Babeş-Bolyai University Cluj-Napoca
Faculty of Biology and Geology
Doctoral School of Integrative Biology

Bioinformatics analysis of nucleic acid sequences generated by structural and functional genomics investigations in evaluating the development of malignant tumors

PhD Thesis Summary

Scientific Coordinators:
Prof. Nicolae Dragoş, PhD
Prof. Dan Dumitraşcu, MD PhD

PhD Candidate:
Roxana M. Cojocneanu (Petric)

Cluj-Napoca, 2015

Table of Contents

Key words: bioinformatics, data analysis, personalized therapy, next generation sequencing, microarray, animal models, extrapolation

1. INTRODUCTION

When science meets high technological progress, great development can take place, and as proof stand the large amounts of data that were generated by the recent technological advances in the field of molecular biology, more precisely in structural and functional genomics. Among these technologies with an impact on medical research is next generation sequencing and microarray, which have the potential to improve clinicians' approach on diagnosis and targeted therapies (1-4). The improvements facilitated molecular profile studies for various illnesses, with the potential to lead to the discovery of suitable and effective targeted therapies, as well as biomarkers for early diagnosis, stratification and prognostic.

The large amounts of information generated by these machines can range from many giga-base pairs (gbp) of raw sequencing data, to large table structured files containing hundreds, or sometimes thousands of rows as in the case of gene or miRNA expression microarray data (5-7). As in the case of any assay, the value of the results resides in the researcher's ability to interpret and integrate the information, in a biologically significant way. Therefore, in order to make high throughput data accessible, it is necessary to organize and decipher it with bioinformatics methods. In short, bioinformatics is a management information system for molecular biology and has many practical applications" (8).

During the past years, the attention of healthcare practitioners and clinical researchers has begun to shift from treating patients already affected by various diseases to what is known as "predictive, preventive, and personalized medicine" (PPPM) or precision medicine, and nowhere is this trend better illustrated than in the field of oncology (9). Trying to match each patient with the best therapeutic plan, caregivers are beginning to rely on more than the histopathologic stratification of the disease. Recent advances in biomedicine and genetic research offer clinicians more precise information such as gene and molecular expression in tumors, mutation status of genes involved in a particular disease, polymorphisms or copy number variations. A good congruence between the patient's molecular profile and their therapeutic

response does not only mean an improvement in their outcome, but it also translates into a step forward towards personalized, precision medicine.

The rationale behind the present thesis was to emphasize the way that "big data" analysis can contribute to the personalization of diagnosis and therapy, by using high throughput bio-molecular profiling technologies to analyze the molecular signature of physiological and pathological conditions. With diminishing costs and more specialized panels being developed all the time – and also with the possibility of designing personalized, custom panels both for microarray and for next generation sequencing assays, these techniques can meet a large variety of clinical needs for many pathologies. By using bioinformatics tools and methods to analyze, interpret and integrate the large amounts of data generated by high throughput technologies, it will soon be possible to develop fast and effective methods applicable especially in the field of oncology. First and foremost, a quick and precise diagnosis will insure that patients will receive the most suitable treatment scheme for their disease. New, precise biomarkers will contribute to more effective and minimally invasive screening methods, which will stratify potential patients and deliver the most targeted and rapid treatment approach. Thus, although it does not promise to be "The Holy Grail" of precision medicine, bioinformatics will most likely change the face of oncological diagnosis and treatment, leading undoubtedly to an improved life and overall survival of patients.

2. THE CURRENT STATE OF KNOWLEDGE

Cancer Statistics

Cancer is among the leading causes of death worldwide, with an average yearly death toll of around 8 million, as stated by the World Health Organization (10). According to the GLOBOCAN 2012 website, there were an estimated 14.1 million new cancer cases around the world at the time of the last global evaluation, of which 7.4 million cases were in men and 6.7 million in women. This number is expected to increase to 24 million by 2035 (11, 12). The same source presents the new number of cases that occurred during the

year 2013 in each type of cancer, as well as the mortality caused by these malignancies during the same year.

Considering the high figures, it is clear that there is an urgent need of finding more targeted and efficient drugs, and the first step must be a more thorough understanding of the *cancer genome* and of molecular profiles specific for each type of tumors.

Conventional Cancer Therapies

For the time being, when referring to oncology, conventional therapy can be described as a treatment that is accepted and used by most healthcare professionals, whether involves surgical procedures, administration of chemotherapeutic drugs, or the use of radiation.

Aside from the possible acquired resistance to therapy, another important aspect which is a cause of concern for both patients and clinicians is the management of the negative side effects of these drugs, which are mainly caused by their nonspecific activity and systemic administration.

Targeted therapies

Because of the downsides of traditional chemotherapy, new targeted approaches are being developed, consisting of drugs which are capable to attack the specific cellular and molecular mechanisms which set transformed cells apart from normal, healthy ones (13). Depending on their particularities, these novel drugs are classified into several groups, like enzyme or angiogenesis inhibitors, monoclonal antibodies and gene therapies.

Although the number of studies conducted on human subjects is still limited, these novel, targeted drugs hold great promise for finding more efficient oncological therapies, and these attempts are supported by the ever increasing amount of knowledge about cancer.

Current Knowledge on Cancer

Generally, malignant tumors are described as a group of diseases which are characterized by intense, monoclonal division of a particular cell type, which invade surrounding tissues and acquires the capacity to colonize tissues

and organs (14, 15). In recent years, cancers are regarded as polygenic, multifaceted diseases in which tumor development is generated by multiple combinations of genetic and epigenetic modifications which counteract DNA damage repair mechanisms and other molecular protection systems and lead to the expression of the malignant phenotype.

The alterations can take place at any step described by the classical dogma of molecular biology. Thus, genes can acquire mutations with the potential to impair their course of becoming translated into the correct proteins, which can have devastating effects in the case of tumor suppressors (16, 17). Analyzing data generated by next generation sequencing applications can shed light on the type of mutations that occur, on their frequency or their likely outcome (18-20). In other situations, alterations can occur at mRNA level and appear as modifications in expression levels, namely up- or down-regulations. With the help of microarray technology, transcription products can be monitored in terms of fold regulation either when studying the particularities of tumor tissue compared to its normal counterpart, or, for example, when assessing the effect of a treatment. Thus, it is possible to capture the expression of a large number of transcripts and, by applying various bioinformatics algorithms, to generate lists of statistically significant differentially expressed genes which can be further interpreted in terms of biological significance (21, 22).

High Throughput Technologies

The development of new high throughput technologies is both a cause and an effect in itself. On the one hand, the new molecular studies instruments, such as next generation sequencers and microarray platforms, have made it possible for researchers to discover and characterize molecular profiles of both tumor types and patients. On the other hand, these discoveries demanded the further refinement of these techniques, to match the growing needs demanded by the quest for personalized medicine.

NGS, or “second generation sequencing” when compared to the Sanger *chain termination* method, offers the possibility to perform massive parallel sequencing of whole genomes, transcriptome, miRnome, targeted amplicone

sequencing, or DNA-protein interactions, with costs that continue to diminish while performances improve.

While NGS is mainly focused on the identification and characterization of structural variations, microarray technology is mostly used to perform functional studies, such as gene expression or microRNA quantification. It offers the possibility to identify genome-wide molecular anomalies which are connected with tumor development and progression.

