Saarland University

Center for Bioinformatics

Integrative Analysis of Genomic Data

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät
der Universität des Saarlandes

von
Vu Ha Tran

Saarbrücken

2018

Tag des Kolloquiums: 1.2.2019

Dekan: Prof. Dr. Guido Kickelbick

Bericherstatter: Prof. Dr. Volkhard Helms

Prof. Dr. Alexandra K. Kiemer

Vorsitz: Prof. Dr. Katrin Philippar

Akad. Mitarbeiter: Dr.-Ing. Karl Nordström

Abstract

This thesis is composed of three different projects, and aims to predict substrates which transported by transmembrane proteins, understand the effects caused by copy number alterations (CNAs) on target proteins of antineoplastic (AN) agents, and on the genes in antineoplastic resistance pathways in cancer patients. In the first project, we propose a computational method to classify membrane transporters from three organisms (*Escherichia coli, Saccharomyces cerevisiae* and *Homo sapiens*) according to their transported substrates. Our method focuses on neighboring genes that show high co-expression with query gene. Then, we identified frequent gene ontology (GO) terms among these co-expressed neighbors and used a support vector machine classifier to annotate the substrate specificity of the query gene. The second project analyses CNAs and clinical data of 31 tumor types from The Cancer Genome Atlas (TCGA). We found that the genome sequences of tumor patients generally contain more recurrently deleted CNAs than recurrently amplified CNAs. We observed certain signs of apparently compensating effects of CNAs. The third project continues the idea of chemoresistance as suggested in the second one. This project utilized TCGA CNAs data from both normal and tumor tissues. We found that the genome sequences of tumor tissues contain more recurrently amplified CNAs of genes in cancer antineoplastic resistance pathways than normal tissues.

Zusammenfassung

Diese Arbeit besteht aus drei verschiedenen Projekten, die darauf abzielen Substrate die von Transmembranproteinen transportiert werden vorherzusagen, die Auswirkungen sog. Kopienzahlvariationen (CNAs) sowohl auf Zielproteine von Antineoplastischen Medikamenten als auch auf die zugehörigen Gene in den entsprechenden Resistenzwegen von Krebspatienten zu verstehen. Im ersten Projekt wird eine computergestützte Methode zur Klassifizierung von Transmembrantransportern dreier Organismen (*Escherichia coli*, *Saccharomyces cerevisiae* und *Homo sapiens*) anhand der von ihnen transportierten Substrate vorgestellt. Im zweiten Projekt wurden CNAs und klinische Daten von 31 Tumorarten die aus dem Cancer Genome Atlas (TCGA) stammen analysiert. Dabei stellte sich heraus, daß die genomischen Sequenzen von Tumorpatienten im allgemeinen mehr wiederkehrend deletierte CNAs aufweisen als wiederkehrend amplifizierte CNAs. Ebenfalls beobachtet wurden bestimmte Anzeichen für offensichtlich kompensatorische Effekte durch CNAs. Wie im vorgehenden Projekt wurde auch im dritten Teil der Arbeit die Idee der Chemoresistenz weiterverfolgt. Hierbei wurden CNA-Daten von normalem Gewebe, als auch von Tumorgewebe aus dem TCGA verwendet. Dabei wurde festgestellt, daß die genomischen Sequenzen von Tumorgewebe mehr wiederkehrend amplifizierte CNAs von Genen aufweisen, welche sich in Resistenzwegen von Antineoplastica befinden, als dies in normalem Gewebe der Fall ist.

Acknowledgements

I would like to thank my supervisor, Prof. Volkhard Helms, for his invaluable guidance throughout my doctoral studies. I would like to thank PD Dr. Michael Hutter, Mrs. Kerstin Gronow-Pudelek and other members of Helms's group for helping me during my study.

I would like to thank German Academic Exchange Service (Deutscher Akademischer Austauschdienst - DAAD) for financial support via a doctoral scholarship.

I would like to thank Department of Computer Science, Faculty of Information Technology, Vietnam National University of Agriculture for their support.

Finally, I would like to thank my parents and my parents-in-law, my sisters and my brothers-in-law. Without their support, I would have not done this work.

To my family

Table of Contents

List of Tables

List of Figures

Abbreviations

ABC	ATP-binding cassette
ACC	Accuracy
AN	Antineoplastic
CNA	Copy number alteration
CNV	Copy number variation
DNA	Deoxyribonucleic acid
DOR	Diagnostic odds ratio
DSB	Double-strand break
EGFR	Epidermal growth factor receptor
FDA	Food and Drug Administration
FDR	False discovery rate
FN	False negative
FNR	False negative rate
FOR	False omission rate
FP	False positive
FPR	False positive rate
GDC	Genomic Data Commons
GISTIC	Genomic Identification of Significant Targets in Cancer
GO	Gene Ontology
HGNC	HUGO Gene Nomenclature Committee
HR	Homologous recombination
KEGG	Kyoto Encyclopedia of Genes and Genome
LR-	Negative likelihood ratio
LR+	Positive likelihood ratio
mRNA	Messenger ribonucleic acid
NPV	Negative predictive value
NSCLC	Non-small-cell lung carcinoma
PPV	Positive predictive value
RBF	Radial basis function
RNA	Ribonucleic acid
SCNA	Somatic copy number alteration
SPC	Specificity
SNP	Single nucleotide polymorphism
SVM	Support vector machine
TCDB	Transporter Classification Database
TCGA	The Cancer Genome Atlas
TN	True negative
TNR	True negative rate
TP	True positive
TPR	True positive rate
tRNA	Transfer ribonucleic acid
TSG	Tumor suppressor gene

Chapter 1 Introduction

1.1 Introduction

Since the discovery of deoxyribonucleic acid (DNA) in 1869 [1], our knowledge about this genetic material has been increasing rapidly. The 1950s can be considered as the start of a digital revolution of genomic data with the appearance of the digital computer [2], and the correct molecule structure of DNA proposed by James D. Watson and Francis Crick [3]. In the 1970s, the sequencing method by Sanger and personal computer accelerated the generation of sequencing data [4], [5]. This required new methods and tools for storing and processing data. Another need for sharing of sequencing data also arose in the 1990s when the Internet became more popular [6]. From this point, more and more database and computational tools have been online available. The next generation sequencing, a rapid large-scale DNA sequencing technology with relatively low cost [7], has made the need of new powerful computational tools become more urgent in the mid of the 2000s. The increasing amount of genomic data, the decreasing cost of data generation, and the success of computational techniques such as machine learning give us an opportunity to understand better genomic diseases and to find out new effective treatments. This thesis serves to improve our understanding of the genes encode transmembrane proteins, and the genomic copy number alterations in cancer patients.

