The best-performing feature in this group compares the ratio of missense mutations with total number of PAMs within OncodriveCLUST peaks (MUTS_clusters_miss_VS_PAM; Fig. 1). Another feature in this group that performs relatively well is the ratio recurrent missense mutations (MUTS_missense_rec_freq).

All features in Group ii are designed to capture the known fact that LoF genes have a tendency to be deleted, whereas Act genes are more frequently affected by amplifications (Davoli *et al.*, 2013). In this case, we found that the ratio of amplifications to deletions across all tumors in the cohort achieved the best separation of the two groups of genes.

3.2 Developing a classifier to differentiate between LoF drivers and Act drivers

Thereafter, we created a feature set that contained non-redundant best-performing features from Groups i, ii and iv, disregarding those of Group iii because of their poor performance resulting in three features: MUTS_trunc_mutfreq, MUTS_clusters_miss_VS_PAM and CNA_cbs_logratio_GvL. We tested six machine learning approaches trained with the trimmed version of the CGC (see Section 2). For each gene, the classifiers produced a score of the likelihood that it belonged to the CGC Dom class. A score of value 0 means that the classifier regards the gene as an LoF beyond all doubt, whereas a score of value 1 means it exactly resembles the model of an Act gene. We assessed the performance of each classifier through the ACC, the MCC and the COV of the driver set (all listed in Supplementary Table S1). ACC and MCC validate the performance of the classifiers on the 76 CGC driver genes by means of a leave-one-out cross-validation approach. We computed these values for different classification probabilities thresholds to select the cutoff that maximize the ACC and MCC, even at the cost of reducing the COV. Then, we used these sets of values to choose the classifier with the best performance and a reasonable COV. Overall, *randomForest* produced the best results

(Supplementary Table S1). We also trained classifiers with different combinations of the three selected features and included MUTS_missense_rec_freq feature for testing purposes. We found that multiple combinations of these features perform similarly (Supplementary Table S2 and Supplemental Text). We decided to use the *randomForest* classifier trained with the three non-redundant features shown in Figure 1B to create OncodriveROLE, under the rationale that features representing the three independent groups could provide more information to classify novel drivers. The method shows an ACC of 0.94, MCC of 0.84 and COV of 88% in the leave-one-out cross-validation. We further tested OncodriveROLE in an independent set of tumor suppressor genes (Zhao *et al*., 2013) that are not present in the CGC. OncodriveROLE accurately classified 91.7% of those genes as LoF drivers (Supplemental Text).

3.3 Applying OncodriveROLE to lists of cancer driver genes

We identified two recent studies in which identified novel cancer driver genes could be classified with OncodriveROLE. The first study detected cancer drivers by integrating four methods that assess different signals of positive selection across samples of the pan-cancer 12 dataset. This analysis resulted in 291 high-confidence cancer drivers (Tamborero *et al*., 2013b). In the second study, MutsigCV was applied in a cohort of about 5000 tumor samples to obtain a cancer driver list composed of 260 genes (Lawrence *et al*., 2013, 2014). The two lists will be referred to as HCD and Cancer5000 further on. Even though both lists have similar sizes, their overlap is only 50%, making the two gene sets different as can be seen in Figure 2. As for the training set, we applied the one-signal filter to only predict the role of genes possibly acting as drivers in the dataset under evaluation resulting in 200 HCD and 144 Cancer5000 genes.

The overall distribution of probabilities of these two groups of genes is roughly bimodal in both driver lists, which allowed us to choose these symmetric cutoff values (Fig. 2 and Supplementary Fig. S2) such as 0.3 and 0.7 for LoF and Act genes, respectively. Other cutoffs may be used for the datasets under analysis depending on how strict a classification the user wants for their list of cancer drivers. Interestingly, we classified three CGC Dom genes as LoF ('Dom?' in Fig. 2). The genes in question are NOTCH1, NPM1 and CEBPA. All three have been implicated in leukemia (Cancer Genome Atlas Research Network, 2013; Liu *et al*., 2013; Ohlsson *et al*., 2014) and both NOTCH1 and NPM1 are annotated in the CGC as partners of translocation events in leukemia. NOTCH1 has been described as an oncogene as well as a tumor suppressor. Its actual role may depend on the tumor type (Licciulli *et al*., 2013; Liu *et al*., 2013; Vogelstein *et al*., 2013). Equally, CEBPA and NPM1 have been characterized as tumor suppressors in the literature (Halmos *et al*., 2002; Sportoletti *et al*., 2008). We cannot be certain of the functional impact of the translocation on the function of the product of the fused gene. It may associate to a new promoter and change its expression accordingly, or it may be truncated as a result of the fusion and thus function as an LoF. For this reason, we had previously excluded all CGC Dom genes that are solely associated to translocation events in the Census. The plot in Figure 2 shows those genes labeled as DomT, and their classification shows no clear resemblance to LoF or Act, which supports our decision to remove them from the training set.

3.4 Comparison of OncodriveROLE with other bioinformatics approaches

The 20-20 rule was created to identify mutational driver genes, both oncogenes and tumor suppressor genes (Vogelstein *et al*., 2013). Therefore, it differs from OncodriveROLE, designed to classify previously identified driver genes into their most probable roles. The simple 20-20 rule reaches a high ACC (Table 2) when applied to the trimmed CGC list. However, it is unable to reach a decision on many drivers where none of its two estimators (see Section 2) surpasses the threshold of 20% (Tables 2 and 3).

We also compared the results obtained by the approach designed by Davoli *et al*. (2013), implemented in a classifier named Tuson. As with the 20-20 rule, Tuson was created to distinguish oncogenes and tumor suppressor genes from genes with passenger mutations, instead of classifying previously identified cancer drivers as is the case of OncodriveROLE. We found OncodriveROLE slightly outperforms Tuson in ACC and MCC on the trimmed CGC dataset. Note that Tuson method was trained with CGC genes, and the performance reported in Table 2 does not remove genes in the training set, as it is done in the leave-one-out cross-validation of OncodriveROLE. We can conclude that well-known cancer genes are classified with a high

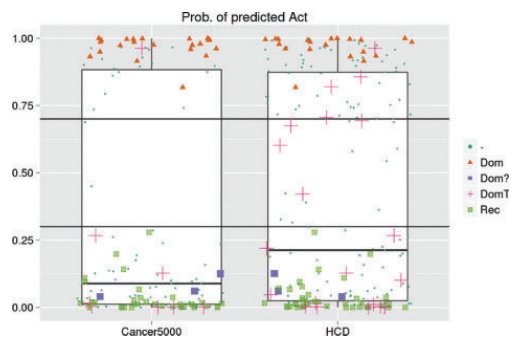

Fig. 2. Classification of 200 (HCD list) and 144 (Cancer5000 list) cancer driver genes into the classes Act and LoF. The training set of OncodriveROLE constitutes of all 'Dom' and 'Rec' labeled data points. 'Dom?' are CGC-annotated dominant genes excluded from the training set because of strong resemblance to the 'Rec' genes and previous literature evidence of this role. 'DomT' genes are CGC-annotated dominant genes only citing translocation events as prove and therefore not included in the training set. All '-' labeled data points are driver genes not annotated in CGC, and whose prediction was the main goal of the study. The thresholds are drawn at 0.3 (as top limit of the LoF class) and 0.7 (as bottom limit of the Act class). Working with classification score thresholds of 0.3 (as top limit of the LoF class) and 0.7 (as bottom limit of the Act class), we classified 109 genes as LoF, 76 as Activating and left 15 genes as unclassified in the HCD list; meanwhile, we classified 97 genes as LoF, 43 as Activating and left 4 genes as unclassified (Fig. 2) in the Cancer5000 list. Genes for which we have observed <12 mutations were directly classified as 'No class' and assigned NA values in the classifications results (see Supplementary Tables S4 and S6)

Table 2. List of approaches and their performance on trimmed CGC dataset

Method	ACC	MCC	COV (%)
OncodriveROLE[a]	0.925	0.848	83
20-20 rule	0.895	0.769	75
Tuson	0.914	0.817	92

[a]Results of leave-one-out cross-validation.

Table 3. List of approaches and their performance on the 290 drivers from the HCD list and 260 drivers from the Cancer5000 list

Method	Act/ Oncogene	LoF/ Tumour suppressor	Unclassified	Coverage (%)
HCD				
Oncodrive ROLE 0.3/0.7	76	109	15	92
Oncodrive ROLE 0.2/0.8	58	96	46	77
20-20 rule	23	96	81	60
Tuson	44	92	64	68
Cancer5000				
Oncodrive ROLE 0.3/0.7	43	97	4	97
Oncodrive ROLE 0.2/0.8	40	91	13	91
20-20 rule	18	90	36	75
Tuson	32	90	22	85

ACC with all approaches. The main difference between the three approaches lies in the COV that can be reached when predicting the role of novel cancer drivers in tumorigenesis.

4 DISCUSSION

Two main rationales to detect LoF and Act driver genes acting across tumor samples exist. The first approach consists in directly detecting genes that exhibit known alterations patterns corresponding to these two classes from mutations and CNA data. This strategy was first conceptualized by Vogelstein *et al*. (2013) to be implemented later on as a machine learning algorithm by Davoli *et al*. (2013). In the second approach, first driver genes acting in tumor samples are detected by combining the signals of positive selection they exhibit (Lawrence *et al*., 2014; Tamborero *et al*., 2013b). Then, in a second step, these drivers are classified into the two aforementioned classes exploiting similar alteration patterns as in the first approach. This second two-step approach has two main advantages. First, genes that do not exhibit clear alterations pattern that define them as LoF or Act can still be detected as drivers if they show clear signals of positive selection. Second, the combination of several signals controls the ratio of false-positive drivers identified (Tamborero *et al*., 2013b), which is unattainable to the direct classification of genes.

This is the reason why we have decided to develop OncodriveROLE, a machine learning classifier, which takes a list of pre-selected driver genes and sorts them according to their mode of action. We first carefully compared and selected a set of features that best captures the differences of alterations patterns of these two groups of drivers. We then used those features to train the classifier, on a carefully trimmed subset of the CGC genes. When applied to two recent lists of drivers, we found that, even under strict classification conditions, OncodriveROLE was able to classify more drivers than the 20-20 rule and the Tuson machine learning algorithm.

The OncodriveROLE validation procedure identified several likely misclassified drivers in the CGC. The most salient examples of these are probably some genes that drive hematopoietic malignancies upon translocation and fusion with other genomic regions, all classified as Dom in the GCG. However, when analyzed using mutational and CNAs data from the pan-cancer 17 dataset, some of them appear as clear LoF drivers. For instance, OncodriveROLE assigns MLL, RUNX1 and SUZ12 classification probabilities under 0.003 (see Supplementary Tables S3–S6 for feature and classification values). These genes could be Act drivers upon fusion to other genes, but LoF upon mutations.

Even though OncodriveROLE is able to classify most of the genes in the two drivers lists as LoF or Act, it still leaves few of them unclassified. Some of these correspond to lowly recurrent drivers whose mutational features are not correctly computed because of the scarcity of their alterations. Sequencing more tumors will certainly improve their classification. Others may not have a clear enough pattern to be classified in one of the two classes, as they could be exhibiting different roles in different contexts. In some rare cases, the method misclassifies known cancer genes. For example, KEAP1 is classified as an Act driver, although it is reported to lose its function upon mutation (Hayes and McMahon, 2009; Shibata *et al*., 2008). A close look at its mutational pattern reveals missense mutations dominate and accumulate in certain regions of the protein. As member of a ubiquitin-mediated proteolysis complex, the function of KEAP1 is probably essential to the cell, and its impairment is likely lethal. Therefore, few truncating mutations may appear in KEAP1, and it is ultimately misclassified by OncodriveROLE. Future finer measurements of the impact of missense mutations may help correcting this problem.

Summing up, in this article, we have described the development and validation of OncodriveROLE, an approach to differentiate between LoF and Act driver genes. The OncodriveROLE classifier is freely available at http://bg.upf.edu/oncodrive-role as an R object that researchers may use to classify the drivers they have detected across a cohort of tumor samples. At the same URL, the pre-computed TCGA pan-cancer 17 mutational and copy number features used for the classification are available for download.

Funding: We acknowledge funding from the Spanish Ministry of Economy and Competitivity (grant number SAF2012-36199) and the Spanish National Institute of Bioinformatics (INB). M.P.S. and C.R.-P. are supported by FPI fellowships.

Conflict of Interest: none declared.

REFERENCES

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Cancer Genome Atlas Research Network. (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.*, **368**, 2059–2074.

Chen,Y. *et al.* (2010) Ensembl variation resources. *BMC Genomics*, **11**, 293.

Davoli,T. *et al.* (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, **155**, 948–962.

Dees,N.D. *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.

Flicek,P. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.

Futreal,P.A. *et al.* (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.

Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res.*, **40**, e169.

Gonzalez-Perez,A. *et al.* (2012) Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med.*, **4**, 89.

Gonzalez-Perez,A. *et al.* (2013a) Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods*, **10**, 723–729.

Gonzalez-Perez,A. *et al.* (2013b) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.

Halmos,B. *et al.* (2002) Down-regulation and antiproliferative role of C/EBPα in lung cancer. *Cancer Res.*, **62**, 528–534.

Hayes,J.D. and McMahon,M. (2009) NRF2 and KEAP1 mutations: permanent activation of an adaptive response in cancer. *Trends Biochem. Sci.*, **34**, 176–188.

Hothorn,T. *et al.* (2006) Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.*, **15**, 651–674.

Hudson,T.J. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Kandoth,C. (2014) MAF files - strictly filtered. http://dx.doi.org/10.7303/syn1729383.2.

Kandoth,C. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.

Kuhn,M. (2008) Building predictive models in R using the caret package. *J. Stat. Softw.*, **28**, 1–26.

Lambert,J.M.R. *et al.* (2009) PRIMA-1 reactivates mutant p53 by covalent binding to the core domain. *Cancer Cell*, **15**, 376–388.

Lawrence,M.S. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.

Lawrence,M.S. *et al.* (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.

Licciulli,S. *et al.* (2013) Notch1 is required for Kras-induced lung adenocarcinoma and controls tumor cell survival via p53. *Cancer Res.*, **73**, 5974–5984.

Liu,N. *et al.* (2013) The emerging roles of Notch signaling in leukemia and stem cells. *Biomark. Res.*, **1**, 23.

Ohlsson,E. *et al.* (2014) Initiation of MLL-rearranged AML is dependent on C/EBPα. *J. Exp. Med.*, **211**, 5–13.

Olshen,L.B. *et al.* (1984) *Classification and Regression Trees*. Wadsworth Int. Group. CHAPMAN & HALL/CRC.

R Core Team. (2013) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Reimand,J. *et al.* (2013) The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.*, **3**, 2651.

Shibata,T. *et al.* (2008) Cancer related mutations in NRF2 impair its recognition by Keap1-Cul3 E3 ligase and promote malignancy. *Proc. Natl Acad. Sci. USA*, **105**, 13568–13573.

Sportoletti,P. *et al.* (2008) Npm1 is a haploinsufficient suppressor of myeloid and lymphoid malignancies in the mouse. *Blood*, **111**, 3859–3862.

Strobl,C. *et al.* (2008) Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307.

Tamborero,D. *et al.* (2013a) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–2244.

Tamborero,D. *et al.* (2013b) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.*, **3**.

Vogelstein,B. *et al.* (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.

Zhao,M. *et al.* (2013) TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res.*, **41**, D970–D976.

3.2 Assessing statistical significance of mutual exclusive patterns amongst cancer driver alterations

In this chapter I present a method which allows to assess the significance of mutual exclusivity in the alteration pattern of several cancer drivers. The implementation of the method to assess the likelihood of the occurrence of somatic alterations in a mutually exclusive manner between samples was required for a manuscript that is currently under revision. Our main interest was that the method would respect the mutational burden for samples and genes. The development and testing of the method has been performed in collaboration with Abel Gonzalez-Perez who has collected the mutational data and classified the driver genes in the cancer modules within which we searched for mutual exclusivity patterns.

Introduction

The selective pressure acting upon certain pathways and functions of the cells that are involved in the tumorigenic process is expected to leave behind a traceable pattern of mutually exclusive alteration between samples (Greaves and Maley 2012). The underlying assumption is that a pathway that is needed to behave differently within a cancer cell which may be achieved targeting genes of the pathway. And once a pathway or module is altered, an alteration of an equivalent target in the same pathway does not increase further the fitness of the tumor and is therefore not selected for. Once one of the possible targets has been altered and the pathway behaves as required for the tumorigenesis, the selective pressure ceases to act upon this pathway and a mutation of a further target becomes more unlikely. Thus, upon screening a cancer sample cohort, it is expected that cancer driver mutations occur in a mutually exclusive manner within pathways.

The somatic mutation datasets available nowadays provide a unique chance to test pathways for mutual exclusive alteration patterns for a big cohort. Mutual exclusive alteration patterns have been reported for Glioblastoma, ovarian and lung cancer in 2011 in approaches to detect *de novo* cancer drivers via their mutual exclusive patterns (Ciriello et al. 2011; Vandin, Upfal, and Raphael 2011). Since then the available alteration data has grown considerably and the detection of cancer mutational drivers has improved substantially (Tamborero et al. 2013), which allows us to work with consolidated cancer drivers. A very important feature to take into account is the heterogenic nature of cancer samples and genes. The alteration burden is not equal for all the samples and genes which has implications for the measure of mutual exclusivity: Gene X and Y both are more likely to be in altered state in a sample group of so-called hypermutators – cancer samples with many alterations/mutations. The probability of both genes being altered at the same time is therefore elevated and if the case, does not necessarily reflect a selection process but be a result of stochastic (passenger) alterations due to genomic instability.

We therefore propose the classification of the cancer drivers with methods that test for traces of positive selection within cancer mutations and classify the drivers in tumorigenic modules that understood as groups of functionally related genes that when altered produce the same cancer phenotype. The method serves to test the integrity of the modules by a hypothesis-driven approach for testing mutual exclusivity and takes into account the observed alteration burden of samples and genes. Note that the test does not identify *de novo* cancer drivers. The detection of mutually exclusive driver alterations helps on the one hand to understand tumorigenesis in different cancer types and discover possible homogeneity underlying the heterogeneity of tumor samples and on the other to discover routes to indirectly target driver alterations with anti-cancer drugs.