Bioinformatics Analyses of High Throughput Data

The development of the previously mentioned techniques made it possible for researchers from different fields to gather knowledge about the way that alterations at various levels influence pathophysiology, with great importance in the field of cancer. While PCR and other common molecular biology techniques give fast results which can be easily interpreted by wet lab scientists, high throughput methods generate large amounts of data which need to undergo many stages of pre-processing, analysis and integration. It is therefore necessary to use bioinformatics methods and algorithms to manipulate these data, from storage to interpretation. This multidisciplinary field combines knowledge from molecular biology, genetics, biochemistry, systems biology and computer science, and it has become mandatory for deciphering “big data” (23, 24).

The Importance of “Big Data”

High throughput data, either produced by in-house experiments or downloaded from public databases, can become a valuable tool in the hands of researchers. From mutations in tumor suppressor genes, to post-transcriptional regulation, gene expression and mRNA inhibition via microRNAs, high throughput technologies produce data that have the potential to improve the management of oncologic patients. As previously mentioned, and as it will be further emphasized, the results of these molecular analyses can reach the clinic in many ways (25, 26).

On the one hand, modern technology leads to the development of new, targeted drugs, more specifically designed for each particular type of cancer.

At the same time, being able to determine the molecular profile of the patient, clinicians will have the means to choose the treatment that best targets the particularities of their tumor, with maximum of efficiency and reduced side effects (27, 28).

Another area of interest for the translational potential of bioinformatics interpretation of high throughput data is prevention, by means of developing more sensitive screening methods. Recent studies have demonstrated the biomarker potential of various molecular structures, including microRNAs, which can become the bases for developing screening tests capable to eventually reduce the incidence, and consequently the mortality, of malignant tumors (29, 30).

3. PERSONAL CONTRIBUTION: Data Management for Various *"omics"* Applications

3.1. NGS STUDY: Next generation sequencing bioinformatics analysis for triple negative breast cancer

(Parts of this chapter have been published in Clujul Medical)(31)

Breast cancer and TNBC – characterization, need for better management

According to the latest centralized reports on cancer occurrence published by GLOBOCAN, in 2012 breast malignancies were the second most common type of cancer worldwide, with an estimated 1.67 million new cases diagnosed in women, making it the most frequent type of female cancer in both developed and developing countries. In Romania, breast cancer is the most common female malignancy, having in 2012 an incidence of 25.22%. Also, it is the main cause of death by cancer in women, with a mortality rate of 16.74%. Regarding its prevalence, EUCAN presents the following values: 12.58% for one year, 34.54% for 3 years, and a 5-year prevalence of 52.88% (11).

Triple negative breast cancer comprises tumors that do not express the estrogen receptor, progesterone receptor and human epidermal growth factor 2 (HER2), and account for approximately 15-20% of total diagnosed breast cancers (32). By not presenting the hormone receptors, triple negative breast cancer tumors are thus missing the main therapeutic targets used in hormone modulation therapy, which comes in addition to the fact that this mammary tumor subtype is more aggressive, has lower overall survival, and occurs at younger ages (33-35).

Each type of cancer displays specific somatic mutations that can influence oncogenesis, and the different subtypes of breast cancer make no exception (36). Several studies present the importance of using next generation sequencing for mutation evaluation of different genes and types of cancer, due to the fact that this technique is more sensitive and able to identify more mutations than Sanger sequencing, being also capable of generating high throughput data (37-40).

The purpose of this study was the application of bioinformatics methods in the analysis of somatic mutations in 46 genes involved in cancer in 31 triple

negative breast cancer tumors evaluated with the help of next generation sequencing, specifically the Ion Torrent PGM (Life Technologies) sequencing platform.

Materials, methods and protocols

Ethical statement

All the experimental protocols were supervised and approved by the Ethical Committee of the Oncology Institute "Prof. Dr. I Chiricuta" and "Iuliu Hatieganu" University of Medicine and Pharmacy Cluj-Napoca. All patients included in the study have read and signed the informed consent forms, complying with the national and European legal requirements.

Patients

This study includes 31 patients that were diagnosed with triple negative breast cancer, diagnosis that was established using universally accepted criteria. For all the patients there is a 5-6 year clinical follow-up.

Ion Torrent technology and workflow

Ion Torrent, the "Personal Genome Machine" (PGM™) released by Life Technologies in 2011, brings a new approach to second generation sequencing. Although the technology is still based on sequencing by synthesis, it does not require fluorescently labeled or chemiluminescent dNTPs, or an image-based detection of incorporated nucleotides. Instead of sensing the pyrophosphate released during nucleotide incorporation, the disposable chip built on the semiconductor technology detects the proton that is also discharged during this process. In other words, the Ion Torrent chip acts as a pH-meter, detecting subtle pH shifts which occur when a phosphodiester bond is formed during elongation of the sequencing strand (41).

The workflow starts with the construction of DNA libraries. Genomic or cDNA is first fragmented into 200-400 bp strands, the ends are repaired, and the library is amplified and purified; then specific adaptors are attached to the ends of the fragments, and the gaps are filled in. When performing multiplex

sequencing, such as sequencing DNA for more than one patient, specific barcodes are added to the fragments. The library is then size-selected and quality-controlled, either by using an Agilent Bioanalyzer or traditional quantitative PCR (qPCR). Next step is amplification via emulsion PCR (1). After amplification, the amplicon library undergoes template bead enrichment. PCR amplification and library enrichment is conducted on the "Ion OneTouch" system, which automates the process and reduces hands-on time a great deal (42). After this step, the DNA library is loaded on the chip and centrifuged, to ensure that each well of the chip is occupied by a template bead, and the sequencing process is initiated (43).

DNA extraction

The DNA was extracted from the formalin fixed paraffin embedded (FFPE) tissues using the PureLink Genomic DNA Mini Kit from Invitrogen, according to the producer's protocol. This kit permits the isolation and purification of genomic DNA in a fast and efficient manner, starting from a variety of biological samples, such as cells, tissues – FFPE or fresh frozen, whole blood, etc. The principle by which the kit functions is the ability of DNA to selectively bind to the silicagel membrane in the spin columns, in the presence of chaotropic salts. Another advantage presented by this kit is the fact that it uses powerful buffer solutions which are able to lyse cells and tissues simply by incubation at 55°C in the presence of the proteinase K enzyme, without the need of employing any mechanical cell membrane disruption methods.

DNA quantification

The quantity of DNA that was obtained during the extraction and purification protocol was quantified with the NanoDrop-1000 (Thermo Scientific) spectrophotometer. The advantage of using this instrument for nucleic acid quantification lays in the fact that it needs a short time to complete, and only 1 µl of undiluted DNA, so the quantity assessment can take place right after conducting the extraction and purification.

Amplicon library preparation

For each sample, the amplicon library was prepared strictly following the manufacturer's protocol, using a series of Ion AmpliSeq Kits from Applied Bioscience. These kits and primer pools have been designed with the purpose of multiplexing PCR reactions for large numbers of genomic target areas and insuring high specificity and a uniform coverage.

Amplification of genomic DNA with specific primers

The amplification of genomic DNA was performed using the "Ion Ampliseq Cancer Panel" kit from Applied Bioscience, which contains primers for 46 genes that are known to present mutations related to cancer (AKT1, BRAF, FGFR1, GNAS, IDH1, FGFR2, KRAS, NRAS, PIK3CA, MET, RET, EGFR, JAK2, MPL, PDGFRA, PTEN, TP53, FGFR3, FLT3, KIT, ERBB2, ABL1, HNF1A, HRAS, ATM, RB1, CDH1, SMAD4, STK11, ALK, SRC, SMARCB1, VHL1, CTNNB1, KDR, FBXW7, APC, CSF1R, NMP1, SMO, ERBB4, CDKN2A, NOTCH1, JAK3, PTPN11). By using only 10 ng of genomic DNA, this kit provides a 97% coverage of the targeted genes, by covering 739 know mutations which are registered in the COSMIC database (Catalogue of Somatic Mutations in Cancer).

