Acknowledgments

I would like to thank my supervisors, Fabrice André and Céline Lefebvre for their support, especially Céline who has devoted a great amount of time and effort to guide me through this thesis, who has been a great advisor and dear friend, without her none of these work would be possible.

Thanks also to my colleagues and friends, Cécile, Hélène, Noémie, Teele, Virginie, Véronique, Semih, Marion, Fred, Aicha, Tony, Bojana, Chloé, Delphine, Sandrine and many others, from whom I've learned a lot of things not only professionally but also personally. Thank you all for your kindness and your support during these years both when things were going well and going rough. The moments that we shared together will always be an unforgettable part of this great journey.

Last but not least, I want to thank my parents and my aunt, who have supported me with unconditionally love and trust in every decision that I've made throughout this incredible long pursuing of academics. Thank you so much and I love you all.

List of figures

List of tables

Abbreviations

Area under the drug response curve (nAUC)

Base excision repair (BER)

Cancer Cell Line Encyclopedia (CCLE)

Cancer Genome Project (CGP)

Chronic myelogenous leukemia (CML)

Circulating tumor DNA (ctDNA)

Comparative genomic hybridization array (CGHa)

Copy number variation (CNV)

ctDNA mutation Detect (cmDetect)

Epidermal growth factor receptor (*EGFR*)

Estrogen receptor (ER)

Expectation Maximization (EM)

False discovery rate (fdr)

Fluorescence In Situ Hybridization (FISH)

Genomics of Drug Sensitivity in Cancer (GDSC)

Hidden Markov Model (HMM

Homologous recombination (HR)

Human Epidermal Growth Factor Receptor-2 (HER2)

ImmunohistoChemistry (IHC)

Kilobase (kb)

Log2 relative ratio (LRR)

Minor allele frequency (MAF)

Next generation sequencing (NGS)

Progesterone receptor (PR)

Progression free survival (PFS)

Random forest (RF)

Support Vector Machine (SVM)

The Cancer Genome Atlas (TCGA)

The half maximal inhibitory concentration (IC50)

The International Cancer Genome Consortium (ICGC)

Tyrosine kinase inhibitors (TKIs)

Introduction

Cancer and Cancer genomics

The human genome carries nearly all the information that one inherited from its ancestors, to form cells with different functions that develop into extremely complicated organisms. The DNA code is translated into protein products, which then form complex interactive networks (also called pathways) that are critical for cell proliferation, cell division, cell migration, programmed cell death or DNA repair.

It took scientists several decades from thinking neoplastic tumors are foreign bodies that have invaded the patients to prove that cancer arise from the acquisition of genomic alterations in the genomes of the patients' cells that fundamentally alter the function of the protein products of key genes transformation these into tumor cells. This established the fundamental idea that genomic alterations have an important implication in both the diagnostic and prognostic of cancer. The idea changing was initially brought up in the late nineteenth century when scientists have observed chromosomal abnormalities in cancer cells under a microscope (Boveri 2008; von Hansemann 1890; Boveri 1914).

A genomic alteration can be either germline and is presented in every cell of the human body, or somatic if it is acquired in a subset of cells during the life time. Both forms of genomic alterations have been observed to be related to cancer. For instance, a germline loss of function mutation in *BRCA1* or *BRCA2* genes gives a predisposition to breast cancer for women by increasing their

risk of developing a breast or ovarian cancer by five as compared to the rest of the women population. Although BRCA1/2 mutations are often associated with a type of aggressive breast cancer called triple negative breast cancer with no effective treatment, patients with this mutation can benefit from a drug targeting PARP, a member of a family of enzymes involved in base excision repair (BER), an important DNA repair mechanism. While BRCA1/2 genes play important roles in the homologous recombination (HR) DNA repair mechanism, the inhibition of PARP leads to the loss of 2 main DNA repair mechanisms in the cancer cells, creating a synthetic lethal (Kaelin 2005) phenomenon leading to cell death (Figure 1).

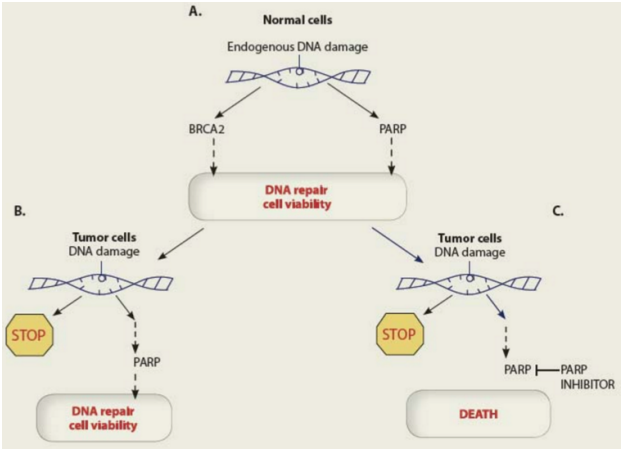


Figure 1. A normal cell with intact BRCA and PARP functions is able to repair DNA normally (A). In a tumor cell with a mutation in BRCA, intact PARP function results in ability to repair DNA and subsequent viability (B). Central to the use of PARP inhibitors in the treatment of patients with malignancy related to BRCA1/2 mutations is the concept of synthetic lethality (Rios and Puhalla 2011).

One other example of cancer related genomic alteration is the amplification of the Human Epidermal Growth Factor Receptor-2 (HER2) in breast cancer patients. HER2 is a protein that activate intracellular signaling pathways in response to extracellular signals, and is implicated in the pathogenesis of human breast cancer by increasing invasiveness, proliferation and tumorigenicity of cells when over-expressed (Slamon, et al. 1987). HER2 is amplified or over-expressed in 20% to 30% of breast cancer patients. Inhibiting HER2 in breast cancer patients with amplification or overexpression has shown significant increase of time to disease progression and higher response rates than chemotherapy alone (Slamon, et al. 1987). These two examples have demonstrated that the genomic alterations are important biomarkers for defining cancer treatment strategy and thus improving drug response.

Cancer related genes

Given the huge number of cells in a human body and the non-stop processes of DNA replication and DNA repair in cells undergoing divisions, the occurrence of somatic alteration is not only frequent but also inevitable (Drake, et al. 1998). A mutation may cause cancer when it affects the activity of a gene involved in key cellular functions such as cell proliferation or cell death. These genes can be classified into the following two groups (Figure 2):

- Oncogenes

Activated oncogenes can transform a healthy cell into a cancerous cell by driving the normal cell to uncontrolled proliferation, defective differentiation or

failure to undergo programmed cell death. The activation of an oncogene is usually caused by genomic alterations including amplification, gain-of-function mutation or over-expression. HER2 in the previous section is an example of oncogene.

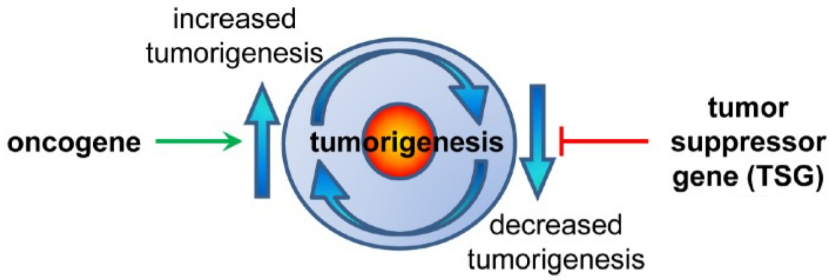


Figure 2. Oncogenes promote tumorigenesis while tumor suppressor genes inhibit cell proliferation and tumor development. An activation of oncogenes or and inactivation of tumor suppressor genes can cause cancer (adapted from (Shah, et al. 2013).

- Tumor suppressor genes

Tumor suppressor genes are the ones that inhibit cell growth, cell division and cell survival. An inactivation of a tumor suppressor gene caused by a genomic alteration such as loss-of-function mutation, deletion or low expression, can also lead to an abnormal growth of the cells. The BRCA1/2 genes are tumor suppressor genes. Another well-known example of tumor suppressor gene is *p53*, which has many anticancer functions such as activation of DNA repair proteins, inhibition of cell growth and initiation of apoptosis in cells with DNA damage. More than 50% of all cancers involve a missing or

damaged *p53* gene, making *p53* the most commonly mutated gene in cancers (Kandoth, et al. 2013).

Types of genomic alterations

DNA level

Different types of alterations occur at the DNA level including translocations, copy number variations such as amplifications or deletions of segments of DNA, and mutations including small insertion and deletions, point mutations.

- **Translocation** is one common form of chromosomal rearrangement that is probably created by failed DNA repair mechanism after DNA double-strand breaks (Griffiths AJF 1999). A translocation can put a gene next to a transcriptionally active promoter or enhancer element of another gene on a different chromosome, thereby leading to an abnormal expression of the translocated gene. A translocation can also result in fusion genes, which in turn codes for an activated form of a protein that affects the normal cellular physiology. The fusion gene BCR-ABL (Figure 3), found in most patients with chronic myelogenous leukemia (CML), results from the translocation of the ABL gene from chromosome 9 and the BCR gene from chromosome 22 and has been shown essential and sufficient for the malignant transformation of the cell (Gambacorti-Passerini and Piazza 2015; Salesse and Verfaillie 2002). CML patients carrying the BCR-ABL fusion gene can benefit from tyrosine-kinase inhibitors such as Imatinib, which has been approved for

the disease treatment in 2001 and is now the first line treatment for CML patients with BCR-ABL fusion (Gambacorti-Passerini and Piazza 2015).

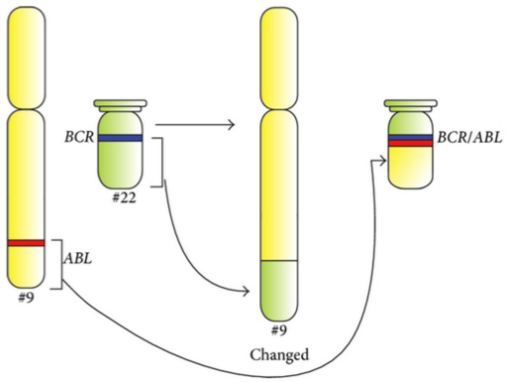


Figure 3. The fusion BCR-ABL gene is formed as a result of the chromosome 9 and chromosome 22 translocation (Trela, Glowacki, and Błasiak 2014).

- **Copy number variation (CNV)** can be a deletion or an amplification of a segment of DNA of a size larger than 1 kilobase (kb) of nucleotides, causing a gene or a part of gene, or even a whole chromosome to have respectively less or more than the normal 2 copies. A copy number variation can have an important influence in the expression of the lost or gained genes. CNVs are frequently detected and have a significant role in tumorigenesis in many tissues, such as colorectal cancer (Leary, et al. 2008), ovarian cancer (Despierre, et al. 2014) and prostate cancer (Chen, et al. 2005). For example, deletion of PTEN, a commonly lost tumor suppressor gene, is estimated to occur in up to 70% of men with prostate cancer at the time of diagnosis. Deletion of PTEN has been associated with improved response and longer progression-free survival

in metastatic castration-resistant prostate cancer patients treated by Everolimus, an anti-mTOR inhibitor, in a phase II clinical trial, and can therefore be used as a predictive biomarker to anti-mTOR inhibitor sensitivity (Templeton, et al. 2013)

- **Mutations** are often designated as genomic alterations at a smaller scale, i.e. affecting one to a few base pairs of nucleotides, and include small insertions and deletions and single base point mutations. Based on the impact of the mutation on the protein sequence, it can be classified as (Figure 4):

  - a nonsense mutation, if it leads to the gain of a stop codon that terminates the translation process;
  - a missense mutation, if it leads to a different amino acid;
  - a silent mutation, if it does not change the protein sequence.

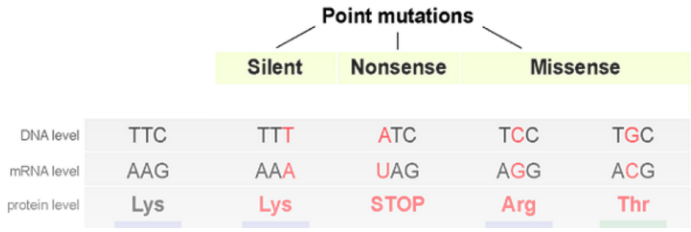


Figure 4. An illustration of nonsense, missense and silent mutation.

For a mutation to cause cancerous consequences, it has to not only occur in the right gene, but also at the right position. Some mutations occur at a position that does not disrupt seriously the cell functions, thus can be tolerated and may be naturally found at a certain frequency in the population. These mutations are

called polymorphism. Mutations are probably the most frequently observed type of genomic alterations in cancer and have an important role, especially somatic mutation which are specific to the tumor. One example is the point mutation V600E in BRAF (Figure 5), that induce an increased kinase activity of the gene leading to an activation of the MEK pathway and is presented in 30% to 50% of melanoma patients. Patients with tumors harboring V600E and V600K BRAF mutations showed better responses to the BRAF and MEK inhibitors than to chemotherapy (Sosman, et al. 2012; Flaherty, et al. 2012).

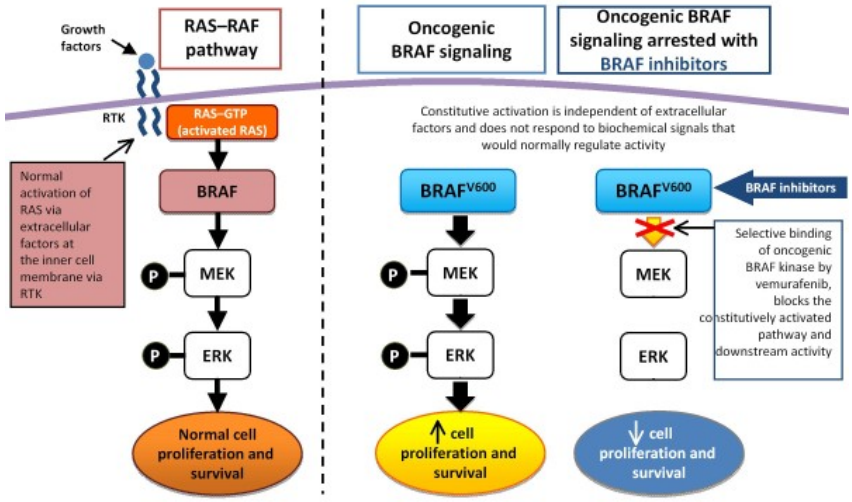


Figure 5. BRAF V600E oncogenic signaling pathway in melanoma. (Ascierto, et al. 2012)

Gene expression level

A genomic alteration that is not expressed in the cell is less probable to have a serious impact on the cell function. On the other hand, an overexpression of

oncogenes caused by other mechanisms such as epigenetic changes can lead to important functional impacts in the cells. Therefore, it is necessary to measure gene expression and protein activity to have a more complete and broader picture of the molecular alterations of cancer cells. For example, overexpression of epidermal growth factor receptor (*EGFR*), which is a receptor tyrosine kinase, can lead to increased cell proliferation, migration, and survival (Sharma, et al. 2007). Although the major mechanism of *EGFR* overexpression is gene amplification, a certain level of overexpression can occur without gene amplification (Flaherty, et al. 2012). Patients who do not have an amplification but a high level of *EGFR* expression can also benefit from an EGFR inhibitor. Gene expression can be affected by its own alteration but also by the alteration of other cell products such as, for example, transcription factors directly regulating it.

Methylation level

Methylation is the process of adding a methyl group at a given site of DNA, RNA or protein which can alter gene expression and protein activity. Methylation can repress gene expression by adding methyl groups in CpG islands located in the gene promoter, the region of the DNA where transcription initiates. There is a negative correlation between DNA methylation and transcriptional activity (Bird 2002). An abnormal methylation pattern such as hypermethylation can lead to inappropriate gene silencing, and can have major impact in cells such as the silencing of tumor suppressor genes in cancer (Baylin 2005). For example, genes such as RARB and APC were found frequently hypermethylated in head and neck squamous cell carcinoma and

were showed to be related to diagnostic and therapeutic markers (Chen, et al. 2007). The DNA methylation can be measured at a genome wide scale or regions specific scale by sequencing technologies.

Genomic alterations detection technologies

There are many different technologies to measure a modification in these different levels. For example, ImmunohistoChemistry (IHC) can be used to detect protein expression levels, Fluorescence In Situ Hybridization (FISH) can be used to detect copy number variations and Sanger sequencing can be used to identify single gene mutations. These techniques are very effective and have shown their capacity in practice but they are limited by the number of detectable alterations. Here we will discuss more in detail high-throughput genomic alterations detecting methods at the whole genome scale.

DNA sequencing

From the discovery of the double helix in 1953 (WATSON and CRICK 1953) to the recent invention of massively parallel DNA sequencing also called next generation sequencing (NGS, Figure 6), the field of DNA sequencing development has a rich and diverse history. Since the sequencing of the first small phage genome, 5,386 bases in length (Hutchison 2007), DNA sequencing has advanced so much that it is possible to sequence the human genome in a few days (Venter, Smith, and Adams 2015). With the great decrease of sequencing cost, NGS makes population-scale sequencing feasible so that the foundation for personalized genomic medicine as a part of

evolutional standard medical care could be established. DNA sequencing has provided us with an efficient way to study somatic mutations by sequencing both the tumor DNA and the normal DNA. The information of somatic mutations can help us understand the development of human cancers or can point to therapeutic treatment strategies. Many collective efforts for gathering cancer genomic information has been made in the last decades such as the Cancer Genome Project (CGP, http://www.sanger.ac.uk/science/groups/cancer-genome-project), The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) and the International Cancer Genome Consortium (ICGC, http://icgc.org/).

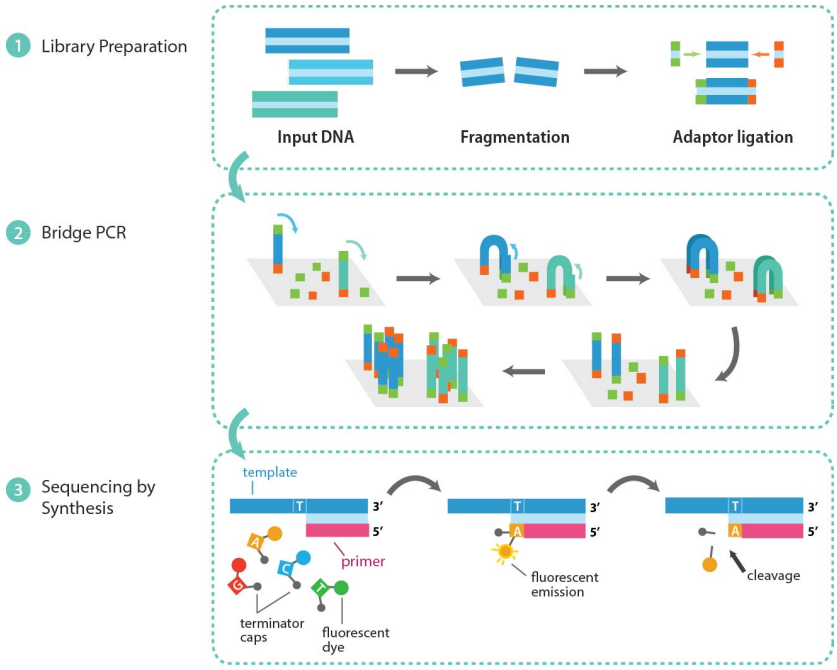


Figure 6. A brief demonstration of NGS process (ref: Illumina)

Based on the region of the sequencing targets, sequencing techniques can be classified into the 3 following widely used methods:

- Whole genome sequencing

Whole genome sequencing refers to the determination of the complete DNA sequence of a human genome at a single time. Whole genome sequence is expensive and time consuming as the aim is to sequence the 6 billion bases of DNA of a human genome. Nonetheless, there are some advantages compared to other less expensive sequencing methods such as whole exome sequencing. First of all, whole genome sequencing allows to not only detect genomic alterations in coding regions but also the non-coding area of the genome, which contains promoters and enhancers that have an important role in gene expression regulation. It also allows the detection of chromosomal structural variation outside of exomes (or at least one end is out of the exome) which cannot be identified if only the exomes are sequenced. Secondly, as whole genome sequencing aims to capture the whole continuous genome, the sequence coverage is often more uniform compared to whole exome sequencing that could generate regions with little or no coverage due to lack of hybridization efficiency of the captured probes. A uniformly distributed sequencing data will facilitate the following variant calling steps.

- Whole Exome sequencing

The exome represents less than 2% of the human genome but contains a majority of known disease-causing variants, making whole exome sequencing a cost-effective alternative to whole genome sequencing. With exome sequencing, the protein-coding portion of the genome is selectively captured

and sequenced. It can efficiently identify variants such as single-nucleotide variants, small insertions and deletions, which open it to a wide range of applications, including population genetics, genetic disease, and cancer genomics. There are much more research projects that focus on whole exome sequencing rather than whole genome sequencing. For example, in the Cancer Genome Altas consortium (TCGA), 10,825 samples were sequenced by whole exome sequencing while only about 10% of these samples were sequenced by whole genome sequencing (http://cancergenome.nih.gov), mainly because for many studies, whole exome sequencing is sufficient for the question asked such as identification of disease biomarkers that are only functionally related.

- Targeted gene sequencing

Another sequencing technique that is widely used is targeted gene sequencing. Targeted gene sequencing only sequence a subset of genes or regions of the genome based on the need of the study, the number of genes of interest can vary from one to hundreds of genes. It allows to focus on specific areas of interest and enables sequencing at much higher coverage levels (>500X) for a gain of time, expense and data analysis while allowing identification of variants with high confidence. In the case of personalized medicine programs, gene panels targeted sequencing is the most widely used technique to discover the genomic alterations of enrolled patients, because it is cost effective to examine only the targetable genes.

CGH arrays

Comparative genomic hybridization array (CGHa) is an effective and popular genome-wide array based technology for characterizing copy number variants (CNVs, Figure 7). CNVs play an important role in generating necessary variation in the population such as common copy-number polymorphisms but also frequently contribute to tumorigenesis while affecting the phenotypes (McCarroll and Altshuler 2007). Characterization of these DNA copy number changes is important for both the understanding of cancer, its diagnosis and its treatment.

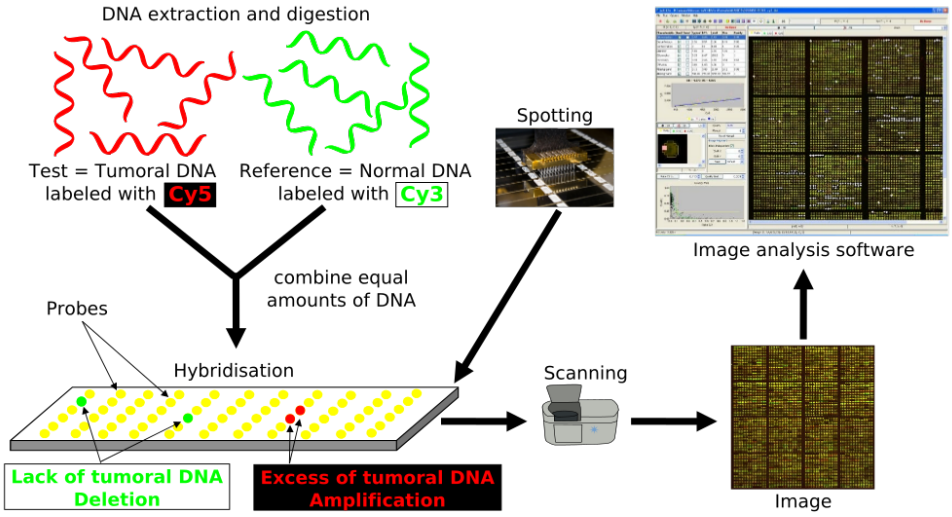


Figure 7. A brief illustration of CGH array process (Wikipedia).

There are two main types of CGH array based on the number of samples to analyze:

- Dual-color array CGH

The most commonly used CGH technique is the dual-color array. The tumor and normal DNA are colored (commonly red and green), mixed and hybridized to the arrays of specific genomic regions to test. A higher intensity for the tumor sample color indicates the gain of DNA in the corresponding regions compared to the normal DNA while a higher intensity for the normal sample color indicates the loss of DNA in the tumor sample in the corresponding regions compared to the normal DNA. The dual-color CGH array is widely used in the medical context to characterize the copy number alterations in the tumor samples compared to the normal samples especially for the enrollment of patients in clinical trials for targeted therapies.

- One-color array

In one-color array CGH experiments, the reference DNA is not tested within the same experiment, but data are compared with a reference dataset either generated in the same laboratory by analyzing a cohort of several normal individuals or provided by the company selling the array(Redon and Carter 2009). One advantage of one-color array is to avoid testing the same reference multiple times when there are more than one tumor sample. But this method is more prone to noise and need more sophisticated bioinformatics software to

interpret the results. This technique is more adequate for the research purpose such as a study of multiple metastatic sites of patients.

RNA sequencing

RNA sequencing is a high throughput sequencing technology at the whole genome level that allows transcriptome profiling of cells. Differing from the existing gene expression microarrays techniques, RNA sequencing method do not require a prior knowledge of the reference genome, therefore can detect non coding RNAs, splice variants and chimeric gene fusions. By doing transcriptome profiling, one can catalogue all types of transcripts, including mRNAs, non-coding RNAs or small RNAs; determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications such as polyadenylation and RNA editing (Marguerat and Bähler 2010); and quantify the change in expression levels of each transcript during development or under different conditions (Wang, Gerstein, and Snyder 2009). RNA sequencing data provided new information and a better characterization of cancer genomics such as, for example, the identification of fusion genes in cancer (Maher, et al. 2009).

Bioinformatics tools

Genomic characterization with modern high throughput technologies has given the cells, the organs and even the individual human beings a brand new appearance. Biological organisms can be translated into numbers and quantifications, which can only be handled by mathematical and computational tools. The application of bioinformatics, in another term computational biology, along with the new high-throughput technologies has helped gaining large amount of information and knowledge at the molecular level about physiological and pathological processes such as cancer. As the size of existing and new sequencing data grow every day, treating, analyzing and modeling genomics data is nowadays one of the most important applications of bioinformatics, especially in the field of genetic diseases. The importance is even largely emphasized when it comes to cancer research because of the extreme complex molecular mechanisms that are involved in the pathology.

Genomic variations detection

To analyze the data produced by high-throughput genomic alteration detection technologies, many bioinformatics tools have been developed based on different technologies and different needs.

Sequencing data

From either DNA or RNA sequence data to the characterization of genomic alterations, 3 main steps are needed: quality control, sequence alignment and variant calling.

- Quality control

High-throughput sequencing machines generate hundreds of millions of sequences in a single run, bad reads can be produced during the sample preparation or the sequencing process such as low-quality reads and contaminating reads or an overrepresentation of certain GC content rich region of the genome due to PCR bias. These errors can cause an inaccurate representation of the genome sequenced if not properly treated and therefore need to be examined and discarded before further use. Tools like FastQC (Andrews 2014) or NGSQC (Dai, et al. 2010) can be used for quality control and tools like FASTX-Toolkit and Trimmomatic (Bolger, Lohse, and Usadel 2014) can be used to discard low-quality reads.