Material and Methods

Identification of mutational drivers: We obtained mutation data for 27 cancer types from TCGA, ICGC and independent studies. We then applied the IntOGen pipeline (Gonzalez-Perez, Perez-Llamas, et al. 2013) in order to obtain a list of mutational cancer drivers for each cancer type and as well the complete cancer samples cohort. The detection of cancer drivers was performed by combining several signals of positive selection as explained in (Tamborero et al. 2013) and (Rubio-Perez et al.). All data sources are listed in Table 2 of this chapter along with the abbreviations of the tumor types.

Modules: The detected mutational driver genes were classified into 41 biological modules (gene sets), which were created based on literature of the gene in question. Additionally we annotated each gene with known implication in one of the hallmarks of cancer. A gene may be included in multiple modules which are listed in Table 1 of this chapter.

Computation: To test the significance of an observed pattern of mutual exclusivity, we compared its signal (the total number of samples where at least one gene in the module bears an alteration) to that of $1x10^5$ randomly generated mutational patterns, respecting the number of alterations observed in each gene and the mutational burden of each sample, following the rationale of the CDCOCA

method (N. Kumar et al. 2011).

See Figure 1, a step by step illustration of the MutEx algorithm.

1. We calculate sample alteration weights by summing up the positive events, in our case mutations, for each sample and divide it by the total alteration count. (Figure 1, step 1).
2. For the gene set in question the total signal (all positive events within the binary matrix) and the coverage (number of samples with at least one mutation) are calculated for later reference (Figure 1, step 2).
3. Then, permutations were performed by a random generation of altered samples per gene in which the number of altered samples per gene is maintained as observed and the overall alteration burden per sample was preserved by using the sample alteration weights as alteration probability in each of the samples (Figure 1, step 3). For each permutation we calculate the coverage.
4. With the array of coverage values from the permutation we can calculate the empirical p-value and Z-score. Formulas are listed in Figure 1, step 4.

MutEx has been implemented and run using the R environment (R Core Team 2013). Results were only computed for modules that have at least 2 driver genes within the cancer type in question. The method has also been implemented for Gitools, see details in Chapter 3.3.

Results

We collected a cancer tumor cohort of almost 7000 samples and identified mutational drivers following the rational described in (Rubio-Perez et al.; Tamborero et al. 2013) for all cancer cohorts of each cancer type and a pan-cancer cohort in order to obtain mutational driver lists. All data sources are listed in Table 1.

Cancer driver genes were manually mapped into functional modules and sub-modules for each cancer type, including both well-known cancer drivers as well as novel cancer genes which are func-

tionally connected through pathways (Minoru Kanehisa and Goto 2000; Minoru Kanehisa et al. 2014). In order to assess the likelihood of each of the modules to function as a driver module in a given cancer type we implemented the MutEx method for the R programming environment and applied it to each of the 27 cancer sample cohort and the pool of all TCGA samples, designated by PAN. Figure 2 shows the Z-scores of each module in for each cancer type and module. The colors of the heatmap cells designate the tendency to mutual exclusivity of the mutations within the module whereas the number within the cells shows the proportion of samples that are mutated within the module.

Many of the non-significant results can be explained with the fact the corresponding modules have very few mutations recorded in that module whereas the cohort of all samples together, the PAN column, gives more statistical power to the driver detection on one hand and the MutEx calculations on the other. This effect is visible in the *Apoptosis*, *HDAC and targets* and *Cadherin prod control* modules. The PAN results let us conclude that many modules indeed may constitute units of the regulatory network that are targeted as a pathway by tumorigenesis. But a significant mutually exclusive pattern within the PAN cohort, does not mean that same selective pressure acts upon a specific cancer type. E.g. the *MAPK-JNK Stress Resp* or the *PI3K-PI3K activation* modules show tendency towards mutual exclusion in some cancer types whereas in others clearly not in spite of a high proportion of samples mutated within the module (See Figure 2 and 3). A possible explanation is that the canonical units that are targeted by the tumorigenesis depend on the tissue of origin and cancer type.

Discussion

As opposed to approaches which use a combination of some prior knowledge and unsupervised combinations (Ciriello, Cerami, et al. 2013; Vandin, Upfal, and Raphael 2011) we chose an entirely hypothesis driven approach for testing the mutual exclusivity of mutational events. This choice is due to our goal which was not to detect driver genes *de novo* but rather understand which cancer drivers

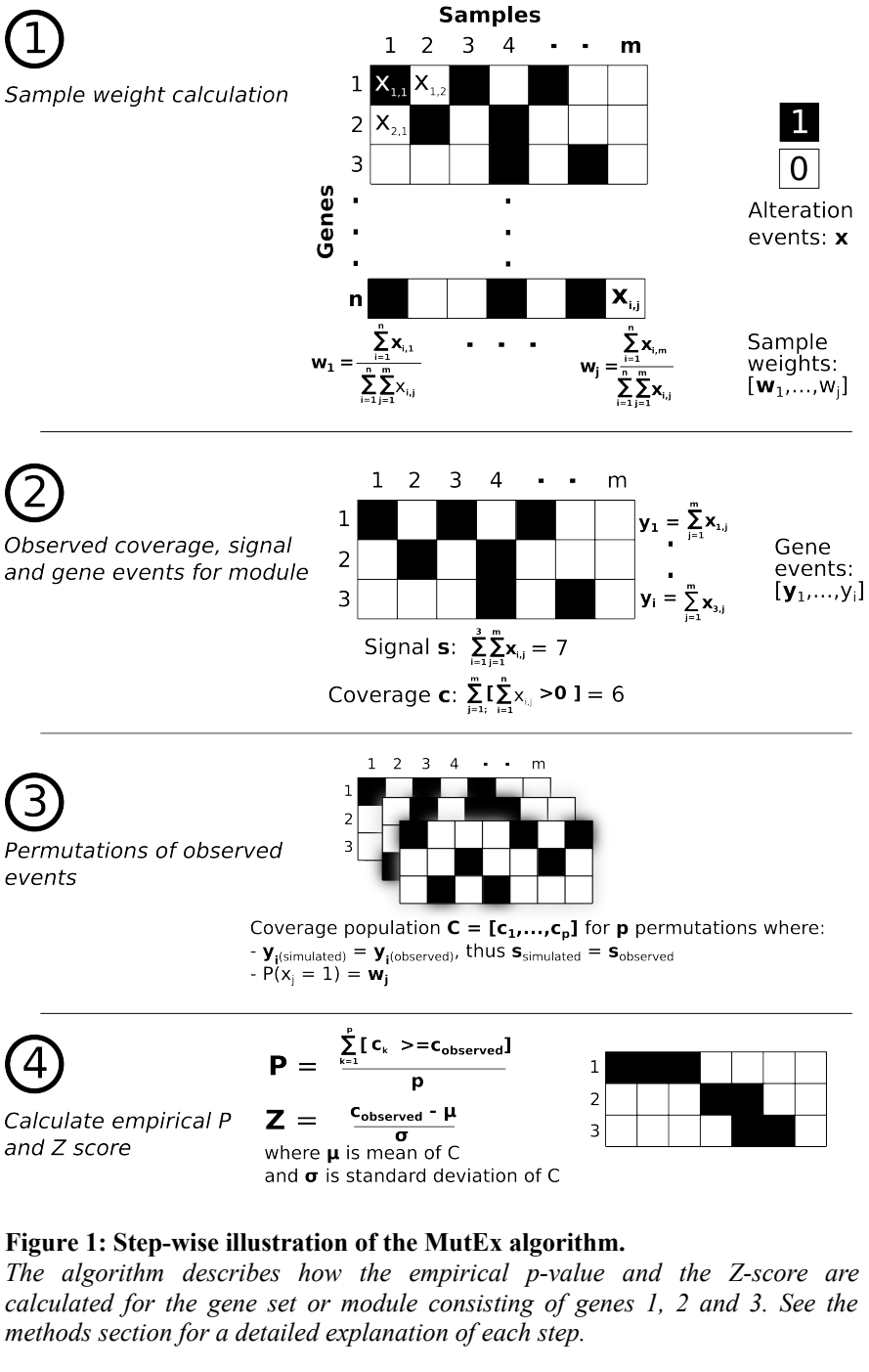

Figure 1: Step-wise illustration of the MutEx algorithm.
The algorithm describes how the empirical p-value and the Z-score are calculated for the gene set or module consisting of genes 1, 2 and 3. See the methods section for a detailed explanation of each step.

may act coordinately in tumorigenesis. Thus, before applying the MutEx algorithm we needed to detect the driver genes and map them into modules according to their known functions. This approach lets us perform the MutEx testing in a computation-efficient manner as we consider only a subset of all possible combinations of genes that bear driver mutations. E.g. the MEMo approach is computationally expensive as no limitation to what functional units to examine is supplied and many possible combinations are explored. A disadvantage of this very property is that we are limited to the available knowledge of interactions, which may be incomplete.

Another important property of MutEx is that the mutational burden of a cancer sample is taken into account, as opposed to conventional statistical tests of distribution imbalances such as the Fisher's exact test or the solution proposed by (Yeang, McCormick, and Levine 2008). The MutEx algorithm has also been implemented in the Java language as a part of the Gitools application. We have thus brought together the visual interactive exploration of the pattern of alterations across tumor samples of genes in a module and its statistical analysis (See chapter 3.3).

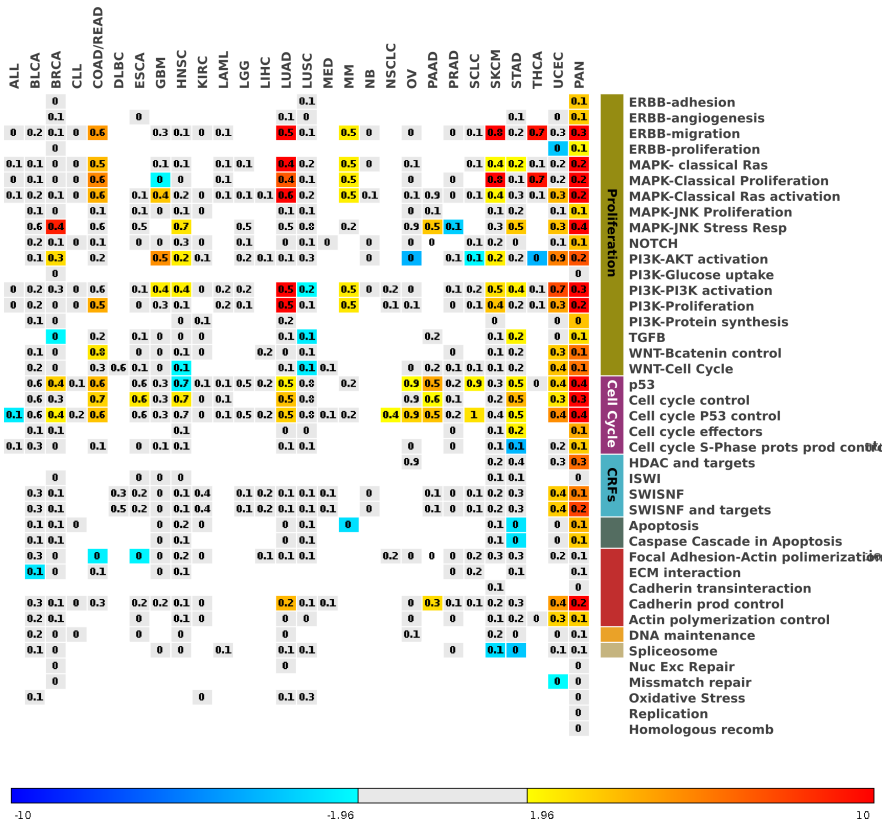

Figure 2: Z-scores for mutual exclusion of mutations within all 41 modules and 27+1 sample cohorts.

Cancer types are listed in columns whereas modules are shown in rows. Modules that show a clear bias towards having a mutually exclusive mutational pattern are colored in dark red in the respective cancer tissue. Grey cells show insignificant results and white cells represent module-cancer type combinations for which not enough mutations were recorded in order to apply MutEx. Blue-shifted colors are modules that have tend towards overlapping mutations amongst the genes. The numbers in the cells designate the proportion of samples that is mutated within the module.

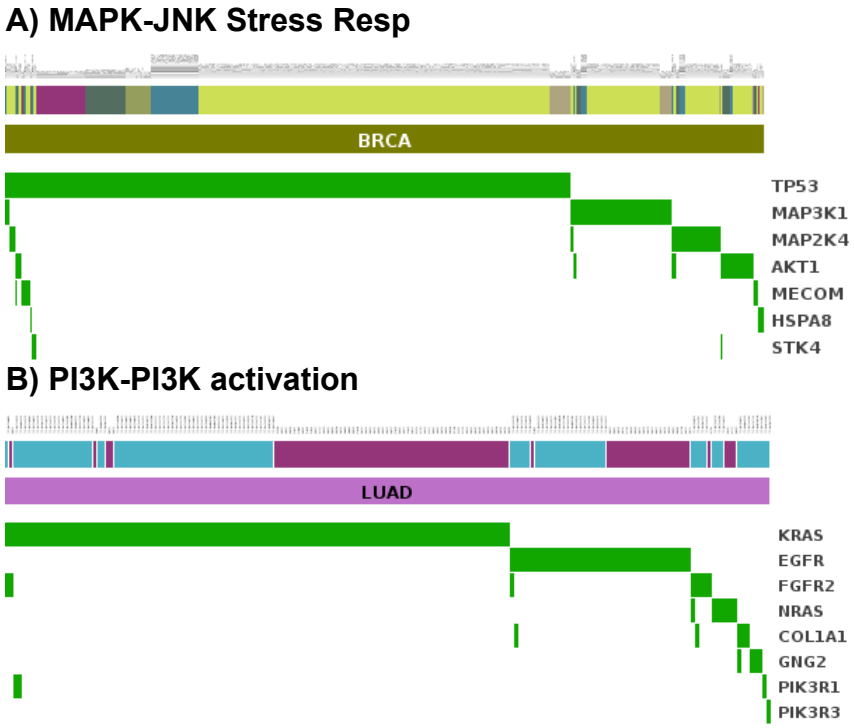

Figure 3: Modules with significant trend towards mutual exclusivity
A and B display the mutational signal for two distinct modules that with high Z-scores (> 9). Genes are aligned in rows and the samples as columns. The column headers show the tissue code and color the project sources in different colors ***A*** *shows the MAPK-JNK Stress Response module for the 510 out of 1148 breast carcinoma (BRCA) samples covered by 540 mutations in the gene set.* ***B*** *displays the data for the PI3K-PI3K activation module in which 183 out of 388 lung adenocarcinoma (LUAD) samples bear 192 mutations. Both modules belong to the proliferation hallmark and the respective modules are do not a significant tendency of mutual exclusive mutational patterns in the other tumor type.*

Table 1: The classification of cancer drivers into modules

This table lists all the modules that have been created in order to test the mutational distribution within their member genes for mutual exclusivity.

Module	Hallmark	Genes
Cell cycle P53 control	Cell Cycle	CDKN2A, MDM2, EP300, ATM, ATR, SF3B1, U2AF1, TP53, RB1, CHEK2, DDX3X
p53	Cell Cycle	ATM, ATR, CDKN1A, CDKN1B, CDKN2A, CHD8, CHEK2, CSNK1E, CSNK1E, PPM1D, TP53, TP53BP1
Cell cycle control	Cell Cycle	TGFBR2, SMAD2, SMAD4, CDKN1A, CUL1, CDKN1B, CDKN1C, TP53
Cell cycle S-Phase prots prod control	Cell Cycle	CCND1, CUL1, TFDP1, TFDP2, RB1, CUL3, FBXW7, UBR5
Cell cycle effectors	Cell Cycle	CDH1, APC, RAD21, RAD23, STAG1, STAG3
Apoptosis	Apoptosis	APAF1, CASP1, CASP8, MAP3K1, SPTAN1, VIM, AKT1, DDX3X
Caspase Cascade in Apoptosis	Apoptosis	APAF1, CASP1, CASP8, MAP3K1, SPTAN1, VIM, AKT1
Actin polymerization control	Cell Adhesion	CDH1, CTNND1, PTPRF, CTNNB1, CSNK2A1, RAC1, TJP1, ACTB, ACTG1
Cadherin transinteraction	Cell Adhesion	RAC1, WASF3,
Cadherin prod control	Cell Adhesion	CDH1, EGFR, ERBB2, FGFR1, FGFR2, TGFBR2, MAPK1, SMAD2, SMAD3, SMAD4, CREBBP, EP300, CTNNB1, TCF7L2
ECM interaction	Cell Adhesion	CD36, FN1, LAMA2, LAMB1
ERBB-proliferation	Proliferation	EGFR, ERBB2, PLCG1
ERBB-adhesion	Proliferation	EGFR, ERBB2, ABL2
ERBB-angiogenesis	Proliferation	EGFR, ERBB2, NCK1, MAP2K4
ERBB-migration	Proliferation	EGFR, ERBB2, ERBB3, SOS1, SOS2, HRAS, KRAS, NRAS, BRAF, MAPK1, MAP2K1, MYC
MAPK-Classical Ras activation	Proliferation	ALK, ACVR1B, ACVR2A, EGFR, FGFR1, FGFR2, NTRK2, SOS1, SOS2, HRAS, KRAS, NRAS, RASGRP1, NF1, NF2, RASA1, RASA2
MAPK- classical Ras	Proliferation	SOS1, SOS2, RASGRP1, NF1, NF2, HRAS, KRAS, NRAS, RASA1, RASA2
MAPK-Classical Proliferation	Proliferation	HRAS, KRAS, NRAS, BRAF, MAPK1, MAP2K1, MYC
MAPK-JNK Stress Resp	Proliferation	RAC1, STK4, AKT1, MAP4K1, MAP3K11, MAP3K1, MAP2K4, MECOM, HSPA8, TP53
MAPK-JNK Proliferation	Proliferation	ALK, ACVR1B, ACVR2A, FAS, TGFBR2, STK4, AKT1, MAP3K4, TAOK1, MAX, MEF2C