1.2 Motivation

Proteins play a vital role in biological processes (e.g. catalyze reactions, transport molecules such as oxygen) [8]. A large portion of proteins are membrane proteins. According to Krogh *et al*. [9], about 21% of the *Escherichia coli* genes encode transmembrane proteins. The corresponding numbers are 21% in *Saccharomyces cerevisiae*, 30% in *Caenorhabditis elegans* and 20% in *Arabidopsis thaliana*. In *Homo sapiens*, membrane transporters comprise the second largest protein family next to G-protein coupled receptors. However, it is experimentally hard to identify their substrate specificities [10]. To address this problem, many computational methods were developed. Previously, substrate specificities of membrane transporters have been predicted, for example, based on sequence homology [11] and amino acid composition [12]–[14]. Meta-methods that combine different features for functional annotation often gave improved performance compared to single-feature methods. For example, Yayun Hu *et al.* used four sequence features including amino acid composition, composition, transition and distribution properties, position-specific scoring matrices, and biochemical properties to annotate the substrate specificity of ATP-binding

cassette (ABC) transporters [15]. They reported an accuracy of 88% to distinguish between four classes of ABC transporters. Still, it is worthwhile to characterize the benefits of individual features before combining them with others.

Transmembrane proteins play important roles, especially in mediating the interaction between cells and their surroundings. Thus, membrane proteins are important targets for drugs (about 60% of all modern medical drugs [16]). These proteins also participate in drug resistance, e.g. MDR1 and MDR2 play important role in increasing drug efflux from cancer cell [17]. The drug targets, in general, can also resist to the drug by being mutated. Because of this relationship between drug targets and drug resistance, we would like to explore the characters of target genes of antineoplastic agents and the genes belong to antineoplastic resistance pathways. We retrieved cancer data from The Genome Cancer Atlas (TCGA), drugs targets from Drugbank, and four antineoplastic resistance pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG). TCGA includes data for more than 30 cancer types. For each cancer type, data was organized into seven categories (Raw Sequencing Data, Transcriptome Profiling, Simple Nucleotide Variation, Copy Number Variation, Deoxyribonucleic acid (DNA) Methylation, Clinical, Biospecimen). In our work, we only used copy number variation and clinical data.

1.3 Contributions

Chapter 3 and 4 of this thesis are based on manuscripts that were already published. Chapter 5 has been prepared as a manuscript for submission.

- Chapter 3: Tran, V.H., Barghash, A., Helms, V., (2018) Journal of Proteomics & Bioinformatics, V. 11, p. 868-874, doi: 10.4172/jpb.1000468. Annotating the function of protein-coding genes based on Gene Ontology terms of neighboring co-expressed genes.
- Chapter 4: Tran, V.H., Kiemer, A., Helms, V., (2018) Cancer Genomics & Proteomics, V. 15, p. 365-378,doi: 10.21873/cgp.20095 . Copy number alterations in tumor genomes deleting antineoplastic drug targets partially compensated by complementary amplifications
- Chapter 5: Tran, V.H., Helms, V., Tumor genomes frequently contain amplified resistance genes prior to treatment (manuscript under preparation)

1.4 Thesis organization

The structure of the thesis as follows:

- Chapter 2 provides a general introduction to biological background (e.g. genome, gene expression, operon, protein) and computational methods (e.g. SVM classification, some statistical tests) used in this thesis.
- Chapter 3 introduces a novel method for annotating the function of transmembrane proteins based on Gene Ontology terms and gene expression data.
- Chapter 4 analyzes the effect of CNAs of target proteins of antineoplastic agents.
- Chapter 5 compares the effects of CNAs in normal tissues and in tumor tissues on genes in four antineoplastic resistant pathways.
- Chapter 6 summarizes the results of three projects and provides conclusions with regard to the aims of the studies and contribution made.

Chapter 2 Biological background and computational methods

2.1 Genes and Genomes

2.1.1 Genome organization

Deoxyribonucleic acid (DNA), the genetic material of a cell, was first isolated by Friedrich Miescher, a Swiss physician, in 1869. He named it as “nuclein” because DNA resided inside the nuclei of eukaryotic cells [1]. More than eighty years after the existence of DNA was discovered, in 1953, James Watson and Francis Crick proposed the first correct structural model of DNA [3]. A macromolecule DNA consists of a long chain of connected nucleotides. Each nucleotide contains a nitrogen-containing nucleobase, a sugar (deoxyribose), and a phosphate group. There are four types of nucleotides discriminated by their nitrogen bases: cytosine (C), guanine (G), adenine (A) or thymine (T). Nucleobases are classified into two types: the purines (A and G), and the pyrimidines (C and T) [3]. As shown in Figure 2.1, a molecule is composed of two chains (made up of nucleotides) which coil around each other to form a double helix. The nucleotides are linked to each other to form a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next. The double-strand DNA are then formed by binding of 2 chains using hydrogen bonds between nitrogenous bases (A with T and C with G) [18].