- Sequence mapping

Due to the limitation of sequencing techniques, short segments of DNA (called reads) are sequenced instead of the entire genome. These reads need to be mapped back to the reference genome or transcriptome (for RNA sequencing) to reconstruct the genome being sequenced. Many tools have been developed using different text matching, indexing techniques and graph theory methods to map the reads to a reference genome such as BWA for DNA sequence (Li and Durbin 2010), TopHat for RNA sequence (Kim, et al. 2013) and Novoalign for both DNA and RNA sequence (http://www.novocraft.com/products/novoalign/).

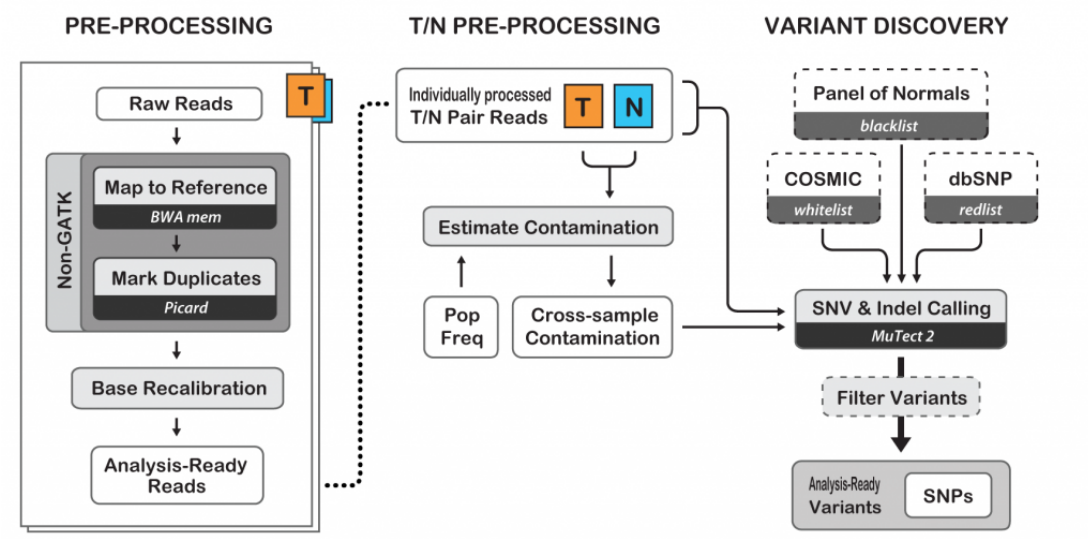


Figure 8. The recommended pipeline for calling somatic mutations using tumor/normal sequencing data (from https://software.broadinstitute.org).

- Variant calling

The most common in silico variant calling application for DNA sequencing in the context of cancer is the identification of somatic mutations. It is therefore essential to identify somatic mutations that presented in tumor cells but not normal cells as they may be drivers of the disease and/or targetable by a drug. Many software are available for the detection of somatic mutations, including Mutect (Cibulskis, et al. 2013), GATK (Figure 8, https://software.broadinstitute.org/gatk/) and VarScan2 (Koboldt, et al. 2012). After the sequences of both normal and tumor DNA have been aligned to the

reference genome, we can detect the positions where the sequenced DNA does not match the reference genome. The variant can either be somatic or germline based on whether or not the tumor sample matches the normal sample. Once a variant is called, a functional annotation is required in order to analyze the probable outcome of the variant. There are databases of known variants such as dbSNP (https://www.ncbi.nlm.nih.gov/SNP/) or polymorphisms and COSMIC (http://cancer.sanger.ac.uk/cosmic) for cancerous somatic mutations. If the variant was not previously documented, it is possible to make a prediction of its functional impact based on the effect of the mutation on the protein sequence. For non-silent mutations, software such as poly-phen (http://genetics.bwh.harvard.edu/pph2/), SIFT(Sim, et al. 2012), or MutationAssessor(Reva, Antipin, and Sander 2011) can be applied to predict the impact of the mutation on the structure and function of the protein product.

CGH array data

As the CGH array methods evaluate the different DNA quantities between the tumor sample and the normal sample (Figure 9), the intensity data representing each probe is first transformed into a log2 relative ratio (LRR) of the intensity signals of the tumor and the normal samples.

- Centralization

To identify copy number variant, the first and critical step is to identify a baseline population of regions that have a copy number of 2. Under the hypothesis that regions that contain copy number changes are relatively small

as compared to the whole genome (except for the cases of extreme unstable genomes where the whole or large part of chromosomes are amplified or lost), the centralization is equivalent to identifying the largest population of regions that have comparable LRR and then set its LRR as the base line of centralization. Method such as EM Expectation Maximization (EM) are effective for this purpose (Commo et al., 2015).

- Segmentation

Segmentation is a method for detecting breakpoints of copy number changes in the genome, which serves as an important step of the CGH array analyses workflow. Segmentation can also reduce the data noise by unifying the copy number of a segment where no abrupt copy changes are detected. Several segmentation methods can be applied such as the Hidden Markov Model (HMM; (Marioni, Thorne, and Tavaré 2006)), the GLAD algorithm (the gain and lost analysis of DNA; (Hupé, et al. 2004)), the CBS algorithm (circular binary segmentation; (Olshen, et al. 2004)) or the HaarSeg algorithm (Ben-Yaacov and Eldar 2008).

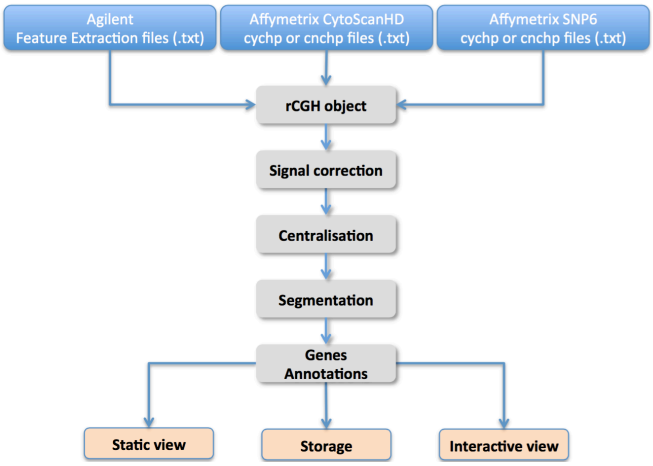


Figure 9. A workflow overview of CGH data analyses (from Commo et al., 2016).

- Copy number variant calling and interpretation

After centralization and segmentation, a threshold has to be set in order to call amplifications or deletions. This threshold is critical because it can generate false positive or false negative calls if improperly chosen, which can have a serious impact on treatment decision in precision medicine trials. For example, only patients with amplification of EGFR can benefit from an anti-EGFR inhibitor (Shen et al., 2014) and a wrong identification of an EGFR amplification may lead to poor drug response. here are many tools to call copy number variants from CGH data as well as visualization for easy interpretation such as CGHnormaliter (Van Houte et al., 2009) or rCGH (Commo et al., 2016).

Modeling genomic alterations in cancer

Somatic mutations in genes can affect the biological functions of their protein products and result in different behaviors in tumors cells from the normal cells. Mathematical modeling is a great tool when the data scale becomes too large to be handled manually. As the high throughput technologies allows the characterization of tumors in large scale, the need of the development and application of mathematical and computational models is greater than ever. The use of mathematical modeling has expended into many different areas of cancer research and has proven of great utility. A few examples are given here.

Cancer phenotype prediction

Accurate prediction of different tumor types has great value in providing better treatment options for cancer patients. The traditional method of cancer classification is based on morphological analysis of the cancer tissue and clinical information of the patients. For some highly heterogeneous cancer such as colorectal cancer (Vermeulen and Snippert 2014), such a classification is difficult to make, which will directly affect the therapeutic strategies and patients' prognostic.

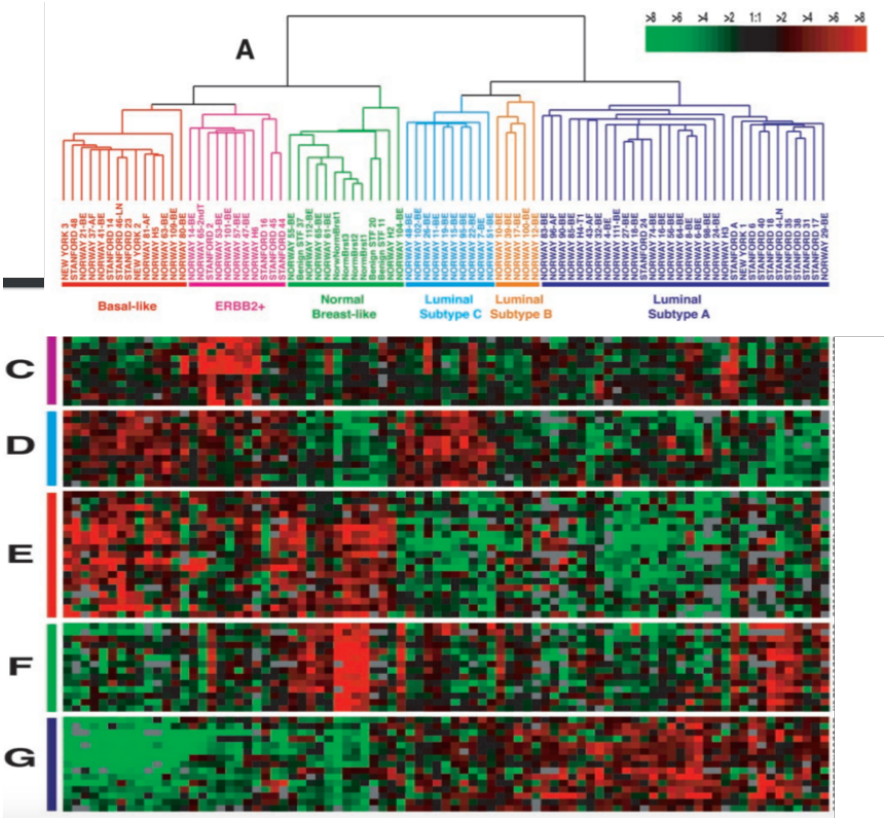


Figure 10. A set of 85 breast cancer samples were divided into five main subtypes by hierarchical clustering based on differences in gene expression. (Sørlie, et al. 2001)

Many studies have shown the utility of classifying tumors using gene expression data and machine learning algorithms such as clustering methods (Sotiriou, et al. 2003; Sørlie, et al. 2001) and tree methods (Dettling and Bühlmann 2003), and also shown their important implication in the medical diagnosis and prognosis. In particular, Sørlie, et al has shown that breast cancer samples can be robustly classified into five or six subtypes with different clinical outcomes based on gene expression data, which is a method that is still wildly used in classifying breast cancer patients (Figure 10).

Driver mutations identification

Driver mutations are mutations that provide an advantage to tumor cell formation and survival. Differ from driver mutation, passenger mutations are the mutations presented in tumors but not necessary for neither the transformation nor the survival of the tumor. In order to identify subgroups of patients who are most likely to benefit from a targeted therapy which a drug designed to inhibit specific molecules, it is critical to distinguish driver mutation from passenger mutations. While an activated mutation in an oncogene or an inactivated mutation in a tumor suppressor gene is probably a driver mutation, the identification of the entire repertoire of driver mutations is not straightforward. First, it is not always easy to predict whether an unknown mutation is an activating or an inactivating mutation based only on the sequence changes. Second, a mutation can be a driver mutation even if it is not inside of a known oncogene, it can either have an impact on some other mechanism involved in cancer or can affect a new oncogene or tumor suppressor gene. Recently, researchers have proposed new methods (Figure 11) to identify driver mutations in a more systematic way using high throughput genomic data such as statistical test for identifying recurrent mutated genes or pathways from an ensemble of somatic mutations (Dees, et al. 2012) or genes with recurrent copy number and structural aberrations (Raphael, et al. 2014). Driver mutations can be used as biomarkers to enable treatment decision and prognosis in cancer.

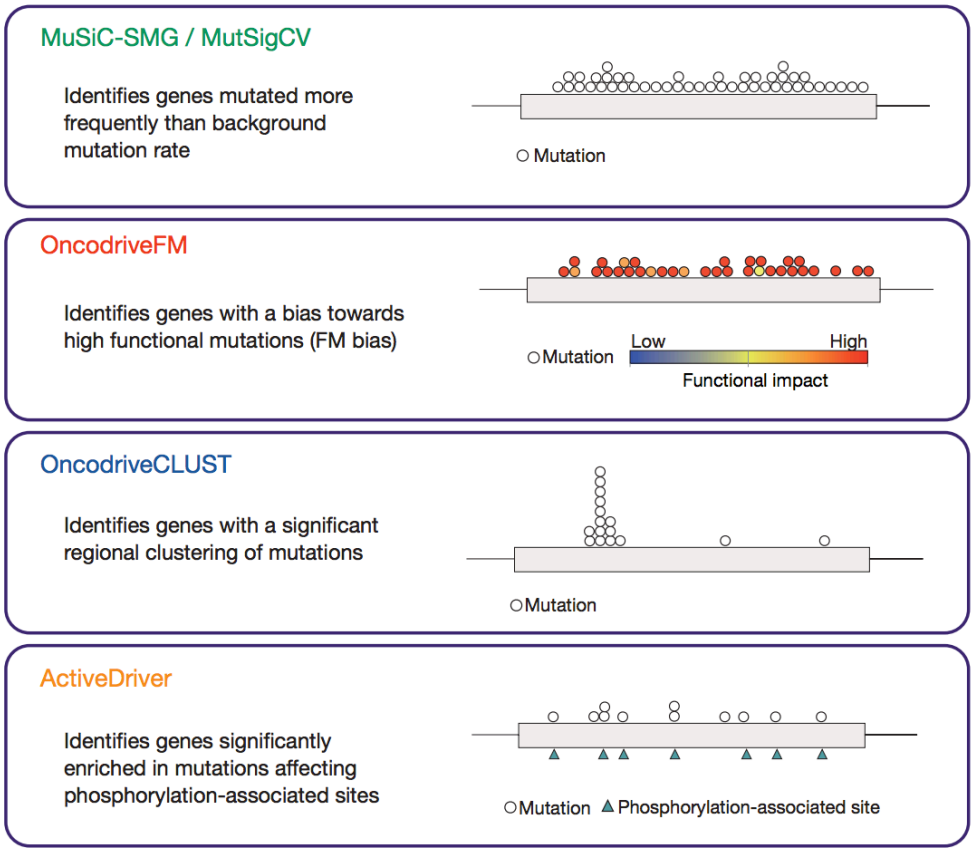


Figure 11. A few popular software for identifying driver mutations from passenger mutations (Marx 2014).

Mutational signature

The detection of mutational signature is a relatively new area that emerged thanks to the accumulation of sequencing data on cohorts of tumors. It is based on a mathematical model for the detection of patterns of somatic mutations in cancer cells that reflect the mutational processes at play in the cancer cells such as for example due to UV exposure. It has been shown that cells acquire preferentially C>T and CC>TT mutations under UV irradiation (HOWARD and TESSMAN 1964). A mutational signature is represented by the distribution of 96 types of mutation (each mutation can take 6 forms: C>A, C>G, C>T, T>A, T>C, T>G, leading to 96 types of mutation when taking into consideration the

nucleotide just before (4 possibilities) and the one just after (4 possibilities) the mutated position) (Figure 12). Studies have shown that by modeling the mutational pattern of a large number of tumors and their clinical information using matrix decomposition method, strong correlations can be observed between some mutational signatures and known mutagens (Alexandrov, et al. 2013). *Alexandrov et al.* identified and validated 21 mutational signatures in human cancer. Most of these signatures have distinct mutational distribution, for example the signature 3 has an even representation of all 96 mutations while the signature 1 is mainly represented by C>T mutations (Figure 12).

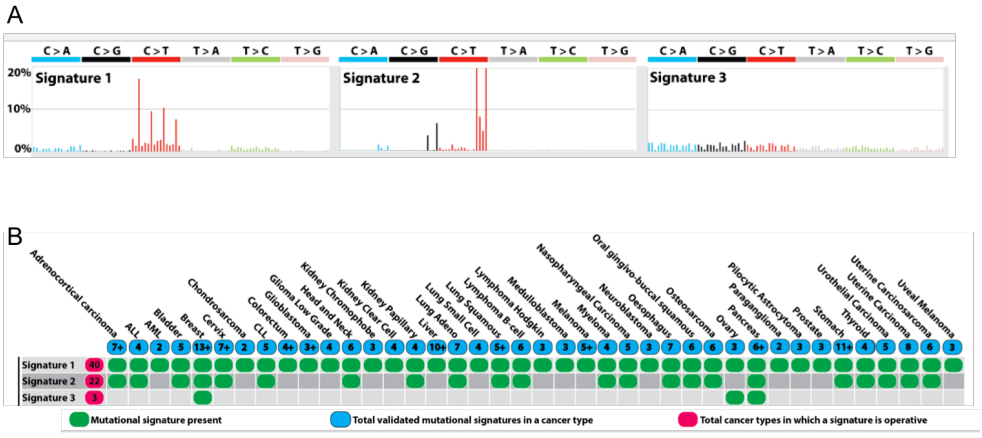


Figure 12. Three mutational signatures illustrated by the contribution of different mutation types (A), the presence of each signature in different cancer types (B) Alexandrov, et al. 2013.

Mutational signatures can obviously reveal the biological processes of the origin of mutation in human but it still remains a research area of the utility of mutational signature in clinical use.

Gene network/interaction discovery

There are about 20,000 genes in a human cell, all these genes work in interaction with other genes to form a network, also called pathways, in order to carry out cellular processes of different functions. It is therefore important to identify various pathways to obtain a coherent global picture of cellular activity. Given the huge number of genes and the complex interactions undergoing in the cells, it is impossible to study pathways at a whole genome scale manually.

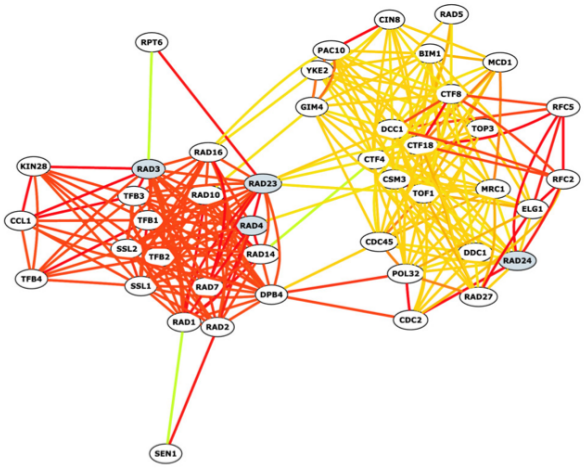


Figure 13. An example of gene network constructed by bioPIXIE using genome-wide data such as microarray expression data and gene-gene co-localization data (Myers, Chiriac, and Troyanskaya 2009).

Computational tools and mathematical models such as Naive Bayes model combined with gene expression data (Segal, Wang, and Koller 2003) or other types of genomic data (Myers, Chiriac, and Troyanskaya 2009) can be used to

discover gene-gene interactions and build functional pathways. In particular, *Myers, Chiriac, and Troyanskaya* have shown that, using their method combine with genome-wide data, they could recover known functional gene networks as well as novel network components that were validated afterward by experiments (Figure 13). This method can help us to better understand cellular biology by identifying unknown components of pathways as well as the relationship between different pathways that involve in different biological changes such as diseases.

Personalized medicine

In all types of cancer, patients with different biological characteristics have different clinical behavior and response to treatment. With the help of large-scale high-throughput detection techniques, we now can collect large amount of molecular data from tumor specimens in genome, transcriptome and proteome level and treat patients based on the molecular characteristics of their tumor with targeted therapies, which is the concept of personalized medicine.

Targeted therapy

Different from chemotherapy that aim at killing cancer cells in the process of replicating their DNA or dividing into new cells, targeted therapy works by modifying the processes that control cell growth, cell division, cell migration and cell signaling. Targeted therapies are small molecules or monoclonal antibodies that can inhibit the activity of the target or prevent it from binding to another protein that it normally activates. Small molecules compounds are typically developed for targets that are located inside the cell because such agents are able to enter cells relatively easily. Monoclonal antibodies, that are relatively large and generally cannot enter cells, are used for targets that are located outside or on the surface of the cell (such as receptors).

There are many targeted therapies under development, in clinical trials or that have been approved by the drug control authorities to treat specific types of cancer. The development of a targeted therapy requires the identification of a

good target that is, a target that play a key role in cancer cell growth and survival.

Tumor profiling

Molecular tumor profiling is a critical step in the context of personalized medicine for identifying and characterizing the unique somatic genomic alterations in patients' tumor so that suitable targeted therapies could be given to the patients.

Based on the targeted therapies available for the corresponding cancer type, different approaches might be appropriate for the profiling. For instance, to test a patient with non-small cell lung cancer for an EGFR inhibitor, a targeted sequencing of the tumor can be suitable to identify a specific activating mutation in EGFR such as exon 19 deletion or L858R mutation in exon 21, in which case the inhibitor Erlotinib has shown better outcomes than traditional chemotherapy (Rosell, et al. 2012). In breast cancer patients, where 3 hormone receptors (the estrogen receptor (ER), progesterone receptor (PR) and erb-b2 receptor tyrosine kinase 2 (ERBB2 or HER2)) are routinely tested for a diagnosis of the cancer type as well as a possible treatment, an IHC staining or FISH assay (for HER2) is often used to test the receptors' expression level in tumor cells. The tumor can then be classified into different subtypes based on the status of these 3 receptors and will help defining the optimal treatment.

In another type of clinical trial where multiple targeted therapies are available, patients can receive a broader test for biomarkers. For example, a subgroup of

patients in clinical trial MOSCATO-01 at Gustave Roussy (Charles Ferté 2014) have received a high throughput genomic analysis (CGH and DNA targeted sequencing of a panel of 74 target genes) to detect biomarkers that finally oriented the patients towards different type of targeted therapies.

In retrospective research studies, the tumor profiling can be done at much larger scale such as whole genome sequencing, whole exome sequencing and RNA-seq to identify new biomarkers.

Matching patients to targeted therapy

The current way of defining the targeted therapy for a cancer patient is rather straightforward: patients are usually given one targeted therapy corresponding to a genomic alteration in a known targetable gene identified in its tumor. In the standard care of certain advanced or metastatic cancer patients, if the patient is a potential benefiter for one targeted therapy, then only the corresponding target is tested and the patient will receive the targeted therapy alone or combined with chemotherapy only if the target is positive, otherwise an alternative treatment is given if available such as immunotherapy. For example, metastatic melanoma patients with a V600E mutation in BRAF are treated with anti BRAF targeted therapy, while patients without this mutation can also benefit from immune therapy (Robert, et al. 2015).

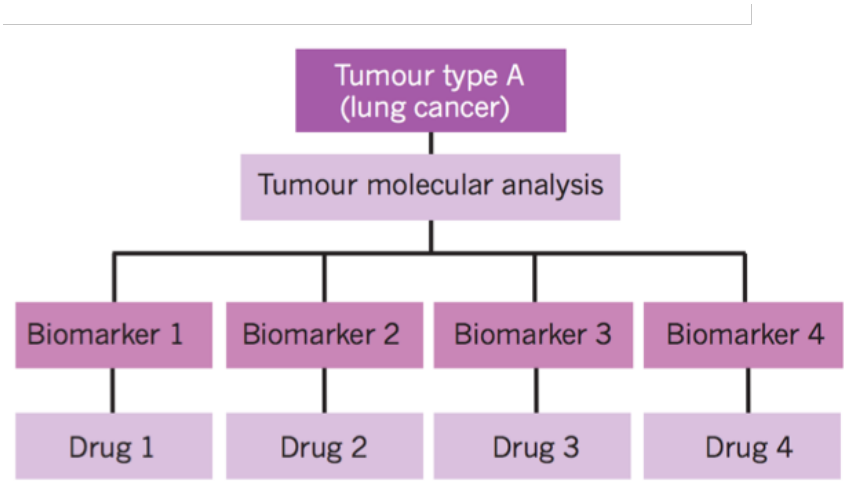


Figure 14. An example of a clinical trial design of one type of cancer in which multiple drugs are available. Patients are tested for biomarker1, biomarker2, biomarker3 and biomarker4, and receive a corresponding drug based on their molecular analysis result (Biankin, Piantadosi, and Hollingsworth 2015).

In personalized medicine programs, multiple targets are tested and patients are redirected towards clinical trials with corresponding targeted therapy (Figure 14). For example, in the clinical trial MOSCATO at Gustave Roussy, multiple targeted therapies are available for patients. The enrolled patients are first tested for mutations in a panel of more than 70 genes and copy number variations of genes to identify biomarkers. Patients are then treated by targeted therapies if they are positive for the corresponding biomarker. Such type of clinical trial offers cancer patients a much bigger opportunity to receive a suitable treatment, but there are still some limitations.

Limitations

Lack of biomarkers/drugs

Despite the promise of personalized cancer treatments, not all types of cancer have personalized treatment options because of lack of promising biomarkers or available approved targeted therapies in the corresponding cancer type (Table 1). Genetic testing for patients and tumor samples can be costly and time-consuming and sometimes disappointing when only a small proportion of patients enrolled are positive for treatable biomarkers (André, et al. 2014).

Table 1. A table of frequently mutated genes in lung cancer and available or in development targeted therapies. Approved targeted therapies are only available for EGFR and ALK alterations; targeted therapies for HER2 and DDR2 showed negative results; targeted therapies for KRAS, BRAF are still under clinical trials; and no targeted therapy's available for targets as PTEN, NRAS(Lovly 2016; Chan and Hughes 2015).

Gene	Alteration	Frequency in NSCLC	Targeted therapies
AKT1	Mutation	0,01	
ALK	Rearrangement	3–7%	Crizotinib (approved), Ceritinib (approved), Alectinib (approved)
BRAF	Mutation	1–3%	Dabrafenib (Phase 2)
DDR2	Mutation	~4%	negative result
EGFR	Mutation	10–35%	Erlotinib (approved), Afatinib (approved), Gefitinib (approved)
FGFR1	Amplification	0,2	Brivanib (Phase 2)
HER2	Mutation	2–4%	Negative result
KRAS	Mutation	15–25%	Selumetinib (Phase 3)
MEK1	Mutation	0,01	
METa	Amplification	2–4%	Onartuzumab (Phase 3), Cabozantinib (Phase 2)
NRAS	Mutation	0,01	
PIK3CA	Mutation	1–3%	Buparlisib (Phase 2)
PTEN	Mutation	4–8%	
RET	Rearrangement	0,01	Cabozantinib (Phase 2)
ROS1 a	Rearrangement	0,01	ASP3026 (Phase 1)

In some cases, even if an oncogenic alteration is identified, drugs for the target are difficult to develop because of the target's structure and/or the way its function is regulated in the cell. One example is RAS, a signaling protein that is mutated in as many as one-quarter of all cancers but it has not been possible to develop inhibitors of RAS signaling with existing drug development technologies (https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/targeted-therapies-fact-sheet). Moreover, a large number of cancer drugs have not been linked to specific genomic alterations that could be used as biomarkers to specify their selective therapeutic effectiveness (McDermott and Settleman 2009). Systematic study is needed to discover the deeper relationships

between biomarkers and drug response in order to optimize the utility of existing drugs as well as the development of new drugs and the identification of new biomarkers.