Module	Hallmark	Genes
PI3K-Protein synthesis	Proliferation	STK11, AKT1, TSC1, RHEB, MTOR
PI3K-Glucose uptake	Proliferation	AKT1, PRKCZ
PI3K-Proliferation	Proliferation	HGF, EGFR, EPHA2, FGFR1, FGFR2, FGFR3, KIT, MET, SOS1, SOS2, HRAS, KRAS, NRAS, MAP2K1, MAPK1, BRCA1
PI3K-PI3K activation	Proliferation	GNG2, HGF, EGFR, EPHA2, FGFR1, FGFR2, FGFR3, KIT, MET, RAC1, SYK, PIK3CA, PIK3CB, PIK3R1, PIK3R3, COL1A1, LAMA2, LAMB1, ITGA9, HRAS, KRAS, NRAS
PI3K-AKT activation	Proliferation	PIK3CA, PIK3R1, PIK3CB, PIK3R3, PTEN, PPP2R1A, PPP2R5C, HSP90AA1, HSP90AA2
Focal Adhesion-Actin polimerization	Cell Adhesion	COL1A1, LAMA2, LAMB1, RHOA, ROCK2, ACTB, ACTG1, ACTG2, ARFGAP1, ARFGAP3, ARFGEF1, ARHGAP29, ARHGAP35, ARHGEF2, ARHGEF6, ITGA9, MYH10, MYH14
ISWI	CRFs	BAP1, BPTF, RBBP7
NOTCH	Proliferation	ADAM10, CREBBP, CUL1, DTX2, FBXW7, MFNG, NCOR1, NCOR2, NOTCH1
Oxidative Stress		KEAP1, NFE2L2
TGFB	Proliferation	ACVR1B, ACVR2A, CUL1, NEDD4L, SMAD2, SMAD4, SMURF2
WNT-Bcatenin control	Proliferation	CSNK1E, CSNK2A1, AXIN1, AXIN2, APC, CTNNB1, CUL1, TBL1XR1, CHD8
WNT-Cell Cycle	Proliferation	CTNNB1, CHD8, RUVBL1, CREBBP, EP300, TCF7L2, MYC, CCND1, SMAD2, SMAD4, SOX17
Spliceosome	SF	AQR, CRNKL1, EFTUD2, HSPA8, PRPF8, SF3B1, U2AF1
SWISNF	CRFs	ARID1A, ARID1B, ARID2, ARID4A, ARID4B, ARID5B, PBRM1, SMARCA1, SMARCA4
DNA repair	DNA damage	RAD23B, ERCC2, RFC4
DNA repair	DNA damage	MLH1, MLH3, RFC4
DNA repair	DNA damage	BRCA2, BLM
DNA repair	DNA damage	MCM3, RFC4
DNA maintenance	DNA damage	RAD23B, ERCC2, RFC4, BRCA1, BRCA2, BLM, MCM3, FANCI, FANCM, POT1

Module	Hallmark	Genes
SWISNF and targets	CRFs	MYC, CDH1, TGFBR, ARID1A, ARID1B, ARID2, ARID4A, ARID4B, ARID5B, PBRM1, SMARCA1, SMARCA4
HDAC and targets	CRFs	HDAC8, HDAC3, HDAC9, TP53

Table 2: Projects of data sources for mutational data

This table lists all the sources from where we have obtained mutational data that has been used in this analysis. The data has been downloaded from either ICGC (Zhang et al. 2011), Synapse repository (Omberg et al. 2013) of the TCGA pan-cancer project (The Cancer Genome Atlas Research Network et al. 2013; syn300013, syn1710431) from (Alexandrov et al. 2013) or supplementary material (SM) of the respective published work.

Tumor project	Tumor type (short)	Tumor type	Project source	Downloaded	Sample size	Source (doi)
ALL_LEU	ALL	Acute lymphocytic leukemia	Leuven University	Alexandrov	29	10.1038/ng.2508
ALL_MEM	ALL	Acute lymphocytic leukemia	Memphis Hospital	Alexandrov	38	10.1038/ng.2532
ALL_WTSI	ALL	Acute lymphocytic leukemia	Sanger Institute	Alexandrov	55	10.1038/nature12477
BLCA_TCGA	BLCA	Bladder carcinoma	TCGA	Synaptic	98	10.1038/nature12965
BRCA_JH	BRCA	Breast carcinoma	Johns Hopkins	ICGC	42	10.1126/science.1145720
BRCA_WTSI	BRCA	Breast carcinoma	Sanger Institute	ICGC	100	10.1038/nature11017
BRCA_BROAD	BRCA	Breast carcinoma	Broad Institute	SM	103	10.1038/nature11154
BRCA_TCGA	BRCA	Breast carcinoma	TCGA	Synaptic	762	10.1038/nature11412
BRCA_WU	BRCA	ER+ Breast carcinoma	Washington University	SM	76	10.1038/nature11143
BRCA_BCU	BRCA	TN Breast carcinoma	British Columbia University	SM	65	10.1038/nature10933
CESC_TCGA	CESC	Cervical squamous carcinoma	TCGA	Synaptic	39	
CLL_ICGC	CLL	Chronic lymphocytic leukemia	MICINN ES	Alexandrov	303	10.1038/nature10113
CLL_DFCI	CLL	Chronic lymphocytic leukemia	Dana Farber Cancer Institute	SM	159	10.1016/j.cell.2013.01.019
CO_JH	COADREAD	Colorectal adenocarcinoma	Johns Hopkins	ICGC	36	10.1126/science.1145720
COADREAD_TCGA	COADREAD	Colorectal adenocarcinoma	TCGA	Synaptic	193	10.1038/nature11252
DLBCL_WTSI	DLBCL	Diffuse B cell lymphoma	Sanger Institute	Alexandrov	23	10.1038/nature12477
ESCA_DFCI	ESCA	Esophageal carcinoma	Dana Farber Cancer Institute	Alexandrov	146	10.1038_ng.2591
GBM_JH	GBM	Glioblastoma multiforme	Johns Hopkins	ICGC	89	10.1126/science.1164382
GBM_TCGA	GBM	Glioblastoma multiforme	TCGA	Synaptic	290	10.1038/nature07385
HNSC_BROAD	HNSC	Head and neck squamous cell carcinoma	Broad Institute	SM	74	10.1126/science.1208130

Tumor project	Tumor type (short)	Tumor type	Project source	Downloaded	Sample size	Source (doi)
HNSC_TCGA	HNSC	Head and neck squamous cell carcinoma	TCGA	Synaptic	301	
KIRC_TCGA	KIRC	Kidney renal clear cell carcinoma	TCGA	Synaptic	417	10.1038/nature12222
KIRP_TCGA	KIRP	Kidney renal papillary carcinoma	TCGA	Synaptic	100	
LAML_TCGA	LAML	Acute myeloid leukemia	TCGA	Synaptic	196	10.1056/NEJMoa1301689
LGG_TCGA	LGG	Lower grade glioma	TCGA	Synaptic	169	
LIHC_ICGC	LIHC	Liver hepatocarcinoma	ICGC	Alexandrov	66	10.1038/nature12477
LIHC_IACR	LIHC	Liver hepatocarcinoma	International Agency for Cancer Research	ICGC	24	10.1038/ng.2256
LUAD_WU	LUAD	Lung adenocarcinoma	Washington University	SM	163	10.1038/nature07423
LUAD_TCGA	LUAD	Lung adenocarcinoma	TCGA	Synaptic	228	
LUSC_TCGA	LUSC	Lung squamous cell carcinoma	TCGA	Synaptic	174	10.1038/nature11404
MB_WTSI	MB	Medulloblastoma	Sanger Institute	Alexandrov	100	10.1038/nature12477
MB_DKFZ	MB	Medulloblastoma	German Cancer Research Centre	ICGC	113	10.1016/j.cell.2011.12.013
MM_WTSI	MM	Multiple myeloma	Sanger Institute	Alexandrov	69	10.1038/nature12477
NB_BROAD	NB	Neuroblastoma	Broad Institute	Alexandrov	210	10.1038/ng.2529
NSCLC_WMC	NSCLC	Non small cell lung carcinoma	Wisconsin medical college	SM	31	10.1093/carcin/bgs148
OV_TCGA	OV	Serous ovarian adenocarcinoma	TCGA	Synaptic	316	10.1038/nature10166
PA_WTSI	PA	Pylocytic astrocytoma	Sanger Institute	Alexandrov	63	10.1038/nature12477
PA_SJ	PA	Pylocytic astrocytoma	St. Jude Hospital	Alexandrov	38	10.1038/ng.2611

Tumor project	Tumor type (short)	Tumor type	Project source	Downloaded	Sample size	Source (doi)
PAAD_JH	PAAD	Pancreas adenocarcinoma	Johns Hopkins	ICGC	114	10.1126/science.1164368
PAAD_OICR	PAAD	Pancreas adenocarcinoma	Ontario Institute for Cancer Research	ICGC	33	
PAAD_QCMG	PAAD	Pancreas adenocarcinoma	Queensland Centre for Genome Medicine	ICGC	67	10.1097/PAT.0b013e3283445e3a
PAAD_TCGA	PAAD	Pancreas adenocarcinoma	TCGA	Synaptic	34	
PRAD_TCGA	PRAD	Prostate adenocarcinoma	TCGA	Synaptic	243	
SCLC_JH	SCLC	Small cell lung carcinoma	Johns Hopkins	SM	42	10.1038/ng.2405
SCLC_UC	SCLC	Small cell lung carcinoma	University of Cologne	SM	27	10.1038/ng.2396
SKCM_BROAD	SKCM	Skin cutaneous melanoma	Broad Institute	Alexandrov	120	10.1016/j.cell.2012.06.024; 10.1038/nature11071
SKCM_TCGA	SKCM	Skin cutaneous melanoma	TCGA	Synaptic	249	
STAD_PFS	STAD	Stomach adenocarcinoma	Pfizer	SM	22	10.1038/ng.982
STAD_TCGA	STAD	Stomach adenocarcinoma	TCGA	Synaptic	139	
THCA_TCGA	THCA	Thyroid carcinoma	TCGA	Synaptic	322	
UCEC_TCGA	UCEC	Uterine corpus endometrioid carcinoma	TCGA	Synaptic	230	

References

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. "Signatures of Mutational Processes in Human Cancer." Nature 500 (7463): 415–21. doi:10.1038/nature12477.

Ciriello, Giovanni, Ethan Cerami, Bulent Arman Aksoy, Chris Sander, and Nikolaus Schultz. 2013. "Using MEMo to Discover Mutual Exclusivity Modules in Cancer." Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis... [et Al.] Chapter 8 (March): Unit 8.17. doi:10.1002/0471250953.bi0817s41.

Ciriello, Giovanni, Ethan G Cerami, Chris Sander, and Nikolaus Schultz. 2011. "Mutual Exclusivity Analysis Identifies Oncogenic Network Modules." Genome Research, September. doi:10.1101/gr.125567.111.

Gonzalez-Perez, Abel, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. 2013. "IntOGen-Mutations Identifies Cancer Drivers across Tumor Types." Nature Methods 10 (11): 1081–82. doi:10.1038/nmeth.2642.

Greaves, Mel, and Carlo C Maley. 2012. "Clonal Evolution in Cancer." Nature 481 (7381): 306–13. doi:10.1038/nature10762.

Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." Nucleic Acids Research 28 (1): 27–30. doi:10.1093/nar/28.1.27.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2014. "Data, Information, Knowledge and Principle: Back to Metabolism in KEGG." Nucleic Acids Research 42 (D1): D199–D205. doi:10.1093/nar/gkt1076.

Kumar, Nitin, Hubert Rehrauer, Haoyang Cai, and Michael Baudis. 2011. "CDCOCA: A Statistical Method to Define Complexity Dependence of Co-Occuring Chromosomal Aberrations." BMC Medical Genomics 4 (1): 21. doi:10.1186/1755-8794-4-21.

Omberg, Larsson, Kyle Ellrott, Yuan Yuan, Cyriac Kandoth, Chris Wong, Michael R. Kellen, Stephen H. Friend, Josh Stuart, Han Liang, and Adam A. Margolin. 2013. "Enabling Transparent and Collaborative Computational Analysis of 12 Tumor Types within The Cancer Genome Atlas." Nature Genetics 45 (10): 1121–26. doi:10.1038/ng.2761.

R Core Team. 2013. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rubio-Perez, Carlota, David Tamborero, Michael P Schroeder, Albert A Antolín, Jordi Deu-Pons, Christian Perez-Llamas, Jordi Mestres, Abel Gonzalez-Perez,

and Nuria Lopez-Bigas. “In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Novel Targeting Opportunities” Under Review.
Tamborero, David, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S. Lawrence, et al. 2013. “Comprehensive Identification of Mutational Cancer Driver Genes across 12 Tumor Types.” Scientific Reports 3. doi:10.1038/srep02650.
Vandin, Fabio, Eli Upfal, and Benjamin J. Raphael. 2011. “De Novo Discovery of Mutated Driver Pathways in Cancer.” Genome Research, June. doi:10.1101/gr.120477.111.
Yeang, Chen-Hsiang, Frank McCormick, and Arnold Levine. 2008. “Combinatorial Patterns of Somatic Gene Mutations in Cancer.” The FASEB Journal 22 (8): 2605–22. doi:10.1096/fj.08-108985.
Zhang, J., J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, et al. 2011. “International Cancer Genome Consortium Data Portal--a One-Stop Shop for Cancer Genomics Data.” Database 2011 (0): bar026–bar026. doi:10.1093/database/bar026.

3.3 Exploring cancer genomics data with interactive heatmaps in Gitools 2

The exploration of multidimensional cancer genomics data imposes numerous challenges in terms of data collection, normalization and harmonization. In this chapter I present a manuscript in preparation that describes Gitools, a data visualization software, together with comprehensive cancer genomics data sets from several TCGA tumor sample cohorts and the IntOGen resource. For the effective navigation of those large multidimensional cancer genomics data sets, we developed a new version of Gitools, version 2, which is able to load and display the data as editable heatmaps. Jointly with Jordi Deu-Pons we have particularly worked on the implementation of new analyses, the improvement of the user's interface, the data access management (loading and saving) as well as interaction with third-party tools.

Schroeder, M.P, Deu-Pons, J., Tamborero, D., Perez-Llamas, C., Gonzalez-Perez, A. and Lopez-Bigas, N. Exploring cancer genomics data with interactive heatmaps in Gitools 2. (in preparation).

Abstract

The increased abundance of profiled genomic, transcriptomic and epigenomic data from multiple tumor types is providing a valuable data resource for cancer research. The complexity and size of such datasets hampers the intuitive exploration of this valuable data by many cancer researchers and clinicians. Data pre-processing and curation as well as intuitive software is needed to provide easy access to this data. We have compiled genomic, transcriptomic, epigenomic and clinical data of several large cohorts of tumor samples, to form multidimensional matrices ready to be explored using Gitools interactive heatmap visualizer and analyzer. In order to efficiently explore these matrices we have developed a new version of Gitools. Gitools 2 is able to load heavy data matrices in a memory efficient way, possesses an improved user's interface and includes new functions and analysis options specifically designed for the study of multidimensional cancer genomics data. Gitools 2 and prepared datasets are available from within the application and at http://www.gitools.org.

Introduction

The complexity of cancer genomics has given rise to large consortia such as The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Network 2008) and The International Cancer Genome Consortium (ICGC) (Hudson et al. 2010), which aim to complete genomic, transcriptomic and epigenomic profiles of at least 500 samples of numerous tumor types. Several independent cancer genomics datasets covering different cancer types have been published (The Cancer Genome Atlas Network, 2008; The Cancer Genome Atlas Network, 2013; The Cancer Genome Atlas Network, 2013a; The Cancer Genome Atlas Research Network, 2011), and one important challenge that arises is how to visually explore these large and complex data efficiently to contribute to the final goal of speeding up cancer research.

There are two main challenges to visually explore large cancer ge-

nomics datasets: i) datasets are big, complex and difficult to manipulate for most researchers and ii) multitude of different questions can be addressed using these datasets. Thus there is a need for visualization software that allows researchers to efficiently and flexibly explore cancer genomics datasets in a user-friendly manner.

Several programs and web portals exists that facilitate cancer genomics data access, such as the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/), Synapse (Omberg et al. 2013), the cBio Cancer Portal (Cerami et al. 2012), the UCSC cancer genome browser (Zhu et al. 2009) and IntOGen (Gonzalez-Perez, Perez-Llamas, et al. 2013; Gundem et al. 2010). Cancer genomics data is often visually represented in one of these three basic visualization types: genomic coordinates, matrix heatmaps and interaction networks (as covered in a recent review (Schroeder, Gonzalez-Perez, and Lopez-Bigas 2013)).

Heatmaps, in particular are useful intuitive graphical representations of matrices frequently used to represent transcriptomics and genomics data. Many existing tools and resources can represent genomics data, especially expression values, as heatmaps (Lex et al. 2012; Pavlidis and Noble 2003; Michael Reich et al. 2006; Saeed et al. 2003). In most cases heatmaps are generated as static images, which fall short to address the myriad of questions that arise from the analysis of these multidimensional cancer data sets, which in principle require that the user is able to flexibly interact with the heatmap.

Three years ago we presented Gitools an interactive heatmap visualization tool accompanied with some common analyzes, such as enrichment and correlation studies (Perez-Llamas and Lopez-Bigas 2011). The heatmaps in Gitools can represent multiple values in each cell, which makes it especially well suited for the representation of multidimensional cancer genomics data. When using Gitools to show cancer genomics data columns and rows normally represent tumor samples and genes respectively, and each cell contains multiple values for the different omic profiles obtained. Its interactive capabilities allow the user to filter, sort, move and hide rows and columns in the heatmap in context of gene and tumor sample annota-

tions and to launch several common exploratory analyses such as correlations, clustering, enrichment and statistical comparisons between groups of samples. A built-in option allows the users to sort the genes and samples within a heatmap following the pattern of the mutual exclusivity of alterations and test if the distribution of mutual exclusivity is expected with the given distribution of events amongst samples and genes. In summary, we have developed the second generation of Gitools which presents fundamental improvements in three main areas of the exploration of oncogenomics datasets. First, data manipulation has been reinforced to allow loading, exploring and analyzing bigger datasets in desktop computers, and data sharing and interoperability with other common tools has been introduced. For example, Gitools 2 is integrated within GenomeSpace (http://www.genomespace.org) and can inter-operate with the Integrative Genomics Viewer (Thorvaldsdóttir, Robinson, and Mesirov 2012), allowing the user to explore the same data in two complementary visualization tools that can communicate with each other. Second, the interface has been enhanced to account for more intuitive heatmaps handling and exploration. The Graphic User Interface (GUI) provides a more streamlined workflow and the gene and sample annotations now play a critical role in the process of exploration. Third, new analyses linking several data dimensions, called data layers in Gitools, in the dataset have been implemented. One example of these are comparisons of the values of one dimension in the matrix in groups formed by values in another dimension.