Amplicon purification

The purification step is necessary in order to remove the leftover genomic DNA and primers and to insure the high quality of the amplicons. It is performed using the magnetic stand DynaMag-2 and the "Agencourt AMPure XP" reagent from Beckman Coulter which contains magnetic particles that bind either the genomic DNA or the amplicons, depending on the volumetric ratio between the magnetic beads and the sample.

Partial digestion of primers

In order to proceed with the sequencing of the amplicons, the primers that were used for the targeted PCR amplification need to be removed.

Amplicon labelling

After a purification step, the amplicons need to be labelled with the barcodes, which permit the high multiplexing capabilities of Ion Torrent sequencing.

Nick-translation and amplification of the barcoded library took place, according to the manufacturer's protocol, using the Platinum PCR SuperMix High Fidelity reagent Library Amplification Primer Mix

Amplicon library quantification

The quantification of the library was conducted using the "Ion Library Quantification Kit" (Life Technologies) and the ViiA7 RT-PCR platform (Life Technologies), following the manufacturer's instructions. It includes the preparation of serial dilutions for the *E coli* DH10B control library, with known initial concentration, which are used to make the standard curve for determining the concentrations of the sample amplicon libraries.

After all the amplicon libraries have been diluted to a uniform concentration, the samples were pooled in groups of 4 in an equimolar manner, for further processing and sequencing. This was possible due to the multiplexing capabilities offered by the use of barcode labelling.

ISP (Ion Sphere Particles) library amplification

The "Ion Sphere Particles" (ISPs) are small polystyrene beads to which the amplicons are attached via the AP adaptors that have previously been ligated to the libraries. With the help of this technology, the libraries are subjected to clonal amplification on the Ion OneTouch system, which is part of the Ion Torrent PGM platform.

ISP enrichment

This step of the library preparation process insures an efficient loading of the sequencing chip, and a high quality of the sequencing itself. The purpose of ISP enrichment is to eliminate the ISP beads that do not have amplicon libraries attached, and this is achieved with the help of the biotin – streptavidin complex.

Loading of the "316" chips

Before loading the ISPs to the sequencing chip, certain reagents have to be added to the libraries, such as control particles, primers and the polymerase which will add the nucleotides by creating phosphodiester bonds. Then, each

pool of 4 barcoded libraries was loaded on a "316" chip, which are used to obtain up to 100 Mb of data. The chip is placed in the Ion Torrent PGM platform, which had previously undergone a series of washing and initialization steps. The Torrent Server which is connected to the PGM machine allows the user to monitor online the sequencing process, including the loading parameters of the chip, the percentage of live, template-positive ISPs and clonal beads, the amount of primer dimers, etc.

The Torrent Suite V4.4 program preinstalled on the server connected to the sequencing platform was used to perform the first steps of data analysis, namely signal processing and base calling. All the sequences were aligned to the Human Genome Build 19 (hg19), and variants were called using the Variant Caller 4.4.0.6 plug-in for detecting somatic mutations, setting the Target Regions parameters for the CP.20131001 AmpliSeq panel.

Table 3.1.6 shows some statistical data for each of the chips used to sequence the 31 triple negative breast cancer samples, information obtained from the report logs generated by the Torrent Suite software at the end of each run.

Table 3.1.6 Statistical data for the 9 chips used for sequencing the TNBC samples

Run (chip number)	Total Number of Bases (Mbp)	Number of Q20 Bases (Mbp)	Total Number of Reads	Mean Length (bp)	Longest Read (bp)	Longest AQ20 Alignment (bp)
1	142.59	123.68	1,661,227	85	197	167
2	180.07	153.88	2,125,563	84	195	170
3	78.93	65.52	983,722	80	202	166
4	14.65	13	173,907	84	195	169
5	27	24.2	327,619	82	190	165
6	208.91	191.85	2,731,127	76	368	270
7	217.34	195.63	2,878,130	75	389	241
8	50.8	47.26	666,223	76	340	252
9	201.16	175.6	2,728,373	73	377	265

Once the variants present in the sequenced samples were identified by applying the plug-ins present in the Torrent Suite software, these nucleotide modifications needed to be characterized. For this, the *.vcf* files generated for

each of the 31 samples were downloaded and transferred to an Ubuntu 10.04 computer, and a command line program, ANNOVAR, was used to analyze the data and annotate the variants(44). This tool uses the information from various databases such as COSMIC, 1000genomes, refGene, dbSNP, or ClinVar, to annotate the mutations discovered in the sequencing experiment, also assigning a PolyPhen-2 (Polymorphism Phenotyping version2) score to the nonsynonymous mutations, thus predicting their damaging, possibly damaging or benign status.

Results and discussions

The sequencing experiment conducted on 31 FFPE tissue samples from patients diagnosed with triple negative breast cancer evaluated the mutational status of 46 oncogenes and tumor suppressor genes, from which mutations were observed in 37 genes. After filtering the variants which had low qualitative parameters, the total number of mutations found in the 37 genes was 165. Synonymous variants were also eliminated, due to the fact these types of nucleotide substitutions do not cause any amino acid modifications.

The clinical assessment of these mutations was done according to the PolyPhen-2 score, which predicts function modifications at protein level as a result of a non-synonymous single nucleotide polymorphism (nsSNP). From the 165 mutations identified in the TNBC samples, 70 were predicted as being damaging, 22 possibly damaging, and 27 were assessed as benign, according to the PolyPhen-2 score for somatic mutations. The other 46 mutations had no clinical assessment based on this algorithm (Figure 3.1.8).

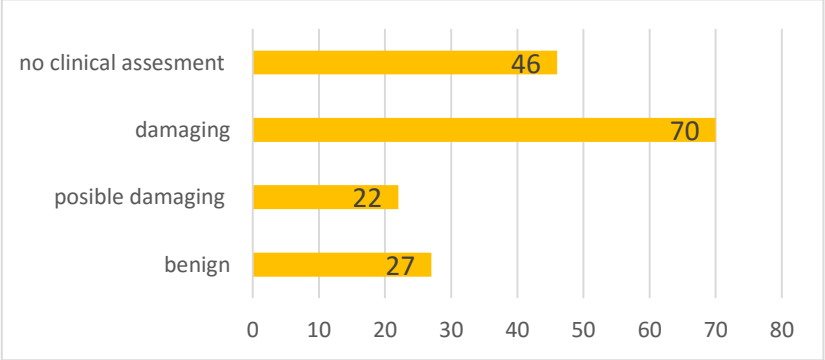

Figure 3.1.8 Clinical assessment of the identified mutations according to the PolyPhen-2 score

Of the total number of variants, 25 were known – of which 16 were already described in the COSMIC database, while the other 140 were new mutations. The 138 exonic mutations were divided in the four known categories: missense mutations – which usually are point mutations which cause an amino acid change; nonsense mutations – in which the nucleotide change leads to a premature stop codon, causing a nonfunctional mRNA; frameshift mutations – which are indels that modify the codon and shift the reading frame, generating a different translation pattern; and no frameshift deletions – which are deletions of multiples of three nucleotides, which to not modify the reading of codons, but lead to the deletions in the corresponding number of amino acids, hence generating a modified protein.