Drug resistance

A subgroup of patients can benefit from therapies targeting cancer driver mutations identified from the characterization of their tumor. For example, from 10% to 30% of patients with non-small cell lung cancer have mutations in EGFR, of which 75% have been shown to respond to the tyrosine kinase inhibitors (TKIs) Gefitinib (Maemondo, et al. 2010). However, almost all of the patients that initially benefit from the EGFR inhibitors eventually develop resistance and relapse, the median progression free survival (PFS) after treatment with Gefitinib in patients has been shown as less than a year. Some mechanisms of resistance have already been identified such as the co-existence of the abnormal activation of HER2, BRAF, MAPK1 or PIK3CA but many of the remaining mechanisms are still unknown *(Stewart, et al. 2015,* Figure 15*)*.

Drug resistance to targeted therapy is generally classified into two types: 1) early intrinsic resistance referring to a lack of treatment response and 2) late acquired resistance referring to disease progression after initial response. The intrinsic resistance may be caused by the differential drug sensitivities across various genomic alterations in the target, for example, cells with different EGFR alterations do not respond equally to the TKI inhibitors, the exon 19 deletion and L858R mutation are associated with high sensitivity to TKIs while exon 20

insertion, which are typically located after the C-helix of the tyrosine kinase domain of EGFR, seem to be resistant to EGFR inhibitors (Yasuda, Kobayashi, and Costa 2012), which demonstrated the importance of identifying the driver mutations not only by gene but also by mutation. Another scenario of an intrinsic resistance can be explained by the co-existence with the sensitizing genomic alteration of a secondary genetic alteration.

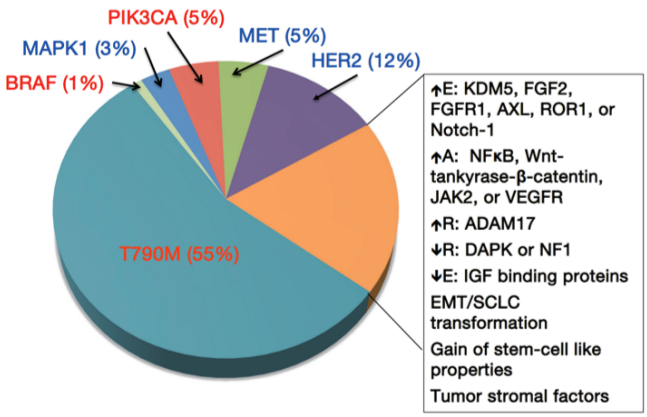


Figure 15. Summary of mechanisms of resistance to first generation EGFR TKIs. Red text represents mutations; blue text represents amplifications. ↑E increased expression; ↑A, increased activation; ↑R, up-regulation; ↓R, down-regulation; ↓E, loss of expression (Stewart, et al. 2015).

The late acquired resistance may be the result of a secondary mutation in the primary oncogene so that targeted therapy no longer interacts well with it. For example, lung cancer patient with EGFR mutation develop resistance to anti EGFR TKI by acquiring a secondary mutation T854A in EGFR (Bean, et al. 2008). The acquired resistance can also be due to the selection of clones harboring resistant mutations initially present in the tumor at low percentage and expanding during treatment. These events might be predicted based on

the genomic profile of the patients through bioinformatics tools. For example, the EGFR T790M mutation has been associated with intrinsic resistance to EGFR TKI therapy in patients with activating EGFR mutations. Even at a very low frequency of the T790M mutation in the tumor, the pressure from targeted agents may select the T790M clones and lead to an overall resistance (Stewart, et al. 2015).

Overcome limitations

In order to overcome these two main limitations of personalized medicine, many efforts have been done in the following directions:

- New drugs and new targets

Great efforts in developing more specific and more efficient inhibitors have been made leading to the development and use of second generation inhibitors. For instance, the first generation of mTOR inhibitors, such as rapamycin, is being replaced by its second generation, such as AZD8055. Meanwhile, countless efforts are being made to develop third generation mTOR inhibitor that could overcome drug resistance (Rodrik-Outmezguine, et al. 2016). At the same time, attempts to identify new mechanisms of response to existing drugs as well as new potential targets are being made constantly from large scale genomic profiling analyses to their in vitro and in vivo validations (Nicolas Goossens 2015).

- Combine multiple inhibitors

One common practice to overcome drug resistance is to combine a targeted therapy with one or more chemotherapy drugs. For example, the targeted therapy trastuzumab has been used in combination with docetaxel, a traditional chemotherapy drug, to treat women with metastatic breast cancer that overexpress the protein HER2/neu (Burris 2001). The other option is to give the patients combinations of targeted therapies based on their genomic profiles. Many recent studies have demonstrated the benefits of combining targeted therapies in overcoming resistance that arises through secondary mutations in the driver genes. The combination can either be 2 or more drugs targeting the same altered signaling pathway or targeting different pathways based on the patients' genomic profiles. For example, treating melanoma patients with a BRAF V600E mutation with the combination BRAF inhibitor (Vemurafenib) / MEK inhibitor (Cobimetinib) as compared to the single BRAF inhibitor, can slow the development of resistance and therefore the disease progression (Larkin, et al. 2014). Some preclinical studies and early-phase clinical trials have also shown potential antitumor activities of the combination of MEK inhibitor and PI3K inhibitor with manageable safety and toxicity for patients with RAF or RAS mutations (Jokinen and Koivunen 2015).

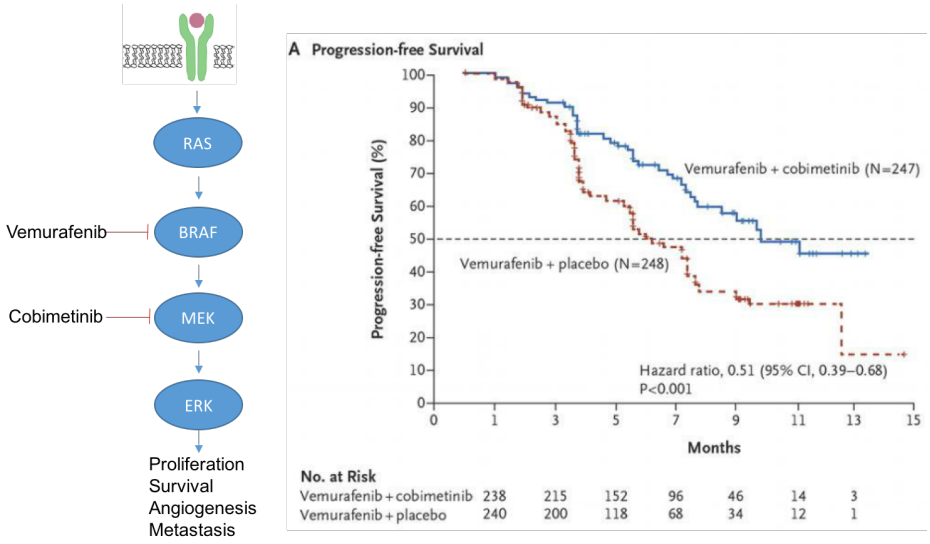


Figure 16. Dual therapy of BRAF inhibitor and MEK inhibitor increased progression-free survival when compared to BRAF inhibitor monotherapy. (Larkin, et al. 2014)

In order to find the right combination of drugs, it is essential to first identify the most prevalent targets as well as their interaction in tumor cells in the response to the drugs. But the contribution of the combinations of genomic alterations to the cancer phenotype and response to treatment is still poorly understood. Plus, most candidate gene association studies typically assess the effects of candidate genes independently of each other making the identification of gene-gene interactions in the discovery of biomarkers to drug response difficult.

Drug testing in cell lines is one of the initial steps in drug development and allows the access to a large number of Drug testing in cell lines is one of the initial steps in drug development, it allows the access to a large number of potential drugs, at different concentrations. It permits to analyze the drugs

mode of action, test the combination of drugs as well as offer the possibility of studying the association of signaling pathways to therapeutic response. The characterization of cancer cell lines with high throughput sequencing technologies is a powerful tool to provide insight into the pharmacogenomics of tumor cells.

Modeling pharmacogenomics data of cell lines

Rationale for studying cancer through cell lines

Human cancer cell lines are biological models that represent a wide range of tumor biology with great diversity. A cancer cell line is a collection of immortalized cancer cells of one tumor. Although there is a debate about whether the cell lines are representative of the original tumor, a high, but not perfect, similarity between the original tumor and the cancer cell lines derived from it can be observed at the level of genomic alterations and gene expression (Masters 2000). We can discover the biological mechanistic and connect the genomic alterations to drug sensitivity through experimental manipulation of cancer cell lines such as deep analysis at the genomic level or drug response at different concentrations so that further studies and eventually rationale clinical trials could be designed to test these hypotheses. Cancer cell lines are widely used in research for studying the biology of cancer and testing cancer treatments for the following reasons: 1) Cell lines provide an almost unlimited supply of cells with similar genotypes and phenotypes; 2) Cell lines are easy to manipulate (cell lines can be genetically or epigenetically altered using demethylation agents, siRNA, expression vectors and drugs) and easily molecularly characterized; 3) experiments can be reproduced in the same condition.

Two pharmacogenomics projects, the “Genomics of Drug Sensitivity in Cancer” (GDSC), (Garnett, et al. 2012) and the “Cancer Cell Line Encyclopedia” (CCLE), (Barretina, et al. 2012), collectively characterized more than 1,000 cell lines tested with more than 150 anti-cancer drugs for their sensitivity, providing

valuable comprehensive molecular characterization and drug response of the cancer cells of different origins. This molecular characterization gives the opportunity to study the relationship between genomic features and drug responses, and identify the genomic biomarkers of drug sensitivity.

- GDSC

“Genomics of drug sensitivity in cancer “characterized 1,074 cell lines of different types of cancers of different origins by whole exome sequencing (Agilent sureselectXT human all exon 50Mb bait set), gene expression (Affymetrix human genome U219 array), copy number alterations (Affymetrix SNP6.0 array), DNA methylation (Illumina human methylation 450 array), gene fusion (targeted PCR sequencing or split probe FISH analysis) and microsatellite instability. Most of these cell lines were also tested at different concentrations for 138 drugs, including chemotherapies and targeted therapies.

- CCLE

“The Cancer Cell Line Encyclopedia” (CCLE) enables predictive modeling of anticancer drug sensitivity, with a collection of gene expression (Affymetrix U133 plus 2.0 arrays), copy number (Affymetrix SNP 6.0 array), point mutations (targeted sequencing) of 947 human cancer cell lines with pharmacologic profiles for 24 anticancer drugs (8-dose drug sensitivity assay) across 479 of the cell lines.

Predictive models to identify biomarkers

One of the most exciting opportunities presented by large scale genomic data of a large number of cancer cell lines is the possibility to build analytic models to study cancer biology. In both original database publications, a predictive model, elastic net regression was applied to identify biomarkers that related to the drug response. The use of this model in both studies allowed interesting discoveries about how genomic features related to the drug response. For example, in the study of GDSC, EWS-FLI1 gene translocation was identified as a sensitivity biomarker to PARP inhibitors in Ewing's sarcoma cells and validated in mouse models.

Elastic net

Elastic net (Zou and Hastie 2005) is a regularized regression method that linearly combines the l_1 penalty (LASSO, least absolute shrinkage and selection operator) and l_2 penalty (ridge). By using a parameter alpha that takes value between 0 and 1, elastic net adjusts the weight of the LASSO and ridge penalization based on the data.

For a given response $y = (y_1, y_2, \dots, y_n)$ and predictors $X = (X_1, X_2, \dots, X_p)$ where $X_j = (x_1, x_2, \dots, x_n)', j = 1, 2, \dots, p$, a regression problem is to estimate the parameters $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ in:

$$y = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p \qquad (1)$$

An elastic net model is to solve equation (1) by solving equation (2) which is a penalized form of equation (1);

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta| \quad (2)$$

where, λ_1, λ_2 are non-negative and $|\beta|^2 = \sum_1^p {\beta_j}^2$, $|\beta| = \sum_1^p |\beta_j|$.

Let $\alpha = \frac{\lambda_1}{(\lambda_1 + \lambda_2)}$, then solving equation (2) equals to optimize:

$$\beta = argmin\ (|y - X\beta|^2)\ on\ \beta, \quad (3)$$
$$subject\ to\ (1 - \alpha)\ |\beta| + \alpha|\beta|^2 \leq t\ for\ some\ t$$

Elastic net has been shown to be suitable for the p>>n problem where the number of predictors are much larger than the number of observations, and for the case where the predictors are highly correlated (Zou and Hastie 2005). Given these two properties of elastic net, it is suitable for large scale genomic data such as gene expression data where the number of predictors can be as large as more than 20,000 genes and some of the genes are strongly correlated.

Random forest

Random forest (Breiman 2001) is a popular supervised machine learning algorithm made of an ensemble of decision trees (Figure 17). Many studies have shown its capability in prediction and classification (Bienkowska, et al. 2009; Riddick, et al. 2011; Stetson, et al. 2014). As the decision tree is a pure machine learning method with no probabilistic assumption about the input or output variables, no particular pretreatment is needed before using the random

forest model. Unlike most of the probabilistic machine learning models, there is no assumption about the relationship between the input variables therefore they can be independent, linearly correlated or non-linearly correlated. It is therefore adapted to the situation where no assumption can be made about the relationship between the input and the output variables, or the relationship between the input variables, which is the case in the genomic alteration-drug response scenario.

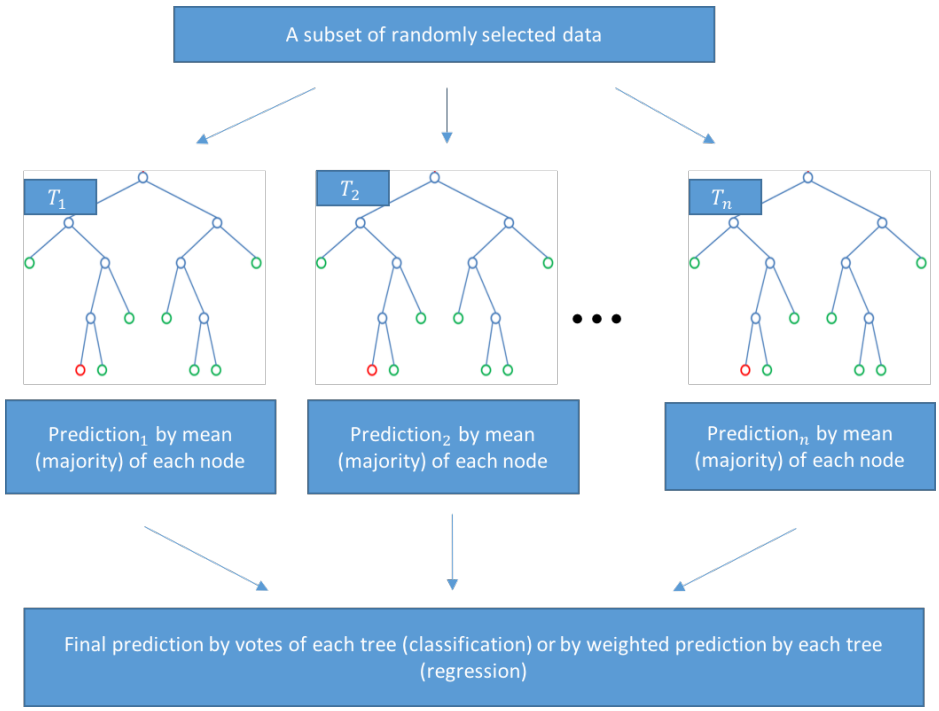


Figure 17. Framework of Random Forest.

There are two main ideas in Random Forest (RF), one is to build trees with bootstrapped data set (also called bagging), the other is to randomly select a subset of predictive variables to test for best split. These two characteristics of RF give it unique qualities in predictive modeling. Indeed, the use of bagging in

the construction of the predictive model allows the predictive error to be estimated simultaneously, thus no additional cross validation is required. By performing random selection of data set to use and predictive variable to split, RF build a set of distinct trees so that overfitting can be avoided. Importantly, RF assesses a variable importance by calculating the difference of predictive accuracy after permuting the variable value, the bigger the decrease of predictive accuracy, the more important the variable is. This important feature is extremely suitable for identifying predictive biomarkers.

Although random forest does not have complex parameters to tune before running, there are still some to fix, such as the number of trees to grow (tradeoff between performance and cost), the number of observations to use to build a tree in the bagging procedure, the number of input variables to use at each split (the bigger the number, the bigger the impact of correlated input variables in the model). Apart from these parameters of RF, a few parameters (ex; matrix for evaluating the split) of the decision tree can vary based on the nature of the data set and the purpose of the model.

Classification and regression trees

Decision tree learning is a nonlinear data mining method that can predict the value of a target variable based on several input variables. There are two main types of tree learning models based on the nature of the target variable: classification tree with a categorical target variable and regression tree with a continuous target variable. Decision tree model is a predictive model that matches each observation to series of decisions based on input variables. It is

one of the predictive modeling approaches most used in statistics, data mining and machine learning. In these tree structures, the ensemble of the observations is called the root, leaves represent categorical labels or observations with a certain range of the split input variable and branches represent the input variables that lead to those splits. The node to split is called the parent node and the nodes after the split are called daughter nodes. A tree learning objet can be built by splitting the data into different subgroups (nodes) based on one input variable at each time. A tree is binary when the parent nodes are split into two daughter nodes at each time based on one variable (for example, a categorical variable with two categories or a continuous variable split at a cutoff value). Binary tree is the most frequently used tree model.

Split criteria

Based on the type of trees and implementations, there are different ways to evaluate a split so that the algorithm can decide which variable to split on and associate an importance to the variable used. The main goal is to split the node into more homogeneous nodes regarding the output variable. For a regression tree with a continuous output variable, the most frequent method to evaluate a split is to evaluate the variance in the node before the split and the one after. One common way is to calculate the difference of the sum of squared error (SSE) between the parent node and the sum of the two daughter nodes and choose the variable that leads to the biggest decrease in SSE. In the case of classification tree, entropy, information gain and Gini index are commonly used.

One other approach is to use corrected statistical tests (i.e. permutation-based significance tests) as splitting criteria by computing a p-value for each input variable in the selection of best variable to split. This approach can reduce bias due to the different nature of the input variable, for example, continuous variables tend to be selected more often than categorical variables and categorical variables with more categories tend to be selected more often than the ones with less categories. The computing cost is however much higher than non-statistical based splitting criteria, thus the statistical based split criteria is less used in practice.

Research objectives

The aim of this thesis is to better interpret cancer pharmacogenomics by identifying predictive genomic biomarkers of targeted therapies drug response using mathematics and informatics tools. To build robust predictive models, we need to ensure that the genomic data used in the model is accurate and that the predictive models are adequate.

The research presented in this thesis therefore covers these two aspects. First, I focused on the development of a pipeline to improve the performance of somatic mutation calling using whole exome sequencing in the context of contaminated normal DNA source (whole blood). It is to our knowledge the first study to provide an analytical solution to correct for this bias.

Next, I focused my research on the modeling of anti-cancer drug response using genomic, pharmacological and clinical information for identifying new biomarkers of therapeutic response. For this purpose, we integrated genomic data including gene expression profiles, mutations and copy number variations

in a two-step predictive model in order to identify single and pairs of predictive biomarkers of targeted therapies response. The proposed model showed an increased sensitivity in identifying the direct targets of the drugs compared to other predictive models tested and discovered new predictive biomarkers strongly related to drug response that could, when validated, be used as new guides for therapeutic decision in precision medicine programs.

Results

cmDetect -- ctDNA mutation calling method for whole exome sequencing of tumor/whole blood samples

Somatic mutation identification using whole exome sequencing data

As discussed previously, identifying somatic mutations in patients' tumor using whole exome sequencing is a useful way to characterize and study the genetics of tumors. In order to identify somatic mutations for cancer patients, both the tumor DNA and germline DNA have to be sequenced and compared. For solid tumors, the source of tumor DNA is often a biopsy of the tumor, while the source of germline DNA can be a biopsy of normal tissue or the peripheral blood. As the peripheral blood is relatively easy to collect, it is often the choice of the source of germline DNA.

False negatives caused by cell-free circulating tumor DNA in whole blood

Normal or tumor cells release DNA in the blood stream when they are under apoptosis. As tumor cells normally have a more active cell reproduction and apoptosis, they can release an important amount of cell-free DNA regardless of the relative size of the tumor as compared to the whole human body, especially for the cancer patients who went through therapies (radiotherapy, chemotherapy, and other cancer treatments cause cell death by apoptosis). A higher than normal concentration of circulating DNA may already be detected even in early stage of cancer where there is little cell death (van der Vaart and Pretorius 2007). As the cancer burden increases, so does the rate of cell death and the amount of proliferating cancer cells, with a concomitant increase in

release of fragments of DNA. This cell-free DNA that came from the tumor is called circulating tumor DNA (ctDNA) (Schwarzenbach, Hoon, and Pantel 2011).

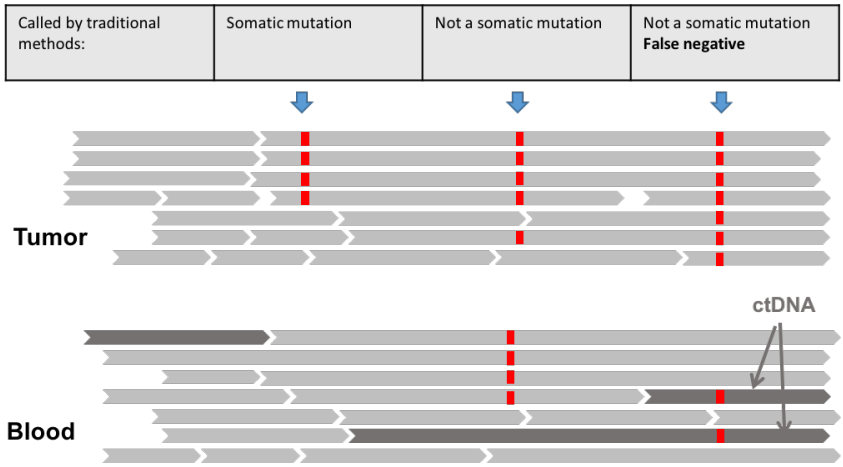


Figure 18. ctDNA in peripheral blood prevent accurate identification of somatic mutations using whole exome sequencing data

Numerous publications have shown ctDNA as a noninvasive biomarker of tumor in clinical medicine (Gordon 2016; Ai, et al. 2016; Cheng, Su, and Qian 2016; Schwarzenbach and Pantel 2015). As the same time, the presence of ctDNA can be seen as a source of contamination of the germline DNA, especially when the DNA extracted from the blood is used as a control. In that case, the detection of somatic mutation using whole blood as the source of germline DNA can cause false negatives, i.e. true somatic mutations filtered out as they are also detected in the germline DNA, with current somatic mutation detection software. Facing this problem, we developed a method called cmDetect (available on GitHub https://github.com/yufu2015/cmDetect) to run as a complementary tool to traditional somatic callers in order to avoid false negatives (Figure 18).

Integrated Systems and Technologies

Cancer Research

Improving the Performance of Somatic Mutation Identification by Recovering Circulating Tumor DNA Mutations

Yu Fu[1,2], Cécile Jovelet[3], Thomas Filleron[4], Marion Pedrero[1,5], Nelly Motté[3], Yannick Boursin[6], Yufei Luo[6], Christophe Massard[7], Mario Campone[8], Christelle Levy[9], Véronique Diéras[10], Thomas Bachelot[11], Julie Garrabey[12], Jean-Charles Soria[1,2,5,7], Ludovic Lacroix[1,3], Fabrice André[1,2,7], and Celine Lefebvre[1]

Abstract

DNA extracted from cancer patients' whole blood may contain somatic mutations from circulating tumor DNA (ctDNA) fragments. In this study, we introduce cmDetect, a computational method for the systematic identification of ctDNA mutations using whole-exome sequencing of a cohort of tumor and corresponding peripheral whole-blood samples. Through the analysis of simulated data, we demonstrated an increase in sensitivity in calling somatic mutations by combining cmDetect to two widely used mutation callers. In a cohort of 93 breast cancer metastatic patients, cmDetect identified ctDNA mutations in 54% of the patients and recovered somatic mutations in cancer genes *EGFR*, *PIK3CA*, and *TP53*. We further showed that cmDetect detected ctDNA in 89% of patients with confirmed mutated cell–free tumor DNA by plasma analyses ($n = 9$) within 46 pan-cancer patients. Our results prompt immediate consideration of the use of this method as an additional step in somatic mutation calling using whole-exome sequencing data with blood samples as controls. *Cancer Res; 76(20); 1–8. ©2016 AACR.*

Introduction

Mutations acquired in a patient's tumor can be identified with deep sequencing with either targeted sequencing, where only the informative mutations for the clinical practice are screened, or whole-exome sequencing, to get a broader picture of the genetics of the tumor. As opposed to targeted sequencing, whole-exome sequencing is applied not only on the tumor DNA but also on the germline DNA obtained from a normal cell population of the patient to differentiate tumor acquired, or somatic, mutations from patient's germline variations or polymorphisms. Current bioinformatics methods for somatic mutation identification from whole-exome sequencing data are designed to accurately identify DNA variations, nucleotide substitutions, and small insertions or deletions, in the tumor DNA that are not detected in the germline DNA (1). However, for solid tumors, the germline DNA is often extracted from peripheral whole blood that may contain cell-free tumor DNA (cfDNA). cfDNA is composed of cell-free fragments of DNA coming from the patient's tumor and circulating in the bloodstream (2) and has been associated to tumor burden (3). Therefore, DNA extracted from cancer patients' blood may contain various levels of tumor DNA also called circulating tumor DNA (ctDNA). If the level of ctDNA in the blood is high enough, there is a possibility for some mutated DNA fragments to be detected by whole-exome sequencing of the DNA extracted from whole blood sample. This becomes a problem when these ctDNA mutations are detected in whole blood samples used as normal controls to distinguish somatic from germline events, therefore preventing the accurate determination of somaticity. The existence of cfDNA in the peripheral blood has already been reported for various cancers, with a variable contribution ranging from <0.1% to >10% of the DNA molecules (4–8). ctDNA is usually considered as the mutated fraction of cfDNA and was recently shown to reach up to 88% of cfDNA in metastatic patients by sequencing analyses (9). However, the impact of this contamination was rarely taken under consideration in somatic variants calling and whole blood samples are still widely used as normal samples in sequencing studies for solid tumors. As an illustration, in a total of 9,202 solid tumor samples with whole-exome sequencing data available in The Cancer Genome Atlas project (TCGA) through the Cancer Genomics Hub (https://cghub.ucsc.edu/), 7,257 cases only had blood sample as normal control (79%). In particular, in a set of 829 breast carcinomas, 715 (86%)

[1]INSERM Unit U981, Gustave Roussy, Villejuif, France. [2]Faculté de Médecine, Kremlin-Bicêtre, Université Paris Sud, France. [3]Department of Medical Biology and Pathology, Translational Research Laboratory and BioBank, Gustave Roussy, Villejuif, France. [4]Biostatistics Department, Institut Claudius Regaud, IUCT-Oncopole, Toulouse, France. [5]Drug Development Department (DITEP), Gustave Roussy, Villejuif, France. [6]Bioinformatics Core Facility, Gustave Roussy, Villejuif, France. [7]Department of Medical Oncology, Gustave Roussy, Villejuif, France. [8]Department of Medical Oncology, Institut de Cancérologie de l'Ouest, Nantes, France. [9]Department of Medical Oncology, Centre François Baclesse, Caen, France. [10]Department of Medical Oncology, Institut Curie, Paris, France. [11]Department of Medical Oncology, Centre Léon Bérard, Inserm U1052, Lyon, France. [12]R&D, UNICANCER, Paris, France.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

cmDetect is available online: https://github.com/yufu2015/cmDetect.