We present here a data repository of compiled genomic, transcriptomic, epigenomic and clinical data of several large cohorts of tumor samples, ready to be explored in the form of interactive heatmaps with Gitools 2. Currently the repository contains 19 heatmaps generated from data obtained from two main sources: TCGA (The Cancer Genome Atlas Research Network et al. 2013) and IntOGen-mutations (Gonzalez-Perez, Perez-Llamas, et al. 2013), which covers more than 5000 tumor samples and contains data from 12 different cancer types. We present specific examples in which browsing this data in the form of Gitools interactive heatmaps allows to easily answer key questions related to cancer biology.

Results

We first describe the cancer genomics datasets prepared to be explored with Gitools and available at www.gitools.org/datasets (Table 1), we then explain Gitools software improvements designed to effectively explore this data, finally we depict some use cases of these data employing Gitools 2.

Oncogenomics datasets ready to be explored with Gitools

TCGA pan-cancer 12 heatmaps

The TCGA pan-cancer 12 dataset consists in the union of twelve cohorts of samples of different tumor types. Each tumor sample has been probed for various genetic, epigenetic and transcriptomic alterations, such as somatic mutations, copy number alterations (CNAs), promoter methylation and mRNA expression levels. Additionally, patients' information has been gathered, yielding a rich dataset of clinical information. We have collated the aforementioned omics data, together with a set of tumor and patient's clinical annotations into a multidimensional Gitools data heatmap, containing information for 5'065 samples and 22'047 protein-coding genes and five data layers (see Table 1 and Table 2 for a detailed explanation of the content of the matrix).

We have prepared a series of meta-data for both samples and genes which can be added as labels to the columns' or rows' headers and can be used to operate with, such as when applying filters or orders based on those annotations or to compare profiles among annotation groups (see case studies below). In Figure 1, the annotations added as column headers are TCGA-Project Id and Sample Id, while gene annotations, added as row headers are CGC (Cancer Gene Census), OncodriveFM q-value and the Gene Symbol. The complete set of sample and gene annotations is listed in Table 2.

Protein expression data measured by RPPA (reverse phase protein array) technology for 3467 tumor samples are also ready to navigate

in a separate heatmap. This heatmap has 131 rows corresponding to proteins probed in the protein array (J. Li et al. 2013).

We have also prepared an excerpt from TCGA pan-cancer heatmap focused on driver genes and in which samples are grouped by COCA subtypes according to (Hoadley et al. 2014). This heatmap contains information only of genes annotated in the cancer gene census (CGC) (Futreal et al. 2004) and those 291 genes detected as high confidence drivers as described in (Tamborero et al. 2013).

TCGA individual projects

In addition to the TCGA 12 pan-cancer matrix we have prepared heatmaps focused on each particular TCGA project, which contains only samples of that tumor type. This avoids downloading large amount of data to those users interested only in a particular tumor type. Some of those heatmaps contain specific annotations only relevant for that tumor type, for example, Breast Cancer dataset contains the annotation of intrinsic subtypes for each sample and the Glioblastoma tumors are annotated with Glioblastoma molecular subtypes Classical, Mesenchymal, Neural and Proneural (Table 1). The expression data has been median-centered for each individual project in order to reflect expression differences relative to the project cohort rather than all the pan-cancer12 dataset.

IntOGen-mutations datasets

IntOGen-mutations is a platform devoted to identify cancer driver mutations, genes and pathways across tumour types on the basis of the analysis of somatic mutations from thousands of tumour re-sequenced genomes (Gundem et al. 2010; Gonzalez-Perez, Perez-Llamas, et al. 2013). We have prepared heatmaps summarizing the information contained in IntOGen-mutations on the analysis of 28 independent cancer genome sequencing projects covering more than 4600 tumors (Table 1).

Improvements of Gitools 2

Gitools was originally developed to visualize high-throughput data in the form of matrix heatmaps and to be able to interact with them and perform some exploratory analyses (Table 3; Perez-Llamas and Lopez-Bigas 2011). Gitools 2 presents major improvements over the original software that can be summarized in three main points: i) improved data manipulation and data storage, ii) more intuitive graphic user interface, iii) introduction of new analysis options.

Improved data manipulation and data storage

Some cancer genomics datasets describe above have large sizes, of the order of several gigabytes (Gb). To be able to explore this data powerfully using regular computers we have specifically improved the efficiency to handle big datasets in Gitools 2. On the basis of the .tabix (H. Li 2011) file indexing for genomic coordinates, we developed a new format for multidimensional matrices and tables (.mtabix) in order to save RAM (Random-Access Memory) when loading large matrices. For example, the TCGA pan-cancer matrix has a file size of 2.2 Gb in the flat text file format, which would increase when loaded into the application's memory. The new .mtabix format indexes and compresses the data matrix down to 380 Mb on disk and only reads and decompresses the data needed for visualization. This allows loading and browsing this heatmap at any desktop computer with only 1 Gb of available RAM.

In order to make it easy to load user's own data, now Gitools can open any text file with tab, comma or semicolon separated data fields. Data files can be either in a matrix format, with column identifiers in the first line and row identifiers in the first column of each line, or a table format with lines containing data for a heatmap cell as well as the column and row identifier. If multiple data sets for the same sample are available, it is possible to integrate them by importing the additional data file as new data layer.

Interoperability with other tools

We have developed an interface for Gitools that can be used for in-

teroperability with other tools or platforms. This interface has facilitated the interoperability of Gitools with Integrative Genomics Viewer (IGV) (Thorvaldsdóttir, Robinson, and Mesirov 2012) and with GenomeSpace (M. Reich et al. 2013), in a way that it is possible to run Gitools directly from the GenomeSpace or open data from IGV directly into Gitools. IGV communicates to a running instance of Gitools and GenomeSpace starts a new Gitools using the Java Web Start available at our server. Both options are available to any developer.

More intuitive graphic user interface

The browsing interface of Gitools has been re-designed and the control panels to the left have been minimized, so that the heatmap occupies more space (see Figure 1B). Other changes make the heatmap browsing and editing easier and quicker for the user: a right-click on gene and sample annotations now reveals a contextual menu which provides many readily accessible options, such as sorting, filtering and searching the rows and columns of the heatmap in function of the given annotation. Rows and columns can now be dragged and dropped to the position desired by the user. Zooming allows to rapidly change the perspective and amount of data displayed on screen. The minimal zoom (broadest view) permitted is 1 pixel height and width per cell. A new button to open the heatmap as a static image in a new tab in order to view its status in full size is available in the toolbar. Following exploration and analysis, heatmaps can be exported as image files either in bitmap (PNG) or in scalable (SVG) format for later use in publications.

Additional annotations to heatmaps are very important to correlate clinical variables with genomics and transcriptomics profiles within the data matrix. Thus Gitools allows the user to add color-coded annotations as headers to rows and columns and to sort them both as text and numerically. Rows and columns can also be annotated with aggregation functions. For example, the mean expression value per tumor type can be added as an annotation track. Specific contextual menus in columns and rows headers help the user to sort, group and filter based on annotations. Thus, the heatmap can easily be sorted

following annotations to samples, such as tumor type and anatomical site to highlight differences in genomics/transcriptomics profiles between clusters or groups of samples.

The color coding of the heatmap and header data is flexible; there are several kinds of color scales to represent different data types. In particular, the new categorical scale allows to visualize and annotate the data points for categorical values, such as the Genomic Alterations data layer in the TCGA pan-cancer matrix (see Figure 1). Each color scale comes with an event classifier. According to the value cut-offs chosen in the color scale, the event classifier can decide if a specific value is an event or a non-event. For example, in the p-value color scale, all values passing below the significance threshold are events. This information is displayed in a selection-specific context box and is used for the mutual exclusivity sorting and weight calculations in the statistical test. Upon selecting rows and/or columns, the selection specific context box displays simple statistics that may be useful to the user. For example when viewing data represented by a categorical scale such, the counts of occurrence of each category are reported. If the selected data is represented by a linear scale, mean and standard deviation are reported as well as a count of adjustable events (above scale values, below scale values).

New analysis options

In addition to exploratory options, Gitools contains some built-in analysis to be done over the heatmaps (Table 3). In Gitools 2 we have included several new analyses, which are especially useful to analyze cancer genomics data, although are not exclusive for this type of data. The mutual exclusivity sorting and testing feature (applied in Fig. 1, 2A, 2B and 2C) helps the researcher to identify gene sets that follow this alteration pattern (case study 1), which may indicate pathways containing cancer genes with driver alterations involved in tumorigenesis (Ciriello et al. 2011). A permutation based approach (see chapter 3.2 for details) allows to test the significance of mutual exclusivity and co-occurrence in the alteration pattern of a gene set. The new Group Comparison analysis allows the user to

perform a Mann-Whitney-Wilcoxon pairwise comparison of two groups of samples, in order to detect significant differences between them. Samples may be grouped using their identifiers, annotations or values in the data layer being compared or any other. One example involves grouping samples with the same copy number status of a given gene, and detect differences of gene expression levels between them, thus measuring the cis- or trans- effect of amplifications and deletions (see case study 2).

The clustering methods for numerical clustering of the heatmap have been rewritten and their default settings are adjusted to provide good clustering results with different data types. In particular the hierarchical clustering opens the hierarchical tree as a bitmap file in a new tab. In the heatmap tab, the hierarchical cluster is represented by color codes at different levels of the hierarchical tree (Figure 2), simplifying the information for better understanding and staying valid upon changing the order of the rows or columns. Upon performing a clustering the new bookmark function is used to store the order rows and columns and visible data layer of the heatmap in a heatmap bookmark. The bookmarks are saved along with the heatmap to the disk and can be restored later on.

Case studies: Visualizing pan-cancer data with Gitools

In the following case studies we describe several questions that may be answered through exploration of the pan-cancer dataset provided with Gitools 2.

Searching for mutational cancer drivers across 12 cancer types

The discrimination between driver – that cause or promote tumorigenesis-- and passenger genes is still a challenge even though intense research in that direction has been done in recent years (Kandoth et al. 2013; Lawrence et al. 2014; Tamborero et al. 2013). In order to be able to help the researcher to spot likely driver genes, we have added to the pan-cancer heatmap the annotations of several bioinformatics methods that identify mutational driver genes: Ac-

tiveDriver (Reimand, Wagih, and Bader 2013), MuSiC (Dees et al. 2012), OncodriveFM and OncodriveCLUST (Tamborero, Gonzalez-Perez, and Lopez-Bigas 2013; Gonzalez-Perez and Lopez-Bigas 2012). We have also annotated genes that are High Confidence Drivers (HCDs) and Candidate drivers (CDs), determined as described elsewhere (Tamborero et al. 2013). Briefly, these two groups were retrieved after combining the results of the aforementioned methods. Additional gene annotations include genes in the Cancer Gene Census (Futreal et al. 2004) and the pan-cancer MutSig results (syn1715784.2; (Lawrence et al. 2014)). Once the PCMD matrix is loaded, annotations can be visualized at the extreme right of each row as indicated in Figure 1. Because researchers may be interested in a small group of driver candidates, it may help to see if the alterations have an overlapping or mutually exclusive alteration pattern in this group, as represented in Figure 3A. There, genes that exhibit a mutually exclusive pattern of alterations in Glioblastoma (Ciriello et al. 2011) samples have been sorted according to the alteration data using the built-in mutual exclusivity sorting capability. In Figure 3A we can see that alterations in these genes cover a large fraction of the cohort. A further exploratory step could be aimed at finding potential drivers in the rest of the cohort (not displayed). This can be achieved by hiding all the samples where these genes are mutated and browse the mutations in the remaining tumor samples.

The cis-impact of CNAs on gene expression

Tumor genomes contain changes in the number of copies of certain chromosomal regions. We included CNA events in the pan-cancer alterations matrix relating each events to all the genes in the affected chromosomal region. CNA events can be browsed in the pan-cancer heatmap at two different data layers: CNA Status represents homozygous losses of a gene in blue and multi-copy amplifications in red, labeled accordingly in Gitools. In Genomic Alterations data layer, the sample CNA status of each gene can be visualized together with protein affecting mutations. The pan-cancer RNA-sequence data can be visualized in the Expression data layer and, then directly compared to the CNA status. A decrease in gene expression

is expected for homozygously or heterozygously lost genes which could help elucidate their implication in tumorigenesis; amplified genes involved in tumorigenesis are expected to show higher expression values. It is therefore useful to test whether the CNA of a gene has a cis-effect on its expression. The "Group Comparison" capability that has been added recently to Gitools allows the user to compare the values of two groups of samples employing the Mann-Withney-Wilcoxon test. We have performed this test for the whole PCMD Matrix and annotated the rows with the corrected p-value for negative or positive cis-effects, as shown in Figure 3B, followed by two column annotations marking which genes are mutational HCDs, and their chromosomal location. The figure contains the heatmaps of the CNA Status and Expression data layers. The genes with the 10 most significant Group Comparison p-values are displayed. Should a user want to perform a Group Analysis for a particular subtype, it can be easily done with the data we distribute following the on-line Tutorial at our documentation website (http://www.gitools.org/documentation).

Exploring molecular differences between cancer subtypes

In a large cohort of different tumor types such as the Pan-cancer data set, clustering and stratification of samples based on genomic features becomes an interesting way to assess similarities or important differences between samples. The PCMD Matrix contains several clinical annotations for each sample that can be employed to stratify patients. In Figure 3C, the integrated alterations data of the pan-cancer heatmap is annotated and sorted by the International Classification of Diseases (ICD-10) (http://www.who.int/classifications/icd/en/) cancer site in a manner that allows to compare genomic alterations between tumors from different sites. The figure shows that mutations in e.g. TP53 or APC are unevenly distributed between sites. It is thus possible to visually recognize differences in the mutational frequency of genes across cancer sites. (This numeric gene annotation has been added by an aggregation function within Gitools.) By means of the Group Comparison analysis the user can test if two groups of samples that seem different in their mutational patterns, also differ in the expression values of certain genes, analo-

gously to the approach of the cis-impact case study.

Sample Level Enrichment Stratification

Changes in the level of expression of individual genes are complex and difficult to interpret and it is often easier to study the transcriptional changes of groups or modules of genes. Gitools has a built-in capability to perform a Sample Level Enrichment Analysis (SLEA) (Gundem et al. 2010). A SLEA result reveals the relative transcriptional status of gene sets in each input sample. Figure 3D depicts the results of a SLEA which represents the relative expression status of eight core KEGG pathways (M. Kanehisa et al. 2010) and a set of chromatin regulatory factors (Gonzalez-Perez, Jene-Sanz, and Lopez-Bigas 2013). The three color-coded column headers represent annotation data (from top downwards): TCGA project, cancer tissue site and ICD-O 3 histology. Some obvious differences in the expression status of modules between tumor types are easy to observe. The samples of the acute myeloid leukemia cohort show very high relative expression of chromatin regulatory factors when compared to other cancer types. On the other hand, kidney tumors show very high relative expression of Cytokine-Cytokine receptors pathway and very low relative expression levels of chromatin regulatory factors and Cell Cycle genes. A SLEA analysis can be carried out directly within Gitools for any gene set submitted by the researcher. Any annotation for rows and columns such as those mentioned above may serve as criteria for stratification. Gitools now allows to sort and filter by any annotation loaded, be it text or numeric, and this option is easily accessible by right clicking on the column or row headers. This empowers the researcher to easily break big data matrices into smaller subsets without the necessity of any coding skills.

Discussion

Although several other heatmap viewers are available (Lex et al. 2012; Pavlidis and Noble 2003; Saeed et al. 2003; Michael Reich et al. 2006), Gitools 2 is currently the only one that allows the described high degree of interactivity in combination with the flexibil-

ity to import any additional user data. This is a key aspect of its conceptual design that makes it very suitable for the type of analyses described in the use cases, particularly, in large datasets like the pan-cancer cohort. Its efficient data management in terms of memory usage permits its employment in almost every personal computer and increases its value in the context of research and clinical settings. With Gitools 2, researchers are able to visually explore and analyze both pre-compiled or ad-hoc datasets. Such exploratory analyses can aid to propose hypotheses on the involvement of candidate driver genes in tumorigenesis in particular cancer types. Furthermore, they help to identify new profiles of genomic alterations that define specific groups of tumor samples.

The new Gitools 2 software as well as all the pre-compiled datasets are freely available at http://www.gitools.org/datasets. For illustrative purposes, we have prepared four use cases of Gitools 2 in the PCMD matrix and a genomics alterations and mutations summary matrix available under the same URL. All examples can be downloaded and then opened within Gitools 2. Additionally to the pan-cancer matrix, the reader may find cohort specific matrices, with annotations of four specific mutational driver detection methods: ActiveDriver, MuSiC, OncodriveCLUST and OncodriveFM.

Downloading the pre-compiled matrices and Gitools 2 enables offline exploration of all mutations, copy number alterations and expression data of the TCGA pan-cancer. Heatmaps configuration can be saved and reopened later and can also be exported as bitmap or vectorial image files for publication, or sharing with collaborators. If the researcher is interested in the data itself, it is possible to export it in a flat file format, both in matrix and in table form.

The software architecture of Gitools 2 allows loading any dataset in a matrix format. Thus, in addition to explore the pre-compiled TCGA data, users can load their own data in Gitools 2: the user may add further information to TCGA data, for example extra annotations to samples or genes, or load a custom dataset of cancer genomics data. This can be done with a flat matrix file or, in case of a multidimensional data set, a flat text file or spreadsheet in table format. Navigation settings are configurable for each data data layer, in

the panels below the details box.

In summary, we have presented a second generation Gitools, optimized to explore large multidimensional oncogenomics heatmaps interactively. We foresee that this new version of Gitools and the datasets pre-compiled for this paper (available at http://www.gitools.org/datasets), will simplify the effective exploration of multidimensional oncogenomics datasets by cancer researchers We could already observe a relative popularity of the second generation Gitools which has seen several public software releases during the last year: We have registered about 4500 events (downloads and launches) for the Gitools 2 application and about 1000 events for the prepared datasets. For both figures, the events counts from Barcelona have been removed.