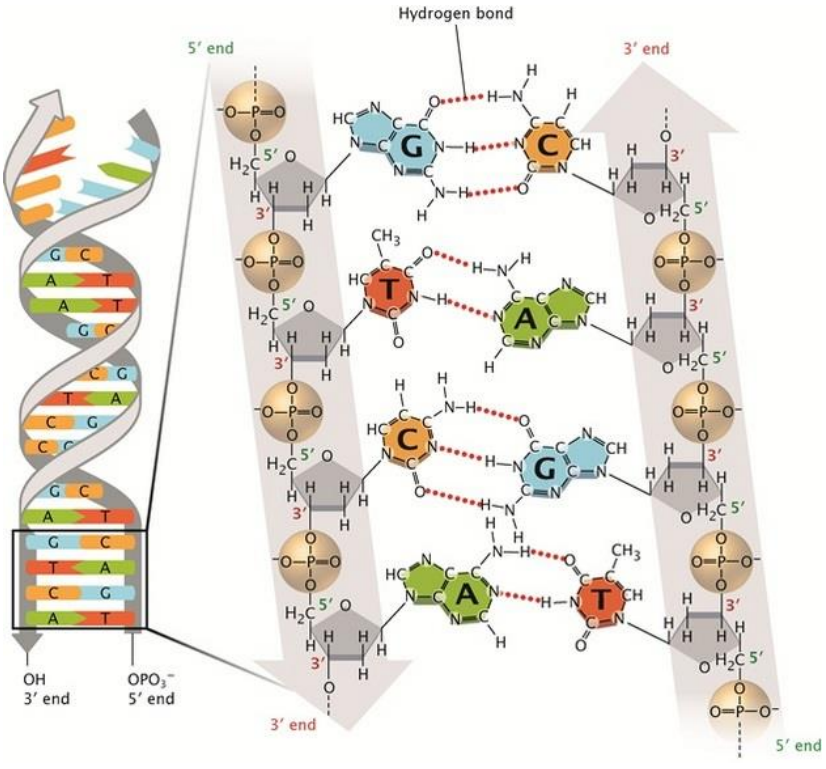

Figure 2.1 DNA structure. Image was taken from [18]

DNA does not only resides inside the nucleus. DNA is also located in mitochondria and is then named mitochondrial DNA. In human, the 16,569 base pairs of mitochondrial DNA encode for only 37 genes [19]. In prokaryotes, the species that have no nuclei, in *Escherichia coli* for example, DNA forms a single circular chromosome packaged within the cell nucleoid [20], [21].

In eukaryotes, 145-147 base pairs of double-stranded DNA may wrap around a histone octamer and form a complex called nucleosome [22], [23]. Nucleosomes are then connected via 10-80 base pairs of linker DNA. Linked nucleosomes are the primary structure of chromatin. Next, this primary structure is coiled into 30-nanometer fibers [24]. Figure 2.2 shows that the higher-order structures of chromatin are formed until finally a chromosome is created. This DNA packing process helps a human cell to store about 2 meters of DNA into its nucleus [24]. A nucleosome is a basic repeating structural unit of chromatin [25] in eukaryotes. Most prokaryotes (except species in the domain Archaea), however, do not have histone proteins. Thus, prokaryotes (e.g. *Escherichia coli*) use supercoiling as a method to compress their DNA into smaller space (see Figure 2.3) [26].

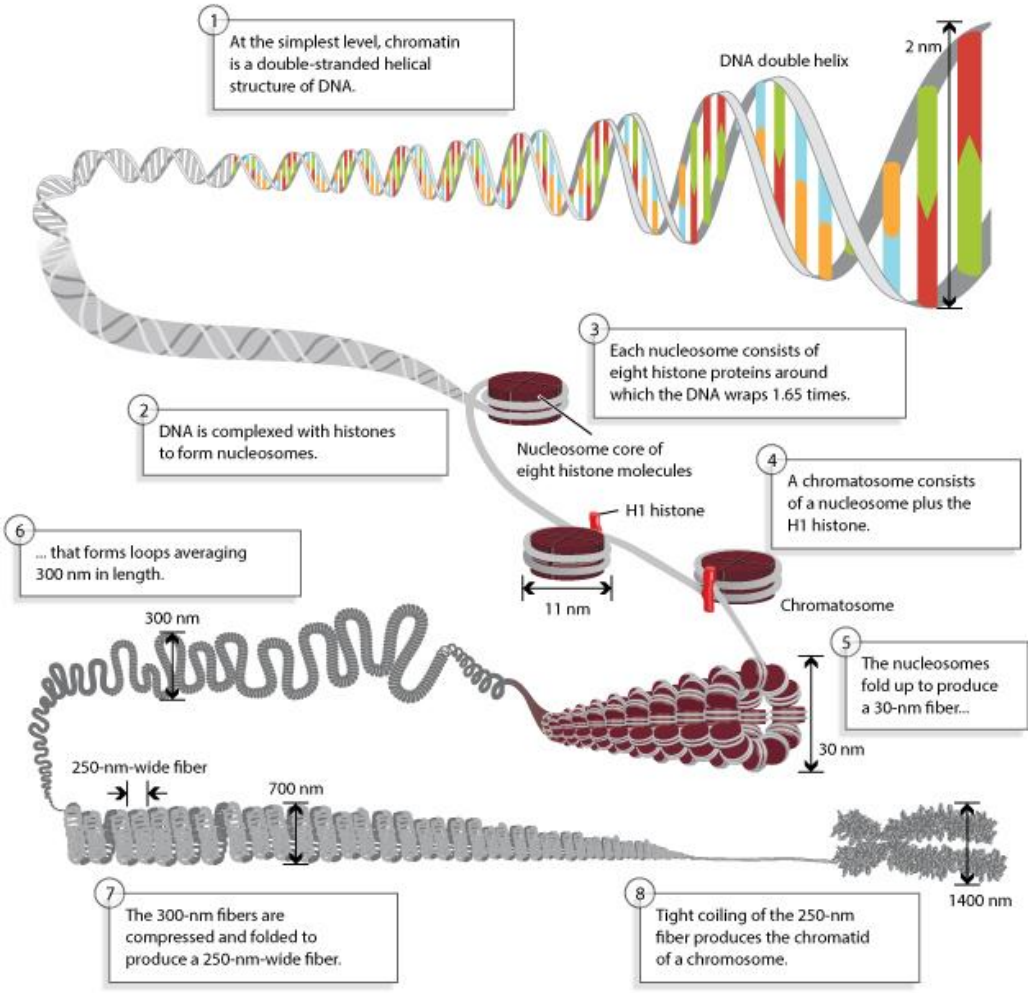

Figure 2.2 Chromosome structure. This image was taken from [24]