Most of the mutations that had no clinical assessment as a result of the data analysis were observed in sample 14. Of these mutations, the one with the highest frequency within the studied group was in gene *AKT1*, a G > A intronic mutation already present in the dbSNP, which was identified in 9 of the samples.

The possibly damaging mutations, which were attributed PolyPhen-2 scores ranging from 0.505 to 0.956, were observed in 13 genes, of which the most frequently mutated was *KDR,* observed in 12 samples. The other mutations were observed mostly in one sample, except for a couple – one in *ATM* and one in *SMAD4*, which were detected in two of the 31. The mutations with the highest PolyPhen-2 scores, in the 0.965 – 1.0 range, were observed in 29 of the genes, and most were found in only one sample.

Out of the total of 37 genes which were found to be mutated in this experiment, the genes which presented the most overall frequent mutations were *TP53* (mutated in 19 samples), *KDR* (in 13 samples), *PIK3CA* (in 12 samples), *AKT* and *ATM* (each mutated in 11 samples), *JAK3* (in 5 samples), *MET*, *SMAD4*, *FGFR2* and *FGFR3* (each mutated in 4 samples), *PTEN*, *ABL1* and *ERBB4* (each found as being mutated in 3 samples). These genes present an interest not only by being the most frequently mutated in our sample population, and each also having the highest numbers of individual mutations, but also because they have been proven by many studies as presenting cancer-related mutations. Involved in many significant signaling pathways and cellular processes such as cell cycle, DNA damage repair, growth and proliferation, any

modifications that occur in these genes have the potential to generate or be involved in modifications connected to tumorigenesis, progression, or response to therapy (45, 46).

The application of bioinformatics methods to discover the mutations present in the oncogenes and tumor suppressor genes from patients with breast cancer, and especially with an emphasis on the more aggressive triple negative breast cancer, gains even more importance when correlating the results with clinical information from the patients, such as the existence of possible metastases. The genes that presented the most mutations in the samples from patients who presented metastases were *TP53, KDR, AKT1*, and *PIK3CA*, which coincided also with the most mutated genes in all 31 samples.

The results generated by this present study were in accordance with the research conducted by other teams. Analyzing tissue breast cancer tissue samples or plasma from patients with mammary malignancies, and using the same sequencing platform, the Ion Torrent Personal Genome Machine, and the same library preparation kits – the AmpliSeq reagents, other teams also discovered that the highest number of somatic mutations were in genes like *PIK3CA, PTEN, AKT1, TP53, SMAD4*, offering a first degree of validation to this study (47, 48).

Validation of the identified mutations

In order to even begin to consider the possibility of translating NGS bioinformatics data analysis into the clinic for diagnostic, prognostic or biomarker purposes, it is mandatory to validate the obtained mutation with a help of a technique based on a different molecular principle than sequencing. In this situation, the method of choice was the TaqMan SNP Genotyping assays from Life Technologies. The activity of the assay is based on two probes which are complementary to the wild type allele, and the mutated one, respectively. By using different fluorochromes for each probe, the assay reveals the presence of the most frequent allele, thus validating or disproving the mutation identified during bioinformatics analysis of the data.

We chose a total number of 9 different mutations found in more than one sample, from the top 13 most frequently mutated genes. The SNP genotyping validation assays was performed for all 31 samples on which next

generation sequencing was conducted. A comparison between the mutation frequencies, expressed as percentages, obtained after performing the two methods is presented in Figure 3.1.11.

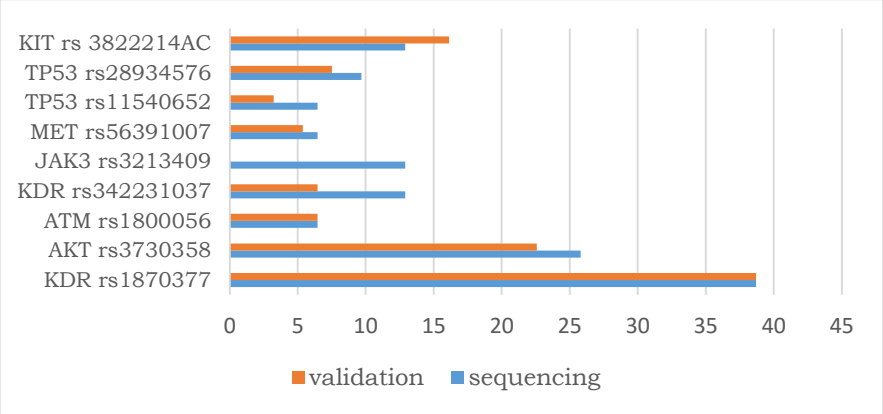

Figure 3.1.11. Frequency of mutations identified in sequencing and in validation assays

The mutations observed after NGS data analysis were validated by the SNP Genotyping Assay with the exception of *JAK3*. For this gene, the mutation that we identified with the NGS technology in several of the studied samples was not found during the validation process probably due to the lack of amplification in the PCR step of the assay. This, in turn, may be due to the insufficient quantity or / and quality of the DNA sample used, since the nucleic acid was extracted directly from sections obtained from the FFPE block, without undergoing prior microdissection.

Nonetheless, the fact that a great majority of the SNPs that were identified during sequencing were validated with the genotyping assay shows that Next Generation Sequencing data, analyzed with suitable bioinformatics tools and with proper interpretation, has the potential to become a useful instrument in the clinic, ultimately leading to a better management of the patients' wellbeing.

Limitations and further considerations

As a limitation of the present study – which is equally a motivation to continue it - is important to mention the fact that it was conducted on a rather small number of patients, the size of the cohort being constrained by the need to have, in addition to suitable biological material, also significant follow-up as

well as and recent information regarding the vital status of the subjects. Another limitation was encountered because of the relatively incomplete clinical information on the patients, which made it difficult to accomplish significant correlations between the mutations observed in the samples and the clinical data and outcome. These limitations must be overcome by further studies on larger groups of patients, with sufficient follow-up and clinical information in order to make the suitable correlations between the mutations and the phenotypes they might determine.

Conclusions

By sequencing 31 samples of FFPE triple negative breast cancer tissue using the next generation sequencing platform Ion Torrent PGM, we obtained both known and unknown mutations in 34 genes involved in various signaling pathways that take part in tumor development and progression. By using bioinformatics methods to analyze the data with the help of successive alignments to different databases, we identified a series of mutations of which some were also found in literature as being involved in breast cancer, which proves, to a certain extent, the efficiency of the method. By using a SNP Genotyping Assay based on PCR technology, we validated some of the mutations present in more than one sample, in eight of the 13 most frequently mutated genes, proving the efficiency of bioinformatics analysis of Next Generation Sequencing data. These preliminary results generated by what is one of the first such studies in Romanian population, show that bioinformatics is a powerful tool that can bridge the gap between the large amounts of data produced through next generation sequencing experiments and the interpretation and integration of meaningful information that can be eventually used in the patients' direct benefit.

3.2. MICROARRAY STUDY: Microarray bioinformatics analysis and pig-to-human extrapolation in relation to Zearalenone and *Escherichia coli* exposure

Animal Models – when and why we use them?