Methods

Pan-cancer data set compilation

All pan-cancer related data have been downloaded from the pan-cancer repository at Synapse (Omberg et al. 2013), id syn300013. Samples with at least one mutation were retrieved from syn1729383, after excluding 71 considered as hypermutated samples according to the criteria described in that repository. Copy number alteration data was retrieved from syn1695369; only multi-copy amplifications or homozygous deletions were considered as copy number changes. Methylation data has been obtained from syn2486658, protein expression from syn1756922. Finally, expression values were retrieved from the RNA-seq data available in syn1695373.

Patient Annotations

All patient annotations were downloaded from the same Synapse pan-cancer collection (syn1446125.3, syn1446151.3, syn1446088.3, syn1446058.3, syn1446167.2, syn1446078.3, syn1446135.3, syn1446109.3, syn1446094.3, syn1446118.3, syn1446065.3, syn1446101.3) with the exception of the PAM50 calls and Glioblas-

toma subtype annotations, which were downloaded from the UCSC Cancer Genomics Browser.

Gene Annotations

General gene information and description were downloaded from the Ensembl Biomart version 71 (Kinsella et al. 2011). The annotations for the driver calling methods ActiveDriver ((Reimand, Wagih, and Bader 2013), MuSiC (Dees et al. 2012), OncodriveFM and OncodriveCLUST (Gonzalez-Perez and Lopez-Bigas 2012; Tamborero, Gonzalez-Perez, and Lopez-Bigas 2013) have been generated from the corresponding outputs generated by the respective authors. The cis-effect annotations were generated with Gitools Group Comparisons for all gain values and all homozygous loss values. The Cancer Gene Census information was downloaded from their web site and the MutSig results for (Lawrence et al. 2013) pan-cancer were obtained from Synapse (syn1715784.2).

Mutual exclusive test

The empirical p-values for the mutual exclusivity mutational pattern test are calculated taking into account the overall distribution of events across genes and samples as explained in (Hoadley et al. 2014). When performing the test for a group of genes (in rows), the overall sample coverage is used as background model for 10'000 weighted permutations for each gene. The weights are calculated according to the mutational burden of the samples, taking into account the entire data matrix, including hidden rows). The p-value reflects the probability of obtaining the overall sample coverage given X mutations in each gene within the group.

Indexing big matrix and tabular data files with .mtabix

We adapted the approach and format of the genomic file indexing .tabix to generic tabular data files and is now used as the standard to access the gitools matrix data sets. Code is available at https://bitbucket.org/bbglab/mtabix and can be freely used for other applications or to prepare big data for loading in Gitools.

Gitools 2

Gitools 2 is written in the Java programming language and is available for major operating systems (Linux, Windows, OS X). The new compressed format for storing the matrix, the use of the .mtabix format, allows to independently read and decompressed the data values when they are needed at runtime.

Figures

Figures 1-3 have been produced with Gitools 2 with and arranged with Inkscape and Gimp.

Acknowledgements

We acknowledge funding from the Spanish Ministry of Science and Technology (grant number SAF2012-36199) and the Spanish National Institute of Bioinformatics (INB). We are grateful to Khademul Islam, Alba Jené-Sanz and Mamun Majumder for acting as beta testers of Gitools and proposing new features and improvements and we want to thank Jim Robinson and Michael Reich for the collaborations on the interoperability improvements with IGV and GenomeSpace.

Author Contributions

JD-P, MPS and NL-B have authored Gitools 2. JD-P developed Gitools 2. MPS participated in the development of Gitools 2, prepared the TCGA data for Gitools, designed the use cases and drafted the manuscript. CP-L developed the first version of Gitools and provided feedback and ideas. DT and AG-P participated in the preparation of TCGA data for Gitools. NL-B supervised the whole project.

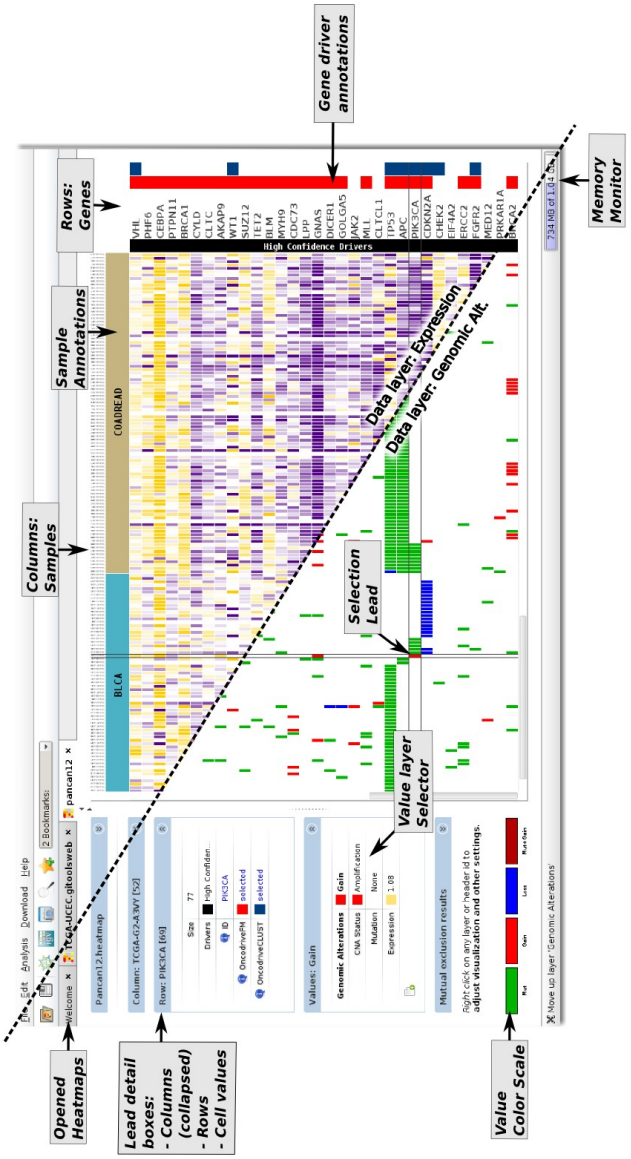

Figure 1: The Gitools 2 interface

Two screenshots of Gitools 2 with TCGA data separated by black diagonal line. In each half of the Figure we can see different value dimensions: Genomic alterations in the lower part and expression in the upper part. The figure shows the distribution of visualization, information and input space of Gitools.

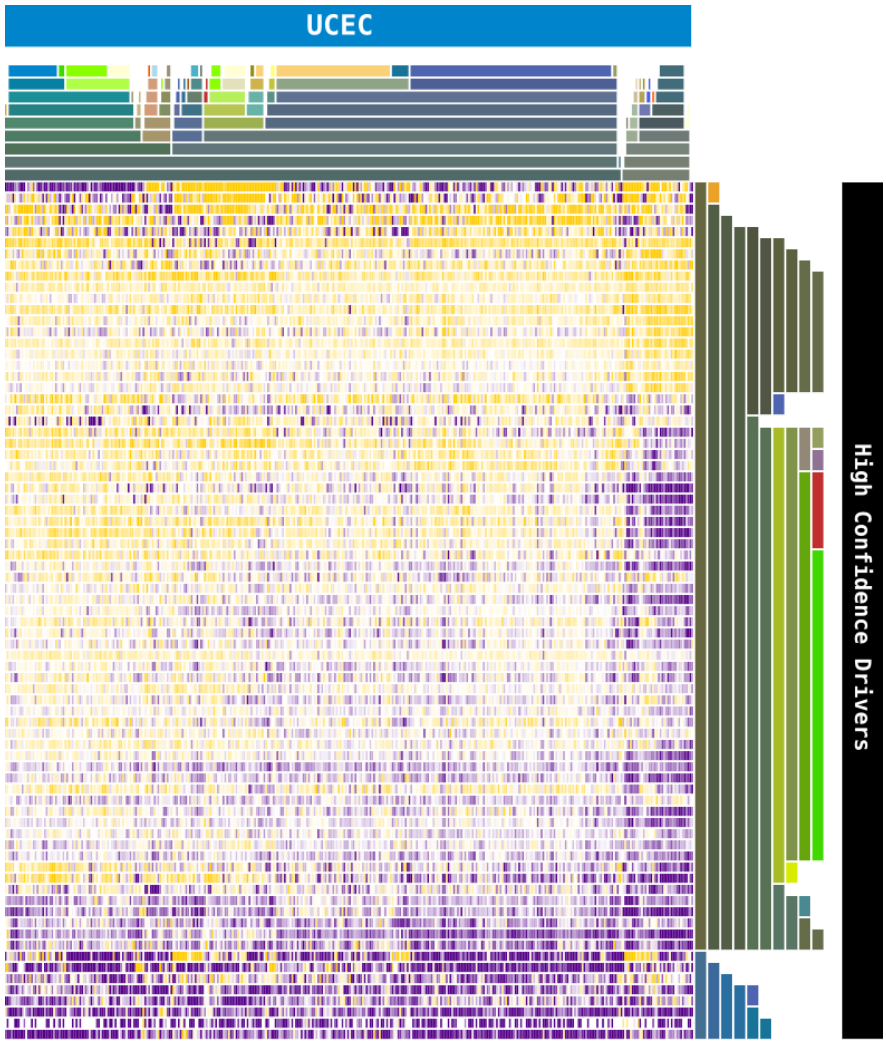

Figure 2: Hierarchical cluster representation

Figure to shows the expression data for the UCEC cancer samples in the genes annotated as cancer drivers. Both sample and gene dimensions are clustered with the hierarchical cluster algorithm, whose hierarchies are displayed at the top and at the right of the heatmap. The default parameters were used: As distance measure we employed the Euclidean distance and used average linking.

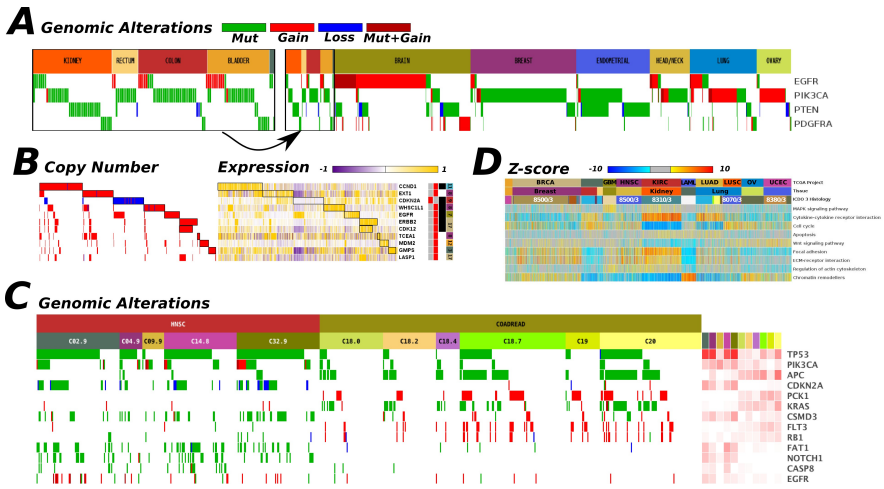

Figure 3: Cancer genomics cohorts in Gitools

*Figure three shows the figures that accompany the different use cases described in the manuscript. **A** displays the genomic alteration for four genes and all samples. The leftmost part is a zoomed-in view of the four TCGA projects with fewer samples. **B** shows the CIS-effect of the CNA on expression status. The CNA data has been ordered by mutual exclusivity. When visualizing the expression values after sorting we can see the mutual exclusive CNA pattern clearly reflected in the expression status. Sub-figure **C** shows the genomic alterations for the HNSC and COADREAD samples, sorted by mutual exclusivity and grouped by the ICD10 cancer site annotation. It is recognizable that both APC and TP53 are not mutated with the same frequencies across different ICD10 cancer site groups. Additionally we used Gitools to add a row header that shows the frequency for each group. **D** Shows the expression status as reflected by SLEA for select KEGG pathways. The samples are grouped by TCGA project and ICDO 3 Histology annotations in order to reveal possible subgroup-specific expression status of the pathways.*

Table 1. List of Cancer Genomics Gitools heat-maps available

Category	Name	Description
TCGA pan-cancer 12	TCGA pan-cancer 12 supermatrix	Complete pan-cancer dataset with 5065 samples and all 22'047 protein coding genes. It has 5 data layers: Gene expression, Mutations, Copy number, Methylation and integrated genomic alterations (which depicts mutations and copy number alterations in the same layer).
	pan-cancer12 protein expression (RPPA)	Complete protein expression dataset for the TCGA pan-cancer 12 project obtained using Reverse Phase Protein Array (RPPA) (J. Li et al. 2013). The heatmap contains 3467 samples and 131 protein expression data points.
	Subytpes of pan-cancer12 driver genes	Excerpt from the TCGA pan-cancer 12 supermatrix focused on driver genes and in which samples are grouped by COCA subtypes according to (Hoadley et al. 2014). Genes from the cancer gene census and detected as high confidence drivers as detected by (Tamborero et al. 2013) are shown.
TCGA Individual Projects	Bladder Urothelial Carcinoma (BLCA)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 138 BLCA samples.
	Acute Myeloid Leukemia (LAML)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 200 LAML samples.
	Glioblastoma multiforme (GBM)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 561 GBM samples. Most GBM samples are additionally annotated with Glioblastoma subtypes information (Neural, Pro-neural, Mesenchymal and Classical).
	Breast invasive carcinoma (BRCA)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 889 BRCA samples. Most BRCA samples are additionally annotated with intrinsic subtypes information (Luminal A, Luminal B, Basal, Her2 and Normal).
	Colorectal adenocarcinoma & Rectum adenocarcinoma (COADREAD)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 584 Colorectal adenocarcinoma and rectum adenocarcinoma samples.
	Head and Neck squamous cell carcinoma (HNSC)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 310 HNSC samples.
	Kidney renal clear cell carcinoma (KIRC)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 501 KIRC samples.
	Lung adenocarcinoma (LUAD) and Lung squamous cell carcinoma (LUSC)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 435 LUAD and 360 LUSC samples.
	Ovarian serous cystadenocarcinoma (OV)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 585 OV samples.
	Uterine Corpus Endometrioid carcinoma (UCEC)	Excerpt from the TCGA pan-cancer 12 supermatrix with information on 500 UCEC samples.

IntOGen	Genes by project	This heatmap contains the information on mutation frequency and putative driver genes per project (across 31 projects from 13 different cancer sites) as reported in IntOGen-mutations (Gonzalez-Perez, Perez-Llamas, et al. 2013). Driver genes were detected using OncodriveFM (Gonzalez-Perez and Lopez-Bigas 2012) and OncodriveCLUST (Tamborero, Gonzalez-Perez, and Lopez-Bigas 2013) using IntOGen-mutations pipeline.
	Genes by cancer site	This heat-map contains the information on mutation frequency and putative driver genes per cancer site (across 13 different cancer sites) as reported in IntOGen-mutations (Gonzalez-Perez, Perez-Llamas, et al. 2013).
	Pathways per project	This heat-map contains the information on mutation frequency and putative driver pathways per cancer site (across 13 different cancer sites) as reported in IntOGen-mutations ((Gonzalez-Perez, Perez-Llamas, et al. 2013).
	Cancer drivers Chromatin Regulatory Factors per project	This heat-map contains the information on mutation frequency and putative driver Chromatin Regulatory Factors per project (across 31 projects from 13 different cancer sites) as reported in (Gonzalez-Perez, Jene-Sanz, and Lopez-Bigas 2013).
	Cancer drivers Chromatin Regulatory Factors per site	This heatmap contains the information on mutation frequency and putative driver genes per cancer site (across 13 different cancer sites) as reported in (Gonzalez-Perez, Jene-Sanz, and Lopez-Bigas 2013). This heatmap reproduces the lower panel of Figure 1 of the manuscript: The mutations landscape of chromatin regulatory factors across 4623 tumor samples. Genome Biology.

Table 2. Content of the in the TCGA pan-cancer 12 supermatrix.

	ID of data dimension or annotation	**Description**
Data Layers (Heatmap Cells)	Expression	The expression values for each sample/gene combination according to the RNA-seq median-centered data.
	Mutation	States whether a protein affecting mutation was observed in the gene/sample. As protein affecting mutation, the following consequence types were considered: frameshift deletion, frameshift insertion, missense mutation, nonsense mutation, nonstop mutation and splice site mutation.
	Copy Number Alteration (CNA)	States whether a homozygous deletion or a multi-copy amplification was observed in the gene/sample.
	Methylation	Reflects if a gene has been detected as hypermethylated.
	Genomic Alterations	An integrated data layer for genomic alterations. It states whether a gene appeared as mutated and/or

		with copy number changes in each sample.
Sample Annotations (Columns)	tumor_tissue_site	The tissue site of the tumor
	CoCA subtype	Subtype classificaion according to an 'integrative analysis using five genome-wide platforms and one proteomic platform'
	icd_o_3_histology	The ICD-O 3 histology code
	icd_o_3_site	The ICD-O 3 site code
	icd_10	The ICD-10 site code
	histological_type	Histological type
	tumor_stage	The stage of the tumour
	ajcc_cancer_staging_handbook_edition	AJCC staging
	person_neoplasm_cancer_status	Tumor Status(at time of last contact or death)
	NRAS:mutation	Mutations have been detected in the NRAS gene in this cancer sample has
	KRAS:mutation	Mutations have been detected in the KRAS gene in this cancer sample has
	prior_diagnosis	If the patient has been diagnosed with another disease prior to study enrolment
	ALK:mutation	Mutations have been detected in the ALK gene in this cancer sample has
	BRAF:mutation	Mutations have been detected in the BRAF gene in this cancer sample has
	ERBB2:mutation	Mutations have been detected in the ERBB2 gene in this cancer sample has
	EGFR:mutation	Mutations have been detected in the EGFR gene in this cancer sample has
	age_at_initial_pathologic_diagnosis	Age of cancer patient at initial pathologic diagnosis
	days_to_last_known_alive	Difference in days from recording the data point and last known date of the patient being alive
	days_to_birth	Difference in days from recording the data to date of birth
	days_to_initial_pathologic_diagnosis	Difference in days from recording the data to initial pathologic diagnosis
	days_to_last_followup	Difference in days from recording the data to last followup
	gender	The gender of the cancer patient
	patient_id	Assigned patient Id
	vital_status	If at last recording the patient was alive or deceased
	pretreatment_history	Previous treatment before entering the study
	year_of_form_completion	Year of recording of clinical data for the patient
	month_of_form_completion	Month of recording of clinical data for the patient
	year_of_initial_pathologic_diagnosis	Year in which the patient was initially diagnosed
	TCGA-project	The TCGA project code
	icluster	iCluster Plus: Pattern discovery and cancer gene identification in integrated cancer genomic data.