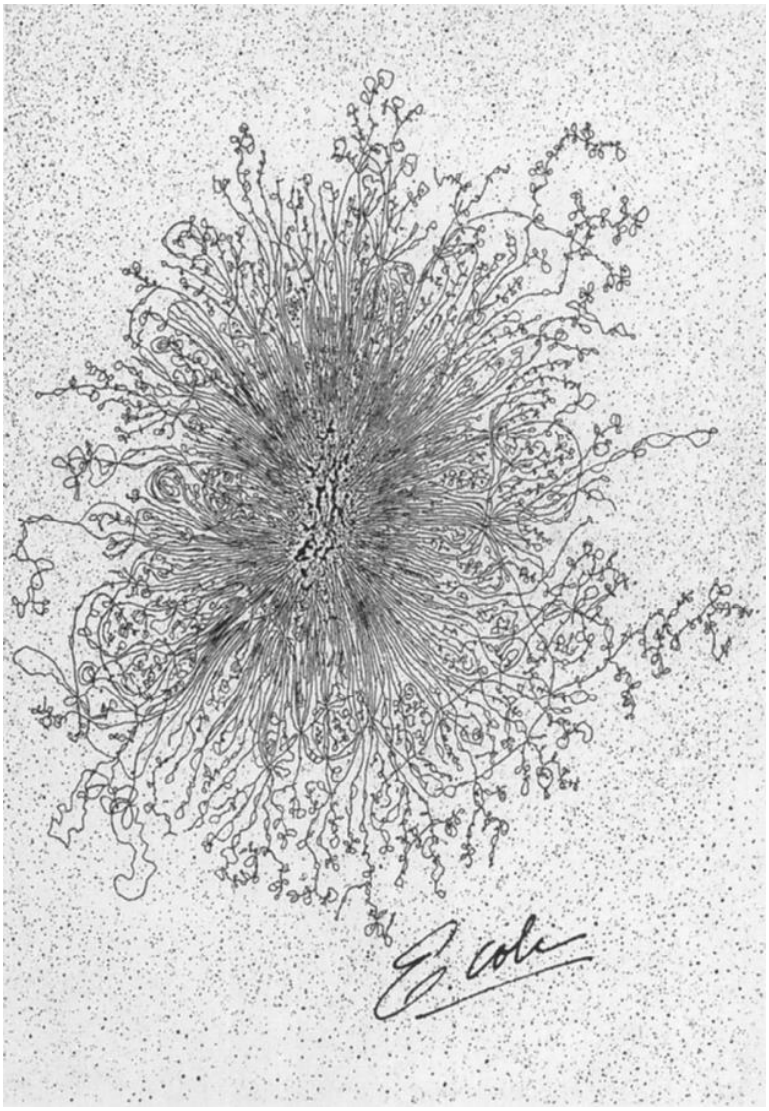

Figure 2.3 Supercoiled chromosome of Escherichia coli. This image was taken from [26]

2.1.2 Copy number variations

In eukaryotes, the genomes of species are replicated (duplicated) through mitosis. The replication can be blocked if the DNA is damaged [27]. The damages may incur due to effects of both endogenously arising compounds (e.g. reactive oxygen species) and by exogenous agents (e.g. mutagenic chemicals, radiation) [28]. One of the most cytotoxic forms of damage is double-strand breaks (DSBs). The good news is that DSBs can be repaired by different mechanisms, including homologous recombination (HR) and nonhomologous end-joining [29]. HR repairs damaged DNA sequence by using another identical sequence from homologous chromosome. However, the replacement sequence may have segment duplications, HR may lead to changes of the chromosome structure [30]. In contrast, nonhomologous recombination mechanisms use only microhomology of

a few complementary base pairs or no homology, and has the possibility of changing the structure of chromosomes [30].

Copy number changes are a type of structural variant involving alterations in the number of copies of specific large regions of DNA (thousands of nucleotides [>1 kb]), which can either be deleted or duplicated [31]. When these changes occur in germline cells, they are referred to as DNA copy number variations (CNV). When they occur in somatic cells, they are termed copy number alterations (CNA) [32]. Copy number changes may affect a large proportion of the human genome. In a study of 270 individuals, Redon *et al*. reported 1447 copy number variable regions covering 360 megabases (12% of the genome) [33]. As discussed by Hastings *et al*., copy number changes are at least as important in determining the differences between individual humans as single nucleotide polymorphisms (SNPs), and appear to be a major driving force in evolution within the human and great ape lineage [30]. Copy number changes also have severe disadvantages. They are involved in many human diseases such as the Down syndrome caused by trisomy of human chromosome 21. Copy number changes caused by submicroscopic genomic deletions were found to be involved in human diseases such as thalassaemia and red-green color blindness [34]. Changes in copy number are also involved in cancer formation and progression [35].

The CNA data that we used in this thesis is TCGA level 3 data files. The process by which these files were generated contains three main steps. First, Affymetrix SNP 6.0 platform generates TCGA level 1 files, which contain original array intensity values. These files are then processed by Birdsuite [36]. Birdsuite first normalizes array intensity values. Then it estimates raw copy number and performs tangent normalization. In the third step, DNAcopy R-package [37] analyses the result files from Birdsuite (TCGA level 2 files) using circular binary segmentation algorithm and generates copy number segment files (TCGA level 3).

2.1.3 Gene expression detected by microarrays

DNA is the basic molecular unit of heredity; it carries the raw genetic information that can be turned into functional products, usually proteins [38]. Proteins are the main actors inside cells [39], they control the functions of the cell. Humans, for example, have over 200 different types of cells [40]. Cell identity is established by transcriptional regulation so that different sets of proteins are synthesized [41]. The expression of genes contains two main steps: transcription, where double–stranded DNA is transcribed into single–stranded messenger RNA (mRNA) [42], and translation, when the mRNA molecule is translated into a protein [43]. Transcription proceeds in the following three phases [42]:

- Initiation: the enzyme RNA polymerase binds to a DNA molecule at the location of the promoter sequence (Figure 2.4-a).
- Elongation: the double-strand DNA unwinds. RNA polymerase moves along template DNA strand and adds nucleotides to the three-prime (3') end of RNA molecule (Figure 2.4-b).
- Termination: transcription is completed when RNA polymerase meets the termination sequence on the DNA template strand. At this point, the mRNA transcript and RNA polymerase are released from the complex (Figure 2.4-c).