Animal models have helped the understanding of many diseases, such as cancer, diabetes, cardiovascular illnesses, osteoporosis, HIV, as well as different disorders of the nervous system. Many comparative studies have led not only to progress in human medicine, but also to fast accumulation of knowledge pertaining to veterinary medicine, helping us to better diagnose and care for our pets and the animals that share our habitats. At the same time, many species of animals are raised by humans to be used in various food-related or pharmaceutical biotechnological processes. This livestock must be kept in optimal health conditions, and the results of the comparative studies are also useful in this type of enterprise (49-51).

Aside from the traditional laboratory animals like mice and rats, other species have become used in various preclinical studies, such as primates, pigs, cows and other domestic animals. First, many of them share the same diseases, and the same response to corresponding medication, as humans. The similarities go deeper, to cellular and subcellular level, sharing equivalent molecular pathways and biochemical processes (52-54). But similarities exist also at a larger scale, since most of the times humans and animals share the same habitats, so they are exposed to the same ecological factors and the same living conditions.

Consequently, the use of animal models for preclinical research, especially in the case of proof-of-concept studies, is necessary, and holds promise for interdisciplinary development. And, as we will demonstrate further, the results generated by these studies can be analyzed and interpreted using bioinformatics tools, and further extrapolated to humans.

Introduction and Motivation for the Study

Zealarenone (ZEA) is a secondary metabolite produced by many species that belong to the genus Fusarium, a common group of fungal species (55, 56). These are common contaminants present is almost all types of harvests,

including those used as fodder for farm animal as well as cereal based products made for humans (57-59). Another source of contamination for both species is *Escherichia coli*, the Gram negative bacillus which displays numerous strains which populate the gastrointestinal tract of many animal species. The intestine is an important interface between the organism and the environment, so it interacts both with the existent microflora and with the pathogenic agents that may be present (60-62).

The animal model used for this study is the pig, due to the resemblance between *Sus scrofa* and *Homo sapiens* which covers many aspects, including both physiological and pathological traits. At the same time, both pigs and humans consume high quantities of maize, a cereal which is predisposed to *Fusarium* mycotoxin infection (63).

The Microarray Technology – General Overview

Microarray can be used for characterizing the gene expression profile of tumors, considering the fact that the expression of genes – especially those involved in oncogenesis, such as oncogenes and tumor suppressors – are directly responsible for the behavior of transformed cells (22, 64). By simultaneously analyzing a very high number of transcripts and consequently producing large amounts of data, this technology has the ability to capture modifications related to tumorigenesis at the whole genome level.

Material, methods and protocols

The biological samples consisted in 15 spleen and 15 duodenum tissue samples collected from newborn pigs. The animals had been previously exposed to an experimental contamination of 100 ppb with zearalenone (ZEA) and *Escherichia coli*, either as single contaminating agents or in combination.

Total RNA extraction with TRI Reagent (Sigma-Aldrich)

The tissue samples were homogenized in the presence of TRI Reagent using a Polytron mixer. Then, total RNA was extracted using the phenol-chloroform method.

RNA purification with RNeasy Mini Kit (Qiagen)

The purification of total RNA was conducted in accordance with the protocol recommended by the manufacturer of the extraction kit.

RNA quantification

Quantitative evaluation of the extracted and purified RNA was performed using the NanoDrop-1000 (Thermo Scientific) spectrophotometer, while the quality of the RNA samples was evaluated using the Bioanalyzer 2100 (Agilent Technologies) which uses an electrophoresis based method, miniaturized to the size of a chip – the technology itself being patented under the name "Lab-on-a-Chip".

The preparation of the microarray probes, the hybridization, scanning and data pre-processing were conducted following the recommended protocol provided by the producer (Agilent).

Probe synthesis

The probes used for the microarray gene expression evaluation, cRNA-Cy3, were generated using the "one color" method with the help of the Agilent Low Input Quick Amp Labeling Kit (5190-2305).

Probe purification

The kit that was used for the purification of the microarray probes was RNeasy Mini Kit from Qiagen.

Probe quantification and dilution

In order to insure suitable parameters of the cRNA probes, quality control was conducted using the NanoDrop-1000 apparatus by selecting the *Microarray measurement* tab, then the yield and dilutions were calculated using the guidelines provided by the producers in their original protocol (Agilent manual: G4140-90040).

Probe hybridization

The samples were hybridized on 8 x 60k Agilent slides containing a custom panel (AMADID 056850) made up of 60mer oligos for over 59,000

probes representative for a high number of *Sus scrofa* transcripts, using the Gene Expression Hybridization Kit from Agilent, according to the protocol.

Data processing

Scanning the microarray slides generates a *.tiff* file for each sample, which can be uploaded in the Feature Extraction software, version 11.0.1.1. This computer program reads, interprets and integrates microarray image files by automatically placing the suitable grid, flagging and excluding the pixels which are found to be outliers, and by calculating the intensity and ratio of each feature. After generating the metrics for the quality control report, the software generates the output *.txt* file with the numeric values for data analysis.

Data analysis

The bioinformatics analysis of the microarray data on the *Sus scrofa* duodenum and spleen samples was performed with the help of the GeneSpring GX software, version 13.0, produced by Agilent Technologies. The *.txt* files corresponding to each sample were imported into the program, and each sample was identified and annotated according to the type of treatment/contaminant to which it was exposed. The tactic in the present experiment was to treat the samples from each tissue type differently, and consequently we created two different experiments: one for duodenum tissue and one for spleen. The differential analysis was conducted after applying filters and statistical tests.

A fold change of 2.0 was used to obtain the entity lists for the three pairs of conditions in each experiment: ZEA vs control, *E coli* vs control, and ZEA+*E coli* vs control, for both Duodenum and Spleen, and the transcripts were considered significant when having a *p*-value smaller than 0.05

Table 3.2.5. The number of statistically significant differentially expressed transcripts in all pairs of conditions

Tissue type	Pairs of conditions	FC	p-value	Nr of statistically significant genes	Nr of statistically significant genes with FDR
Duodenum	E coli_vs_ctrl	2.0	0.05	3,058	2,875
	Zea_vs_ctrl	2.0	0.05	4,023	0
	Zea_E coli_vs_ctrl	2.0	0.05	804	316
Spleen	E coli_vs_ctrl	2.0	0.05	3,677	3,446
	Zea_vs_ctrl	2.0	0.05	467	23
	Zea_E coli_vs_ctrl	2.0	0.05	570	141

Of these transcripts whose expression has been considerably altered as a result of being exposed to the contaminants, some are upregulated, while for other the transcription is significantly reduced.

Rationale for the extrapolation pipeline

The shortcoming that we encountered was caused by the fact that the gene panel that we used for this experiment was not a commercial one but a custom one. The use of such panels for structural and functional genomic analyses has both advantages and disadvantages. The most significant weakness was the absence of thorough annotations which caused difficulties in interpreting the data. This was caused by the fact that only part of the probes had a corresponding gene name from a commonly used taxonomy system.

Additionally, the custom panel that was used contained probe identification names from different groupings, most of which were not present in usual repositories.

Extrapolation pipeline

Consequently, it was compulsory to develop a practical bioinformatics method in order to correctly recognize the human equivalent genes for the transcripts that were differentially expressed in the treated samples compared

to the controls. For this, we performed a succession of alignments and file format changes using different computer tools and programs, mostly under the Galaxy suite and in command line, in Linux.

Results and Discussions

As a result of this method, we obtained a file containing a list of the original probe IDs in one column and the corresponding human gene name in another. With the help of an Excel macro, we further annotated each of the differentially expressed original pig transcripts which were generated by GeneSpring analysis with their human equivalent gene name.