	RPPA-cluster	Expression data for 131 proteins and 3467 pan-cancer tumor samples, measured by RPPA (reverse phase protein array) technology
	mutation_data	Mutation data in matrix for the sample
	cna_data	CNA data in matrix for the sample
	rnaseq_data	RNA-seq data in matrix for the sample
Gene Annotations (Rows)	EnsemblGeneID	The gene Id from the Ensembl database
	Description	The gene description from the Ensembl database
	ChromosomeName	The chromosome information from the Ensembl database
	GeneStart	Gene start base pair position (from the Ensembl database)
	GeneEnd	Gene end base pairu position (from the Ensembl database)
	Strand	Plus or minus strand (from the Ensembl database)
	Band	Cytochromic band (from the Ensembl database)
	%GCcontent	The GC content in percentage (from the Ensembl database)
	GeneBiotype	The biotype with which the gene is annotated in the Ensembl data base
	GeneStatus	Ensembl gene status
	MuSiC	Gene has been detected as being mutated more frequently than the expected by a background model.
	OncodriveFM	Gene has been detected to have to have a bias towards the accumulation of functional mutations (and hence its likelihood of being a driver) in tumors from this anatomical site.
	OncodriveCLUST	Gene has been detected as having mutations with a bias towards the clustering within specific protein regions
	ActiveDriver	Identifies genes whose mutations tend to concentrate in active sites of the protein.
	CGC	Cancer Gene Census' information about the gene
	MutSig	Gene has been selected as driver candidate according to this method. Originally, this was based in identifying genes frequently mutated, but at present it als o takes into account additional signals of positive selection.
	Drivers	Whether the gene has been selected as driver candidate by combining the results of various methods aimed to detect signals of positive selection. They can be high-confident drivers (HCD) which requires to pass stringent criteria aimed to favour the specificity of the HCD list; or candidate drivers (CD), which extends the list of HCD with genes passing less stringent criteria aimed to reduce the overall false negatives of the analysis.
	OncodriveFM-qval	Corrected p-value of the FM bias computed for the gene in the pan-cancer data set or a specific tissue. It indicates how biased is the gene towards the accumulation of functional mutations (and hence its likelihood of being a driver) in tumors from this anatomical site.
	OncodriveCLUST-qval	Corrected p-value of the CLUSTbias of a gene in the pan-cancer data set or a specific tissue. Low

		CLUSTbias indicates that the mutations in the gene tend to be clustered in particular protein regions, which is a sign of positive selection during tumor development and thus a sign that those genes may be cancer drivers. Computed with OncodriveCLUST.
	Freq_Mut	Frequency of samples mutated within the pan-cancer dataset (Only samples with mutation data have been considered).
	Freq_CNA	Frequency of CNA events within the pan-cancer dataset (Only samples with copy number data have been considered).
	CIS-effect-loss-qval	Corrected p-value for CIS effect of samples with homozygous loss.
	CIS-effect-gain-qval	Cerrected p-value for CIS effect of samples with amplification.

References

Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, et al. 2012. "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data." *Cancer Discovery* 2 (5): 401–4. doi:10.1158/2159-8290.CD-12-0095.

Ciriello, Giovanni, Ethan G Cerami, Chris Sander, and Nikolaus Schultz. 2011. "Mutual Exclusivity Analysis Identifies Oncogenic Network Modules." *Genome Research*, September. doi:10.1101/gr.125567.111.

Dees, Nathan D, Qunyuan Zhang, Cyriac Kandoth, Michael C Wendl, William Schierding, Daniel C Koboldt, Thomas B Mooney, et al. 2012. "MuSiC: Identifying Mutational Significance in Cancer Genomes." *Genome Research* 22 (8): 1589–98. doi:10.1101/gr.134635.111.

Futreal, P. Andrew, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. 2004. "A Census of Human Cancer Genes." *Nature Reviews Cancer* 4 (3): 177–83. doi:10.1038/nrc1299.

Gonzalez-Perez, Abel, Alba Jene-Sanz, and Nuria Lopez-Bigas. 2013. "The Mutational Landscape of Chromatin Regulatory Factors across 4,623 Tumor Samples." *Genome Biology* 14 (9): r106. doi:10.1186/gb-2013-14-9-r106.

Gonzalez-Perez, Abel, and Nuria Lopez-Bigas. 2012. "Functional Impact Bias Reveals Cancer Drivers." *Nucleic Acids Research*, August. doi:10.1093/nar/gks743.

Gonzalez-Perez, Abel, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. 2013. "IntOGen-Mutations Identifies Cancer Drivers across Tumor Types." *Nature Methods* 10 (11): 1081–82. doi:10.1038/nmeth.2642.

Gundem, Gunes, and Nuria Lopez-Bigas. 2012. "Sample Level Enrichment Analysis (SLEA) Unravels Shared Stress Phenotypes among Multiple Cancer Types."

Gundem, Gunes, Christian Perez-Llamas, Alba Jene-Sanz, Anna Kedzierska, Abul Islam, Jordi Deu-Pons, Simon J Furney, and Nuria Lopez-Bigas. 2010. "IntOGen: Integration and Data Mining of Multidimensional Oncogenomic Data." *Nat Meth* 7 (2): 92–93. doi:10.1038/nmeth0210-92.

Hoadley, Katherine A., Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max D. M. Leiserson, et al. 2014. "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin." *Cell* 158 (4): 929–44. doi:10.1016/j.cell.2014.06.049.

Hudson, Thomas J, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, M K Bhan, et al. 2010. "International Network of Cancer Genome Projects." *Nature* 464 (7291): 993–98. doi:10.1038/nature08987.

Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, et al. 2013. “Mutational Landscape and Significance across 12 Major Cancer Types.” *Nature* 502 (7471): 333–39. doi:10.1038/nature12634.

Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. 2010. “KEGG for Representation and Analysis of Molecular Networks Involving Diseases and Drugs.” *Nucleic Acids Research* 38 (Database issue): D355–D360.

Kinsella, R. J., A. Kahari, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, et al. 2011. “Ensembl BioMarts: A Hub for Data Retrieval across Taxonomic Space.” *Database* 2011 (0): bar030–bar030. doi:10.1093/database/bar030.

Lawrence, Michael S, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. 2014. “Discovery and Saturation Analysis of Cancer Genes across 21 Tumour Types.” *Nature* 505 (7484): 495–501. doi:10.1038/nature12912.

Lawrence, Michael S, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, et al. 2013. “Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes.” *Nature* 499 (7457): 214–18. doi:10.1038/nature12213.

Lex, A., M. Streit, H.-J. Schulz, C. Partl, D. Schmalstieg, P.j. Park, and N. Gehlenborg. 2012. “StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization.” *Computer Graphics Forum* 31 (3pt3): 1175–84. doi:10.1111/j.1467-8659.2012.03110.x.

Li, Heng. 2011. “Tabix: Fast Retrieval of Sequence Features from Generic TAB-Delimited Files.” *Bioinformatics* 27 (5): 718–19. doi:10.1093/bioinformatics/btq671.

Li, Jun, Yiling Lu, Rehan Akbani, Zhenlin Ju, Paul L. Roebuck, Wenbin Liu, Ji-Yeon Yang, et al. 2013. “TCPA: A Resource for Cancer Functional Proteomics Data.” *Nature Methods* 10 (11): 1046–47. doi:10.1038/nmeth.2650.

Omberg, Larsson, Kyle Ellrott, Yuan Yuan, Cyriac Kandoth, Chris Wong, Michael R. Kellen, Stephen H. Friend, Josh Stuart, Han Liang, and Adam A. Margolin. 2013. “Enabling Transparent and Collaborative Computational Analysis of 12 Tumor Types within The Cancer Genome Atlas.” *Nature Genetics* 45 (10): 1121–26. doi:10.1038/ng.2761.

Pavlidis, Paul, and William Stafford Noble. 2003. “Matrix2png: A Utility for Visualizing Matrix Data.” *Bioinformatics* 19 (2): 295–96. doi:10.1093/bioinformatics/19.2.295.

Perez-Llamas, Christian, and Nuria Lopez-Bigas. 2011. “Gitools: Analysis and Visualisation of Genomic Data Using Interactive Heat-Maps.” *PLoS ONE* 6 (5): e19541. doi:10.1371/journal.pone.0019541.

Reich, Michael, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P. Mesirov. 2006. “GenePattern 2.0.” *Nature Genetics* 38 (5): 500–501. doi:10.1038/ng0506-500.

Reich, M., J. Liefeld, H. Thorvaldsdottir, M. Ocana, T. Tabor, D. Jang, and J. P. Mesirov. 2013. “GenomeSpace: An Environment for Frictionless Bioinformatics.” *Cancer Research* 73 (8 Supplement): 5141–5141. doi:10.1158/1538-7445.AM2013-5141.

Reimand, Jüri, Omar Wagih, and Gary D Bader. 2013. “The Mutational Landscape of Phosphorylation Signaling in Cancer.” *Scientific Reports* 3: 2651. doi:10.1038/srep02651.

Saeed, A. I., V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, et al. 2003. “TM4: A Free, Open-Source System for Microarray Data Management and Analysis.” *BioTechniques* 34 (2): 374–78.

Schroeder, Michael P., Abel Gonzalez-Perez, and Nuria Lopez-Bigas. 2013. “Visualizing Multidimensional Cancer Genomics Data.” *Genome Medicine* 5 (1): 9. doi:10.1186/gm413.

Tamborero, David, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. 2013. “OncodriveCLUST: Exploiting the Positional Clustering of Somatic Mutations to Identify Cancer Genes.” *Bioinformatics (Oxford, England)* 29 (18): 2238–44. doi:10.1093/bioinformatics/btt395.

Tamborero, David, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S. Lawrence, et al. 2013. “Comprehensive Identification of Mutational Cancer Driver Genes across 12 Tumor Types.” *Scientific Reports* 3. doi:10.1038/srep02650.

The Cancer Genome Atlas Network. 2008. “Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways.” *Nature* 455 (7216): 1061–68. doi:10.1038/nature07385.

———. 2013a. “Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia.” *New England Journal of Medicine* 368 (22): 2059–74. doi:10.1056/NEJMoa1301689.

———. 2013b. “Integrated Genomic Characterization of Endometrial Carcinoma.” *Nature* 497 (7447): 67–73. doi:10.1038/nature12113.

The Cancer Genome Atlas Research Network. 2011. “Integrated Genomic Analyses of Ovarian Carcinoma.” *Nature* 474 (7353): 609–15. doi:10.1038/nature10166.

Thorvaldsdóttir, Helga, James T Robinson, and Jill P Mesirov. 2012. “Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration.” *Briefings in Bioinformatics*, April. doi:10.1093/bib/bbs017.

Zhu, Jingchun, J. Zachary Sanborn, Stephen Benz, Christopher Szeto, Fan Hsu, Robert M. Kuhn, Donna Karolchik, et al. 2009. “The UCSC Cancer Genomics Browser.” *Nature Methods* 6 (4): 239–40. doi:10.1038/nmeth0409-239.

3.4 jHeatmap: an interactive heatmap viewer for the web

The idea of visualizing and exploring data by means of interactive heatmaps is as valid for the desktop as it is for the web. The main limitation of web-native technologies are the web browsers and their JavaScript engine which determine the possibilities of computation within the browser. In any case, huge advances in web technologies allow for ever more complex computation in the browser and has enabled us to translate the interactive heatmap idea from Gitools to jHeatmap. jHeatmap is a JavaScript library that can load multidimensional datasets into the web browser and integrates user interactions such as sorting, filtering, visibility toggling which makes it an ideal component for web-platforms looking to communicate big data sets to fellow researchers. jHeatmap was developed jointly with Jordi Deu Pons who implemented the jHeatmap core aided by my contribution in form of test and repairing (debugging) with the biological examples. Furthermore I wrote the manuscript published in Bioinformatics.

Deu-Pons, J., Schroeder, M.P., and Lopez-Bigas, N. (2014). jHeatmap: an interactive heatmap viewer for the web. Bioinformatics btu094.

Deu-Pons J, Schroeder MP, Lopez-Bigas N. jHeatmap: an interactive heatmap viewer for the web. Bioinformatics. 2014 Jun 15; 30(12): 1757-8. DOI: 10.1093/bioinformatics/btu094

3.5 SVGMap: configurable image browser for experimental data

In order to communicate data efficiently and easily to decode it is sometimes best mapped onto what the data is representing: the original model. It is common for geographical information such as per-region data points as for example temperature. The same concept can be applied to biological data as for example gene expression in different tissue or cell types or even cell compartment-specific data points. For this reason we have developed SVGMap, a data browser that translates the data points into colored regions of an SVG graphic. The initial idea came from a collaboration with a wet-lab group. Thereafter Xavier Rafel-Plaou and me developed a generic tool that allows researchers to load any of their own data and SVG graphics in order to map the data onto the custom figure. Besides contributing code to the application, I have created various use cases described in the article and written the manuscript which was published in Bioinformatics.

Rafael-Palou, X., Schroeder, M.P., and Lopez-Bigas, N. (2011). SVGMap: configurable image browser for experimental data. Bioinformatics *28*, 119–120.

Rafael-Palou X, Schroeder MP, Lopez-Bigas N. SVGMap: configurable image browser for experimental data. Bioinformatics. 2012 Jan 1; 28(1): 119-20. DOI: 10.1093/bioinformatics/btr581

4 DISCUSSION

The work that I have done during my PhD training can be divided into two areas: cancer genomics and data visualization. Firstly, I will discuss cancer genomic studies that have been realized in the past years and what I was able to contribute with OncodriveROLE and MutEx. Secondly I will discuss my contributions to data visualization, particularly of multidimensional cancer genomics data in a cohort of cancer patients with Gitools and jHeatmap, as well as the SVGMap browser, a technology to create high-quality customized figures onto which experimental data is mapped.

4.1 Cancer genomics

As every day new bits and pieces are being published in many scientific journals, the complexity of the disease seems to be boundless and many aspects of the molecular mechanisms of cancer and the tumorigenesis remain unknown to this day. Nevertheless the orchestrated efforts of TCGA and ICGC plus many independent efforts have provided the scientific community with data that lets us look at a more complete picture of cancer than ever before, from a genomic point of view.

The genomic characterization of the pan-cancer samples has revealed the common and distinct patterns between the different cancer types and has lead to proposals of new classification of cross-tissue cancer types (Zack et al. 2013; Ciriello, Miller, et al. 2013; Hoadley et al. 2014) which may be of great importance for the planning and organizations of upcoming clinical trials. Furthermore, some distinctions between tumor sample groups indicate different oncogenic processes driven by distinct alteration types as main cause (Ciriello, Miller, et al. 2013).

The identification of mutational drivers from cancer genomics data has flourished during last years and has helped to unmask many novel cancer driver candidates, particularly including lowly frequently mutated genes. Studies of the TCGA pan-cancer12 cohort propose up to 290 candidates for driving tumorigenesis (Tamborero et al. 2013; Lawrence et al. 2014). Our research group, in particular, provided the IntOGen-mutations pipeline for online and offline use (Gonzalez-Perez, Perez-Llamas, et al. 2013; see annex), a powerful

tool to identify mutational candidate cancer drivers.

All these efforts to try to understand tumorigenesis and identifying cancer drivers and molecular patterns should eventually translate into a more focused development of new drugs on new targets and improved survival chances for cancer patients as a whole.

Drug development particularly does not only depend on the identification of novel cancer drivers but also on the way they act upon tumorigenesis. Predicting the mode of action of cancer drivers could only be assessed by thorough wet-lab studies. The availability of large re-sequenced tumor exomes and genomes gave us the opportunity to approach this question in a computational way by analyzing the pattern of somatic alterations observed in the tumoral DNA sequences. Differences between oncogenes and tumor suppressor genes have been exploited to develop **OncodriveROLE**, a classifier that can separate identified cancer driver genes into possible activating and loss of action roles within tumorigenesis. As discussed in the chapter of OncodriveROLE, our approach differs from Tuson (Davoli et al. 2013) and the 20/20-rule (Vogelstein et al. 2013) in that OncodriveROLE is not identifying cancer drivers, but classifying already identified cancer drivers into their mode of action. Thus, the cancer drivers that do not have a clear pattern of either role are not discarded as cancer drivers.

With OncodriveROLE we are the first group to provide a public classifier that can fulfill this kind of task. Another available method is Paradigm-Shift (Ng et al. 2012) which may give an indication of whether the gene alteration has an activating effect on downstream pathways. Even though similar, Paradimg-Shift differs from OncodriveROLE. Paradigm-Shift is a sample-based analysis, that returns a value for each sample that reflects a deregulation of the downstream signaling of the gene in question whereas OncodriveROLE returns a value per gene reflecting the driver class. Another important difference is that Paradigm-Shift is a rather elaborate method depending on the network model and is costly in terms of computation.

As for OncodriveROLE, we have assessed a total of 30 features. Besides the chosen features for the OncodriveROLE classifier, we

have also reported many other features in order to avoid that other research groups have to go through repetitive work. The chosen features assess mutational and copy number patterns, but it is possible that new evidences may be incorporated in the future. Abundant cancer genomics and epigenomics data will be released by TCGA and ICGC such that as f. ex. more knowledge about methylation patterns or expression patterns is being generated and could be used make the OncodriveROLE classifier more accurate. Particularly hypermethylation patterns are interesting as recurrent hypermethylation of tumor suppressors gene promoters have been reported (Baylin and Jones 2011). Expression data is somewhat more difficult to use as a matched normal is not always available or it may not be clear if it was affected by the neighboring cancer tissue. The dynamic nature of expression statuses makes it even more complicated to call expression alteration with confidence. The solution to this problem would be a comprehensive catalog of expression data from all human tissues where cancers are known to arise. This catalog should cover several hundred individuals per tissue, age, ethnicity, sex, etc. in order to be able to establish a *normal* gene expression quantity and variability for every gene and possible patient group and most importantly reflect the different tissues of the human body with high accuracy.