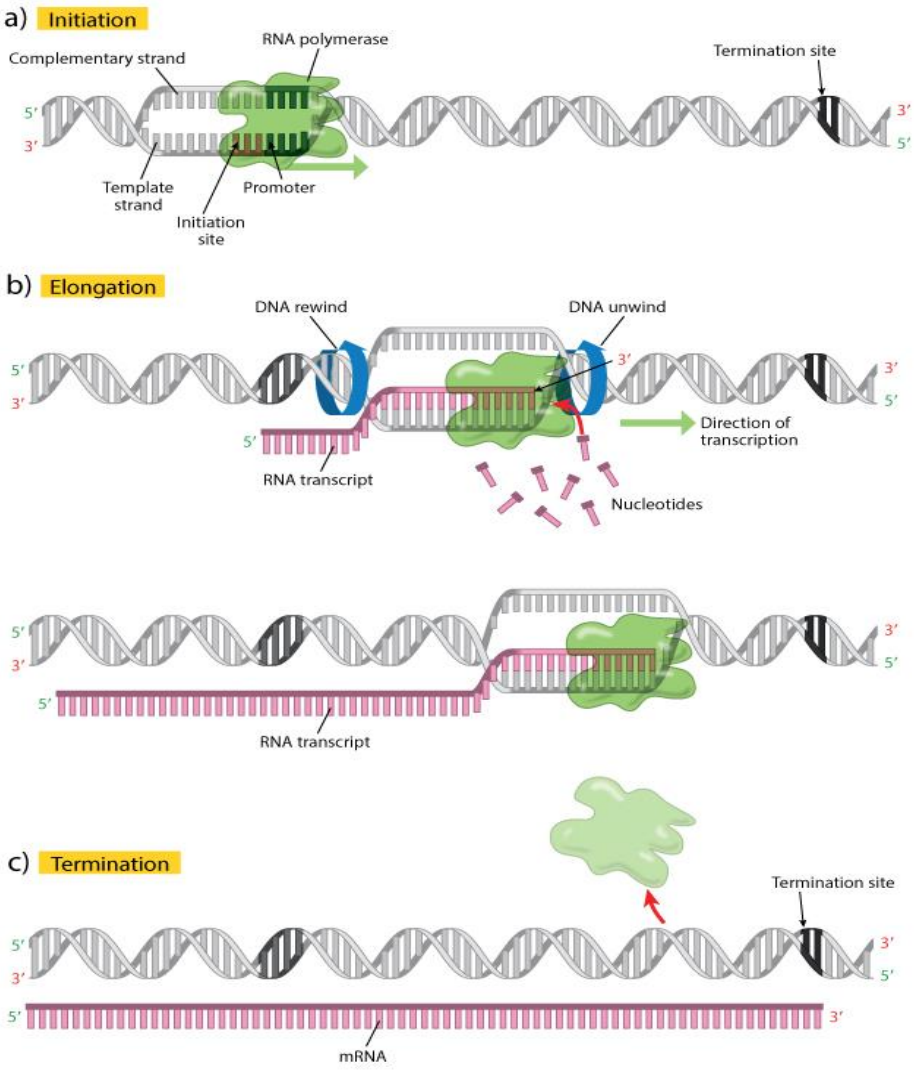

Figure 2.4 Steps in transcription process. This image was taken from [42]

During translation, the ribosome decodes the mRNA in blocks of three non-overlapping nucleotides, or codons, that each specifies an amino acid [44]. Table 2.1 lists all the possible codons and their corresponding amino acids. Translation proceeds in three phases [43]:

- Initiation: The two ribosome subunits bind to mRNA molecule. Normally, the first methionine-carrying tRNA is attached at the start codon (AUG). See Figure 2.5-A.
- Elongation: The tRNA corresponding to the next codon transfer an amino acid to the ribosome. After a peptide bond is formed between amino acids, the ribosome moves to the next mRNA codon to continue the process. See Figure 2.5-B.
- Termination: When the ribosome reaches a stop codon (UAA/UAG/UGA), it releases the polypeptide, and the translation is completed.