Partial conclusions for the extrapolation method

We were able to identify the equivalent human genes which are differentially expressed and statistically significant, increasing our understanding of how co-contamination might affect both species, which share many anatomical, functional, nutritional and environmental resemblances. In this day and age, when similar situations can be often encountered by researcher using different in vivo models, our new animal-to-human gene name extrapolation method gives an original answer for effective data analysis and integration.

The extrapolated effect of the co-contamination in humans

After the extrapolation process was complete, it was possible to infer the potential effects that the mycotoxin zearalenone has, either alone or when it encounters pathogenic strains of *E coli*, on important parts of the digestive system of humans. The genes were presented according to their status of being either upregulated or downregulated, and also taking into consideration the *p*-value.

Gene expression evaluation pattern in the duodenum experiment

Because of the large amount of genes whose alterations influence an even larger number of signaling pathways, as well as because of space constrains, the interpretation of the results was focused on the effect that zearalenone had on the duodenum samples. The rationale behind choosing this organ rests in the fact that the duodenum is the first segment of the small intestine, hence being the interface between the organism and potential pathogens, the first intestinal barrier that protects the integrity of the body.

Looking at the transcripts that presented significant alteration, the pattern that emerges is the implication of these transcripts in signaling pathways that involve molecular processes that are connected to cellular events such as apoptosis, cell cycle, replication and differentiation. The correlation to several of the hallmarks of cancer confirmed our hypothesis that the effects of zearalenone contribute to malignant transformation, both alone and in co-contamination with *E coli*.

Network analysis

Using the Ingenuity Pathway Analysis (IPA) software (Qiagen), we further investigated the molecular and cellular implications of the genes whose expressions were significantly altered as a response to ZEA exposure at duodenum level. Thus, we identified the most important canonical signaling pathways, as well as the top biological and pathological functions of the most dysregulated genes, by inputting into the IPA software the lists of genes generated after applying a 2.0 fold change.

Table 3.2.7. Top 5 canonical pathway modulated by ZEA at duodenum level

	Canonical pathway	p-value	Overlap
1	Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	0.0000000181	18.1 % 54/298
2	TREM1 Signaling	0.00000145	26.7 % 20/75
3	Protein Kinase A Signaling	0.00000163	15.3% 59/385
4	Breast Cancer Regulation by Stathmin1	0.00000429	18.3 % 35/191
5	Dopamine-DARPP32 Feedback in cAMP Signaling	0.00000522	19.3 % 31/161

As observed, most of the molecules with an altered expression pattern belong to signaling pathways involved in inflammation, cell growth and differentiation, apoptosis, cytoskeleton formation, etc. processes which are all connected to malignant transformation (65-67).

Conclusions

Many of the transcripts that presented altered expression levels are part of signaling pathways involved in key cellular and molecular processes, and also in the promotion of inflammation which, as shown by various studies, is directly connected to the development and progression of malignant tumors. Therefore, it can be concluded that the prolonged exposure to zearalenone, a common mycotoxin found as a contaminant in cereals, might be connected to the activation of the tumor-promoting inflammatory molecular processes (68).

Although the information in itself is important, the added value resides in the fact that, because of the novel bioinformatics pipeline that was developed for this experiment, the results can be extrapolated to humans, since the two species share many characteristics, from anatomy, physiology and pathology, to environmental factors and contaminants. Thus, the use of bioinformatics tools to decipher the high throughput information generated by modern molecular technologies, and the responsible studies conducted on animal models can ultimately contribute to the improvement of human life.

4. GENERAL CONCLUSIONS

The aim of the present thesis was to emphasize the way that "big data" analysis can contribute to the personalization of diagnosis and therapy, by using high throughput bio-molecular evaluation technologies to analyze the molecular profile of physiological and pathological conditions. By using bioinformatics tools and methods to analyze, interpret and integrate the large amounts of data generated by high throughput technologies, it will soon be possible to develop fast and effective methods applicable especially in the field of oncology.

The analysis, integration and interpretation of the results that were obtained during this research add up to the following general conclusions:

For the Next Generation Sequencing experiment, by using bioinformatics methods to analyze the data with the help of successive alignments to different databases, we identified a series of mutations of which some were also found in literature as being involved in breast cancer, which proves, to a certain extent, the efficiency of the method. Some other mutations that were found could not be correlated directly to breast cancer based on literature, but nonetheless, they have been previously described as affecting tumorigenic signaling pathways. The SNP Genotyping Assay based on PCR technology contributed to the validation of some of the identified mutations, *proving the efficiency of bioinformatics analysis of Next Generation Sequencing data*. The new mutations that were discovered still need to be further characterized, but these preliminary results, among the first of this type in Romania, show that *bioinformatics is a powerful tool that can bridge the gap between the large amounts of data produced through next generation sequencing experiments and the interpretation and integration of meaningful information that can be eventually used in the patients' direct benefit.*

Regarding the microarray study, many of the transcripts that presented altered expression levels are part of signaling pathways involved in key cellular and molecular processes connected to the development and progression of malignant tumors. Therefore, it can be stated that the prolonged exposure to zearalenone, a common mycotoxin found as a contaminant in cereals, might be connected to the activation of the tumor-promoting inflammatory molecular

processes. Although the information in itself is important, the added value resides in the fact that, because of the novel bioinformatics pipeline that was developed for this experiment, these results obtained on an animal model can be extrapolated to humans, since the two species share many characteristics, from anatomy, physiology and pathology, to environmental factors and contaminants. Thus, t*he use of bioinformatics tools to decipher the high throughput information generated by modern molecular technologies, and the responsible studies conducted on animal models can ultimately contribute to the improvement of human life*. In the "omics" era, when similar situations can be often encountered by researcher using various animal models, our new animal-to-human gene name extrapolation method offers an innovative solution for efficient data analysis and interpretation.

Publication List

ISI Journals

1. **Cojocneanu Petric R**, Braicu C, Raduly L, Zanoaga O, Dragos N, Monroig P, Dumitrascu D, Berindan-Neagoe I. Phytochemicals modulate carcinogenic signaling pathways in breast and hormone-related cancers. Onco Targets Ther. 2015 Aug 6;8:2053-66. doi: 10.2147/OTT.S83597
2. **Roxana Cojocneanu Petric**, Cornelia Braicu, Cristian Bassi, Laura Pop, Ionelia Taranu, Nicolae Dragos, Dan Dumitrascu, Massimo Negrini, Ioana Berindan-Neagoe. Interspecies Gene Name Extrapolation – A New Approach. Under second revision in PLOS ONE
3. Zaharie F*, **Cojocneanu-Petric R***, Muresan M, Frinc I, Dima D, Petrushev B, Tanase A, Berce C, Chitic M, Berindan-Neagoe I, Pileczki V, Irimie A, Tomuleasa C. Small molecules against B-RAF (BRAF) Val600Glu (V600E) single mutation. Int J Nanomedicine. 2015 Jul 31;10:4897-9. doi: 10.2147/IJN.S87405
4. Tudoran O, Soritau O, Balacescu L, Visan S, Barbos O, **Cojocneanu-Petric R**, Balacescu O, Berindan-Neagoe I. Regulation of stem cells-related signaling pathways in response to doxorubicin treatment in Hs578T triple-negative breast cancer cells. Mol Cell Biochem. 2015 Jul 18
5. Cornelia Braicu, Valentina Pileczki, Laura Pop, **Roxana Cojocneanu Petric**, Sergiu Chira, Eve Pointiere, Patriciu Achimas-Cadariu, Ioana Berindan-Neagoe, Dual targeted therapy with p53 siRNA and epigallocatechingallate in a triple negative breast cancer cell model, PLoS ONE 04/2015; 10(4). DOI:10.1371/journal.pone.0120936
6. Cornelia Braicu, **Roxana Cojocneanu-Petric**, Sergiu Chira, Anamaria Truta, Alexandru Floares, Patriciu Achimas-Cadariu, Ioana Berindan-Neagoe. Clinical and pathological implications of miRNA in bladder cancer. International Journal of Nanomedicine 2015:10 1–10
7. Muresan M, Zaharie F, Bojan A, Frinc I, Dima D, Selicean S, Gafencu GA, Petrushev B, **Cojocneanu-Petric R**, Tefas C, Cioca A, Irimie A, Berce C, Berindan-Neagoe I, Tomuleasa C, Achimas-Cadariu P. MicroRNAs in liver malignancies. Basic science applied in surgery. J BUON. 2015 Mar-Apr;20(2):361-75