As more cancer samples are being sequenced, it should be the goal to maximize the number for each tissue, such that rare tissue-specific variants can be detected and an approach such as OncodriveROLE can be applied at a tissue level with confidence. Given that OncodriveROLE is able to gather enough information for most, but not all in a cohort of 4327 from 17 different cancer types, one may estimate that this same number per tissue may yield close to comprehensive results. A similar number has been estimated in (Lawrence et al. 2014). Assuming we would want 5000 cancer samples per cancer type, and take over the goal from ICGC to gather 50 different cancer types, this would mean that we'd need the sequence 250'000 cancer samples another 250'000 matched normal sequences. Ideally thus and if the available data allows enough statistical power, a thorough genomics study has to be conducted for each cancer type separately, considering each cancer a separate disease.

For OncodriveROLE this is particularly important as the classifier could detect cases where the protein products of the same gene in different tissues may have opposite as is the case with NOTCH1 (Licciulli et al. 2013; Liu, Zhang, and Ji 2013).

In the end, these numbers would not only help approaches such as OncodriveROLE, but also as before-mentioned the detection of lowly frequent drivers and also the identification of intrinsic cancer pathways would be facilitated and an important step would be taken towards the often-mentioned goal of personalized therapy for cancer patients. The knowledge gained with thorough genomic, epigenomic, transcriptomic and proteomic studies for every cancer type would not only allow the more precise development of anti cancer drugs, but if somatic alterations are available from hundreds of thousands of cancer patients, we could also generate precise knowledge about somatic alterations that may serve as subtype and treatment markers in order to give the best treatment possible to a patient. All in all, TCGA and ICGC had had ambitious goals when the projects were announced, but it seems that it is time to step up and numerically increase the decided goals to drive the thorough study of the cancer disease. Increased efforts of sequencing would not only give more power to cancer driver detection, but also pinpoint which genes and genomic regions must be understood better because for many novel cancer drivers that are being identified, thorough knowledge about functions and interactions is still lacking.

Nevertheless, not all depends on consortia such as TCGA and ICGC and primary cancer sample sequences. On the one hand, many independent efforts are generating valuable data sets of their own that can be combined with available datasets from ICGC and TCGA in order to gain more statistical (Gonzalez-Perez, Perez-Llamas, et al. 2013). Some may even deserve some special attention as they may represent specific populations and their specific affinities and predispositions for certain cancer types and mutations. On the other hand, we need to keep in mind that understanding tumorigenesis and understanding therapy resistance and disease recurrence are two separate pair of shoes. Cancer is difficult to treat because it is a multi-clonal disease. An existing clone of the cancerous tissue may not be affected by the treatment or new clones emerge upon the

new selective pressure exerted by anti-cancer drugs. The creation of cohorts featuring the follow-up history of cancer patients throughout the treatment may elucidate which paths the cancer cells choose in order to gain resistance. Furthermore the choice of what compounds are to used to treat a cancer patient may also depend on *ex vivo* performance of the drug as has been exemplified in a study where 28 AML samples from 18 patients were obtained and the *ex vivo* drug sensitivity and resistance has been assessed (Pemovska et al. 2013). This study shows two future trends mentioned above: time-series that portray the disease and possible relapses plus additionally to genomic assays a drug-screening that may already give a hint of what drugs may work for the patient. A caveat today is that the various genomic assays may take several weeks until the results are available.

Independent efforts can set future trends and the involved members and institutes gain valuable specific knowledge. But also, one must consider that they may not dispose of the same overall know-how as institutes that are involved in the macro-projects from TCGA, such as the Broad Institute or the Wellcome Trunst Sanger Center which is affiliated to the Cancer Genome Project. Thus, more modest cancer research institutes and hospitals all over the world depend on available methods and software. This is why we chose to develop not only approaches, but aim to make them available as we have done with OncodriveROLE, the rest of the Oncodrive family (Tamborero, Gonzalez-Perez, and Lopez-Bigas 2013; Gonzalez-Perez, Perez-Llamas, et al. 2013; Tamborero, Lopez-Bigas, and Gonzalez-Perez) and also with the MutEx approach.

Private cancer genomics data can easily be readied for loading in Gitools and apply the **MutEx** method, which may serve as testing tool if a predefined gene set of cancer drivers is acting as a tumorigenic function within the cohort in question. As the tumorigenic functions can be different in function of the cancer type or even subtype it is important to confirm those as such before the a treatment can be developed and applied.

MutEx is therefore an approach that is easily accessible and also open to novel cancer genomics data: The MutEx approach may

equally be applied to expression or methylation data with the only condition that data must be interpretable as a binary matrix (data events in Gitools).

4.2 Data Visualization

The importance of data visualization becomes ever bigger as the amount of data that is being generated is increasing daily. New approaches need to be developed in order to grasp all the data available in the least visualizations possible. Nevertheless, one must never loose touch with singular results and their visualization. In order to communicate biological data in an intuitive way it is helpful to map the data onto a cartoon that represents the model which is being studied. In collaboration in 2011 with a wet-lab back working with *Arabidposis* root expression data we come up with an approach, called **SVGMap**, that let them map the data onto a Scalable Vector Graphic (SVG) image containing labeled regions of the model corresponding to the different measurements taken. The colors that are projected onto the SVG model are customizable via the web-interface.

The use of SVG has since been popularized by libraries such as D3 (Bostock, Ogievetsky, and Heer 2011) and derivatives, affirming our choice of using SVG. The main difference between D3 and the SVGMap is that the SVGMap server lets the user load images and tabulated data files via it's interface and does not require any scripting knowledge whereas D3 provides a scripting interface that enables the web-creator to modify and animate SVG images within the web browser. However, given the popularity of D3, if SVGMap would have to be implemented today it would be convenient to use D3 as SVG manipulation instead of the letting the Java back-end server manipulate the SVG image. Nevertheless, SVGMap brings together the ability of generating specific precise figures of singular results with the access and management to all results at once. This idea could be developed further, as both SVG graphics of biological models such as organism, specific tissues etc. could be created and released in public domain as has been done with the figures used for the examples of SVGMap.

As for the consultation and exploration of large data sets such multidimensional cancer genomics datasets that are being generated by TCGA and ICGC, quite tedious preparation tasks and advanced scripting skills may be needed to finally be able to create intuitive and evident. This creates a certain barrier between the data generation and the exploration by researchers as not all groups may dispose of the required knowledge. The data needs to be normalized, prepared and put into easily accessible and intuitive tools and platforms which cancer researchers then may access and query the data. Data portals such as Cosmic, cBio cancer portal, the ICGC data portal or IntOGen database (Cerami et al. 2012; Gundem et al. 2010; Gonzalez-Perez, Perez-Llamas, et al. 2013; Zhang et al. 2011; Forbes et al. 2010) already try to give some aggregated views of the data in form of customized plots.

With Gitools and jHeatmap we have contributed two quite unique solutions to the scientific community which enable users to browse their data in interactive heatmaps. The interactivity is the key for continuous and flexible study of the data.

jHeatmap is the interactive heatmap solution for the web and is designed for easy re-use and incorporation into existing projects. The goal is to lessen the effort that is required by creating data portals, which is rather big as different aspects from data storage over user interface (UI) and plot generation have be well implemented. But all web-portals share the same programming environment: HTML5 documents with JavaScript capabilities. This shared foundation allows for energy-efficient development and re-use of components such as jHeatmap. jHeatmap was initially developed for our own needs of visualizing data from multidimensional cancer genomics cohorts in the IntOGen platform. The heatmap viewer written in JavaScript, can handle relatively large datasets and multiple data dimensions which are properties to handle large genomic cancer data.

As mentioned, jHeatmap is already in use in IntOGen, but it has also been incorporated by the GenomeSpace (M. Reich et al. 2013) platform and Achilles project (Cheung et al. 2011).

To further increase the usefulness of jHeatmap, we have included

the jHeatmap component into the BioJS code base. According to the project description of BioJS, it aims to “create a library of graphical components easy to reuse to represent biological information” and already disposes more than forty components which may be in any website and portal by simply installing BioJS (Gómez et al. 2013). Apart from bringing together many JavaScript visualization solutions in a registry, BioJS also supports a common events framework, that let the different components talk to each other.

Another application of jHeatmap is the use of the JavaScript library within the IPython Notebook (Pérez and Granger 2007). IPython Notebooks are data-driven documents backed by powerful libraries for data manipulation, analysis and documentation tools and therefore combines scripting, documentation and plotting into a sequential document which can be saved as such and therefore allows researchers to create easily reproducible analyses. Any data that is loaded in a so called DataFrame structure, as also known from the R programming environment, can be visualized with a couple of commands as an interactive heatmap by using the jHeatmap IPython package. The code is available for download at GitHub along with instructions and an example IPython Notebook: https://github.com/jheatmap/jheatmap-ipython.

The future will probably bring even more improved environments for web-based computation and a full port of **Gitools** capabilities may be starting to make sense. But for the time being Gitools is providing advanced interactive heatmap manipulation, browsing and data analysis for large multidimensional genomic cancer cohorts on the desktop.

Over the years we have maintained Gitools for both own and public use and converted it into a software application apt for browsing large multidimensional cancer genomics cohorts. On the one hand it is possible to load private data and on the other we have tried to tackle the barrier between data generation and exploration by preparing an array of heatmaps with cancer genomics data sets from TCGA and IntOGen that are available for download and exploration.

In a collaboration with the developers of the IGV genome browser

and GenomeSpace from the Broad institute we have added interactivity and data transport from both applications and platforms respectively. As for the future, more external control functions can be added to Gitools and interactions as well as data sending and receiving can be extended with whatever tool is used in combination with Gitools. The open source code basis would allow for other developers to add new functions, as the code basis is moving towards a plug-in design.

We know that Gitools is able to handle the current datasets even on machines with 2 Gb of RAM (Random-access memory). Nevertheless, we can be certain that the amount of cancer genomics data is ever increasing and will reach levels where a single machine is not able to cope anymore with an entire samples cohort dataset. Technologies available in Java, such as Hazelcast (https://github.com/hazelcast/hazelcast), allow for using distributed memory within a cluster of computers. The idea is to make the Hazelcast technology available in Gitools in order to be prepared for the future datasets. In combination with Webswing (http://www.webswing.org), another technology which allows to execute a server and interact with Java applications via the web browser, Gitools is not far away from being executed on the cluster and accessed from whatever terminal available.

As for functionality, a logical step for the application would be to allow not only visual manipulation of the underlying data, but enter formulas and generate new data layers or aggregation heatmaps. Such capabilities would make Gitools even more interesting for users that do not posses advanced scripting skills as a median-centering may easily performed after loading the data within the Gitools application and turn the application into what one may call an Excel for bioinformatics.

Last, but not least we know from download and usage statistics that the improvements over the years have lent Gitools a certain popularity. Between April 1st and July 31st we have registered 1245 usages of Gitools from 364 users from all continents, excluding all sessions from Barcelona in order to exclude ourselves. The first Gitools publication has been cited over 50 times and we hope that the

upcoming publication of the Gitools 2 will gain further visibility for the application. All in all we have the feeling that we have contributed a useful tool for data exploration and figure generation to the scientific community.

5 CONCLUSIONS

The major points of my work can be summarized as follows:

1. We have developed and published an approach, named OncodriveROLE, to classify cancer driver genes into Loss of Function and Activating roles.
2. We have developed a method to measure the tendency of a mutual exclusive relationship between somatic alteration events, called MutEx
3. We have incorporated MutEx into Gitools application in order to make the approach easily available for the scientific community
4. We have improved Gitools in many aspects such that it is fit to browse multidimensional genomics datasets, particularly large cancer cohorts.
5. We have complemented the available analyses possibilities with “Group Comparisons Analysis“, “MutEx & Co-occurrence Analysis” and substantially improved the “Clustering analysis” with an innovative representation model for dendrograms.
6. We have prepared cancer genomics heatmaps for Gitools that on one hand showcase the possibilities of the application and on the other hand are an easy manner to obtain integrated TCGA datasets with genomic and clinical annotations.
7. We have developed an interactive heatmap JavaScript library, called jHeatmap, for the web which is easy to integrate on any platform.
8. We have developed SVGMap, a data browser that maps experimental values onto SVG graphics in order to provide a tool to generate intuitive and high-quality figures.

6 APPENDIX

6.1 IntOGen-mutations identifies cancer drivers across tumor types

Abstract: The IntOGen-mutations platform summarizes somatic mutations, genes and pathways involved in tumorigenesis. It identifies and visualizes cancer drivers, analyzing 4,623 exomes from 13 cancer sites. It provides support to cancer researchers, aids the identification of drivers across tumor cohorts and helps rank mutations for better clinical decision-making.

My main contribution to this manuscript was the involvement in general design questions of the platform and more particularly in the employment of jHeatmap.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods 10, 1081–1082.

Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013 Nov; 10(11): 1081-2. DOI: 10.1038/nmeth.2642

7 BIBLIOGRAPHY

Advani, Anjali S., and Ann Marie Pendergast. 2002. “Bcr–Abl Variants: Biological and Clinical Aspects.” *Leukemia Research* 26 (8): 713–20. doi:10.1016/S0145-2126(01)00197-7.

Adzhubei, Ivan A, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. 2010. “A Method and Server for Predicting Damaging Missense Mutations.” *Nature Methods* 7 (4): 248–49. doi:10.1038/nmeth0410-248.

Baserga, Renato. 1985. *The Biology of Cell Reproduction*. Harvard University Press.

Baylin, Stephen B., and Peter A. Jones. 2011. “A Decade of Exploring the Cancer Epigenome — Biological and Translational Implications.” *Nature Reviews Cancer* 11 (10): 726–34. doi:10.1038/nrc3130.

Bostock, M., V. Ogievetsky, and J. Heer. 2011. “D3; Data-Driven Documents.” *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2301–9. doi:10.1109/TVCG.2011.185.

Braig, Melanie, Soyoung Lee, Christoph Loddenkemper, Cornelia Rudolph, Antoine H. F. M. Peters, Brigitte Schlegelberger, Harald Stein, Bernd Dörken, Thomas Jenuwein, and Clemens A. Schmitt. 2005. “Oncogene-Induced Senescence as an Initial Barrier in Lymphoma Development.” *Nature* 436 (7051): 660–65. doi:10.1038/nature03841.

Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, et al. 2012. “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data.” *Cancer Discovery* 2 (5): 401–4. doi:10.1158/2159-8290.CD-12-0095.

Cheung, H. W., G. S. Cowley, B. A. Weir, J. S. Boehm, S. Rusin, J. A. Scott, A. East, et al. 2011. “Systematic Investigation of Genetic Vulnerabilities across Cancer Cell Lines Reveals Lineage-Specific Dependencies in Ovarian Cancer.” *Proceedings of the National Academy of Sciences* 108 (30): 12372–77. doi:10.1073/pnas.1109363108.

Ciriello, Giovanni, Ethan Cerami, Bulent Arman Aksoy, Chris Sander, and Nikolaus Schultz. 2013. “Using MEMo to Discover Mutual Exclusivity Modules in Cancer.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* Chapter 8 (March): Unit 8.17. doi:10.1002/0471250953.bi0817s41.

Ciriello, Giovanni, Ethan G Cerami, Chris Sander, and Nikolaus Schultz. 2011. “Mutual Exclusivity Analysis Identifies Oncogenic Network Modules.” *Genome Research*, September. doi:10.1101/gr.125567.111.

Ciriello, Giovanni, Martin L. Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. 2013. “Emerging Landscape of Oncogenic Signatures across Human Cancers.” *Nature Genetics* 45 (10): 1127–33. doi:10.1038/ng.2762.

Collado, Manuel, and Manuel Serrano. 2010. “Senescence in Tumours: Evidence from Mice and Humans.” *Nature Reviews Cancer* 10 (1): 51–57. doi:10.1038/nrc2772.

Czech, Andreas, Ivan Fedyunin, Gong Zhang, and Zoya Ignatova. 2010. "Silent Mutations in Sight: Co-Variations in tRNA Abundance as a Key to Unravel Consequences of Silent Mutations." *Molecular bioSystems* 6 (10): 1767–72. doi:10.1039/c004796c.

Davoli, Teresa, Andrew Wei Xu, Kristen E. Mengwasser, Laura M. Sack, John C. Yoon, Peter J. Park, and Stephen J. Elledge. 2013. "Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome." *Cell* 155 (4): 948–62. doi:10.1016/j.cell.2013.10.011.

Dees, Nathan D, Qunyuan Zhang, Cyriac Kandoth, Michael C Wendl, William Schierding, Daniel C Koboldt, Thomas B Mooney, et al. 2012. "MuSiC: Identifying Mutational Significance in Cancer Genomes." *Genome Research* 22 (8): 1589–98. doi:10.1101/gr.134635.111.

Deu-Pons, Jordi, Michael P. Schroeder, and Nuria Lopez-Bigas. 2014. "jHeatmap: An Interactive Heatmap Viewer for the Web." *Bioinformatics*, February, btu094. doi:10.1093/bioinformatics/btu094.

Dunn, Gavin P., Allen T. Bruce, Hiroaki Ikeda, Lloyd J. Old, and Robert D. Schreiber. 2002. "Cancer Immunoediting: From Immunosurveillance to Tumor Escape." *Nature Immunology* 3 (11): 991–98. doi:10.1038/ni1102-991.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30 (1): 207–10. doi:10.1093/nar/30.1.207.

Eilbeck, Karen, Suzanna E. Lewis, Christopher J. Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. 2005. "The Sequence Ontology: A Tool for the Unification of Genome Annotations." *Genome Biology* 6 (5): R44. doi:10.1186/gb-2005-6-5-r44.

Fiona, on behalf e! Variation. 2014. "Variation Consequences." *Ensembl Blog*. Accessed August 28. http://www.ensembl.info/blog/2012/08/06/variation-consequences/.

Forbes, Simon A, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, et al. 2010. "COSMIC: Mining Complete Cancer Genomes in the Catalogue of Somatic Mutations in Cancer." *Nucleic Acids Research*, October. doi:10.1093/nar/gkq929.

Futreal, P. Andrew, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. 2004. "A Census of Human Cancer Genes." *Nature Reviews Cancer* 4 (3): 177–83. doi:10.1038/nrc1299.

Gómez, John, Leyla J. García, Gustavo A. Salazar, Jose Villaveces, Swanand Gore, Alexander García, Maria J. Martín, et al. 2013. "BioJS: An Open Source JavaScript Framework for Biological Data Visualization." *Bioinformatics*, February, btt100. doi:10.1093/bioinformatics/btt100.