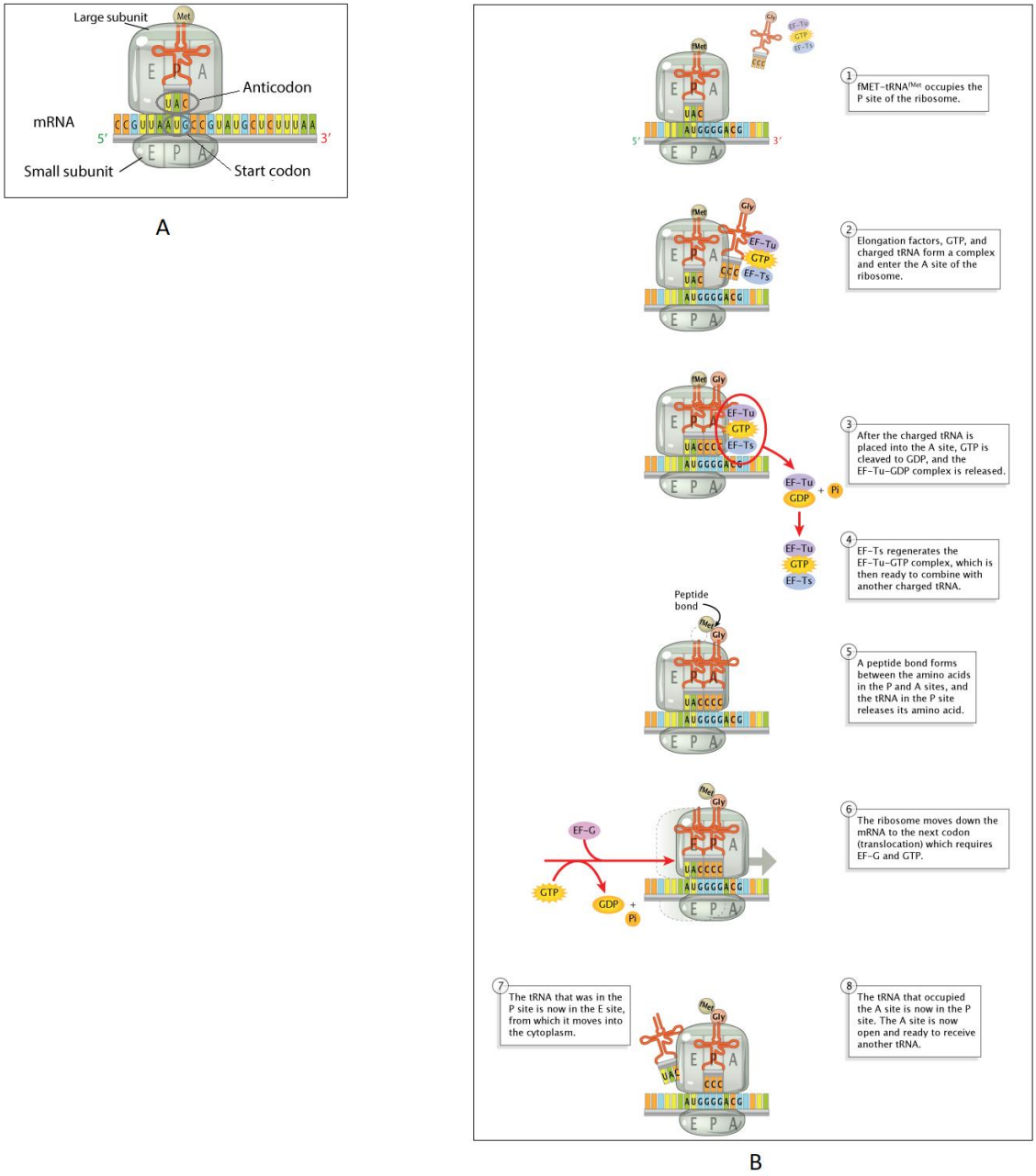

Figure 2.5 Translation from mRNA to protein. This image was taken from [43]

Table 2.1 Genetic code. The genetic code is a three–letter code that defines the translation from three sequential nucleotides into an amino acid [45]

	2nd base								
1st base	U		C		A		G		3rd base
U	UUU	Phenylalanine (Phe/F)	UCU	Serine (Ser/S)	UAU	Tyrosine (Tyr/Y)	UGU	Cysteine (Cys/C)	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine (Leu/L)	UCA		UAA	Stop (Ochre)	UGA	Stop (Opal)	A
	UUG		UCG		UAG	Stop (Amber)	UGG	Tryptophan (Trp/W)	G
C	CUU		CCU	Proline (Pro/P)	CAU	Histidine (His/H)	CGU	Arginine (Arg/R)	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine (Gln/Q)	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine (Ile/I)	ACU	Threonine (Thr/T)	AAU	Asparagine (Asn/N)	AGU	Serine (Ser/S)	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine (Lys/K)	AGA	Arginine (Arg/R)	A
	AUG	Methionine(Met/M)	ACG		AAG		AGG		G
G	GUU	Valine (Val/V)	GCU	Alanine (Ala/A)	GAU	Aspartic acid (Asp/D)	GGU	Glycine (Gly/G)	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamic acid (Glu/E)	GGA		A
	GUG		GCG		GAG		GGG		G

Proteins are the most functional macromolecules in living organisms, they play an important role in essentially all biological processes [46]. For this reason, measuring the expression of all genes in a cell is warranted. Nowadays, this is possible by various laboratory tests that identify all the genes in a cell or tissue that are making messenger RNA (https://www.cancer.gov/publications/dictionaries/cancer-terms/def/gene-expression-profile. Access date: September 10, 2018). One established technology for gene expression profiling are DNA microarrays. Microarrays consists of a large number of microscopic reaction volumes. Each volume contains a short segment of (mostly) linear DNA to which target cDNA labeled with a fluorescent tag can hybridize. The amount of fluorescence output is then of quantitative (gene expression) or qualitative (diagnostic) nature [47]. There are several microarray platforms including printed microarrays, in situ-synthesized oligonucleotide microarrays, high-density bead arrays, electronic microarrays, suspension bead arrays [47]. Besides being used to detect gene expression profiles, microarrays have been used in determining the binding sites of a transcription factor or as genotyping platforms to detect single nucleotide polymorphisms (SNP) [48]. Despite the fact that microarrays have been widely used, they still have limitations, e.g. a DNA array can only detect sequences that it was designed to detect. In case several genes have significant sequence homology, microarrays may detect all of them but they cannot distinguish these genes [48]. Recently, the microarray technique is being superseded by RNA-seq sequencing technologies [49].

2.1.4 Operon concept in prokaryotes

As first mentioned by Monod and co-workers in 1960, an operon is a group of genes for which the expression is coordinated by a single promoter [50]. That paper characterized the *lac* operon in *Escherichia coli* [50], see Figure 2.6. The first element of this operon is a promoter, a nucleotide sequence that enables a gene to be transcribed. Transcription is initiated when this sequence is bound by RNA polymerase. The second element in the operon is termed operator. This is the place where the repressor (lacI regulator protein) can bind. The binding of the repressor to the operator stops transcription and makes the expression of genes fail. The third main element of the *lac* operon is a group of genes (lacZ, lacY, lacA). Because these genes are controlled by a single promoter [50], they are either expressed together or not at all. Osbourn and Field reported that genes in the same operon are usually related in function [51].