8. Claudia Gherman, Matea Cristian Tudor, Bele Constantin, Tabaran Flaviu, Razvan Stefan, Bindea Maria, Sergiu Chira, Cornelia Braicu, Laura Pop, **Roxana Cojocneanu Petric**, and Ioana Berindan-Neagoe Pharmacokinetics Evaluation of Carbon Nanotubes Using FTIR Analysis and Histological Analysis. J. Nanosci. Nanotechnol. 15, 2865-2869 (2015)
9. Calin Ionescu, Cornelia Braicu, Roxana Chiorean, **Roxana Cojocneanu Petric**, Emilian Neagoe, Laura Pop, Sergiu Chira, Ioana Berindan-Neagoe. TIMP-1 expression in human colorectal cancer is associated with SMAD3 gene expression levels: a pilot study. J Gastrointestin Liver Dis 2014 Dec;23(4):413-8
10. Laura-Ancuța Pop, Emil Puscas, Valentina Pileczki, **Roxana Cojocneanu-Petric**, Cornelia Braicu, Patriciu Achimas-Cadariu, Ioana Berindan-Neagoe. Quality control of ion torrent sequencing library. Cancer Biomark 2014 ;14(2-3):93-101
11. Cornelia Braicu, Ioana Berindan-Neagoe, Valentina Pileczki, **Roxana Cojocneanu-Petric**, Laura-Ancuța Pop, Emil Puscas, Alexandru Irimie, Rares Buiga. Breast tumor bank: an important resource for developing translational cancer research in Romania. Cancer Biomark 2014 ;14(2-3):119-27
12. A Irimie, C Braicu, **R Cojocneanu Petric**, I Berindan Neagoe, RS Campian , Novel technologies for oral squamous carcinoma biomarkers in diagnostics and prognostics. Acta Odontologica Scandinavica. 2014;73(3):161-8. doi: 10.3109/00016357.2014.986754
13. Berindan-Neagoe I, Braicu C, Pileczki V, **Cojocneanu Petric R**, Miron N, Balacescu O, Iancu D, Ciuleanu T. 5-Fluorouracil potentiates the anti-cancer effect of oxaliplatin on Colo320 colorectal adenocarcinoma cells. J Gastrointestin Liver Dis. 2013 Mar;22(1):37-43

BDI Journals

1. **Roxana Cojocneanu Petric**, Laura-Ancuta Pop, Ancuta Jurj, Lajos Raduly, Dan Dumitrascu, Nicolae Dragos, Ioana Berindan Neagoe. Next Generation Sequencing Application for Breast Cancer Research. Clujul Medical, [S.l.], v. 88, n. 3, p. 278-287, jul. 2015. ISSN 2066-8872
2. C. Gherman, V. Pileczki, **R. Cojocneanu Petric**, S. Rapuntean, S. Gherman, I. Berindan Neagoe, Molecular mechanisms of action and prediction of response to oxaliplatin in colorectal cancer cells. Annals of the Romanian Society for Cell Biology 2012; 17(1):194-200.
3. Gherman C, Pileczki V, **Cojocneanu Petric R**, Braicu C, Răpuntean S, Berindan Neagoe I. In vitro studies for evaluation the antitumoral and immunomodulator effect of EGCG on Ehrlich Ascites. Archiva Zootechnica 15:2, 79-87, 2012

Selected bibliography

1. Casey G, Conti D, Haile R, Duggan D. Next generation sequencing and a new era of medicine. Gut. 2013;62(6):920-32.
2. Metzker ML. Sequencing technologies - the next generation. Nature reviews Genetics. 2010;11(1):31-46.
3. Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, McDonagh PD, et al. Experimental annotation of the human genome using microarray technology. Nature. 2001;409(6822):922-7.
4. Trevino V, Falciani F, Barrera-Saldana HA. DNA microarrays: a powerful genomic tool for biomedical and clinical research. Molecular medicine. 2007;13(9-10):527-41.
5. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome research. 2011;21(5):734-40.
6. Batley J, Edwards D. Genome sequence data: management, storage, and visualization. BioTechniques. 2009;46(5):333-4, 6.
7. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. Journal of genetics and genomics = Yi chuan xue bao. 2011;38(3):95-109.
8. Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods of information in medicine. 2001;40(4):346-58.
9. Bodrova TA, Kostyushev DS, Antonova EN, Slavin S, Gnatenko DA, Bocharova MO, et al. Introduction into PPPM as a new paradigm of public health service: an integrative view. The EPMA journal. 2012;3(1):16.
10. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. The Lancet Oncology. 2012;13(6):607-15.
11. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer Journal international du cancer. 2015;136(5):E359-86.
12. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International journal of cancer Journal international du cancer. 2010;127(12):2893-917.
13. Slavicek L, Pavlik T, Tomasek J, Bortlicek Z, Buchler T, Melichar B, et al. Efficacy and safety of bevacizumab in elderly patients with metastatic colorectal cancer: results from the Czech population-based registry. BMC gastroenterology. 2014;14:53.
14. Valastyan S, Weinberg RA. Tumor metastasis: molecular insights and evolving paradigms. Cell. 2011;147(2):275-92.

15. Guise T. Examining the metastatic niche: targeting the microenvironment. Seminars in oncology. 2010;37 Suppl 2:S2-14.
16. Studer RA, Dessailly BH, Orengo CA. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. The Biochemical journal. 2013;449(3):581-94.
17. Hochman J, Insel PA, Bourne HR, Coffino P, Tomkins GM. A structural gene mutation affecting the regulatory subunit of cyclic AMP-dependent protein kinase in mouse lymphoma cells. Proceedings of the National Academy of Sciences of the United States of America. 1975;72(12):5051-5.
18. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010;7(4):248-9.
19. Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. Genome research. 2009;19(9):1562-9.
20. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009;4(7):1073-81.
21. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. American journal of obstetrics and gynecology. 2006;195(2):373-88.
22. Slonim DK, Yanai I. Getting started in gene expression microarray analysis. PLoS computational biology. 2009;5(10):e1000543.
23. Altobelli G. Bioinformatics applied to gene transcription regulation. Journal of molecular endocrinology. 2012;49(2):R51-9.
24. Viceconti M, Hunter P, Hose R. Big Data, Big Knowledge: Big Data for Personalized Healthcare. IEEE journal of biomedical and health informatics. 2015;19(4):1209-15.
25. Idris SF, Ahmad SS, Scott MA, Vassiliou GS, Hadfield J. The role of high-throughput technologies in clinical cancer genomics. Expert review of molecular diagnostics. 2013;13(2):167-81.
26. Tran B, Dancey JE, Kamel-Reid S, McPherson JD, Bedard PL, Brown AM, et al. Cancer genomics: technology, discovery, and translation. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2012;30(6):647-60.
27. Hansen AR, Bedard PL. Clinical application of high-throughput genomic technologies for treatment selection in breast cancer. Breast cancer research : BCR. 2013;15(5):R97.
28. Hoelder S, Clarke PA, Workman P. Discovery of small molecule cancer drugs: successes, challenges and opportunities. Molecular oncology. 2012;6(2):155-76.
29. Ren A, Dong Y, Tsoi H, Yu J. Detection of miRNA as non-invasive biomarkers of colorectal cancer. International journal of molecular sciences. 2015;16(2):2810-23.