Corresponding Author: Celine Lefebvre, Gustave Roussy, 114 rue edouard vaillant, Villejuif 94800, France. Phone: 33142114827; Fax: 33142114827; E-mail: celine.lefebvre@gustaveroussy.fr

doi: 10.1158/0008-5472.CAN-15-3457

had only blood sample as control, 32 (3%) had both blood and tissue as normal controls, and 82 (9%) had only tumor-adjacent tissue as normal control (Supplementary Table S1). To understand the level of contamination of ctDNA in whole-exome sequencing of cancer patients' whole blood DNA and the extent to which it affects somatic mutation calls, we developed a method for the systematic identification of ctDNA mutations from a set of tumor/blood samples. The method we propose should be executed as an additional step to classic somatic mutation analysis for recovering bona fide somatic mutations for which the allele read count or frequency in the blood is higher than expected.

Materials and Methods

Simulated data and performance assessment

We first retrieved the whole exome sequencing data of 99 individuals of CEU population [Utah Residents (CEPH) with Northern and Western European Ancestry] in 1000 Genomes project phase III dataset (Supplementary Table S2). The bam files were then remapped to the reference hg19 (using bwa). To create a set of samples with comparable coverage, we included 66 samples with a mean coverage between 80 and 200.

Simulated tumor samples. We randomly assigned a number of variants ranging from 50 to 500 (SNPs and indels) to each sample with a mutant allele frequency following a Gaussian distribution (10). Because of the consideration of tumor heterogeneity and tumor sample purity, the mean of the Gaussian distribution was set between 0.2 and 0.45. We then chose all the variants from the COSMIC database with CNT ≥ 2 and added the variants to the bam files using Bamsurgeon (11).

Simulated contaminated whole blood samples. We first estimated ctDNA concentration in whole blood DNA from a set of 117 pan-cancer patients, where the cfDNA concentration in the plasma ($C_{\text{cfDNA,plasma}}$) and the allele frequency of somatic mutations identified in the cfDNA ($AF_{\text{somaticM,plasma}}$) were available (Supplementary Table S14; ref. 9). The ctDNA concentration in whole blood was calculated as: $C_{\text{ctDNA, whole blood}} = \frac{C_{\text{ctDNA, plasma}} \times 1\text{ mL}}{Q_{\text{DNA, whoole blood}}}$, where $C_{\text{ctDNA,plasma}} = \max(AF_{\text{somaticM,plasma}}) \times C_{\text{cfDNA,plasma}}$ and $C_{\text{DNA,whole blood}} \sim 15 - 60\ \mu\text{g/mL}$ (DNA/RNA Mini Kit protocol, Qiagen), $C_{\text{plasma,whoole blood}} = 55\%$ and $Q_{\text{DNA, whoole blood}} = C_{\text{DNA,whoole blood}} \times (1\text{ mL}/C_{\text{plasma, whoole blood}})$. On the basis of this estimation, we found that the maximum concentration of ctDNA in whole blood was above 2% (Supplementary Table S14). We then set the percentage of contamination to a Gaussian distribution of mean equals to 1.38% (for final values ranging from 0.08% to 2.99%). To simulate the reality where the normal samples are usually sequenced at a lower coverage compared with the tumor samples, and to avoid having the tumor sample greatly identical to the simulated whole blood sample, we mixed the tumor and the normal samples at half of the corresponding proportion to create a mixed sample of lower mean coverage (0.5 $\times$ mean coverage of tumor sample).

Performance. Somatic variants were identified with Mutect and Varscan2 with different sets of filters (maximum variant allele frequency ranging from 0 to 0.15 and maximum variant supporting reads ranging from 0 to 10; Supplementary Table S5). We then compared the performance of traditional somatic variant callers alone versus traditional somatic variant callers plus cmDetect by evaluating the Recall, Precision, and *F*-score (11). False negatives were defined as true somatic mutations incorrectly filtered out by mutations callers because of high detection level in the blood sample. The confidence interval (CI) of each call with different filters was calculated on the basis of a binomial distribution using Clopper–Pearson method (12).

Metastatic samples

We used 86 tumor–normal sample pairs from patients included in the SAFIR01 trial (NCT01414933; ref. 13), and 67 tumor–normal pairs from the MOSCATO trial (NCT01566019). Of these, 93 samples were from metastatic breast cancer and were sequenced on a Hiseq sequencer, whereas the remaining 60 were from a panel of different cancers and were sequenced on a NextSeq sequencer. Tumor DNA was extracted from frozen tissue from a biopsy sample taken in the context of the corresponding trial. A surgical pathologist reviewed the samples for diagnosis purpose and assessed tumor cell content (Supplementary Table S6) before whole-exome sequencing. The average tumor cell content was 61% for the 93 metastatic breast cancer samples and 52% for the 60 pan-cancer samples. Normal DNA was extracted from whole blood, taken at the same time as the biopsy. All patients gave their informed consent for translational research and genetic analyses of their germline DNA.

Whole exome sequencing. Genomic DNA was captured using Agilent in-solution enrichment methodology with their biotinylated oligonucleotides probes library (SureSelect Human All Exon v5–50 Mb; Agilent), followed by paired-end 75 bases massively parallel sequencing on Illumina HiSeq 2500 or NextSeq 500 sequencer.

HiSeq 2500. Sequence capture, enrichment, and elution are performed according to manufacturer's instruction and protocols (SureSelect; Agilent) without modification. Briefly, 600 ng of each genomic DNA are fragmented by sonication and purified to yield fragments of 150 to 200 bp. Paired-end adaptor oligonucleotides from Illumina are ligated on repaired, A-tailed fragments then purified and enriched by four to six PCR cycles. Five hundred nanograms of these purified Libraries are then hybridized to the SureSelect oligo probe capture library for 24 hours. After hybridization, washing, and elution, the eluted fraction is PCR amplified with 10 to 12 cycles, purified and quantified by qPCR to obtain sufficient DNA template for downstream applications. Each eluted-enriched DNA sample is then sequenced on an Illumina HiSeq 2500 as paired-end 75b reads. Image analysis and base calling is performed using Illumina Real Time Analysis Pipeline version 1.12.4.2 with default parameters.

NextSeq 500. Sequence capture, enrichment, and elution are performed according to manufacturer's instruction and protocols (SureSelect; Agilent) without modification except for library preparation performed with NEBNext Ultra Kit (New England Biolabs). For library preparation, 600 ng of each genomic DNA are fragmented by sonication and purified to yield fragments of 150 to 200 bp. Paired-end adaptor oligonucleotides from the NEB Kit are ligated on repaired, A-tailed fragments then purified and enriched by eight PCR cycles. A total of 1,200 ng of these purified libraries are then hybridized to the SureSelect oligo probe capture library for 72 hours. After hybridization, washing, and elution, the

eluted fraction is PCR amplified with nine cycles, purified, and quantified by qPCR to obtain sufficient DNA template for downstream applications. Each eluted-enriched DNA sample is then sequenced on an Illumina NextSeq 500 as paired-end 75b reads. Image analysis and base calling was performed using Illumina Real Time Analysis (RTA 2.1.3) with default parameters.

Somatic mutations calling. Fastq files were aligned to the reference genome hg19 with the Burrows-Wheeler Alignment tool (BWA) 0.7.5a mem algorithm (14). After alignment, the BAM files were treated for PCR duplicate removal then sorted and indexed with samtools (15) version 0.1.19 (options rmdup, sort and index) for further analyses. Base recalibration and local realignment around indels was done with GATK. For defining somatic mutations, we used the Mutect version 1.1.4 algorithm for identifying substitutions and the IndelGenotyper (IndelGenotyper.36.3336-GenomeAnalysisTK.jar) algorithm for identifying small insertions and deletions (indels). We defined the final list of somatic mutations with the following filters: frequency of the reads with the altered base in the tumor $\geq$0.1; number of reads with the altered base in the tumor $\geq$5; frequency of the reads with the altered base in the normal <0.03; number of reads with the altered base in the normal <2; not in dbSNP database except for variants that are also in COSMIC with a variant allele frequency in 1000G < 0.001 or not reported. The resulting somatic mutations were annotated with the snpEff 4.1c algorithm (16).

TCGA cases

Whole exome sequencing data of 60 primary solid tumors of breast invasive carcinoma and corresponding blood derived normal were randomly selected from TCGA. To avoid the bias of breast cancer subtypes, we selected 20 samples from each subtype (Her2 amplified, triple negative, hormone receptor positive with Her2 not amplified). A list of 4,286 curated somatic mutations was also retrieved from TCGA (genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.1.1.0.curated.somatic.maf).

ctDNA mutation identification workflow

The method first retrieves heterozygous variants using GATK HaplotypeCaller in each tumor sample using hard filters (Quality By Depth QD > 2.0, strand bias FS < 60.0, mapping quality MQ > 40.0, MappingQualityRankSum > −12.5, ReadPosRankSum > −8.0, GenotypeQuality > 30 for SNPs and QD > 2.0, FS < 30, ReadPosRankSum > −20.0 for INDELs). The variants in coding regions and satisfying the following filters are then selected: $\geq$5 reads supporting the variant; total base depth $\geq$ 10 and variant allele frequency $\geq$ 0.1. For each position with a variant detected with these filters, the number of reads supporting the reference and the variant allele in the BAM files of the tumor and corresponding blood sample are retrieved using samtools mpileup (15) with a minimum score of 20 for both base and mapping qualities. Only the variants with at least one supporting read for the variant in the BAM file of the blood sample are kept. Patient polymorphisms are then filtered out by three strategies: (i) the difference of allele frequency in the tumor and corresponding normal samples is tested with a one-tailed Fisher exact test (FDR > 0.01); (ii) the probability of each variant to be germline is computed by comparing the variant read frequency in the normal sample to the read frequency distribution of germline heterozygous SNPs per patient (empirical P > 0.01), and (iii) the minor allele frequency (MAF) of each variant in the population under study is calculated as the number of patients with an allele frequency >0.1 in both the tumor and the blood divided by the total number of patients in the population (observed MAF > 0.01). Finally, we filtered out known polymorphisms as defined in dbSNP database (version 138) after excluding positions with variants present in COSMIC (version 67) unless the observed 1,000 Genomes MAF was higher than 0.001.

Sequencing bias. To distinguish a ctDNA mutation from a sequencing bias, we retrieved the variant-supporting reads in each blood sample for all patients for each of the positions detected in the previous step. We hypothesized that, in the absence of a true variant, the variant-supporting read counts N_v should follow a binomial distribution $B(N_t,\ p_{\text{error}})$, where N_t is the total number of reads at the position of interest and p_{error} is the sequencing error rate, estimated by the variant read frequency in the mixture of the blood samples of all the patients. A variant with n_v supporting reads for a total depth of n_t was

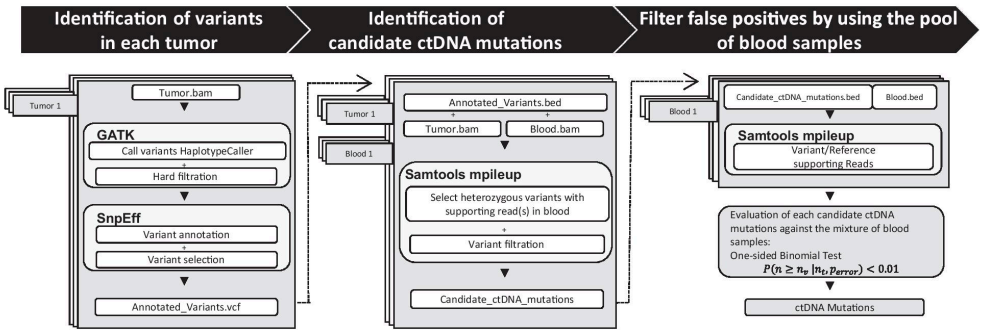


Figure 1.

cmDetect workflow. The method consists of three main steps including: (i) tumor variant identification based on GATK and SnpEff; (ii) selection of non-germline variants with some evidence support in the corresponding blood sample; and (iii) filtration of candidate variants based on sequencing error from the pool of blood samples.

considered as a true variant if $P(n \geq n_v \mid n_t, p_{error}) < 0.05$ using a one-sided binomial test (17).

Plasma mutations

Details of plasma DNA extraction and sequencing can be found in the publication by Jovelet and colleagues (9). Fastq files were treated with the Torrent Suite BaseCaller version 4.0 or 4.2. We retrieved hotspot variants using GATK HaplotypeCaller in each plasma sample and satisfying the following filters: strand bias FS < 30; variant supporting reads >4; total base depth >50; and variant allele frequency ≥0.1. We filtered out variants present in polymorphism databases (ESP, the Exome Sequencing Project or 1000G, from European samples in 1000 genome project; ref. 18) with a minor allele frequency >0.001.

Survival analyses

Overall survival was estimated by the Kaplan–Meier methods. Correlation between the number of ctDNA mutations and survival was assessed using Cox-proportional hazard models. Univariate analysis was performed using Log-rank test for categorized variables. Multivariate analysis was assessed using Cox proportional Hazard Modeling. All factors with $P < 0.10$ in univariate analysis were evaluated on multivariate analysis. All P values reported are two-sided. For all statistical tests, differences were considered significant at the 5% level. Stata 13.0 was used for all statistical analyses.

Results

ctDNA mutation detection workflow

We introduce a method, cmDetect (ctDNA mutation detection), for the systematic identification of ctDNA mutations by leveraging information from the tumor and blood samples (Fig. 1). cmDetect consists of three steps and is described in details in the method section. Briefly, the proposed workflow first retrieves heterozygous variants in gene coding regions in each tumor independently using the Genome Analysis ToolKit (GATK; ref. 19) and selects those variants with supporting read(s) in the corresponding blood sample. The patient's germline variants and common polymorphisms are then filtered out to obtain a set of mutations that are identified with high confidence in the tumor sample and lower incidence in the blood sample. At this step, as the read frequency of the selected variants in the blood may be very small (<0.02), it is important to be able to distinguish ctDNA mutations from sequencing biases. Therefore, the frequency of each selected variant is tested against the observed frequencies of

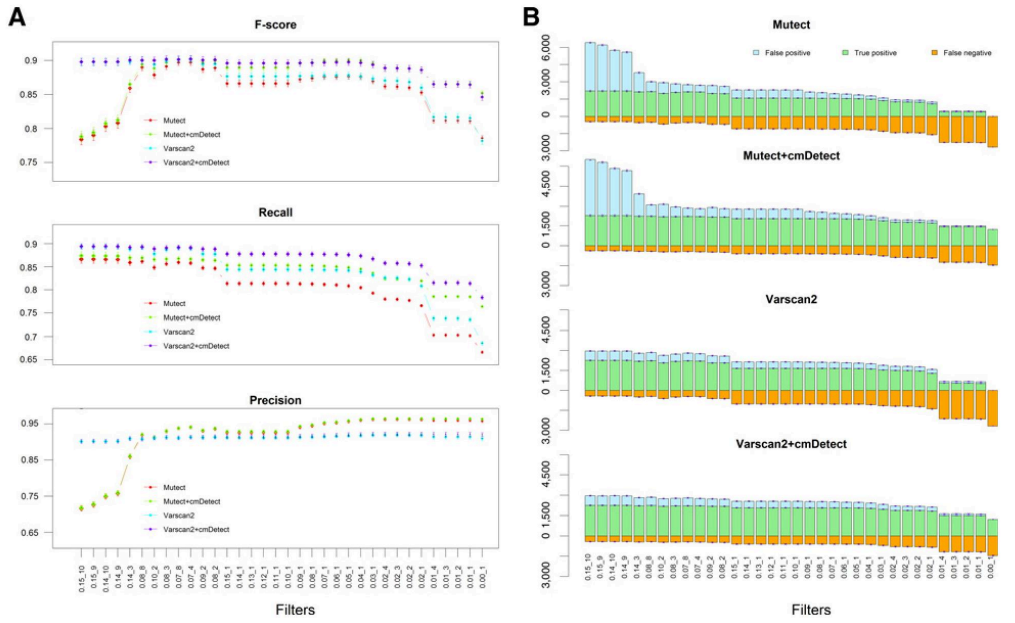


Figure 2

A, combining cmDetect with somatic mutation caller(s) improves the performance for somatic mutation identification. *F*-score, recall, and precision (*y*-axis) between Mutect (red), Mutect+cmDetect (green), Varscan2 (blue), and Varscan2+cmDetect (violet) at different filters (*x*-axis) are shown. Error bars correspond to 95% CIs. **B,** sensitivity of somatic mutation calling by adding cmDetect to Mutect(Varscan2) for different filters. Bar plots showing the numbers of false positives (blue), true positives (green), and false negatives (yellow) called by Mutect only, Mutect+cmDetect, Varscan2 only, and Varscan2+cmDetect with different filters for somatic mutations with at least one supporting read in the blood. The vertical axis shows the number of mutations, true positives are shown above 0, and false negatives are shown below 0. The horizontal axis corresponds to the different filters applied for somatic mutation calling, from the most stringent filter (left) to the least stringent filter (right).

the same variant in the mixture of blood samples to estimate its probability of being a false positive (due to sequencing bias or bad alignment).

Benchmarking

To estimate the statistical power of the cmDetect workflow, a set of 66 tumor/blood pairs of samples was simulated on the basis of real whole-exome sequencing data from the 1000 Genomes project (Supplementary Table S2; ref. 19). Briefly, for each sample, we derived a tumor sample by introducing cancer-causing mutations (COSMIC; ref. 20) at different read frequencies and a normal blood sample contaminated with tumor DNA at various levels (see Materials and Methods; Supplementary Table S2). This way, we introduced a total of 12,469 somatic mutations including 2,676 ctDNA mutations (Supplementary Table S3). We identified ctDNA mutations with cmDetect and retrieved somatic mutations with Mutect (21) and Varscan2 (22) for a range of filters on maximum detection level of the mutation in the normal sample. Application of cmDetect identified 1,219 ctDNA mutations, all true positives (Supplementary Table S4). In addition, 1,458 (54%) ctDNA mutations were not identified by cmDetect for the following reasons: (i) the mutations did not have enough support evidence in the normal sample to be called a ctDNA mutation; (ii) the read frequency in the tumor and corresponding normal was comparable; and (iii) the read frequency in the normal could not be distinguished from a polymorphism (see the polymorphism filtering section in Methods and Supplementary Fig. S1). However, a large proportion (>50%) of the ctDNA mutations missed due to a very low coverage in the normal were correctly identified by the somatic mutation callers [749 (51%) by Mutect and 830 (57%) by Varscan2]. By evaluating the performance of the somatic mutation callers with and without cmDetect at the various filters, we show that without decreasing the sensitivity, adding ctDNA mutations to the results of traditional somatic variant callers decreased the number of false negatives for all the different filters tested (Supplementary Table S5; Fig. 2A). Importantly, the observed recall was not dependent on the initial filters applied in Mutect or Varscan2 (Fig. 2B), indicating that cmDetect can be efficiently combined with somatic mutation callers. This also demonstrated that the gain in sensitivity cannot be achieved by relaxing the initial mutation caller filters, for example by increasing the maximum allowed allele frequency of the variant in the blood, which will have dramatic effects on specificity.

ctDNA is detectable in whole-exome sequencing of metastatic breast cancer patients' blood

We applied our method to a panel of 93 whole-exome sequenced pairs of breast cancer metastases/blood samples from the SAFIR01 (NCT01414933; ref. 13) and MOSCATO (NCT01566019) clinical trials (see Supplementary Tables S6 and S7, for clinical information and sequence quality metrics). We first applied Mutect and indelGenotyper and identified

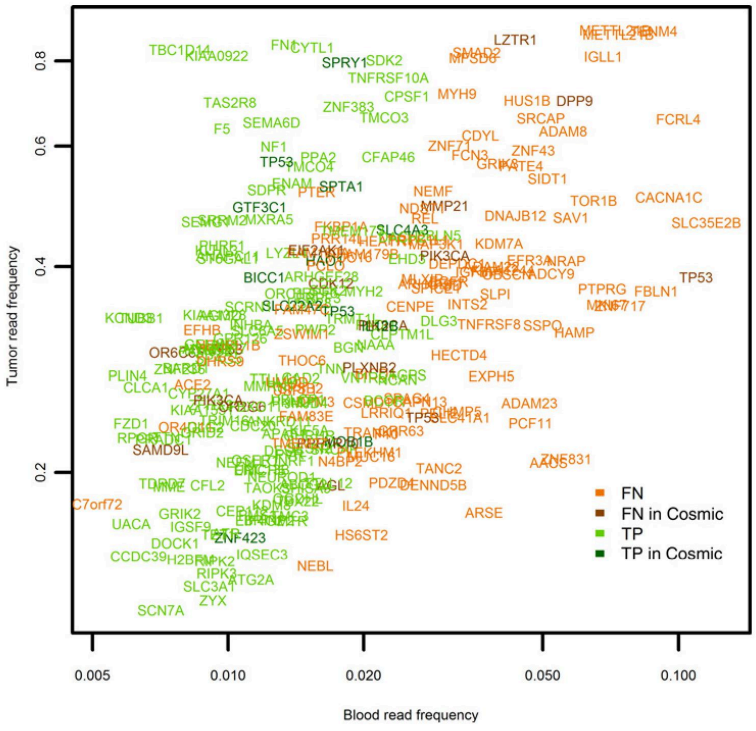


Figure 3.
Observed read frequencies of ctDNA mutations identified by cmDetect in breast cancer metastatic samples. The color of the gene corresponds to the status of the call of the mutation according to Mutect applied with default filters as described in Materials and Methods. Shown is the allele frequency in the blood (*x*-axis) and in the corresponding tumor (*y*-axis). False negative (FN; orange) corresponds to the ctDNA mutations identified by cmDetect but not by Mutect. True positive (TP; green) corresponds to the ctDNA mutations identified by cmDetect and Mutect. The mutations that are documented in COSMIC are shown in darker color.

7,334 somatic mutations (Supplementary Table S8) for the 93 pairs of metastasis/blood samples (see Materials and Methods). The application of cmDetect identified 263 ctDNA mutations (Supplementary Table S9) distributed in 50 patients (54%). Among these 263 mutations, 141 were correctly identified by the somatic mutation analysis whereas 122 were false negatives, defined as true somatic mutations incorrectly filtered out by the mutation callers due to the high level of detection in the control blood sample, representing 1.67% of the total number of somatic mutations. Importantly, among the false negatives we found 12 mutations reported in clinical databases (COSMIC) including two *PIK3CA* missense mutations (V344M, H1047R) and two stop-gain *TP53* mutations (R306*, E349*), with a read frequency as high as 0.11 with seven supporting reads in the corresponding blood sample (E349*; Fig. 3). Other mutations of interest among the false negatives included one *EGFR* missense mutation (G322S) and one small insertion in *TP53* (N288fs). In the following, we consider a total of 7,457 somatic mutations among the 93 pairs of metastases/blood samples.

The 50 patients had in average 5.26 and a maximum of 38 ctDNA mutations, most of them (82%) having less than 10 ctDNA mutations identified. Patients with detectable ctDNA in the blood, identified as patients with at least one ctDNA mutation, had more somatic mutations in their tumor than patients with no detectable ctDNA (t-test $P = 0.0003$; Supplementary Fig. S2). However, there was not a direct correlation between number of ctDNA mutations and total number of mutations (Pearson $\rho = 0.12$; $P = 0.24$), indicating that the mutational load of the tumor does not reflect the amount of ctDNA detectable in the whole blood sample. Importantly, ctDNA mutations represented up to 53% of the total number of somatic mutations whereas false negative rates ranged from 0% to 35% of the total number of mutations per patient (Fig. 4). We also confirmed in the 86 SAFIR01 cases that patients with detectable ctDNA mutations were associated to poor outcome, marginally when only considering the presence/absence of ctDNA mutations ($P = 0.068$; Fig. 5), but very significantly when considering ctDNA mutations quantitatively ($P < 0.001$). Indeed, the number of mutations were highly significant in both univariate analysis (HR = 1.14; $P < 0.001$) and in multivariate analysis (HR = 1.15; 95% CI, 1.07–1.23; $P < 0.001$). Finally, we found that patients who received a prior chemotherapy (Supplementary Table S6) presented more ctDNA mutations (mean = 3.07) than the patients who did not (mean = 0.56; t test $P = 0.002$; Supplementary Fig. S3).

To further evaluate the extent of ctDNA contamination in early-stage disease, we applied cmDetect to a set of 60 primary tumor–blood pairs of whole-exome sequencing data from the TCGA breast cancer collection (see Materials and Methods; ref. 23).

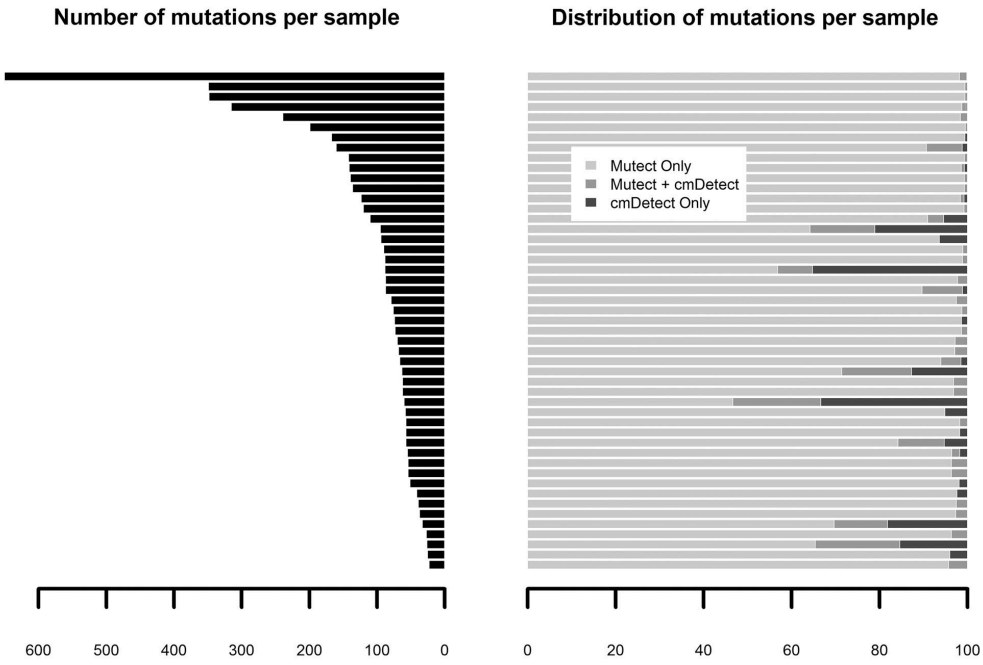


Figure 4.
Somatic mutation profile of breast cancer metastatic patients with detectable ctDNA in whole blood. Left, total number of somatic mutations per sample; right, percentage of somatic mutation types. Mutect only, mutations identified by Mutect but not cmDetect; Mutect+cmDetect, ctDNA mutations identified by Mutect and cmDetect; cmDetect only, ctDNA mutations identified by cmDetect but not Mutect.