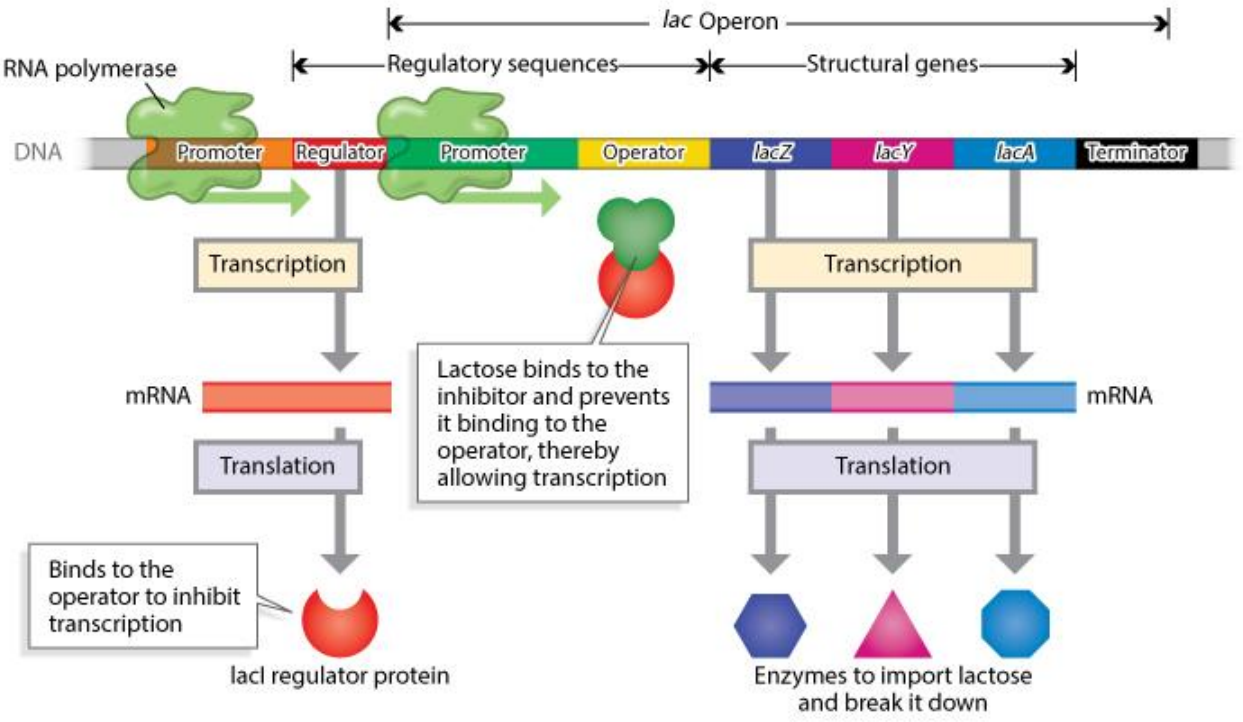

Figure 2.6 Lac operon in E.coli. This image was taken from [52]

2.1.5 Functional Annotation of Genes (Gene Ontology)

The Gene Ontology (GO) was established by the Gene Ontology Consortium [53] by joining three databases: the Saccharomyces Genome Database [54], FlyBase [55], and Mouse Genome Informatics [56], [57]. This project was motivated by the observation that there exists large-scale functional conservation of genes in eukaryotic cells. In different eukaryotic genomes, many genes code for proteins having a role in "core biological processes" that are common to all eukaryotic cells, such as transcription, translation, DNA replication, and metabolism [53]. This conservation motivates the idea of automated transfer of biological annotations from well-studied organisms to other organisms [53]. To this end, GO provides a controlled vocabulary [53], [58] for the description of:

- *Cellular components* – which refer to the place in the cell where a gene product is active.
- *Molecular functions* – which are defined as the job or the "ability" of a gene product.
- *Biological processes* – which refer to a specific objective that the gene or gene product aim to achieve.

In principle, this vocabulary can be used to all eukaryotes regardless the accumulating and the changing of our knowledge about genes and roles of proteins in cells [53]. However, there are

certainly caveats since only a small fraction of these annotations is based on real, direct biological assays, whereas most annotations are "inferred based on electronic annotation" which is termed IEA in GO terminology [59].

Figure 2.7 shows an example of a GO term (data was retrieved from http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology.1_2.obo. Access date: September 12, 2018)

```
[Term]
id: GO:0000054
name: ribosomal subunit export from nucleus
namespace: biological_process
def: "The directed movement of a ribosomal subunit from the nucleus into the cytoplasm." [GOC:ai]
subset: goslim_yeast
synonym: "ribosomal subunit export from cell nucleus" EXACT [GOC:mah]
synonym: "ribosomal subunit export out of nucleus" EXACT [GOC:mah]
synonym: "ribosomal subunit transport from nucleus to cytoplasm" EXACT [GOC:mah]
synonym: "ribosomal subunit-nucleus export" EXACT [GOC:mah]
synonym: "ribosome export from nucleus" RELATED [GOC:mah, GOC:rb]
is_a: GO:0033750 ! ribosome localization
is_a: GO:0051656 ! establishment of organelle localization
is_a: GO:0071428 ! rRNA-containing ribonucleoprotein complex export from nucleus
relationship: part_of GO:0042254 ! ribosome biogenesis
```

Figure 2.7 An example GO term

The Gene Ontology consortium provides annotations through associations between GO terms and entries for genes or gene products. Annotation data files are available at http://www.geneontology.org/page/download-go-annotations. Another way to search and browse the GO database is provided by AmiGO 2 (http://amigo.geneontology.org/amigo/landing) [60].

2.1.6 Hallmarks of cancer

In the year 2000, Hanahan *et al*. published a very influential review article on the "hallmarks of cancer" [61] where they organized the complexities of cancer biology into six major hallmarks: self-sufficiency in growth signals, insensitivity to anti-growth signals, evading apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. A decade later, an updating review [62] adjusted the six original hallmarks to sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis. The authors also added four new hallmarks: reprogramming energy metabolism, evading immune response, genome instability and mutation, and tumor-promoting inflammation. In 2014, Suzuki *et al*. assigned 2050 genes to the 10 cancer hallmarks [63] based on Gene Ontology annotations (Table 2.2).

Table 2.2 GO terms assigned to the hallmarks of cancer. This table was adapted from [63]

Hallmark	GO term id
Activating Invasion and Metastasis	GO:0045216, GO:0034329, GO:0045217, GO:0034334, GO:0016477, GO:0010718, GO:0007155
Resisting Cell Death	GO:0060548 ,GO:0012501, GO:0010941
Evading Growth Suppressors	GO:0007049, GO:0008283
Avoiding Immune Destruction	GO:0002507, GO:0001910, GO:0019882, GO:0002767
Inducing Angiogenesis	GO:0001525
Deregulating Cellular Energetics	GO:0006091
Genome Instability and Mutation	GO:0006281, GO:0051383, GO:0007062, GO:0000819, GO:0051988, GO:0030997, GO:0046605, GO:0060236, GO:0090169, GO:0043146, GO:0031577
Tumor Promoting Inflammation	GO:0006954, GO:0045321
Enabling Replicative Immortality	GO:0032202, GO:0000723, GO:0090398, GO:0090399
Sustaining Proliferative Signaling	GO:0007166, GO:0070848

2.2 Proteins

As mentioned in the previous section, proteins are products of gene expression. In the translation phase of a gene, multiple amino acids are linked together by peptide bonds to form a long chain called polypeptide [64]. A polypeptide can be folded into repeating structures called the alpha (α) helix, the beta (β) pleated sheet [65],the beta (β) turn [66], and omega (Ω) loop [67], [68]. The next level of complexity in polypeptide folding is the formation of tertiary structure. This is the complete three-dimensional structure of a protein [69]. Proteins have various functions. For example, they are the main component of antibodies like immunoglobulins [70], proteins termed enzymes can accelerate chemical reaction [71], proteins can be messengers (a hormone for example) when used to communicate between organs and tissues [72]. Proteins also provide support to protect and maintain cell shape [73]. In the following, we will introduce a specific class of proteins that work as membrane transporters.