Gonzalez-Perez, Abel, and Nuria Lopez-Bigas. 2012. "Functional Impact Bias Reveals Cancer Drivers." *Nucleic Acids Research*, August. doi:10.1093/nar/gks743.

Gonzalez-Perez, Abel, Ville Mustonen, Boris Reva, Graham R S Ritchie, Pau Creixell, Rachel Karchin, Miguel Vazquez, et al. 2013. "Computational

Approaches to Identify Functional Genetic Variants in Cancer Genomes." *Nature Methods* 10 (8): 723–29. doi:10.1038/nmeth.2562.

Gonzalez-Perez, Abel, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. 2013. "IntOGen-Mutations Identifies Cancer Drivers across Tumor Types." *Nature Methods* 10 (11): 1081–82. doi:10.1038/nmeth.2642.

Greaves, Mel, and Carlo C Maley. 2012. "Clonal Evolution in Cancer." *Nature* 481 (7381): 306–13. doi:10.1038/nature10762.

Grimwade, David, Robert K. Hills, Anthony V. Moorman, Helen Walker, Stephen Chatters, Anthony H. Goldstone, Keith Wheatley, Christine J. Harrison, Alan K. Burnett, and National Cancer Research Institute Adult Leukaemia Working Group. 2010. "Refinement of Cytogenetic Classification in Acute Myeloid Leukemia: Determination of Prognostic Significance of Rare Recurring Chromosomal Abnormalities among 5876 Younger Adult Patients Treated in the United Kingdom Medical Research Council Trials." *Blood* 116 (3): 354–65. doi:10.1182/blood-2009-11-254441.

Guidi, Cynthia J., Timothy M. Veal, Stephen N. Jones, and Anthony N. Imbalzano. 2004. "Transcriptional Compensation for Loss of an Allele of the Ini1 Tumor Suppressor." *The Journal of Biological Chemistry* 279 (6): 4180–85. doi:10.1074/jbc.M312043200.

Gundem, Gunes, Christian Perez-Llamas, Alba Jene-Sanz, Anna Kedzierska, Abul Islam, Jordi Deu-Pons, Simon J Furney, and Nuria Lopez-Bigas. 2010. "IntOGen: Integration and Data Mining of Multidimensional Oncogenomic Data." *Nat Meth* 7 (2): 92–93. doi:10.1038/nmeth0210-92.

Haber, Daniel A., and Jeff Settleman. 2007. "Cancer: Drivers and Passengers." *Nature* 446 (7132): 145–46. doi:10.1038/446145a.

Hanahan, Douglas, and Robert A. Weinberg. 2011. "Hallmarks of Cancer: The Next Generation." *Cell* 144 (5): 646–74. doi:10.1016/j.cell.2011.02.013.

Hanahan, D., and R. A Weinberg. 2000. "The Hallmarks of Cancer." *Cell* 100 (1): 57–70.

Heider, Andreas, and Rüdiger Alt. 2013. "virtualArray: A R/bioconductor Package to Merge Raw Data from Different Microarray Platforms." *BMC Bioinformatics* 14: 75. doi:10.1186/1471-2105-14-75.

Heyn, Holger, and Manel Esteller. 2012. "DNA Methylation Profiling in the Clinic: Applications and Challenges." *Nature Reviews Genetics* 13 (10): 679–92. doi:10.1038/nrg3270.

Hoadley, Katherine A., Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max D. M. Leiserson, et al. 2014. "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin." *Cell* 158 (4): 929–44. doi:10.1016/j.cell.2014.06.049.

Hudson, Thomas J, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, M K Bhan, et al. 2010. "International Network of Cancer Genome Projects." *Nature* 464 (7291): 993–98. doi:10.1038/nature08987.

Jay, Ernest, Robert Bambara, R. Padmanabhan, and Ray Wu. 1974. “DNA Sequence Analysis: A General, Simple and Rapid Method for Sequencing Large Oligodeoxyribonucleotide Fragments by Mapping*.” *Nucleic Acids Research* 1 (3): 331–53.

Jones, Peter A., and Peter W. Laird. 1999. “Cancer-Epigenetics Comes of Age.” *Nature Genetics* 21 (2): 163–67. doi:10.1038/5947.

Kim, Ryungsa, Manabu Emi, and Kazuaki Tanabe. 2007. “Cancer Immunoediting from Immune Surveillance to Immune Escape.” *Immunology* 121 (1): 1–14. doi:10.1111/j.1365-2567.2007.02587.x.

Kim, Su Y., and Terence P. Speed. 2013. “Comparing Somatic Mutation-Callers: Beyond Venn Diagrams.” *BMC Bioinformatics* 14 (1): 189. doi:10.1186/1471-2105-14-189.

Koren, Amnon, Paz Polak, James Nemesh, Jacob J. Michaelson, Jonathan Sebat, Shamil R. Sunyaev, and Steven A. McCarroll. 2012. “Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation.” *The American Journal of Human Genetics* 91 (6): 1033–40. doi:10.1016/j.ajhg.2012.10.018.

Kumar, Prateek, Steven Henikoff, and Pauline C Ng. 2009. “Predicting the Effects of Coding Non-Synonymous Variants on Protein Function Using the SIFT Algorithm.” *Nat. Protocols* 4 (8): 1073–81. doi:10.1038/nprot.2009.86.

Laing, Ken. 2011. “NewsLetter Transcriptomics & Functional Genomics.” http://www.ipc.nxgenomics.org/newsletter/no11.htm.

Lambert, Jeremy M R, Petr Gorzov, Dimitry B Veprintsev, Maja Söderqvist, Dan Segerbäck, Jan Bergman, Alan R Fersht, Pierre Hainaut, Klas G Wiman, and Vladimir J N Bykov. 2009. “PRIMA-1 Reactivates Mutant p53 by Covalent Binding to the Core Domain.” *Cancer Cell* 15 (5): 376–88. doi:10.1016/j.ccr.2009.03.003.

Lane, D. P. 1992. “p53, Guardian of the Genome.” *Nature* 358 (6381): 15–16. doi:10.1038/358015a0.

Lawrence, Michael S, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. 2014. “Discovery and Saturation Analysis of Cancer Genes across 21 Tumour Types.” *Nature* 505 (7484): 495–501. doi:10.1038/nature12912.

Lawrence, Michael S, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, et al. 2013. “Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes.” *Nature* 499 (7457): 214–18. doi:10.1038/nature12213.

Leek, Jeffrey T., Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. 2010. “Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data.” *Nature Reviews Genetics* 11 (10): 733–39. doi:10.1038/nrg2825.

Licciulli, Silvia, Jacqueline L. Avila, Linda Hanlon, Scott Troutman, Matteo Cesaroni, Smitha Kota, Brian Keith, et al. 2013. “Notch1 Is Required for Kras-

Induced Lung Adenocarcinoma and Controls Tumor Cell Survival via p53." *Cancer Research* 73 (19): 5974–84. doi:10.1158/0008-5472.CAN-13-1384.

Lieber, Michael R. 2010. "The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End Joining Pathway." *Annual Review of Biochemistry* 79: 181–211. doi:10.1146/annurev.biochem.052308.093131.

Li, Heng, and Nils Homer. 2010. "A Survey of Sequence Alignment Algorithms for next-Generation Sequencing." *Briefings in Bioinformatics* 11 (5): 473–83. doi:10.1093/bib/bbq015.

Liu, Na, Jingru Zhang, and Chunyan Ji. 2013. "The Emerging Roles of Notch Signaling in Leukemia and Stem Cells." *Biomarker Research* 1 (1): 23. doi:10.1186/2050-7771-1-23.

Lowe, Scott W., Enrique Cepero, and Gerard Evan. 2004. "Intrinsic Tumour Suppression." *Nature* 432 (7015): 307–15. doi:10.1038/nature03098.

McLendon, Roger, Allan Friedman, Darrell Bigner, Erwin G. Van Meir, Daniel J. Brat, Gena M. Mastrogianakis, Jeffrey J. Olson, et al. 2008. "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways." *Nature* 455 (7216): 1061–68. doi:10.1038/nature07385.

Mermel, Craig H., Steven E. Schumacher, Barbara Hill, Matthew L. Meyerson, Rameen Beroukhim, and Gad Getz. 2011. "GISTIC2.0 Facilitates Sensitive and Confident Localization of the Targets of Focal Somatic Copy-Number Alteration in Human Cancers." *Genome Biology* 12 (4): R41. doi:10.1186/gb-2011-12-4-r41.

Meyer, L. R., A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, et al. 2012. "The UCSC Genome Browser Database: Extensions and Updates 2013." *Nucleic Acids Research* 41 (D1): D64–D69. doi:10.1093/nar/gks1048.

Minamoto, Toshinari, Masayoshi Mai, and Ze'ev Ronai. 1999. "Environmental Factors as Regulators and Effectors of Multistep." *Carcinogenesis* 20 (4): 519–27. doi:10.1093/carcin/20.4.519.

Mullaney, Julienne M., Ryan E. Mills, W. Stephen Pittard, and Scott E. Devine. 2010. "Small Insertions and Deletions (INDELs) in Human Genomes." *Human Molecular Genetics* 19 (R2): R131–R136. doi:10.1093/hmg/ddq400.

Ng, Sam, Eric A. Collisson, Artem Sokolov, Theodore Goldstein, Abel Gonzalez-Perez, Nuria Lopez-Bigas, Christopher Benz, David Haussler, and Joshua M. Stuart. 2012. "PARADIGM-SHIFT Predicts the Function of Mutations in Multiple Cancers Using Pathway Impact Analysis." *Bioinformatics* 28 (18): i640–i646. doi:10.1093/bioinformatics/bts402.

Nielsen, Rasmus, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. 2011. "Genotype and SNP Calling from next-Generation Sequencing Data." *Nature Reviews Genetics* 12 (6): 443–51. doi:10.1038/nrg2986.

Pak, Theodore R., and Frederick P. Roth. 2013. "ChromoZoom: A Flexible, Fluid, Web-Based Genome Browser." *Bioinformatics* 29 (3): 384–86. doi:10.1093/bioinformatics/bts695.

Pérez, Fernando, and Brian E. Granger. 2007. “IPython: A System for Interactive Scientific Computing.” *Computing in Science and Engineering* 9 (3): 21–29. doi:10.1109/MCSE.2007.53.

Pemovska, Tea, Mika Kontro, Bhagwan Yadav, Henrik Edgren, Samuli Eldfors, Agnieszka Szwajda, Henrikki Almusa, et al. 2013. “Individualized Systems Medicine Strategy to Tailor Treatments for Patients with Chemorefractory Acute Myeloid Leukemia.” *Cancer Discovery* 3 (12): 1416–29. doi:10.1158/2159-8290.CD-13-0350.

Rafael-Palou, X., M. P. Schroeder, and N. Lopez-Bigas. 2011. “SVGMap: Configurable Image Browser for Experimental Data.” *Bioinformatics* 28 (1): 119–20. doi:10.1093/bioinformatics/btr581.

Reich, M., J. Liefeld, H. Thorvaldsdottir, M. Ocana, T. Tabor, D. Jang, and J. P. Mesirov. 2013. “GenomeSpace: An Environment for Frictionless Bioinformatics.” *Cancer Research* 73 (8 Supplement): 5141–5141. doi:10.1158/1538-7445.AM2013-5141.

Reimand, Jüri, Omar Wagih, and Gary D Bader. 2013. “The Mutational Landscape of Phosphorylation Signaling in Cancer.” *Scientific Reports* 3: 2651. doi:10.1038/srep02651.

Reva, B. 2013. “Revealing Selection in Cancer Using the Predicted Functional Impact of Cancer Mutations. Application to Nomination of Cancer Drivers.” *BMC Genomics* 14 (Suppl 3): S8. doi:10.1186/1471-2164-14-S3-S8.

Richards, Eric J. 2006. “Inherited Epigenetic Variation — Revisiting Soft Inheritance.” *Nature Reviews Genetics* 7 (5): 395–401. doi:10.1038/nrg1834.

Rodríguez-Paredes, Manuel, and Manel Esteller. 2011. “Cancer Epigenetics Reaches Mainstream Oncology.” *Nature Medicine*, March, 330–39. doi:10.1038/nm.2305.

Sørlie, Therese, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, et al. 2001. “Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications.” *Proceedings of the National Academy of Sciences* 98 (19): 10869–74. doi:10.1073/pnas.191367098.

Sauna, Zuben E., and Chava Kimchi-Sarfaty. 2011. “Understanding the Contribution of Synonymous Mutations to Human Disease.” *Nature Reviews Genetics* 12 (10): 683–91. doi:10.1038/nrg3051.

Schroeder, Michael P., Abel Gonzalez-Perez, and Nuria Lopez-Bigas. 2013. “Visualizing Multidimensional Cancer Genomics Data.” *Genome Medicine* 5 (1): 9. doi:10.1186/gm413.

Schroeder, Michael P, Carlota Rubio-Perez, David Tamborero, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. “OncodriveROLE Classifies Cancer Driver Genes in Loss of Function and Activating Mode of Action.” *Bioinformatics* 30 (17).

Shackleton, Mark, Elsa Quintana, Eric R. Fearon, and Sean J. Morrison. 2009. “Heterogeneity in Cancer: Cancer Stem Cells versus Clonal Evolution.” *Cell* 138 (5): 822–29. doi:10.1016/j.cell.2009.08.017.

Speleman, F., B. Cauwelier, N. Dastugue, J. Cools, B. Verhasselt, B. Poppe, N. Van Roy, et al. 2005. “A New Recurrent Inversion, inv(7)(p15q34), Leads to Transcriptional Activation of HOXA10 and HOXA11 in a Subset of T-Cell Acute Lymphoblastic Leukemias.” *Leukemia* 19 (3): 358–66. doi:10.1038/sj.leu.2403657.

Stephens, Philip J., Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, et al. 2011. “Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development.” *Cell* 144 (1): 27–40. doi:10.1016/j.cell.2010.11.055.

Stratton, Michael R., P. Andrew Futreal, and Richard Wooster. 2004. “Genome-Wide Searches for Mutations in Human Cancer.” In *Oncogenomics*, edited by Charles Brenner and David Duggan, 15–35. John Wiley & Sons, Inc. http://onlinelibrary.wiley.com/doi/10.1002/047147665X.ch2/summary.

Supek, Fran, Belén Miñana, Juan Valcárcel, Toni Gabaldón, and Ben Lehner. 2014. “Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers.” *Cell* 156 (6): 1324–35. doi:10.1016/j.cell.2014.01.051.

Tamborero, David, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. 2013. “OncodriveCLUST: Exploiting the Positional Clustering of Somatic Mutations to Identify Cancer Genes.” *Bioinformatics (Oxford, England)* 29 (18): 2238–44. doi:10.1093/bioinformatics/btt395.

Tamborero, David, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S. Lawrence, et al. 2013. “Comprehensive Identification of Mutational Cancer Driver Genes across 12 Tumor Types.” *Scientific Reports* 3. doi:10.1038/srep02650.

Tamborero, David, Nuria Lopez-Bigas, and Abel Gonzalez-Perez. “Oncodrive-CNA: A Method to Reveal Likely Driver Genes Based on the Impact of Their Copy Number Changes on Expression (in Press).”

The Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. “The Cancer Genome Atlas Pan-Cancer Analysis Project.” *Nature Genetics* 45 (10): 1113–20. doi:10.1038/ng.2764.

Vandin, Fabio, Eli Upfal, and Benjamin J. Raphael. 2011. “De Novo Discovery of Mutated Driver Pathways in Cancer.” *Genome Research*, June. doi:10.1101/gr.120477.111.

Van ’t Veer, Laura J., Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, et al. 2002. “Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer.” *Nature* 415 (6871): 530–36. doi:10.1038/415530a.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. 2013. “Cancer Genome Landscapes.” *Science* 339 (6127): 1546–58. doi:10.1126/science.1235122.

Wang, Qingguo, Peilin Jia, Fei Li, Haiquan Chen, Hongbin Ji, Donald Hucks, Kimberly Brown Dahlman, William Pao, and Zhongming Zhao. 2013. “Detecting Somatic Point Mutations in Cancer Genome Sequencing Data: A

Comparison of Mutation Callers." *Genome Medicine* 5 (10): 91. doi:10.1186/gm495.

Westesson, Oscar, Mitchell Skinner, and Ian Holmes. 2012. "Visualizing next-Generation Sequencing Data with JBrowse." *Briefings in Bioinformatics*, March, bbr078. doi:10.1093/bib/bbr078.

Wheeler, Richard. 2007. *English: By Richard Wheeler (Zephyris) 2007. The Two Major Two Chromosome Mutations; Insertion and Translocation.* Originally from en.wikipedia; description page is/was here. http://commons.wikimedia.org/wiki/File:Two_Chromosome_Mutations.png.

Wikimedia Commons. 2013. *English: Drawing Showing What Happens in Gene Duplication.* http://commons.wikimedia.org/wiki/File:Gene-duplication.png.

Wikipedia. 2014. "DNA Codon Table." *Wikipedia, the Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=DNA_codon_table&oldid=576071692.

Yu, Zuoren, Timothy G. Pestell, Michael P. Lisanti, and Richard G. Pestell. 2012. "Cancer Stem Cells." *The International Journal of Biochemistry & Cell Biology* 44 (12): 2144–51. doi:10.1016/j.biocel.2012.08.022.

Zack, Travis I., Steven E. Schumacher, Scott L. Carter, Andrew D. Cherniack, Gordon Saksena, Barbara Tabak, Michael S. Lawrence, et al. 2013. "Pan-Cancer Patterns of Somatic Copy Number Alteration." *Nature Genetics* 45 (10): 1134–40. doi:10.1038/ng.2760.

Zech, L., G. Gahrton, L. Hammarström, G. Juliusson, H. Mellstedt, K. H. Robèrt, and C. I. E. Smith. 1984. "Inversion of Chromosome 14 Marks Human T-Cell Chronic Lymphocytic Leukaemia." *Nature* 308 (5962): 858–60. doi:10.1038/308858a0.

Zhang, J., J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, et al. 2011. "International Cancer Genome Consortium Data Portal--a One-Stop Shop for Cancer Genomics Data." *Database* 2011 (0): bar026–bar026. doi:10.1093/database/bar026.