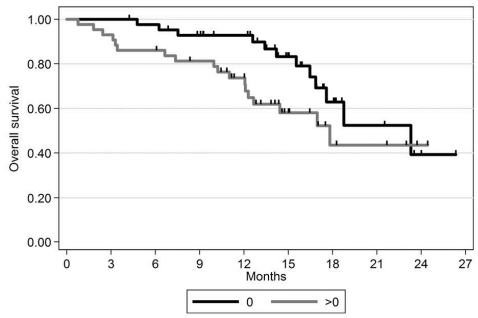


Figure 5.
Patients survival in the SAFIR01 trial according to detectable ctDNA status. The black line (0) represent patients with no detectable ctDNA, whereas the gray line (>0) contains the patients with at least one ctDNA mutation identified.

We identified 41 ctDNA mutations in 18 patients (30%) harboring from 1 to 14 ctDNA mutations with an average of two mutations per patient (Supplementary Table S10). Among these 41 mutations, 10 were present in the somatic mutation results file from TCGA whereas 31 (0.7% of the total number of somatic mutations) were false negatives including nine that were reported as clinical variants (COSMIC; Supplementary Table S10).

Validation with plasma samples

We applied cmDetect to a set of 60 pairs of metastasis/blood samples from the MOSCATO clinical trial sequenced on a NextSeq500, including cancers of different tissues of origin (Supplementary Table S7). For 46 MOSCATO patients, 43 in this cohort and three in the breast cancer metastasis cohort, targeted sequencing data of the plasma was also available (Supplementary Table S11) and reported 12 COSMIC mutations (Supplementary Table S12), validating the presence of ctDNA for a total of nine patients. In parallel, cmDetect identified ctDNA mutations from the whole-exome sequencing data for eight of these nine (89%) patients, confirming the sensitivity of our approach (Supplementary Table S13). We found that two of the 12 COSMIC mutations identified in the plasma were also identified by cmDetect from the whole-blood sample: one *TP53* stop-gained mutation (E349*) in patient BC93 with an allele frequency of 0.76 in the plasma, 0.39 in the tumor, and 0.12 in total blood, and one *TP53* stop-gained mutation (R213*) in patient PCAN39 with an allele frequency of 0.2 in the plasma, 0.35 in the tumor, and 0.013 in total blood. It is interesting to note that BC93 presented with the highest mutated DNA fraction (0.76) in the plasma and also had the highest number of ctDNA mutations (38) detected by cmDetect.

Discussion

We introduced the first method for identifying somatic mutations from whole-exome sequencing data that takes into account the possible presence of ctDNA. We propose to use cmDetect as an additional step to traditional somatic mutations analysis pipelines when analyzing whole-exome sequencing of large sets of pairs of tumor/whole blood samples but it can also be used as a standalone workflow for the identification of ctDNA mutations. We showed that cmDetect is very sensitive but presented some limitations. First, cmDetect will not identify ctDNA mutations with blood read frequencies comparable to polymorphisms' observed read frequencies. Indeed, the maximum blood read frequency for a ctDNA mutation detected by cmDetect in the breast cancer metastasis cohort was 0.11 (Fig. 3). Second, cmDetect will also miss ctDNA mutations having comparable tumor and blood read frequencies as it uses a Fisher exact test to differentiate frequencies in tumor and blood samples, a filtering step that is also applied in traditional somatic mutation callers. Finally, ctDNA mutations with very low coverage in the blood sample will usually be missed by cmDetect, as they will not be distinguishable from sequencing errors. However, we demonstrated that most of these mutations will be identified by somatic mutation callers with default filters for the variant detection in the blood. We demonstrated that metastatic breast cancer patients' whole blood may contain ctDNA mutations detectable at low frequencies by whole-exome sequencing, even at relatively low coverage (mean coverage was 70× in the blood samples). Indeed, some of the variants identified at low frequency in the blood samples were known cancer mutations and were validated in the plasma of two patients, demonstrating that our method was sensitive enough to retrieve somatic mutations in the blood at a frequency as low as 0.013. Importantly, we demonstrated that, in combination with cmDetect, traditional mutation callers can be used with stringent filters on blood read count and frequency supporting the alternate allele, therefore reducing the number of false positives while increasing the number of true positive. Although it is always possible to recover known cancer-related mutations by using databases such as COSMIC, other not-documented mutations may be completely missed if they present read counts or frequencies in the whole blood higher than expected. An alternative approach to limit the contamination of normal germline DNA by ctDNA consist in using either blood cells pellet after discarding plasma, or purified peripheral blood mononuclear cell (PBMC). The use of normal tissues such as skin biopsies or fibroblast culture is also possible but appears to be a heavy procedure in standard biological samples' collection, whereas the bioinformatics method proposed may overcome these problems with an analytical solution and provided additional prognostic information associated to presence of ctDNA. Indeed, the number of ctDNA mutations per patient was strongly associated to survival in metastatic breast cancer patients and may reflect the amount of mutated DNA in circulation. This is consistent with previous analyses that have shown that increasing levels of ctDNA and CTC counts were associated with inferior survival in metastatic breast cancer (24, 25). Although we found a significant number of ctDNA mutations in metastatic cases, we also showed that ctDNA mutations may be identified in early-stage disease, but to a lower extent. The proposed approach was developed for the purpose of identifying tumor DNA contamination in germline DNA from whole blood, but it might also be useful for detecting contamination in blood samples from Heme-malignancies with minimum residual disease (MRD) or in DNA extracted from tumor-adjacent normal tissue.

Disclosure of Potential Conflicts of Interest

M. Campone has received speakers bureau honoraria from Novartis, AstraZeneca, Pfizer, and is a consultant/advisory board member for Pfizer and AstraZeneca. T. Bachelot has received speakers bureau honoraria from

Roche and Novartis and is a consultant/advisory board member for Roche, Novartis, and AstraZeneca. J.-C. Soria is a consultant/advisory board member for AstraZeneca. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: Y. Fu, J. Garrabey, J.-C. Soria, L. Lacroix, F. André, C. Lefebvre

Development of methodology: Y. Fu, Y. Luo, L. Lacroix, C. Lefebvre

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): C. Massard, M. Campone, C. Levy, V. Diéras, T. Bachelot, J. Garrabey, J.-C. Soria, L. Lacroix

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): Y. Fu, T. Filleron, M. Pedrero, N. Motté, Y. Boursin, Y. Luo, V. Diéras, L. Lacroix, F. André, C. Lefebvre

Writing, review, and/or revision of the manuscript: Y. Fu, T. Filleron, N. Motté, Y. Boursin, C. Massard, M. Campone, V. Diéras, T. Bachelot, J. Garrabey, J.-C. Soria, F. André, C. Lefebvre

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): Y. Fu, C. Jovelet, N. Motté, Y. Boursin

Study supervision: L. Lacroix, C. Lefebvre

Acknowledgments

We thank Marta Jimenez from UNICANCER; Marc Deloger and Guillaume Meurice of the bioinformatics core facility of Gustave Roussy and Mélanie Letexier and Emmanuel Martin from Integragen for their assistance. We thank the NIH TCGA project for granting us access to the sequencing data (under project #9082).

Grant Support

This work was supported by Breast Cancer Research Foundation (F. André), Fondation Lombard-Odier "Philanthropia" (F. André), Odyssea (F. André), Operation Parrains Chercheurs (F. André), Dassault Foundation (F. André), French NCI: INCa-DGOS-INSERM 6043 (F. André), SIRIC Socrate (J.C. Soria), and Fondation ARC (to UNICANCER).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 24, 2015; revised June 30, 2016; accepted July 28, 2016; published OnlineFirst August 17, 2016.

References

1. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med 2013;5:91.
2. Ignatiadis M, Dawson SJ. Circulating tumor cells and circulating tumor DNA for precision medicine: dream or reality? Ann Oncol 2014; 25:2304–13.
3. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014;6:224ra24.
4. Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, Li M, et al. Circulating mutant DNA to assess tumor dynamics. Nat Med 2008;14: 985–90.
5. Spindler KL, Pallisgaard N, Andersen RF, Brandslund I, Jakobsen A. Circulating free DNA as biomarker and source for mutation detection in metastatic colorectal cancer. PLoS One 2015;10:e0108247.
6. Haber DA, Velculescu VE. Blood-based analyses of cancer: circulating tumor cells and circulating tumor DNA. Cancer Discov 2014;4:650–61.
7. Leary RJ, Kinde I, Diehl F, Schmidt K, Clouser C, Duncan C, et al. Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med 2010;2:20ra14.
8. McBride DJ, Orpana AK, Sotiriou C, Joensuu H, Stephens PJ, Mudie LJ, et al. Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. Genes Chromosomes Cancer 2010;49:1062–9.
9. Jovelet C, Ileana E, Le Deley MC, Motté N, Rosellini S, Romero A, et al. Circulating cell-free tumorDNA analysis of 50 genes by next-generation sequencing in the prospective MOSCATO trial. Clin Cancer Res 2016; 22:2960–8.
10. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415–21.
11. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods 2015;12:623–30.
12. Clopper S, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26:404–13.
13. Andre F, Bachelot T, Commo F, Campone M, Arnedos M, Dieras V, et al. Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). Lancet Oncol 2014;15:267–74.
14. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589–95.
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078–9.
16. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80–92.
17. Bansal V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 2010; 26:i318–24.
18. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
19. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20:1297–303.
20. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015;43:D805–11.
21. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213–9.
22. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568–76.
23. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.
24. Dawson SJ, Tsui DW, Murtaza M, Biggs H, Rueda OM, Chin SF, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med 2013;368:1199–209.
25. Bidard FC, Peeters DJ, Fehm T, Nolé F, Gisbert-Criado R, Mavroudis D, et al. Clinical validity of circulating tumour cells in patients with metastatic breast cancer: a pooled analysis of individual patient data. Lancet Oncol 2014;15:406–14.

The use of cmDetect in small cohorts

In order to improve mutation calling accuracy for advanced cancer patients, we developed a method to be applied as a complementary tool to traditional somatic mutation callers in the presence of ctDNA in patients' whole blood. Somatic mutation identification is an important step in personalized medicine because of its direct implication on patients' prognostic. We showed the essential need of applying cmDetect in the context of ctDNA contamination of the whole blood sample used as source of germline DNA, especially for patients with metastatic cancer. However, cmDetect is not suitable for analyzing a small cohort of patients (ex N<10) mainly due to two reasons.

First, cmDetect estimates a sequencing bias rate (error rate) based on the total observations at a given position. As the sequencing techniques are well mastered, the sequencing error rate is expected to be quite low (for Illumina Hiseq machine, 96% of the total reads have a base quality score of more than 30, which means 1 error read out of 1000). For example, in a cohort of 10 patients sequenced at 100X (a total of 1,000 reads at each position), only one read with sequence error can be observed at a given position considering a sequencing error rate of 0.1%, and any lower rate is impossible to estimate. Second, a cohort specific Minor allele frequency (MAF) is needed for filtering polymorphisms as the polymorphisms documented in dbSNP may not be sufficient as they are estimated based one the general population and might not be representative in a specific cohort of patients. Again, a small set of patients is not suitable for this estimation (ex: a variant observed in one patient out of ten gives a MAF=10%, which is more likely due to chance than a correct

statistical estimation). We therefore recommend to use cmDetect with caution in the case of small patient cohorts.

A possible solution in the case of ctDNA contamination in a small cohort of patients is to collect a set of sequencing data generated under the same protocol with the same machine for the estimation of sequencing bias and use this estimation in the cmDetect pipeline. Another solution to avoid the contamination of ctDNA in the blood sample is to remove the plasma that contains circulating DNA from the whole blood sample before sequencing.

Integrated analysis of genomic and pharmacological data to better predict anti-cancer drug response

Co-existing alterations of different pathways are related to drug resistance

Precision medicine trials in oncology evaluate the benefit of a therapeutic strategy based on a molecular profile identifying point mutations and/or genomic rearrangements harbored by the tumor cells of the patient. However, the decision making is usually based on the identification of singular actionable aberrations and remains suboptimal. If multiple targets are identified in the tumor, the rules governing the selection of the optimal target for treatment are usually nonexistent. In this work, we investigated the effect of the presence of multiple actionable targets in tumor cells on their response to single-target treatments.

We used pharmacogenomics data from GDSC and CCLE to study the drug response to targeted therapies in cell lines carrying multiple alterations as compared to cell lines with a single alteration focusing on a subset of 7 groups of targeted therapies and 29 gene targets (Figure 19). The 7 groups of targeted therapies were selected as they have been assessed by the 2 cell line projects (GDSC and CCLE) and for their use in clinical practice. The 29 gene targets are the direct targets of the 7 groups of targeted therapies and their upstream genes, up to 2 levels of regulation, in the same pathway. For example, RAS, RAF and MEK were defined as the targets of anti-MEK inhibitors.

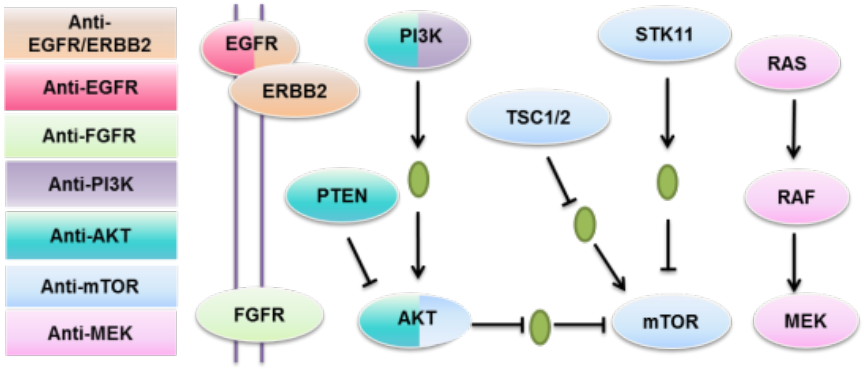


Figure 19. Inhibitors and pathways selected for the study. Genes are defined as target/targets of the inhibitor of the same color, for example, RAS, RAF and MEK are targets of anti-MEK inhibitors.

We studied the drug response to targeted therapies in the 3 groups defined as: "One alteration" cell lines, "Multiple alterations" cell lines and "Wild type" cell lines. "One alteration" cell lines are the cell lines carrying only alteration of one target among the 7 groups of targeted genes; "Multiple alterations" cell lines are the cell lines carrying not only alteration in the group of targeted gene but also in at least one other group of target gene; "Wild type" cell lines are cell lines not carrying any alteration in the gene targets of the study. For example, to an anti-EGFR inhibitor, cell lines with an amplification of EGFR and wild type of the rest of the 28 gene targets are classified in the "One alteration" group; cell lines with a mutation of EGFR and a mutation of PIK3CA are in the "Multiple alterations" and cell lines without any alteration of these 29 genes are in the "Wild type" group (Figure 20).

The measurement of the cell line response to drugs is defined by the normalized area under the drug response curve (nAUC) or the half maximal inhibitory concentration (IC50). Here we define an alteration in a gene as the presence of a mutation or a copy number strictly greater than 4 (amplification) or lower than 2 (loss)(Table 2). We selected a list of 126 drug-gene-alteration pairs for 19 drugs and 29 targetable genes and classified the cell lines according to the number of alterations in these 29 genes.

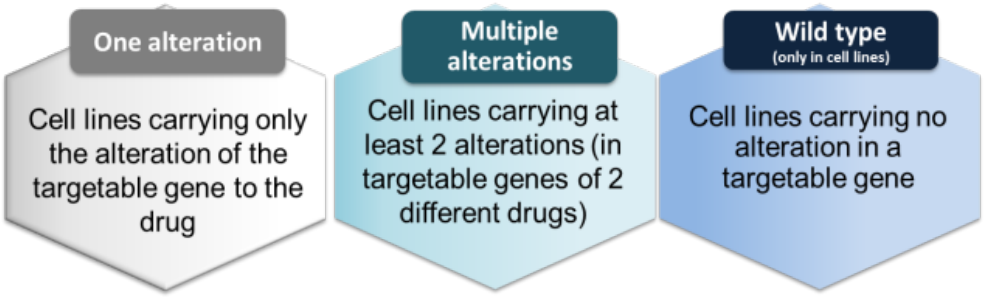


Figure 20. Classification of cell lines according to the mutational status of 29 genes.

Table 2. Data information from two public cell lines databases and patients used in the study.

	In Cell lines	In patients
Source	• **GDSC** (Genomics of Drug Sensitivity in Cancer) 126 drug-gene-alteration pairs for 19 drugs and 29 targetable genes • **CCLE** (Cancer Cell Line Encyclopedia) 16 drug-gene-alteration pairs for 4 drugs and 5 targetable genes	• **MOSCATO** (clinical trial Molecular Screening for Cancer Treatment Optimization at Gustave Roussy Villejuif, France).
Gene alteration	• **Mutation** • **Amplification (CNV > 4)** • **Loss (CNV < 2)**	• **Mutation** • **Amplification** • **Deletion**
Drug response	**nAUC** (Scaled AUC)	**PFS, overall survival**

The comparison of the drug response in these 3 groups (Figure 20) demonstrated that cell lines carrying only alterations in the targets of the drug are more sensitive than cell lines carrying alterations in the targets of the drug plus alterations in at least one group of other targetable gene (Figure 21, Wilcoxon signed-rank test, p = *3e-30* for GDSC and Wilcoxon signed-rank test, p = *4e-3* for CCLE)

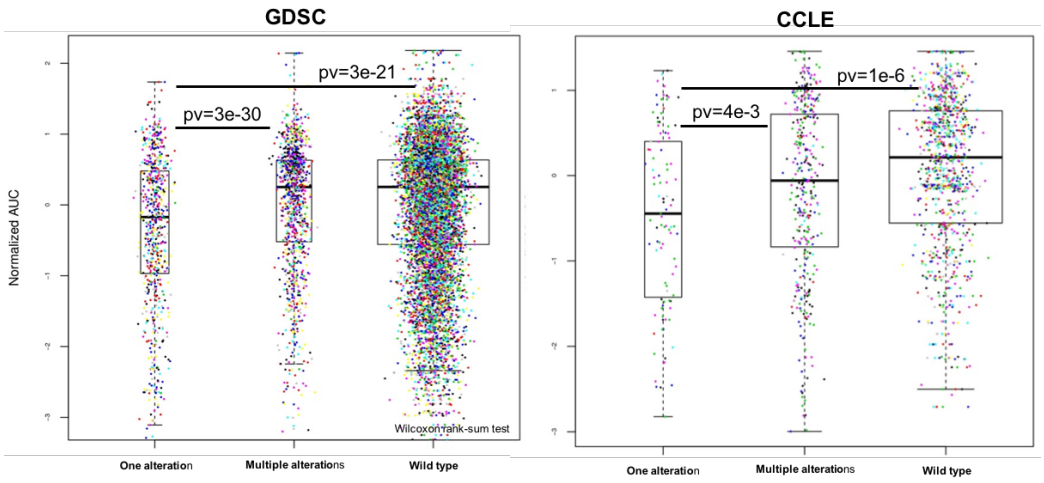


Figure 21. Cell lines with co-existing alterations of multiple targetable genes are more resistant to treatments than those with an alteration in one targetable gene only.

To extend this finding, we were interested in the details about the "plus alteration" and wanted to test if alterations in different pathways generate different levels of resistance. We further investigated the impact of each

pathway as a second alteration on drug resistance in cell lines. We split the group “Multiple alterations” into 7 subgroups according to the second pathway altered, not-targeted by the drug of interest. By comparing the drug response of these subgroups, we showed that different pathways have different impact on drug response (Figure 22; ANOVA, p= 6.42x10-6). In particular, pathway MEK and pathway AKT have the most significant impact on drug resistance (Wilcoxon signed-rank test, p = 1.9x10-10 and p= 2.9x10-5 respectively).

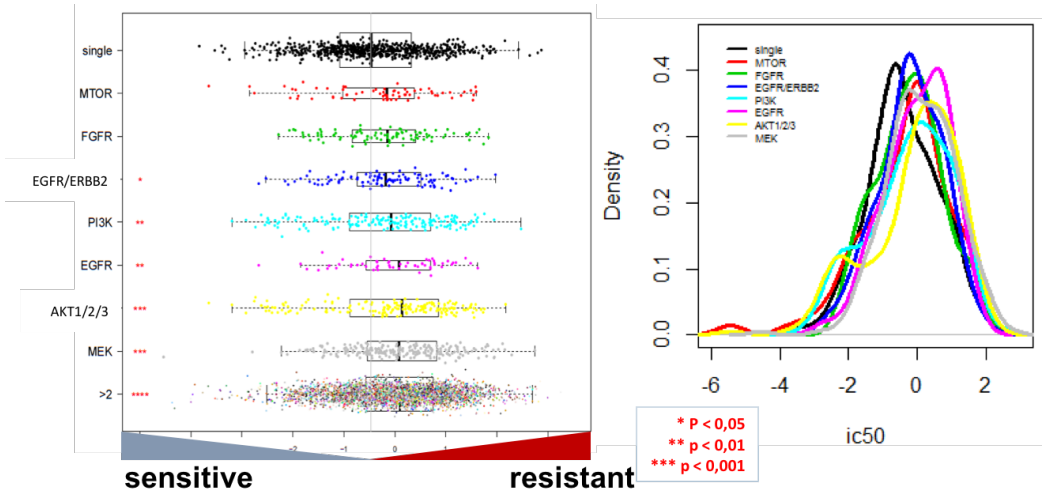


Figure 22. Effect on the drug response of an alteration in a non-targeted pathway.

We then examined clinical data from the MOSCATO clinical trial at Gustave Roussy (MOlecular Screening for CAncer Treatment Optimization) to validate these findings for cancer patients treated with targeted therapies matching an alteration of their tumor. MOSCATO is a clinical trial where genomic characterization of cancer patients’ tumor is investigated in order to direct the

patients to a suitable targeted therapy. Genomic and clinical data for 95 patients that have received one of the 7 targeted therapies were retrieved. Mutation status (whole exome sequencing and targeted sequencing) and copy number variations (CGH array) were available for all these patients (Jovelet, et al. 2016). As patients in the MOSCATO clinical trial receive targeted therapies when the matching targetable genes are identified as altered, the "wild type" group does not exist. The patients were classified into "One alteration" group if they presented an alteration only in the actionable target of their corresponding treatment and "Multiple alterations" group if they also presented alterations in the actionable targets of another drug from Table 3.

Table 3. list of targeted therapies from the MOSCATO trial considered in the analysis.

Treatment	Targets	Actionable Targets
Anti-AKT	AKT	AKT1, AKT2, AKT3, PTEN, PIK3CA, PIK3CB
Anti-EGFR	EGFR	EGFR
Anti-EGFR,ERBB2	EGFR,ERBB2	EGFR, ERBB2
Anti-FGFR	FGFR	FGFR1, FGF3, FGF4
Anti-MEK	MEK	MAP2K1, MAP2K2, BRAF, KRAS, NRAS
Anti-MTOR	MTOR	MTOR, TSC1, TSC2, STK11, AKT1, AKT2, AKT3
Anti-PI3K	PI3K	PIK3CA, PIK3CB

We performed a log-rank test to evaluate the difference of overall survival between the "One alteration" group and the "Multiple alterations" group (R package "survival"). The results showed that the patients in the "One alteration" group had a better overall survival rate as compared to the patients in the "Multiple alterations" group (log-rank test p-value: 0.038, Figure 23). However, the total number of mutations was not found to be associated with the survival

as a continuous variable (Cox proportional-hazards regression model, likelihood ratio test p-value=0.92). These results demonstrated that, as observed in the cell lines, the number of targetable alterations in a tumor is an important factor to consider when selecting a drug treatment for a cancer patient and may affect overall survival.

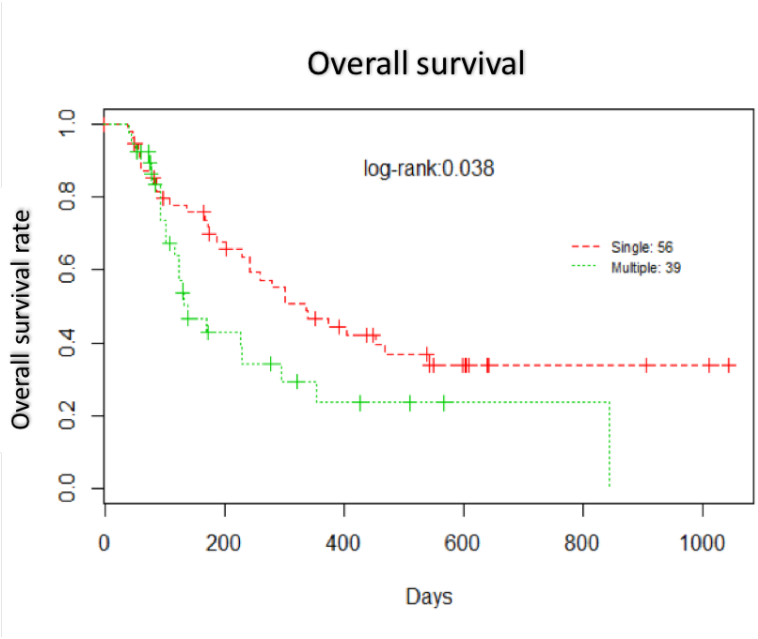


Figure 23. Overall survival comparison of MOSCATO patients treated with a therapy targeting an alteration of their tumor. The 2 groups were built according to the number of targetable alterations the tumor harbored, only 1 alteration (the targeted alteration, group in red, N=56) and >1 alteration (the targeted alteration + another targetable alteration, group in green, N=39).

Conclusion

We have shown in this section that the co-existence of alterations of different pathways have an impact on drug response, both in cell lines and patients. These results demonstrated the limitation of the therapeutic decision making based on the identification of singular actionable aberrations of a tumor and the

potential risk of drug resistance. Therefore, more complex decision roles/prediction algorithms are needed when multiple targets are identified. We then developed an integrated predictive mode to identify more complex patterns of prediction.

An integrated predictive model to identify predictive biomarkers

Regenerating drug response using gene expression data

One study has compared the data available in GDSC and CCLE and has shown an inconsistency in the drug response data for the drugs and cell lines shared by the two large-scale cell line datasets (Haibe-Kains, et al. 2013). They have shown that while the genomic data, especially the gene expression data, of the shared cell lines was highly correlated, the drug response data of the 15 common drugs (Table 4) was poorly correlated. The inconsistency of the drug response in the two data sets could be a result of different drug screenings and measuring methods. Meanwhile numerous studies have shown that gene expression data is one of the most accurate data set to predict drug response in cell lines and in patients (Geeleher, Cox, and Huang 2014; Dong, et al. 2015; Costello, et al. 2014).

Table 4. Drugs in common between GDSC and CCLE

Drug	Available Target
X17AAG	HSP90
AZD0530	SRC ABL1
AZD6244	MEK1 MEK2
ERLOTINIB	EGFR
LAPATINIB	EGFR ERBB2
NILOTINIB	ABL
NUTLIN3A	MDM2
PD0325901	MEK1 MEK2
PD0332991	CDK4 CDK6
PF02341066	MET ALK
PHA665752	MET
PLX4720	BRAF
SORAFENIB	PDGFRA PDGFRB KDR KIT FLT3
NVPTAE684	ALK
VX680	AURKA FLT3 ABL1 JAK2

In order to select predictive biomarkers accurately, we need to overcome the inconsistency problem between the two data sets as well as the problem of a high number of missing values in the drug response data. To this end, we used the gene expression data of the 2,000 genes with the highest variance as predictors to regenerate the drug response (AUC) for each cell line and each drug. As one of our goal is to identify biomarkers related to the sensitivity or resistance to anti-cancer drugs, the continuous drug response data was also transformed into to 3 categories based on their data statistic (Figure 24):

- Sensitive: AUC < mean(AUC) – sd(AUC);
- Resistant: AUC > mean(AUC) + sd(AUC);
- Intermediate: other.