30. Sozzi G, Boeri M, Rossi M, Verri C, Suatoni P, Bravi F, et al. Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2014;32(8):768-73.
31. Cojocneanu Petric R, Pop LA, Jurj A, Raduly L, Dumitrascu D, Dragos N, et al. Next Generation Sequencing Applications for Breast Cancer Research. Clujul Medical. 2015;88(3):278-87.
32. Cleere DW. Triple negative breast cancer: a clinical update. Community Oncology. 2010;7(5):8.
33. Lawrence RT, Perez EM, Hernandez D, Miller CP, Haas KM, Irie HY, et al. The proteomic landscape of triple-negative breast cancer. Cell reports. 2015;11(4):630-44.
34. Lehmann BD, Pietenpol JA. Identification and use of biomarkers in treatment strategies for triple-negative breast cancer subtypes. The Journal of pathology. 2014;232(2):142-50.
35. Boyle P. Triple-negative breast cancer: epidemiological considerations and recommendations. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2012;23 Suppl 6:vi7-12.
36. Stratton MR. Exploring the genomes of cancer cells: progress and promise. Science (New York, NY). 2011;331(6024):1553-8.
37. Buttitta F, Felicioni L, Del Grammastro M, Filice G, Di Lorito A, Malatesta S, et al. Effective assessment of egfr mutation status in bronchoalveolar lavage and pleural fluids by next-generation sequencing. Clinical cancer research : an official journal of the American Association for Cancer Research. 2013;19(3):691-8.
38. Tripathy D, Harnden K, Blackwell K, Robson M. Next generation sequencing and tumor mutation profiling: are we ready for routine use in the oncology clinic? BMC medicine. 2014;12:140.
39. Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Frontiers in bioengineering and biotechnology. 2015;3:92.
40. Shen T, Pajaro-Van de Stadt SH, Yeat NC, Lin JC. Clinical applications of next generation sequencing in cancer: from panels, to exomes, to genomes. Frontiers in genetics. 2015;6:215.
41. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. Analytical chemistry. 2011;83(12):4327-41.
42. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology. 2012;30(5):434-9.
43. Stranneheim H, Lundeberg J. Stepping stones in DNA sequencing. Biotechnology journal. 2012;7(9):1063-73.

44. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164.
45. Balko JM, Giltnane JM, Wang K, Schwarz LJ, Young CD, Cook RS, et al. Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets. Cancer discovery. 2014;4(2):232-45.
46. Zhong X, Yang H, Zhao S, Shyr Y, Li B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. BMC genomics. 2015;16 Suppl 7:S7.
47. Bai X, Zhang E, Ye H, Nandakumar V, Wang Z, Chen L, et al. PIK3CA and TP53 gene mutations in human breast cancer tumors frequently detected by ion torrent DNA sequencing. PloS one. 2014;9(6):e99306.
48. Liu S, Wang H, Zhang L, Tang C, Jones L, Ye H, et al. Rapid detection of genetic mutations in individual breast cancer patients by next-generation DNA sequencing. Human genomics. 2015;9:2.
49. Gauthier C, Griffin G. Using animals in research, testing and teaching. Revue scientifique et technique. 2005;24(2):735-45.
50. Festing S, Wilkinson R. The ethics of animal research. Talking Point on the use of animals in scientific research. EMBO reports. 2007;8(6):526-30.
51. Eliasof S, Lazarus D, Peters CG, Case RI, Cole RO, Hwang J, et al. Correlating preclinical animal studies and human clinical trials of a multifunctional, polymeric nanoparticle. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(37):15127-32.
52. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC. Cross-species sequence comparisons: a review of methods and available resources. Genome research. 2003;13(1):1-12.
53. Yang CS, Wang X, Lu G, Picinich SC. Cancer prevention by tea: animal studies, molecular mechanisms and human relevance. Nature reviews Cancer. 2009;9(6):429-39.
54. Cui X, Vinar T, Brejova B, Shasha D, Li M. Homology search for genes. Bioinformatics. 2007;23(13):i97-103.
55. Caldwell RW, Tuite J, Stob M, Baldwin R. Zearalenone production by Fusarium species. Applied microbiology. 1970;20(1):31-4.
56. Hidy PH, Baldwin RS, Greasham RL, Keith CL, McMullen JR. Zearalenone and some derivatives: production and biological activities. Advances in applied microbiology. 1977;22:59-82.
57. Yazar S, Omurtag GZ. Fumonisins, trichothecenes and zearalenone in cereals. International journal of molecular sciences. 2008;9(11):2062-90.
58. Fazekas B, Tar A. Determination of zearalenone content in cereals and feedstuffs by immunoaffinity column coupled with liquid chromatography. Journal of AOAC International. 2001;84(5):1453-9.
59. Iqbal SZ, Rabbani T, Asi MR, Jinap S. Assessment of aflatoxins, ochratoxin A and zearalenone in breakfast cereals. Food chemistry. 2014;157:257-62.

60. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli. Nature reviews Microbiology. 2010;8(3):207-17.
61. Sekirov I, Russell SL, Antunes LC, Finlay BB. Gut microbiota in health and disease. Physiological reviews. 2010;90(3):859-904.
62. Huffnagle G, Noverr MC. GI microbiota and regulation of the immune system. Preface. Advances in experimental medicine and biology. 2008;635:v-vi.
63. Oswald IP, Desautels C, Laffitte J, Fournout S, Peres SY, Odin M, et al. Mycotoxin fumonisin B1 increases intestinal colonization by pathogenic Escherichia coli in pigs. Applied and environmental microbiology. 2003;69(10):5870-4.
64. Berretta R, Moscato P. Cancer biomarker discovery: the entropic hallmark. PloS one. 2010;5(8):e12262.
65. Rakoff-Nahoum S. Why cancer and inflammation? The Yale journal of biology and medicine. 2006;79(3-4):123-30.
66. Lu H, Ouyang W, Huang C. Inflammation, a key event in cancer development. Molecular cancer research : MCR. 2006;4(4):221-33.
67. Kiraly O, Gong G, Olipitz W, Muthupalani S, Engelward BP. Inflammation-induced cell proliferation potentiates DNA damage-induced mutations in vivo. PLoS genetics. 2015;11(2):e1004901.
68. West AC, Jenkins BJ. Inflammatory and Non-Inflammatory Roles for Toll-Like Receptors in Gastrointestinal Cancer. Current pharmaceutical design. 2015;21(21):2968-77.