  (sd = standard deviation)

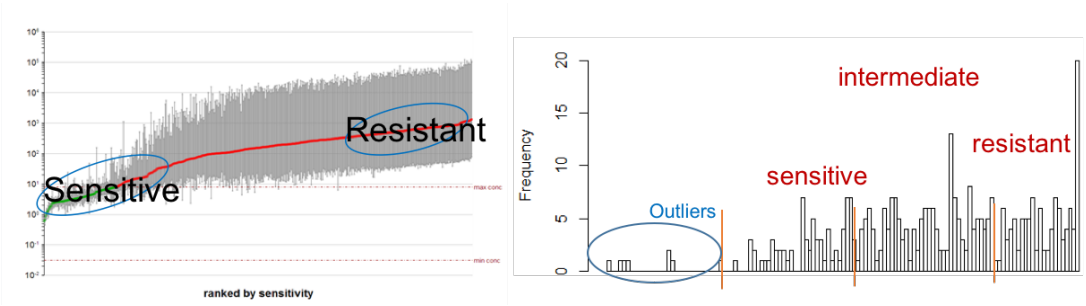


Figure 24. Distribution of the drug response of cell lines and transformation of continuous data to categorical (data from GDSC).

Four popular machine learning methods namely Elastic net (Zou and Hastie 2005), Support Vector Machine (Steinwart and Christmann 2008), Random

Forest (Breiman 2001) and Adaboost (Freund and Schapire 1999) were tested for the prediction model. The machine learning algorithms were performed using R packages (Elastic net: 'glmnet', Random Forest: 'RamdomForest', SVM: 'e1071', Adaboost: 'ada') and the parameters were tuned using 10-time cross validation.

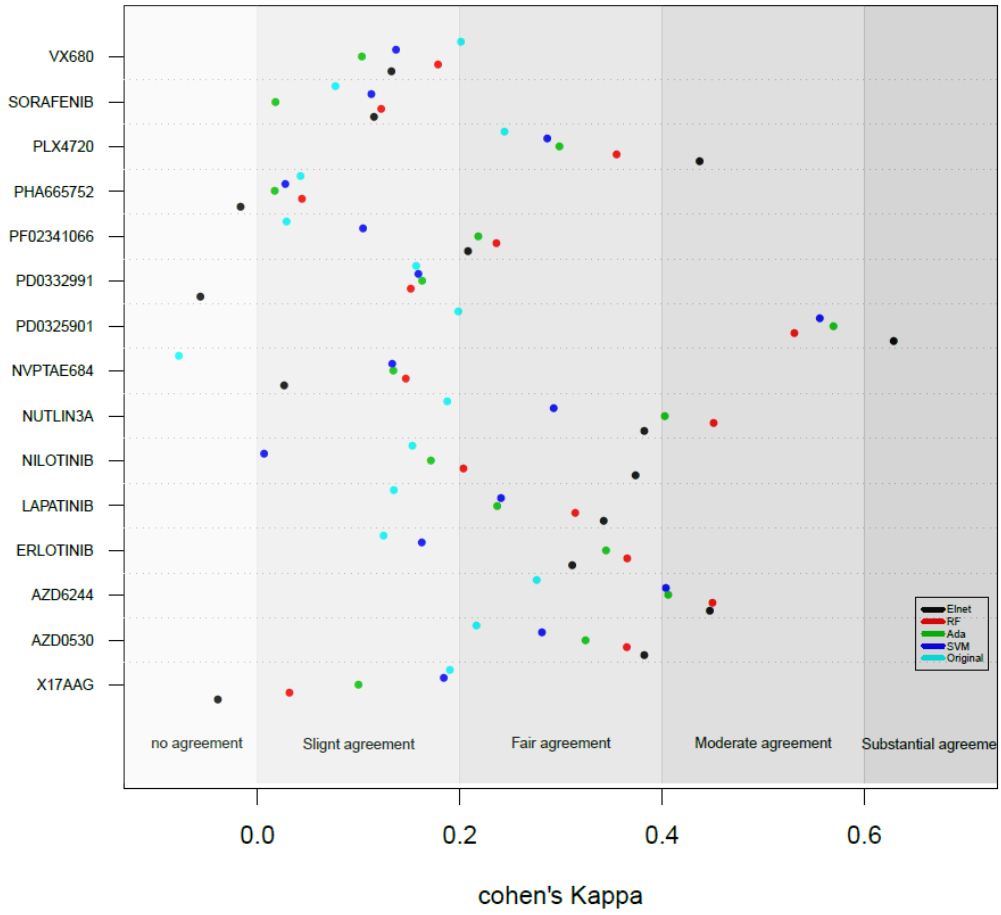


Figure 25. The consistency of drug response between GDSC and CCLE were improved by generating new drug response using gene expression data

As shown in Figure 25, the consistency of the drug response for cell lines was improved between the two data sets for the predicted drug response using gene expression data. We then used the new drug response generated by

SVM, Elastic Net, Adaboost and RF and a set of genomic alterations as predictors in a Random Forest model to identify predictive biomarkers. As discussed in the introduction, Random Forest has many advantages in predictive modeling. RF is easy to use with few parameters to tune; there is no prior assumption about the relationship between the predictors and the variable to predict and the feature importance provided by RF model is extremely useful for predictive biomarker selection. We therefore chose RF as the main predictive model. Only mutations and copy number variations were included as predictors in the RF model in order to generate “easy to use and interpret” treatment decision guiding rules. We observed that adding a step of predicting the drug response data using elastic net improved the random forest model for the identification of direct targets of the inhibitors as predictors of the response of the corresponding drug, as compared to SVM, RF, Adaboost and the original drug response for both GDSC and CCLE database (Table 5).

Table 5. More direct drug targets were identified by Random Forest using drug response generated by the predictive models and gene expression data. Elastic Net outperformed the others. A red case represents that the direct target of the inhibitor in row was correctly identified by the method in column in both GDSC and CCLE.

Drug	SVM + RF	Elastic net + RF	Ada + RF	RF + RF	Original + RF
NO. target / drug found	5	7	5	5	3
ERLOTINIB					
LAPATINIB					
NUTLIN3A					
PD0332991					
PF02341066					
PHA665752					
PLX4720					
SORAFENIB					
NVPTAE684					
VX680					

Drug that at least one target was identified

Finally, we designed an integrated predictive model to identify predictive biomarkers with 2 steps. The first step of the model is to predict the drug response of the cell lines using an Elastic net model with the expression data of 2,000 genes with the highest variance. Parameters of elastic net were tuned using 10-times cross validation. The second step of the model is to identify predictive biomarkers using Random Forest. Each predictor in the Random Forest model was evaluated with an associated importance calculated based on two methods. One method is to calculate an accuracy drop of the drug response prediction averaged on the ensemble of trees while permuting the predictor's value randomly. The other method is to calculate the gain of purity (for categorical data) or the decrease of variance (for continuous data) based on the variable chosen to split the node. A p-value associated to each importance was also calculated using a 1,000-time permutation of the drug

response to evaluate predictor-drug response relationship independently, p values are corrected using False Discovery Rate (FDR) (Figure 26).

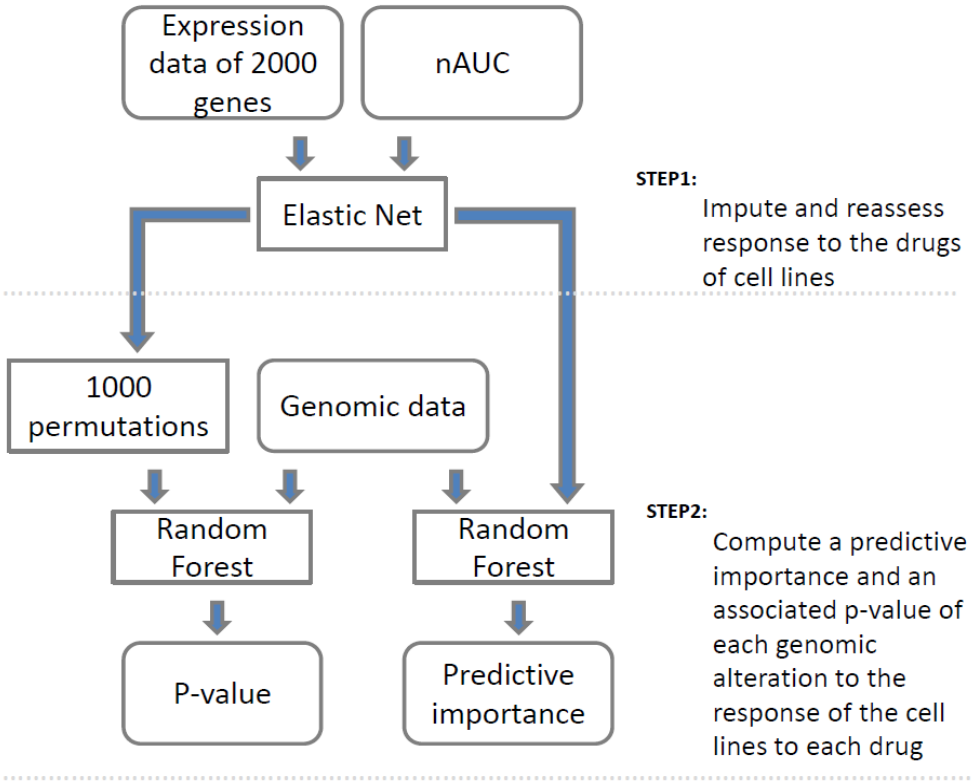


Figure 26. An integrated predictive model to identify predictive biomarkers. A first step of regenerating drug response using elastic net model and gene expression data and a second step of identifying predictive biomarkers using random forest model combine with 1000-time permutation of the drug response.

To identify predictive biomarkers for a larger set of drugs, we focused our efforts on GDSC in which more drugs were available. We applied this model to a set of 111 targeted therapies and confirmed that using reassessed drug response in the Random Forest model improves its ability to identify direct targets of the targeted therapies (Figure 27). We also observed that models using continuous drug response data performed better than those using categorical drug response data, probably due to the information loss while transforming the continuous data to categorical. We decided therefore to focus the further analyses on continuous drug response data.

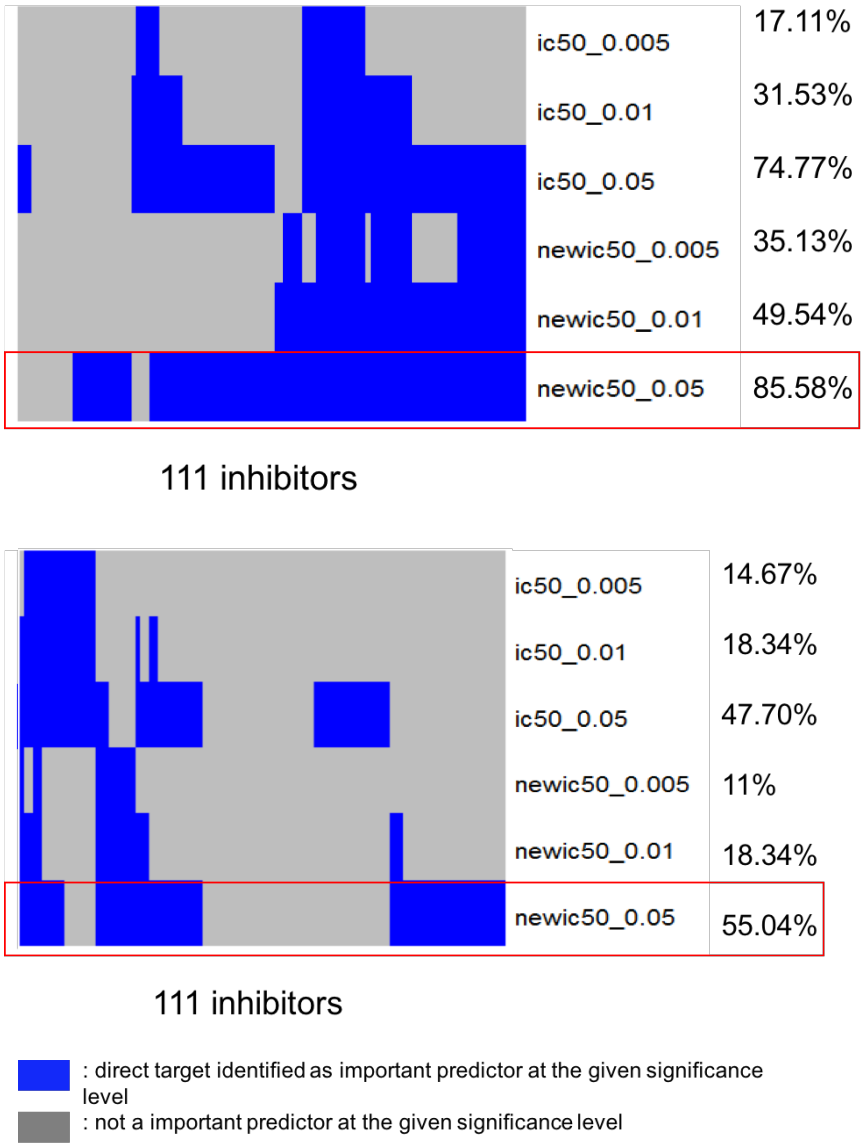


Figure 27. The use of regenerated drug response improved the identification of the direct target of drugs using random forest for a set of 111 targeted therapies available in GDSC. Each column represents a targeted therapy; each row represents a condition, for example, "newic50_0.05 55.04%" means using new drug response generated by elastic net, 55.04% of the direct target of the total 111 drugs are identified by random forest at significance of 0.05. top: continuous drug response data, bottom: categorical drug response data.

As expected, we also identified known predictive biomarkers other than the direct target. For example, mutations of KRAS were identified as a predictive

biomarker for resistance to the anti-AKT drug AZD5363 while mutations of PTEN where predictive biomarkers for sensitivity to AZD5363 (Figure 28) (Davies, et al. 2012).

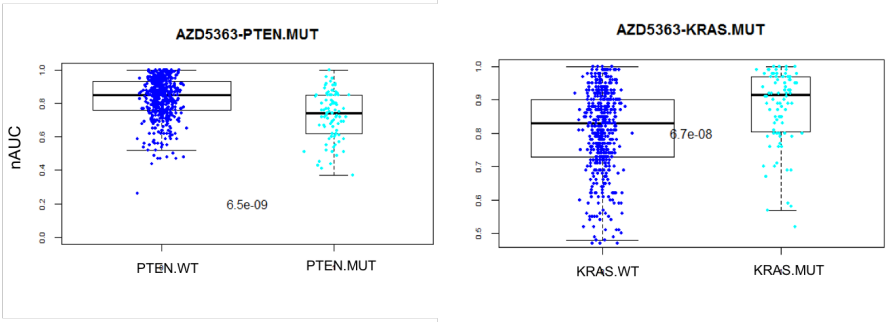


Figure 28. : Cell lines with muted KRAS are more resistant to AZD5363 than cell lines with wild type KRAS (left). Cell lines with PTEN mutation are more sensitive to AZD5363 than cell lines with wild type PTEN (right).

To further investigate the additive predictive power of these two predictors, we built a predictive tree model and observed that the cell lines with KRAS mutation and PTEN wild type are the most resistant subgroup of cells to AZD5363 (Figure 29). When combining the two predictors of resistance together (PTEN wild type and KRAS mutation are biomarkers of resistance to AZD5363), we were able to identify the most resistance subgroup of cells, which indicate the potential of combining multiple biomarkers to better predict the drug response.

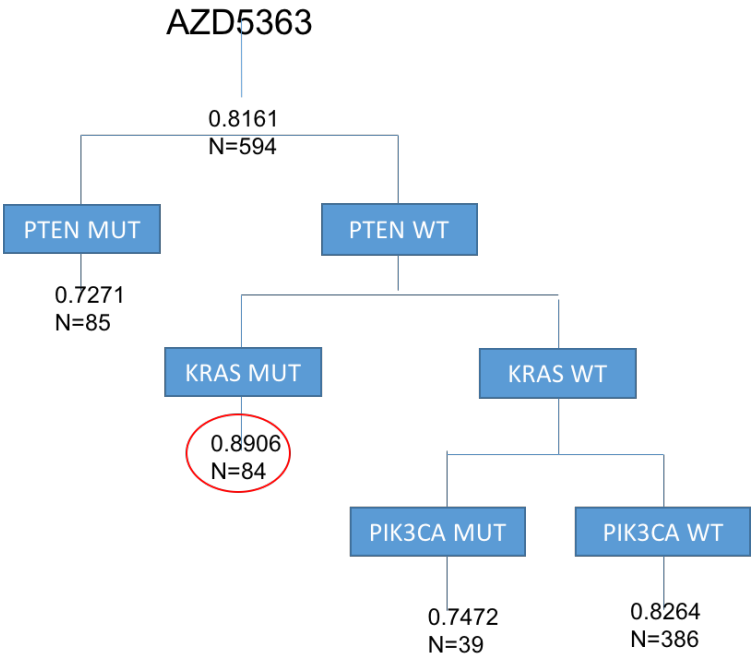


Figure 29. cell lines are classified into different subgroups based on their mutational status of 3 genes. Each node is marked by the mean drug response to AZD5363 and the number of cell lines in the node. Cell lines that are mutated in KRAS but not mutated in PTEN are the more resistant to AZD5363, an anti-AKT inhibitor than the rest of cell lines.

In another example, we identified the mutation of NOTCH1 as the most predictive alterations (p-value=0.003, FDR=0.026) associated to the response to AZD8055, an anti-mTOR inhibitor. The cells with NOTCH1 mutations were significantly more resistant to the drug than cells with wild type NOTCH1 (Figure 30). NOTCH1 gene encodes a class I transmembrane protein functioning as a ligand-activated transcription factor, which plays an important role in a number of cellular functions. NOTCH1 has been considered as an oncogene in several lymphoid malignancies, but recent evidence has also

suggested that Notch signaling could be a tumor suppressor in myeloid malignancies (Lobry, et al. 2014). One study has shown that NOTCH1 knockdown reduced the level of mechanistic target of rapamycin (mTOR) protein in the monoblastic leukemia cell line (Okuhashi, et al. 2013). The reason why cell lines with NOTCH1 mutations are more resistant to AZD8055 is still not clear, further *in vitro* studies are needed to better understand the underlying mechanistic explanation.

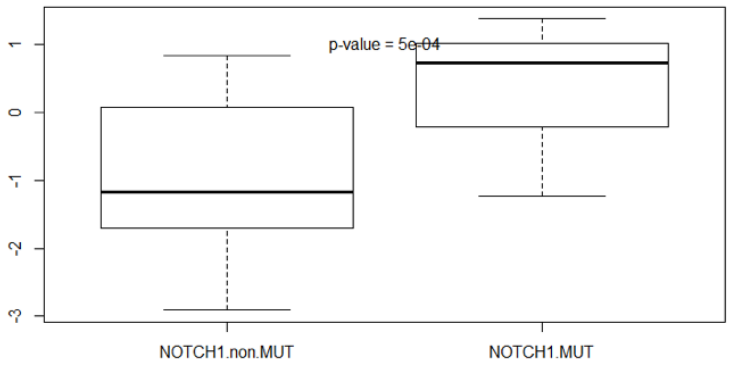


Figure 30. The cells with NOTCH1 mutations are significantly more resistant to AZD8055, an anti-mTOR inhibitor, than cells with normal NOTCH1.

CDKN2A mutations (p-value=0.004, FDR=0.033) and CDKN2A amplification (p-value=0.001, FDR=0.008) were identified as top predictors for JW7521, another anti-mTOR inhibitor. CDKN2A mutation was identified as predictive of sensitivity to JW7521 while amplification of CDKN2A was identified as predictive of resistance to JW7521. We hypothesized that the activation of CDKN2A is predictive of the drug response to JW752 and examined the drug response between cell lines with CDKN2A inactivation (cell lines with the

inactivating mutation or/and with a deletion) and cell lines with CDKN2A activation (cell lines without the mutation but an amplification). We showed that cell lines that are CDKN2A activated were significantly more resistant to JW7521 than cell lines that were CDKN2A inactivated (Figure 31). CDKN2A, frequently mutated or deleted in cancer, is a protein coding gene coding for tumor suppressors p16 and p14arf (McWilliams, et al. 2011). One study has shown that inactivating mutations of CDKN2A result in elevated mTORC1/2 signaling (Souroullas and Sharpless 2015) and this might offer an explanation for identifying the cell lines with inactivated CDKN2A as more sensitive to an anti-mTOR inhibitor.

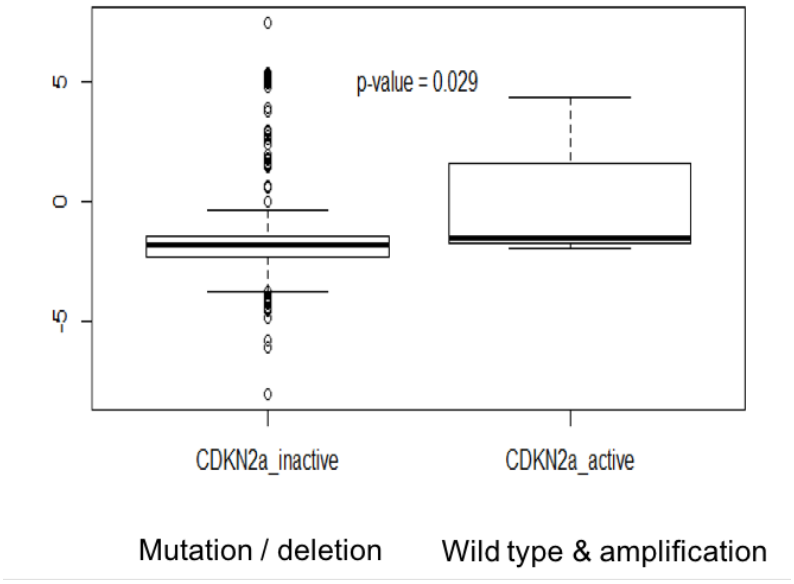


Figure 31. Cell lines that are CDKN2A activated (defined as cell lines with an amplification and not mutated) were significantly more resistant to JW7521 than cell lines that were CDKN2A inactivated (defined as cell lines with a mutation or/and a deletion).

The deletion of PHLPP2 was identified to be associated with drug response to CI-1040, an anti-MEK inhibitor (Figure 32). PHLPP2 is an important regulator of AKT serine-threonine kinases, which inactive AKT1 and AKT3 by dephosphorylation (Brognard, et al. 2007). A loss of PHLPP2 in the cells could lead to increased activity of AKT and ERK pathways, which promote cell survival.

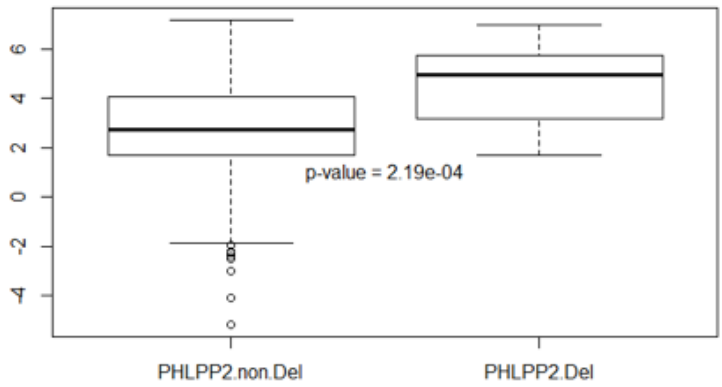


Figure 32. Cells with PHLPP2 deletion are significantly more resistant to CCI-1040, an anti-MEK inhibitor than cells with no deletion of PHLPP2.

Conclusion

In this part, we developed a 2-step predictive model that combines an Elastic net model to impute and adjust the drug response using gene expression data, and a Random Forest model to identify predictive biomarkers. We first showed that our model has better capacity of identifying the direct targets of the targeted therapies as compared to other models tested. We then used our model to identify new biomarkers that have never been identified. Some evidence was found in the literature to support these finding which makes these

biomarkers good candidates to further investigation for drug target development.

Evaluating combination of predictive biomarkers using Random Forest

Although many studies have shown the feasibility and efficiency in identifying biomarkers with mathematical models combined with high throughput genomic data, there are still some unresolved issues. One for instance is that most models identify single biomarkers. Even for models that take into consideration the correlated genomic features such as elastic net, the relevance of these features to the outcome to predict is still evaluated separately. Although biomarkers based on single gene alteration are the most commonly used in clinical applications today, driver events can be due to a combination of multiple alterations given the complexity of cancer biology. It is therefore necessary to identify combinations of predictive biomarkers, which paired with multiple targeted therapies have the potential to play important roles as biomarkers in the future.

In order to identify combination of predictors, we developed a method to evaluate the importance of coupled predictors in Random Forest thanks to the structure of the tree predictors. In random forest model, the importance of each variable is computed by two methods. One is computed from permuting the out-of-bag data, where the out-of-bag data is the observations not used in the construction of the tree model. For each tree, the prediction error on the out-of-bag data is computed before and after permuting each variable. The difference between the two are then averaged and normalized over all trees. The other method is to compute the total decrease in node impurities while splitting based on each variable, then averaged over all trees. studies have been made to evaluate the importance of the combination of variables mainly based on the out-of-bag importance where the importance of paired variables are computed by measuring the prediction error from permuting the two

variables jointly (Bureau, et al. 2005; Ishwaran 2007; Ishwaran, et al. 2010). Here we propose a method to compute the importance of paired variables based on the total decrease of node impurities while splitting. The utility of these two methods of evaluating paired importance should be compared in simulated and real data sets in further study.

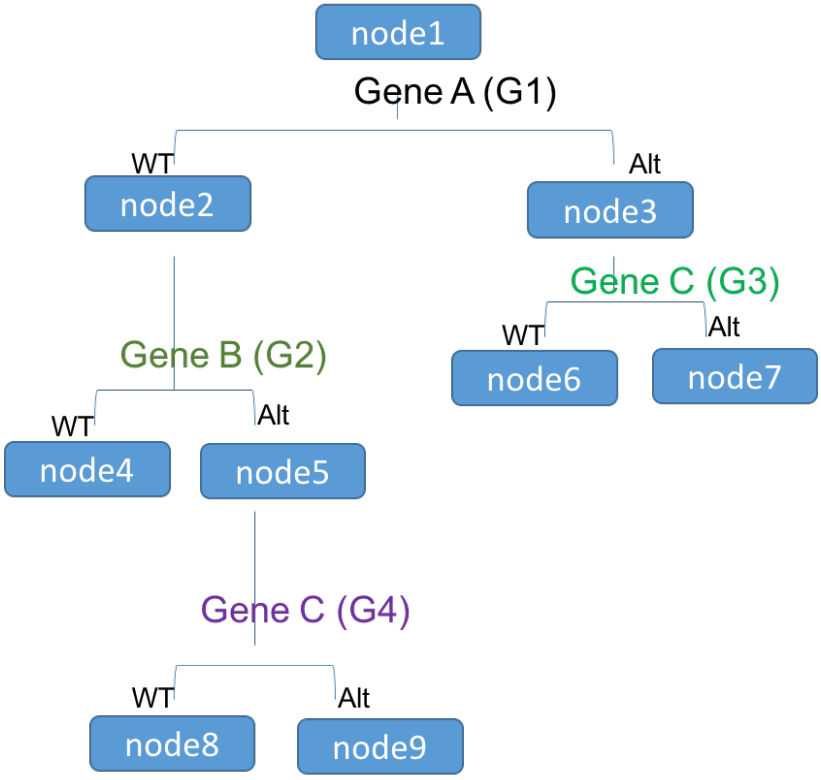
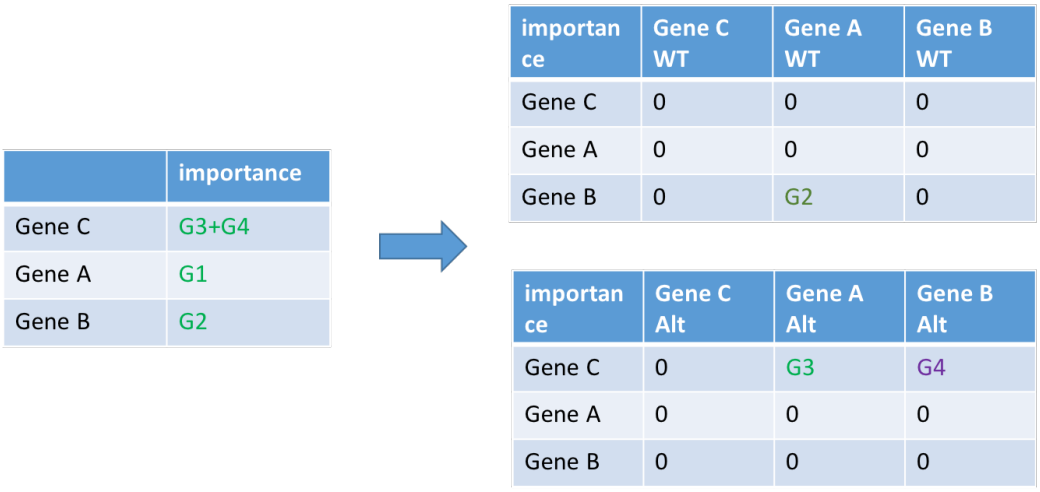


Figure 33. A decision tree model with binary predictors.

The tree model we used in Random forest framework is illustrated in Figure 33, where the parent nodes are split into daughter nodes based on the status of the selected genomic alterations. For example, the node1 is split into node2 and node3 based on the status of the gene A (with a gain of purity of G1), then the node3 is split into node6 and node7 based on the status of gene C (with a gain

of purity of G3). If this pattern (gene C is always chosen to split the node while the parent node contains only altered gene A) is observed repeatedly in the ensemble of tree, we assume that there is a predictive power of the combination of gene A and gene C.

To evaluate the predictive power of all the paired predictors in a Random Forest algorithm, we made a modification based in the R package "RamdomForest" implementation to output not only the predictor to split but also the predictor used to generate the parent node to split as shown in Figure 34.



	importance
Gene C	G3+G4
Gene A	G1
Gene B	G2

importance	Gene C WT	Gene A WT	Gene B WT
Gene C	0	0	0
Gene A	0	0	0
Gene B	0	G2	0

importance	Gene C Alt	Gene A Alt	Gene B Alt
Gene C	0	G3	G4
Gene A	0	0	0
Gene B	0	0	0

Figure 34. Matrix for evaluating predictors in combination.

We calculated the importance of paired predictors for a set of 19 targeted therapies of 7 targets (Table 6) and identified combinations of predictors.

Table 6. List of drugs for which combinations of predictive biomarkers were evaluated.

Treatment	Targets	Drug
Anti-AKT	AKT	A443654, AKTINHIBITORVIII, MK2206
Anti-EGFR	EGFR	ERLOTINIB, GEFITINIB
Anti-EGFR,ERBB2	EGFR,ERBB2	BIBW2992, LAPATINIB
Anti-FGFR	FGFR	PD173074
Anti-MEK	MEK	AZD6244, CI1040, PD0325901, RDEA119
Anti-MTOR	MTOR	AZD8055, RAPAMYCIN, TEMSIROLIMUS
Anti-PI3K	PI3K	AZD6482, GDC0941, JW7521, NVPBEZ235

For example, the top 6 combinations of predictive biomarkers for AKT inhibitor VIII, an anti-AKT inhibitor are shown in Figure 35. We observed that cell lines with mutation of PIK3CA were more sensitive than the wild type cell lines (p-value= 1.78e-6) while cell lines with FGF5 deletion had similar drug respond than cell lines with no deletion of FGF5. Surprisingly, the combination of PIK3CA mutation and FGF5 deletion in cell lines leaded to drug resistance (p-value=0.09) while none of these two predictors in isolation showed such an effect (Figure 35).

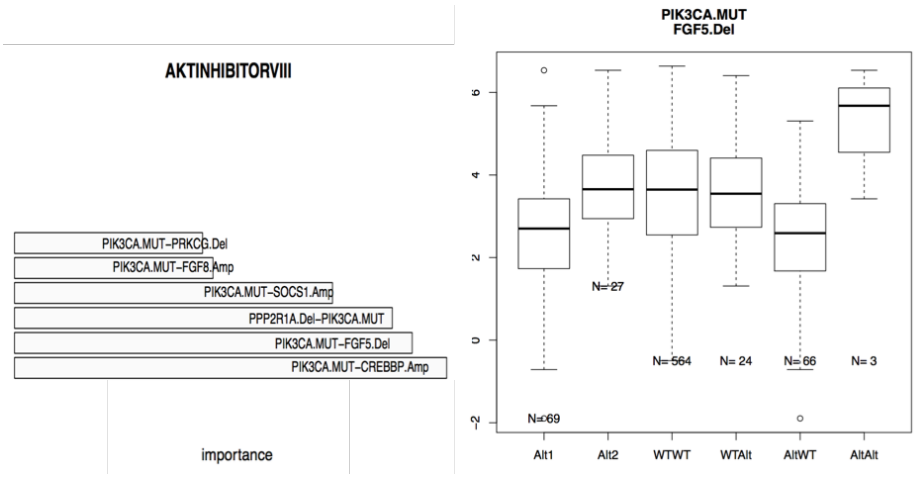


Figure 35. Left: top 6 combinations of predictive biomarkers identified for ATK inhibitor VIII (pv<0.005), an anti-AKT inhibitor. Right: Drug response to AKT inhibitor VIII of cell lines based on their mutational status of PIK3CA and copy number of FGF5. Boxes from the left to right represent the drug response of cell lines: PIK3CA mutated, FGF5 deleted, PIK3CA wild type and FGF5 not deleted, PIK3CA wild type and FGF5 deleted, PIK3CA mutated and FGF5 not deleted, PIK3CA mutated and FGF5 deleted.

We showed that the co-existence of PIK3CA mutation and FGF5 deletion is predictive to resistance to ATK inhibitor VIII. Clinical evidence suggested that patients with PIK3CA-mutant tumors might benefit from treatment with Akt inhibitors (U, et al. 2013), Our results, that need to be validated, suggested that the alteration status of FGF5 should be verified before giving an anti-AKT inhibitor to patients with PIK3CA mutation.

Combinations of predictive biomarkers were also identified for CI1040, an anti-MEK inhibitor (Figure 36). We observed that cells with both BRAF mutation and FGF9 deletion showed significant resistance to CI1040 (p-value=0.04) while cells with BRAF mutation alone were sensitive (p-value=3.5e-13) and

cells with FGF9 deletion alone showed no predictive effect (p-value=0.93). Similar results were observed for BRAF mutation and PIK3CG mutation, cells with both BRAF mutation and PIK3CG mutation showed significant resistance to CI1040 (p-value=0.04) while cells with BRAF mutation alone were sensitive (p-value=2.8e-13) and cells with PIK3CG mutation alone showed no difference in drug response compared to WT cells (p-value=0.55).

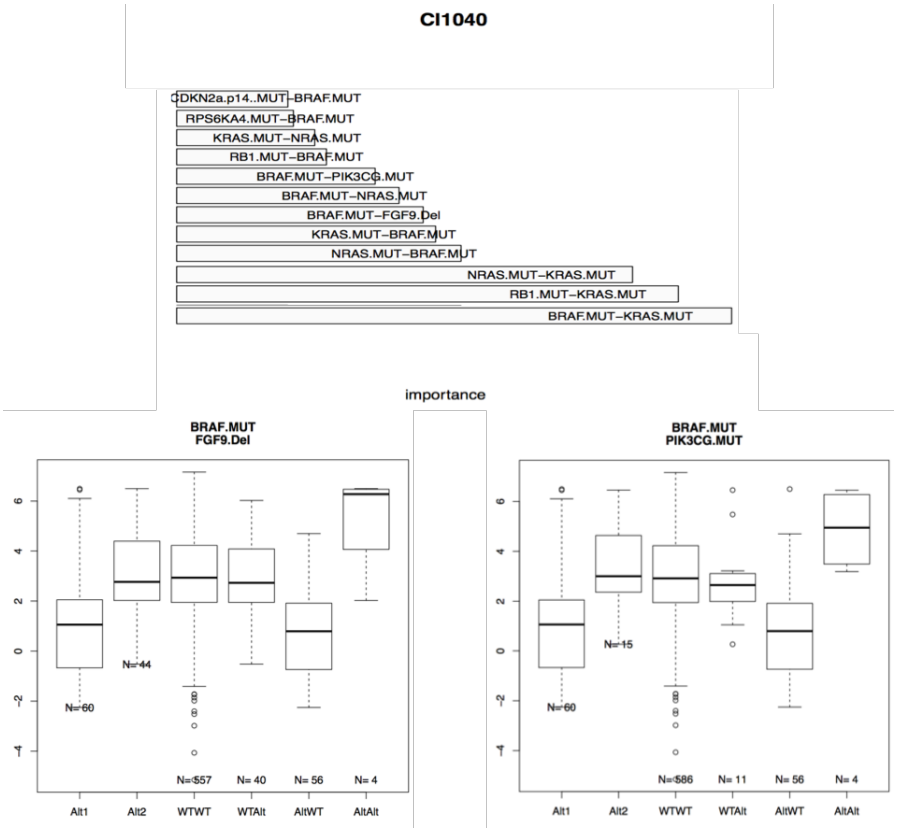


Figure 36. top: top 12 predictive combinations of biomarkers identified for CI1040, an anti-MEK inhibitor; bottom left: cells with the combination of BRAF mutation and FGF9 deletion showed significant resistance to CI1040 while cells with BRAF mutation alone were sensitive and cells with FGF9 deletion alone showed no difference in drug response compared to WT cells; bottom right: cells with BRAF mutation and PIK3CG mutation are more resistant than the other cells.

For AZD8055, an anti-mTOR inhibitor, combinations of biomarkers predictive of sensitivity and of resistance were identified. Cell lines with both NRAS mutation and CDKN2A mutation were more sensitive to AZD8055 than cell lines that had neither NRAS mutation nor CDKN2A mutation (p-value=0.01, Figure 37, top right); cell lines with BRCA2 amplification and KRAS mutation were more resistant to AZD8055 than cells with neither BRCA2 amplification nor KRAS mutation (p-value=0.004, Figure 37, bottom left); the combination of FGF9 amplification and KRAS mutation was significantly associated to drug resistance to AZD8055 (p-value=0.0006, Figure 37, bottom right).

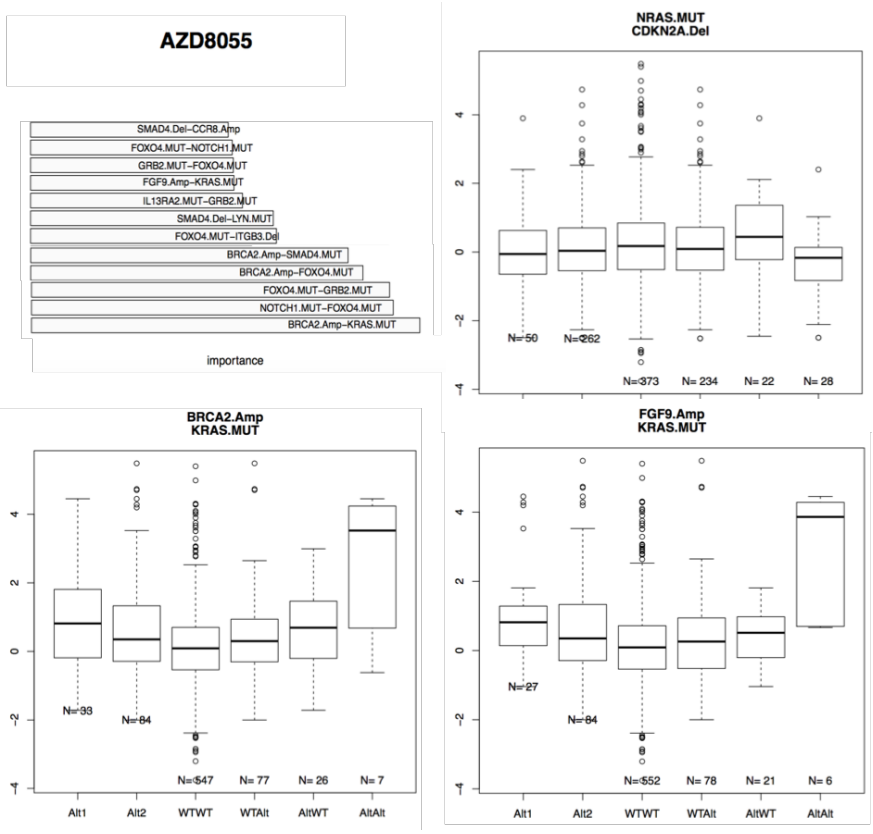


Figure 37. top left: Top 12 predictive biomarkers identified for AZD8055, an anti-mTOR inhibitor; top right: cells with the combination of NRAS mutation and CDKN2A are more sensitive to AZD8055 than other cells; bottom left: cells with BRCA2 amplification and KRAS mutation are more resistant than the other cells; bottom right: cells with FGF deletion and KRAS mutation are more resistant than the other cells.

The combination of CDKN2A mutation with either the deletion of SMAD4 (p-value=0.001) or MAP3K7 (p-value=0.002) was identified as predictive to the resistance of cell lines to TEMSIROLIMUS, an anti-mTOR inhibitor (Figure 38, top left and top right). Interestingly, SMAD4 and MAP3K7 are not located in the same region of the genome (SMAD4 is located at 18q21.2 while MAP3K7 is located at 6q15), but are both involved in the MAPK pathway (Figure 38 bottom), suggesting that CDKN2A mutation combined with alteration of the MAPK pathway could lead to drug resistance to TEMSIROLIMUS.

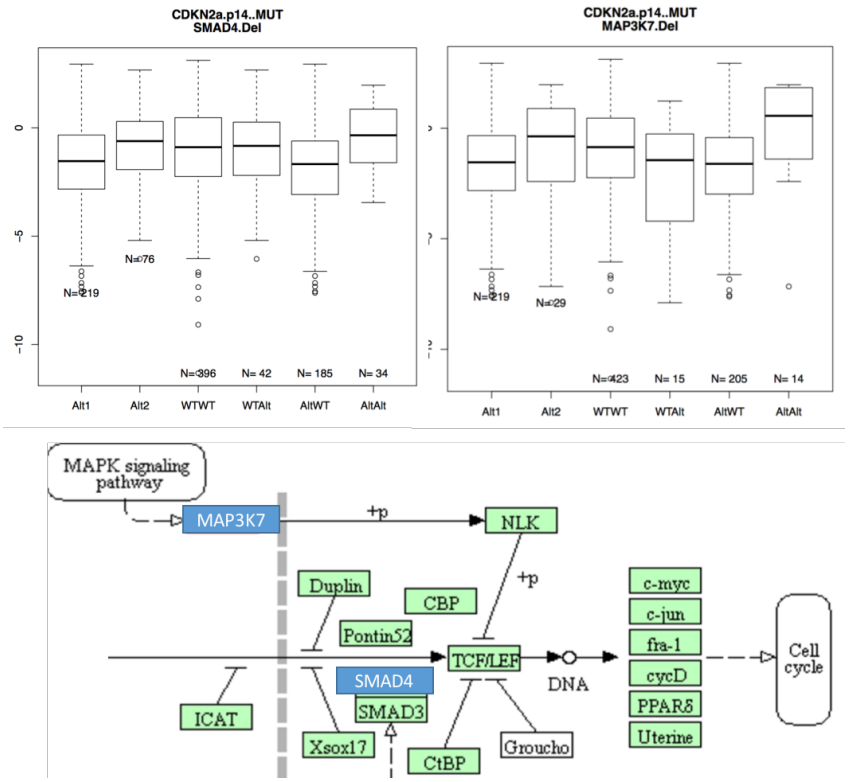


Figure 38. The combinations of CDKN2A mutation of both SMAD4 deletion and MAP3K7 deletion were predictive to drug resistance of cell lines to TEMSIROLIMUS, an anti-mTOR inhibitor; the gene SMAD4 and MAK3K7 are both involve in the MAPK pathway indicating that the alteration of the pathway MAPK combined with an CDKN2A mutation has some impact on the drug response of cell lines to TEMSIROLIMUS.

Conclusion

In this part, we developed a method to evaluate the predictive power of combinations of predictors based on Random Forest model. To our knowledge, this is the first method to select combinations of predictors instead of single ones. The identification of predictive combinations of predictors has an important implication in the study of drug resistance. It can help us better understand the mechanisms of drug resistance and eventually prevent it by identifying complex genomic patterns that are predictive of specific targeted therapy response.

Discussion

In this thesis, we discussed mainly two mathematical and informatics applications in the context of personalized medicine for cancer patients.

The first application is cmDetect, a method that we propose to use as a complementary tool to traditional somatic mutation callers to recover false negatives due to ctDNA contamination of the germline DNA extracted from whole blood using whole exome sequencing data. We validated the accuracy of cmDetect with both simulated data and patients' data. By addressing this issue, we not only demonstrated that ctDNA is detectable in sequencing data of whole blood DNA and has an impact on somatic mutation identification, but also the importance of the accurate identification of somatic mutations that may have a direct impact on treatment decision in the context of personalized medicine for cancer patients.

The second application is a mathematical model to identify predictive biomarkers of targeted therapies drug response. We first showed that co-existing alterations in multiple driver genes can cause drug resistance in both cell lines and patients. We then identified single as well as combinations of predictive biomarkers for a set of targeted therapies using cell line data. The identified biomarkers were supported by statistical tests but further *in vitro* and *in vivo* validation is still needed. We offered an effective method for identifying complex predictive patterns of targeted therapies in a systematic manner, which gives an insight of the underlying biological mechanism of drug response. The predictive biomarkers identified could not only guide further research for identifying new targets to targeted therapies in preclinical studies, but also help designing better cancer clinical trials.

There are still some issues to address in order to improve the predictive models:

Further validation is needed for the biomarker identified

We have shown that co-existing alterations in multiple driver genes can lead to drug resistance and validated the results in a cohort of cancer patients in the MOSCATO clinical trial. *In vitro* validation is still needed to better understand the underlying explanation of such an observation. This is also true for the predictions of the single and combined predictive biomarkers identified in our 2-step model.

Genomic data of real tumors is needed to build the predictive model

Cell lines are useful tools for characterizing genomic profiles of cancer cells and measuring cancer cell drug response to a large set of drugs but the gap between cell lines and real tumors is not negligible. One of the reasons is that tumors are heterogeneous and often formed of multiple sub clones of different genetic backgrounds. As cell lines are derived from one cancer cell, they do not capture the global picture of the tumor. Under the selection pressure of a targeted therapy, a sub clone of cancer cells with a driver mutation in a gene other than the target of the drug can develop rapidly and lead to the drug resistance of the tumor. This type of acquired drug resistance cannot be observed in cell lines thus cannot be learned by a predictive model built using

cell lines data. Although a large number of patients' sequencing data is available in datasets such as TCGA, the clinical response of these patients are often under chemotherapies but not targeted therapies because these patients are mostly with primary cancer. The clinical trials where the cancer patients are treated with targeted therapies based on their genomic profile such as SAFIR01, SAFIR02 and MOSCATO, offer us a unique opportunity to study cancer genomics in advanced cancer patients.

The p>>n problem in the predictive model

The p>>n problem refers to a problem in predictive or regression models where the number of predictors is much bigger than the number of samples. Genomic data usually contains hundreds or even thousands of genes that are used as predictors while the number of available samples is often small (ex: 20,000 genes vs 1,000 cell lines). Although random forest used in our 2-step predictive model has been shown to have a relative good performance in making prediction and identifying predictors facing p>>n problem, methods of reducing predictor dimension can still be helpful. Many methods have been proposed for such purpose such as principal component analysis (Jolliffe 1986), factor analysis (Cattell 1952) and feature selection such as embedded.

Multitask predictive model

Given the fact that some drugs target the same gene and ought to have similar effect in cell lines, multi-task models can be used to improve predictive performance as compared to learning models separately. Multi-task model is a

machine learning method that learns multiple models together by sharing the information in order to improve predictive accuracy. For example, Bayesian multitask multiple kernel learning (MKL, (Gönen and Alpaydin 2011)) has been used for drug response prediction using gene expression data of cell lines and outperformed all the other methods in a collective drug response prediction challenge (Costello, et al. 2014). We can also combine multi-task method with random forest to build multi-task random forest models (Simm, Abril, and Sugiyama 2014; Haider, et al. 2015; Yao, Yang, and Zhan 2013), which have shown good performance in predictive modeling.

Annex

supplementary data—cmDetect

Filter patient specific polymorphism

Most of the mutations identified in the tumor sample and in the normal sample are either germline mutations or polymorphisms. To filter out these mutations that are not caused by ctDNA contamination, we used 3 methods shown in Figure 39. We first hypothesized that a germline mutation would have comparable allele frequency in the tumor sample and the normal sample and removed all the mutations with a non-significant difference of allele frequency between the tumor and the normal sample (pv > 0.01). We then compute the MAF based on our cohort of patients in the analysis and removed all the mutations with a MAF > 0.01. Finally, we generated, for each patient, a distribution of the allele frequency in the normal sample of all the heterozygous mutations and removed the mutations with a high probably of being a heterozygous mutation in the normal sample.

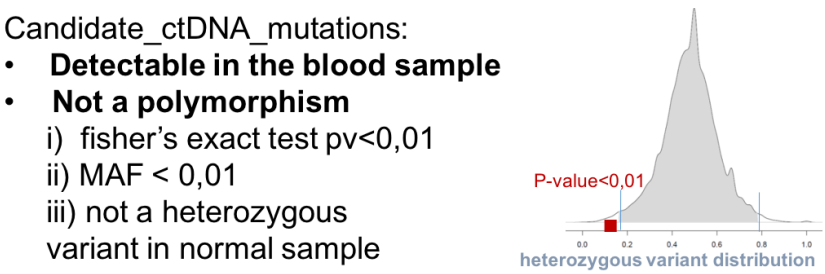


Figure 39. Methods to filter out polymorphisms.

Filter false positives from the pool of blood samples

As the quantity of ctDNA in the whole blood sample is small, it is difficult to distinguish a variant due to ctDNA contamination from a sequencing bias (sequencing error). Here we hypothesized that a variant caused by sequencing bias has an equal probability to occur in all the samples sequenced by the same method while a variant caused by ctDNA contamination will only be present in the patients where the same variant in detected in their tumor. Therefore, we computed the p-value that a given variant identified in the blood sample is a true ctDNA mutation using a one sided binomial test (Figure 40).

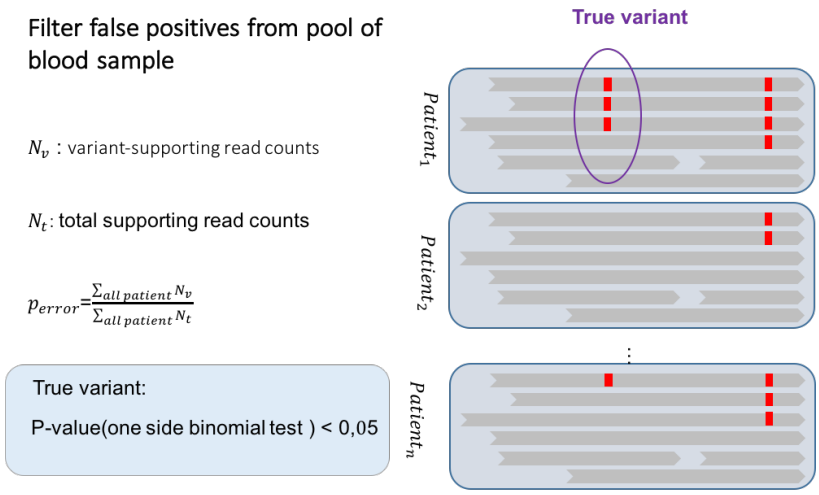


Figure 40. Filter false positives using the pool of blood samples.

Simulation data for cmDetect performance evaluation

To evaluate the performance of cmDetect combined with traditional somatic mutation callers, we built a set of simulated tumor/normal whole exome sequencing data using real sequencing data from 1000 Genomes data set as shown in Figure 41. Whole exome sequencing data of individuals from phase 3

with an average of coverage of 80-200 were retrieved. In order to implement bamsurgeon (Ewing, et al. 2015) to simulate tumor sequencing data, we remapped the data to the reference genome hg19 using samtools (Li, et al. 2009) and bwa (Li and Durbin 2010). A set of somatic mutations and small insertions and deletions were then retrieved from COSMIC data set, only mutations that have been identified in more than 2 samples were used. Bamsurgeon was used to insert the preselected mutations and indels into the whole exome sequencing data of normal individuals to simulate tumor samples. Contaminated whole blood sequencing samples were generated using Samtools by mixing randomly the reads from the normal sample and the simulated tumor sample at a random rate.

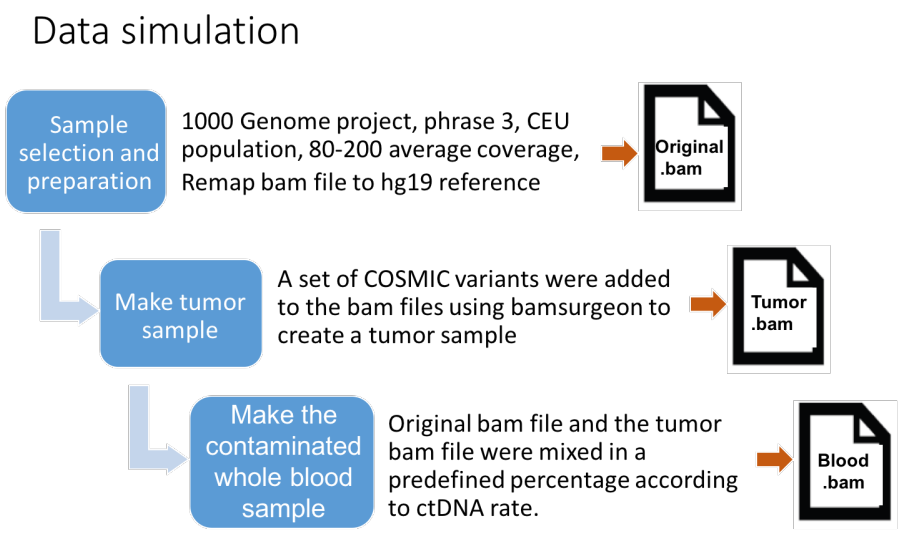


Figure 41. Data simulation for cmDetect performance evaluation.

Reference

Ai, B., et al. 2016. "Circulating Cell-Free Dna as a Prognostic and Predictive Biomarker in Non-Small Cell Lung Cancer." *Oncotarget* (Jun). http://dx.doi.org/10.18632/oncotarget.10069.

Alexandrov, L. B., et al. 2013. "Signatures of Mutational Processes in Human Cancer." *Nature* 500, no. 7463 (Aug): 415-21. http://dx.doi.org/10.1038/nature12477.

Andrews, Simon. 2014. *Fastqc: A Quality Control Tool for High Throughput Sequence Data.*: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

André, F., et al. 2014. "Comparative Genomic Hybridisation Array and Dna Sequencing to Direct Treatment of Metastatic Breast Cancer: A Multicentre, Prospective Trial (Safir01/Unicancer)." *Lancet Oncol* 15, no. 3 (Mar): 267-74. http://dx.doi.org/10.1016/S1470-2045(13)70611-9.

Ascierto, P. A., et al. 2012. "The Role of Braf V600 Mutation in Melanoma." *J Transl Med* 10 (Jul): 85. http://dx.doi.org/10.1186/1479-5876-10-85.

Barretina, J., et al. 2012. "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity." *Nature* 483, no. 7391 (Mar): 603-7. http://dx.doi.org/10.1038/nature11003.

Baylin, S. B. 2005. "Dna Methylation and Gene Silencing in Cancer." *Nat Clin Pract Oncol* 2 Suppl 1 (Dec): S4-11. http://dx.doi.org/10.1038/ncponc0354.

Bean, J., et al. 2008. "Acquired Resistance to Epidermal Growth Factor Receptor Kinase Inhibitors Associated with a Novel T854a Mutation in a Patient with Egfr-Mutant Lung Adenocarcinoma." *Clin Cancer Res* 14, no. 22 (Nov): 7519-25. http://dx.doi.org/10.1158/1078-0432.CCR-08-0151.

Ben-Yaacov, E., and Y. C. Eldar. 2008. "A Fast and Flexible Method for the

Segmentation of Acgh Data." *Bioinformatics* 24, no. 16 (Aug): i139-45. http://dx.doi.org/10.1093/bioinformatics/btn272.

Biankin, A. V., S. Piantadosi, and S. J. Hollingsworth. 2015. "Patient-Centric Trials for Therapeutic Development in Precision Oncology." *Nature* 526, no. 7573 (Oct): 361-70. http://dx.doi.org/10.1038/nature15819.

Bienkowska, J. R., et al. 2009. "Convergent Random Forest Predictor: Methodology for Predicting Drug Response from Genome-Scale Data Applied to Anti-Tnf Response." *Genomics* 94, no. 6 (Dec): 423-32. http://dx.doi.org/10.1016/j.ygeno.2009.08.008.

Bird, A. 2002. "Dna Methylation Patterns and Epigenetic Memory." *Genes Dev* 16, no. 1 (Jan): 6-21. http://dx.doi.org/10.1101/gad.947102.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30, no. 15 (Aug): 2114-20. http://dx.doi.org/10.1093/bioinformatics/btu170.

Boveri, T. 1914. *Zur Frage Der Entstehung Maligner Tumoren* . Gustav Fischer.

Boveri, T. 2008. "Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and Annotated by Henry Harris." *J Cell Sci* 121 Suppl 1 (Jan): 1-84. http://dx.doi.org/10.1242/jcs.025742.

Breiman, L. 2001. *Random Forests*. Vol. **45** (1): 5–32.doi:10.1023/A:1010933404324. *Machine Learning*.

Brognard, J., et al. 2007. "Phlpp and a Second Isoform, Phlpp2, Differentially Attenuate the Amplitude of Akt Signaling by Regulating Distinct Akt Isoforms." *Mol Cell* 25, no. 6 (Mar): 917-31. http://dx.doi.org/10.1016/j.molcel.2007.02.017.

Bureau, A., et al. 2005. "Identifying Snps Predictive of Phenotype Using Random

Forests." *Genet Epidemiol* 28, no. 2 (Feb): 171-82. http://dx.doi.org/10.1002/gepi.20041.

Burris, H. A. 2001. "Docetaxel (Taxotere) Plus Trastuzumab (Herceptin) in Breast Cancer." *Semin Oncol* 28, no. 1 Suppl 3 (Feb): 38-44.

Cattell, R. B. 1952. *Factor Analysis*: Harper.

Chan, B. A., and B. G. Hughes. 2015. "Targeted Therapy for Non-Small Cell Lung Cancer: Current Standards and the Promise of the Future." *Transl Lung Cancer Res* 4, no. 1 (Feb): 36-54. http://dx.doi.org/10.3978/j.issn.2218-6751.2014.05.01.

Charles Ferté, Christophe Massard, Ecaterina Ileana, Antoine Hollebecque, Ludovic Lacroix, Samy Ammari, Maud Ngo-Camus, Rastislav Bahleda, Anas Gazzah, Andrea Varga,Sophie Postel-Vinay, Yohann Loriot, Nathalie Auger, Valerie Koubi-Pick, Bastien Job, Thierry De Baere, Frederic Deschamps, Philippe Vielh, Vladimir Lazar, Marie-Cécile Le Deley, Catherine Richon, vincent ribrag, eric deutsch, eric angevin, gilles vassal, Alexander Eggermont, Fabrice André and Jean-Charles Soria. 2014. *Abstract Ct240: Molecular Screening for Cancer Treatment Optimization (Moscato 01): A Prospective Molecular Triage Trial; Interim Analysis of 420 Patients*. **DOI:** 10.1158/1538-7445.AM2014-CT240 ed.

Chen, K., et al. 2007. "Methylation of Multiple Genes as Diagnostic and Therapeutic Markers in Primary Head and Neck Squamous Cell Carcinoma." *Arch Otolaryngol Head Neck Surg* 133, no. 11 (Nov): 1131-8. http://dx.doi.org/10.1001/archotol.133.11.1131.

Chen, Z., et al. 2005. "Crucial Role of P53-Dependent Cellular Senescence in Suppression of Pten-Deficient Tumorigenesis." *Nature* 436, no. 7051 (Aug): 725-30. http://dx.doi.org/10.1038/nature03918.

Cheng, F., L. Su, and C. Qian. 2016. "Circulating Tumor Dna: A Promising Biomarker in the Liquid Biopsy of Cancer." *Oncotarget* (May).

http://dx.doi.org/10.18632/oncotarget.9453.

Cibulskis, K., et al. 2013. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples." *Nat Biotechnol* 31, no. 3 (Mar): 213-9. http://dx.doi.org/10.1038/nbt.2514.

Costello, J. C., et al. 2014. "A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms." *Nat Biotechnol* 32, no. 12 (Dec): 1202-12. http://dx.doi.org/10.1038/nbt.2877.

Dai, M., et al. 2010. "Ngsqc: Cross-Platform Quality Analysis Pipeline for Deep Sequencing Data." *BMC Genomics* 11 Suppl 4: S7. http://dx.doi.org/10.1186/1471-2164-11-S4-S7.

Davies, B. R., et al. 2012. "Preclinical Pharmacology of Azd5363, an Inhibitor of Akt: Pharmacodynamics, Antitumor Activity, and Correlation of Monotherapy Activity with Genetic Background." *Mol Cancer Ther* 11, no. 4 (Apr): 873-87. http://dx.doi.org/10.1158/1535-7163.MCT-11-0824-T.

Dees, N. D., et al. 2012. "Music: Identifying Mutational Significance in Cancer Genomes." *Genome Res* 22, no. 8 (Aug): 1589-98. http://dx.doi.org/10.1101/gr.134635.111.

Despierre, E., et al. 2014. "Somatic Copy Number Alterations Predict Response to Platinum Therapy in Epithelial Ovarian Cancer." *Gynecol Oncol* 135, no. 3 (Dec): 415-22. http://dx.doi.org/10.1016/j.ygyno.2014.09.014.

Dettling, M., and P. Bühlmann. 2003. "Boosting for Tumor Classification with Gene Expression Data." *Bioinformatics* 19, no. 9 (Jun): 1061-9.

Dong, Z., et al. 2015. "Anticancer Drug Sensitivity Prediction in Cell Lines from Baseline Gene Expression through Recursive Feature Selection." *BMC Cancer* 15: 489. http://dx.doi.org/10.1186/s12885-015-1492-6.

Drake, J. W., et al. 1998. "Rates of Spontaneous Mutation." *Genetics* 148, no. 4 (Apr): 1667-86.

Ewing, A. D., et al. 2015. "Combining Tumor Genome Simulation with Crowdsourcing to Benchmark Somatic Single-Nucleotide-Variant Detection." *Nat Methods* 12, no. 7 (Jul): 623-30. http://dx.doi.org/10.1038/nmeth.3407.

Flaherty, K. T., et al. 2012. "Improved Survival with Mek Inhibition in Braf-Mutated Melanoma." *N Engl J Med* 367, no. 2 (Jul): 107-14. http://dx.doi.org/10.1056/NEJMoa1203421.

Freund, Yoav, and Robert E. Schapire. 1999. *A Short Introduction to Boosting*. Vol. 14(5):771-780: Journal of Japanese Society for Artificial Intelligence.

Gambacorti-Passerini, C., and R. Piazza. 2015. "Imatinib--a New Tyrosine Kinase Inhibitor for First-Line Treatment of Chronic Myeloid Leukemia in 2015." *JAMA Oncol* 1, no. 2 (May): 143-4. http://dx.doi.org/10.1001/jamaoncol.2015.50.

Garnett, M. J., et al. 2012. "Systematic Identification of Genomic Markers of Drug Sensitivity in Cancer Cells." *Nature* 483, no. 7391 (Mar): 570-5. http://dx.doi.org/10.1038/nature11005.

Geeleher, P., N. J. Cox, and R. S. Huang. 2014. "Clinical Drug Response Can Be Predicted Using Baseline Gene Expression Levels and in Vitro Drug Sensitivity in Cell Lines." *Genome Biol* 15, no. 3: R47. http://dx.doi.org/10.1186/gb-2014-15-3-r47.

Gordon, L. I. 2016. "Precision Monitoring by Next-Generation Sequencing in Lymphoma: Circulating Tumor Dna as a New Biomarker." *Oncology (Williston Park)* 30, no. 8 (Aug).

Griffiths AJF, Gelbart WM, Miller JH. 1999. *Chromosomal Rearrangements*. New York:

W. H. Freeman: Modern Genetic Analysis.

Gönen, M., and E. Alpaydin. 2011. *Multiple Kernel Learning Algorithms*. 12 vols.: Mach. Learn. Res.

Haibe-Kains, B., et al. 2013. "Inconsistency in Large Pharmacogenomic Studies." *Nature* 504, no. 7480 (Dec): 389-93. http://dx.doi.org/10.1038/nature12831.

Haider, S., et al. 2015. "A Copula Based Approach for Design of Multivariate Random Forests for Drug Sensitivity Prediction." *PLoS One* 10, no. 12: e0144490. http://dx.doi.org/10.1371/journal.pone.0144490.

HOWARD, B. D., and I. TESSMAN. 1964. "Identification of the Altered Bases in Mutated Single-Stranded Dna. Ii. In Vivo Mutagenesis by 5-Bromodeoxyuridine and 2-Aminopurine." *J Mol Biol* 9 (Aug): 364-71.

Hupé, P., et al. 2004. "Analysis of Array Cgh Data: From Signal Ratio to Gain and Loss of Dna Regions." *Bioinformatics* 20, no. 18 (Dec): 3413-22. http://dx.doi.org/10.1093/bioinformatics/bth418.

Hutchison, C. A. 2007. "Dna Sequencing: Bench to Bedside and Beyond." *Nucleic Acids Res* 35, no. 18: 6227-37. http://dx.doi.org/10.1093/nar/gkm688.

Ishwaran, Hemant. 2007. *Variable Importance in Binary Regression Trees and Forests*. Vol. Vol. 1 (2007) 519-537: Electronic Journal of Statistics.

Ishwaran, hemant, et al. 2010. *High-Dimensional Variable Selection for Survival Data*. Vol. 105: Journal of the American Statistical Association.

Jokinen, E., and J. P. Koivunen. 2015. "Mek and Pi3k Inhibition in Solid Tumors: Rationale and Evidence to Date." *Ther Adv Med Oncol* 7, no. 3 (May): 170-80. http://dx.doi.org/10.1177/1758834015571111.

Jolliffe, I.T. 1986. *A Tutorial on Principal Components Analysis*: Springer-Verlag.

Jovelet, C., et al. 2016. "Circulating Cell-Free Tumor Dna Analysis of 50 Genes by Next-Generation Sequencing in the Prospective Moscato Trial." *Clin Cancer Res* 22, no. 12 (Jun): 2960-8. http://dx.doi.org/10.1158/1078-0432.CCR-15-2470.

Kaelin, W. G. 2005. "The Concept of Synthetic Lethality in the Context of Anticancer Therapy." *Nat Rev Cancer* 5, no. 9 (Sep): 689-98. http://dx.doi.org/10.1038/nrc1691.

Kandoth, C., et al. 2013. "Mutational Landscape and Significance across 12 Major Cancer Types." *Nature* 502, no. 7471 (Oct): 333-9. http://dx.doi.org/10.1038/nature12634.

Kim, D., et al. 2013. "Tophat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions." *Genome Biol* 14, no. 4: R36. http://dx.doi.org/10.1186/gb-2013-14-4-r36.

Koboldt, D. C., et al. 2012. "Varscan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing." *Genome Res* 22, no. 3 (Mar): 568-76. http://dx.doi.org/10.1101/gr.129684.111.

Larkin, J., et al. 2014. "Combined Vemurafenib and Cobimetinib in Braf-Mutated Melanoma." *N Engl J Med* 371, no. 20 (Nov): 1867-76. http://dx.doi.org/10.1056/NEJMoa1408868.

Leary, R. J., et al. 2008. "Integrated Analysis of Homozygous Deletions, Focal Amplifications, and Sequence Alterations in Breast and Colorectal Cancers." *Proc Natl Acad Sci U S A* 105, no. 42 (Oct): 16224-9. http://dx.doi.org/10.1073/pnas.0808041105.

Li, H., and R. Durbin. 2010. "Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 26, no. 5 (Mar): 589-95.

http://dx.doi.org/10.1093/bioinformatics/btp698.

Li, H., et al. 2009. "The Sequence Alignment/Map Format and Samtools." *Bioinformatics* 25, no. 16 (Aug): 2078-9. http://dx.doi.org/10.1093/bioinformatics/btp352.

Lobry, C., et al. 2014. "Notch Signaling: Switching an Oncogene to a Tumor Suppressor." *Blood* 123, no. 16 (Apr): 2451-9. http://dx.doi.org/10.1182/blood-2013-08-355818.

Lovly, C., L. Horn, W. Pao. 2016. *Molecular Profiling of Lung Cancer*: *My Cancer Genome* https://www.mycancergenome.org/content/disease/lung-cancer/

Maemondo, M., et al. 2010. "Gefitinib or Chemotherapy for Non-Small-Cell Lung Cancer with Mutated Egfr." *N Engl J Med* 362, no. 25 (Jun): 2380-8. http://dx.doi.org/10.1056/NEJMoa0909530.

Maher, C. A., et al. 2009. "Transcriptome Sequencing to Detect Gene Fusions in Cancer." *Nature* 458, no. 7234 (Mar): 97-101. http://dx.doi.org/10.1038/nature07638.

Marguerat, S., and J. Bähler. 2010. "Rna-Seq: From Technology to Biology." *Cell Mol Life Sci* 67, no. 4 (Feb): 569-79. http://dx.doi.org/10.1007/s00018-009-0180-6.

Marioni, J. C., N. P. Thorne, and S. Tavaré. 2006. "Biohmm: A Heterogeneous Hidden Markov Model for Segmenting Array Cgh Data." *Bioinformatics* 22, no. 9 (May): 1144-6. http://dx.doi.org/10.1093/bioinformatics/btl089.

Marx, Vivien. 2014. *Cancer Genomes: Discerning Drivers from Passengers*. **11** vols.: Nature Methods.

Masters, J. R. 2000. "Human Cancer Cell Lines: Fact and Fantasy." *Nat Rev Mol Cell Biol* 1, no. 3 (Dec): 233-6. http://dx.doi.org/10.1038/35043102.

McCarroll, S. A., and D. M. Altshuler. 2007. "Copy-Number Variation and Association Studies of Human Disease." *Nat Genet* 39, no. 7 Suppl (Jul): S37-42. http://dx.doi.org/10.1038/ng2080.

McDermott, U., and J. Settleman. 2009. "Personalized Cancer Therapy with Selective Kinase Inhibitors: An Emerging Paradigm in Medical Oncology." *J Clin Oncol* 27, no. 33 (Nov): 5650-9. http://dx.doi.org/10.1200/JCO.2009.22.9054.

McWilliams, R. R., et al. 2011. "Prevalence of Cdkn2a Mutations in Pancreatic Cancer Patients: Implications for Genetic Counseling." *Eur J Hum Genet* 19, no. 4 (Apr): 472-8. http://dx.doi.org/10.1038/ejhg.2010.198.

Myers, C. L., C. Chiriac, and O. G. Troyanskaya. 2009. "Discovering Biological Networks from Diverse Functional Genomic Data." *Methods Mol Biol* 563: 157-75. http://dx.doi.org/10.1007/978-1-60761-175-2_9.

Nicolas Goossens, Shigeki Nakagawa, Xiaochen Sun and Yujin Hoshida. 2015. *Cancer Biomarker Discovery and Validation*: Transl Cancer Res. 2015 Jun; 4(3): 256–269.

Okuhashi, Y., et al. 2013. "Notch Knockdown Affects the Proliferation and Mtor Signaling of Leukemia Cells." *Anticancer Res* 33, no. 10 (Oct): 4293-8.

Olshen, A. B., et al. 2004. "Circular Binary Segmentation for the Analysis of Array-Based Dna Copy Number Data." *Biostatistics* 5, no. 4 (Oct): 557-72. http://dx.doi.org/10.1093/biostatistics/kxh008.

Raphael, B. J., et al. 2014. "Identifying Driver Mutations in Sequenced Cancer Genomes: Computational Approaches to Enable Precision Medicine." *Genome Med* 6, no. 1: 5. http://dx.doi.org/10.1186/gm524.

Redon, R., and N. P. Carter. 2009. "Comparative Genomic Hybridization: Microarray

Design and Data Interpretation." *Methods Mol Biol* 529: 37-49. http://dx.doi.org/10.1007/978-1-59745-538-1_3.

Reva, B., Y. Antipin, and C. Sander. 2011. "Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics." *Nucleic Acids Res* 39, no. 17 (Sep): e118. http://dx.doi.org/10.1093/nar/gkr407.

Riddick, G., et al. 2011. "Predicting in Vitro Drug Sensitivity Using Random Forests." *Bioinformatics* 27, no. 2 (Jan): 220-4. http://dx.doi.org/10.1093/bioinformatics/btq628.

Rios, J., and S. Puhalla. 2011. "Parp Inhibitors in Breast Cancer: Brca and Beyond." *Oncology (Williston Park)* 25, no. 11 (Oct): 1014-25.

Robert, C., et al. 2015. "Nivolumab in Previously Untreated Melanoma without Braf Mutation." *N Engl J Med* 372, no. 4 (Jan): 320-30. http://dx.doi.org/10.1056/NEJMoa1412082.

Rodrik-Outmezguine, V. S., et al. 2016. "Overcoming Mtor Resistance Mutations with a New-Generation Mtor Inhibitor." *Nature* 534, no. 7606 (Jun): 272-6. http://dx.doi.org/10.1038/nature17963.

Rosell, R., et al. 2012. "Erlotinib Versus Standard Chemotherapy as First-Line Treatment for European Patients with Advanced Egfr Mutation-Positive Non-Small-Cell Lung Cancer (Eurtac): A Multicentre, Open-Label, Randomised Phase 3 Trial." *Lancet Oncol* 13, no. 3 (Mar): 239-46. http://dx.doi.org/10.1016/S1470-2045(11)70393-X.

Salesse, S., and C. M. Verfaillie. 2002. "Bcr/Abl: From Molecular Mechanisms of Leukemia Induction to Treatment of Chronic Myelogenous Leukemia." *Oncogene* 21, no. 56 (Dec): 8547-59. http://dx.doi.org/10.1038/sj.onc.1206082.

Schwarzenbach, H., D. S. Hoon, and K. Pantel. 2011. "Cell-Free Nucleic Acids as

Biomarkers in Cancer Patients." *Nat Rev Cancer* 11, no. 6 (Jun): 426-37. http://dx.doi.org/10.1038/nrc3066.

Schwarzenbach, H., and K. Pantel. 2015. "Circulating Dna as Biomarker in Breast Cancer." *Breast Cancer Res* 17, no. 1: 136. http://dx.doi.org/10.1186/s13058-015-0645-5.

Segal, E., H. Wang, and D. Koller. 2003. "Discovering Molecular Pathways from Protein Interaction and Gene Expression Data." *Bioinformatics* 19 Suppl 1: i264-71.

Shah, N. M., et al. 2013. "Understanding the Role of Nrf2-Regulated Mirnas in Human Malignancies." *Oncotarget* 4, no. 8 (Aug): 1130-42. http://dx.doi.org/10.18632/oncotarget.1181.

Sharma, S. V., et al. 2007. "Epidermal Growth Factor Receptor Mutations in Lung Cancer." *Nat Rev Cancer* 7, no. 3 (Mar): 169-81. http://dx.doi.org/10.1038/nrc2088.

Sim, N. L., et al. 2012. "Sift Web Server: Predicting Effects of Amino Acid Substitutions on Proteins." *Nucleic Acids Res* 40, no. Web Server issue (Jul): W452-7. http://dx.doi.org/10.1093/nar/gks539.

Simm, Jaak, Ildefons Magrans de Abril, and Masashi Sugiyama. 2014. *Tree-Based Ensemble Multi-Task Learning Method for Classification and Regression*. Vol. E97-D. 6 vols.: IEICE Transactions on Information and Systems.

Slamon, D. J., et al. 1987. "Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the Her-2/Neu Oncogene." *Science* 235, no. 4785 (Jan): 177-82.

Sosman, J. A., et al. 2012. "Survival in Braf V600-Mutant Advanced Melanoma Treated with Vemurafenib." *N Engl J Med* 366, no. 8 (Feb): 707-14.

http://dx.doi.org/10.1056/NEJMoa1112302.

Sotiriou, C., et al. 2003. "Breast Cancer Classification and Prognosis Based on Gene Expression Profiles from a Population-Based Study." *Proc Natl Acad Sci U S A* 100, no. 18 (Sep): 10393-8. http://dx.doi.org/10.1073/pnas.1732912100.

Souroullas, G. P., and N. E. Sharpless. 2015. "Mtor Signaling in Melanoma: Oncogene-Induced Pseudo-Senescence?" *Cancer Cell* 27, no. 1 (Jan): 3-5. http://dx.doi.org/10.1016/j.ccell.2014.12.005.

Steinwart, Ingo, and Andreas Christmann. 2008. *Support Vector Machines*: Springer Publishing Company, Incorporated.

Stetson, L. C., et al. 2014. "Computational Identification of Multi-Omic Correlates of Anticancer Therapeutic Response." *BMC Genomics* 15 Suppl 7: S2. http://dx.doi.org/10.1186/1471-2164-15-S7-S2.

Stewart, E. L., et al. 2015. "Known and Putative Mechanisms of Resistance to Egfr Targeted Therapies in Nsclc Patients with Egfr Mutations-a Review." *Transl Lung Cancer Res* 4, no. 1 (Feb): 67-81. http://dx.doi.org/10.3978/j.issn.2218-6751.2014.11.06.

Sørlie, T., et al. 2001. "Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications." *Proc Natl Acad Sci U S A* 98, no. 19 (Sep): 10869-74. http://dx.doi.org/10.1073/pnas.191367098.

Templeton, A. J., et al. 2013. "Phase 2 Trial of Single-Agent Everolimus in Chemotherapy-Naive Patients with Castration-Resistant Prostate Cancer (Sakk 08/08)." *Eur Urol* 64, no. 1 (Jul): 150-8. http://dx.doi.org/10.1016/j.eururo.2013.03.040.

Trela, E., S. Glowacki, and J. Błasiak. 2014. "Therapy of Chronic Myeloid Leukemia: Twilight of the Imatinib Era?" *ISRN Oncol* 2014: 596483.

http://dx.doi.org/10.1155/2014/596483.

U, Banerji, et al. 2013. *Results of Two Phase I Multicenter Trials of Azd5363, an Inhibitor of Akt1, 2 and 3: Biomarker and Early Clinical Evaluation in Western and Japanese Patients with Advanced Solid Tumors*. Vol. 73:abstr LB-66.: Cancer Res.

van der Vaart, M., and P. J. Pretorius. 2007. "The Origin of Circulating Free Dna." *Clin Chem* 53, no. 12 (Dec): 2215. http://dx.doi.org/10.1373/clinchem.2007.092734.

Venter, J. C., H. O. Smith, and M. D. Adams. 2015. "The Sequence of the Human Genome." *Clin Chem* 61, no. 9 (Sep): 1207-8. http://dx.doi.org/10.1373/clinchem.2014.237016.

Vermeulen, L., and H. J. Snippert. 2014. "Stem Cell Dynamics in Homeostasis and Cancer of the Intestine." *Nat Rev Cancer* 14, no. 7 (Jul): 468-80. http://dx.doi.org/10.1038/nrc3744.

von Hansemann, D. 1890. *Ueber Asymmetrische Zelltheilung in Epithel Krebsen Und Deren Biologische Bedeutung* . Vol. 119: Virchows Arch. Path. Anat.

Wang, Z., M. Gerstein, and M. Snyder. 2009. "Rna-Seq: A Revolutionary Tool for Transcriptomics." *Nat Rev Genet* 10, no. 1 (Jan): 57-63. http://dx.doi.org/10.1038/nrg2484.

WATSON, J. D., and F. H. CRICK. 1953. "Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid." *Nature* 171, no. 4356 (Apr): 737-8.

Yao, Dengju, Jing Yang, and Xiaojuan Zhan. 2013. *A Novel Method for Disease Prediction: Hybrid*
of Random Forest and Multivariate Adaptive
Regression Splines. Vol. 8. 1 vols.: JOURNAL OF COMPUTERS.

Yasuda, H., S. Kobayashi, and D. B. Costa. 2012. "Egfr Exon 20 Insertion Mutations in Non-Small-Cell Lung Cancer: Preclinical Data and Clinical Implications." *Lancet Oncol* 13, no. 1 (Jan): e23-31. http://dx.doi.org/10.1016/S1470-2045(11)70129-2.

Zou, H, and T Hastie. 2005. *Regularization and Variable Selection Via the Elastic Net*: J. R. Statist. Soc. B (2005).