2.2.1 Transmembrane proteins

Membrane proteins are associated with the membranes of a cell. In prokaryotes, these proteins play important roles in mediating the interaction between cell and surroundings [74]. Moreover, in eukaryotes, these proteins also catalyze transport processes into and out of intracellular compartments such as mitochondria or the endoplasmic reticulum [75], [76]. Figure 2.8 shows three ways how proteins can attach to the membrane. The Transporter Classification Database (TCDB) organizes transporter proteins into the following classes: channels/pores, electrochemical potential-driven transporters, primary active transporters, group translocators, transmembrane electron

carriers, accessory factors involved in transport, and incompletely characterized transport systems [77].

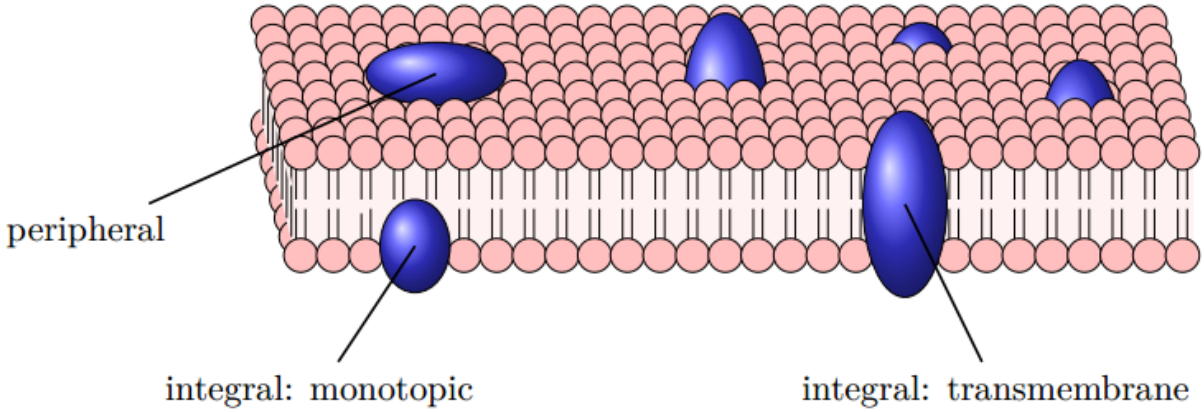

Figure 2.8 Fluid mosaic model introduced by Singer–Nicholson. Image was taken form [78]

2.2.2 Target proteins of antineoplastic drugs

Cancer is a disease in which cells divide uncontrollably. Division of cells, in turn, depends on DNA replication, transcription, and translation. This makes DNA a major target for drug development against cancer [79]. Other important targets for anticancer drug development include RNA, enzymes, and other proteins [80]. Kumar *et al*. reviewed some anticancer drug mechanisms which are listed in Table 2.3 [79].

Table 2.3 Anticancer drug mechanisms and their targets

Antineoplastic mechanism	Targets of antineoplastic agents
Angiogenesis inhibitors	Angiogenin, growth factor such as transforming growth factor-β (TGF-β), vascular endothelial growth factor (VEGF), and fibroblast growth factor (FGF).
DNA Intercalators and Groove Binding Agents	Proteins associated with recognition and function of DNA (e.g. transcription factors, polymerases, DNA repair systems, and topoisomerases).
DNA Synthesis Inhibitors	Folic acid plays an important role in de novo synthesis of purines, thymidylate, and polyamines. This in turn affects de novo synthesis of DNA in mammalian cells.
Transcription Regulators	Transcription factors.
Enzyme Inhibitors	Metabolic enzymes (e.g. pyruvate kinase M2, glucose transporters, hexokinase, fatty acid synthase, lactate dehydrogenase A, and pyruvate dehydrogenase kinase) when inhibited may induce apoptotic death in cancer cells.
Gene Regulation	Histone deacetylases are responsible for the deacetylation of histones in cells. This is important for transcriptional regulation.
Microtubule Inhibitors	Microtubules, components of the cytoskeleton, are involved in many biological processes such as cell intracellular transport, cytokinesis, signaling, maintenance of cell shape, and polarity.

2.3 Pathways

The KEGG pathway map is a network diagram of molecular interaction/reaction. This map is represented in terms of the KEGG Orthology (KO) groups. This allows the experimental evidence in specific organisms can be transferred to other organisms [81]. Each map contains graphics objects that are linked to KEGG objects. Basic graphics objects in the reference KEGG pathway maps are:

- boxes - ortholog (KO) groups identified by K numbers (KO identifiers). In metabolic maps, boxes represent reactions that are identified by R numbers.
- circles - other molecules identified by C numbers. They are usually chemical compounds.
- lines - reactions identified by R numbers in metabolic maps. In global metabolism maps, lines represent ortholog (KO) groups.

While reference KEGG pathway maps are drawn manually, organism specific pathway maps are computationally generated. In the latter ones, boxes contain genes or gene products. Each pathway map has an identifier made up by the combination of a 2-4 letter code and a 5-digit number (e.g. hsa01521). The prefix letter code can be one of the following:

- ko - Reference pathway (KO)
- map - Reference pathway
- rn - Reference pathway (Reaction)
- ec - Reference pathway (EC)
- org - Organism-specific pathway map (this prefix for *Homo sapiens* is *hsa*, for *Escherichia coli* K-12 MG1655 is *eco* …the full list of organisms is available at https://www.genome.jp/kegg/catalog/org_list.html)

A collection of pathway maps are stored in the KEGG PATHWAY database and represent knowledge on the molecular interactions, reactions and relation networks for metabolism, cellular processes, environmental information processing, genetic information processing, human diseases, organismal systems, and drug development [82].

2.3.1 Antineoplastic resistance pathways

KEGG PATHWAY contains four pathway maps showing mechanisms of resistance for four categories of anticancer drugs including epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor, platinum, antifolate, and endocrine.

EGFR tyrosine kinase inhibitor resistance (hsa01521) - Most outstanding resistant mechanisms to EGFR tyrosine kinase inhibitor treatment are “the secondary EGFR mutation (T790M), aberrance of the downstream pathways (K-RAS mutations, loss of PTEN), activation of alternative pathways (c-Met, HGF, AXL), histologic transformation, and impairment of the EGFR-TKIs-mediated apoptosis pathway (BCL2-like 11/BIM deletion polymorphism)” [83].

Endocrine resistance (hsa01522) - Cells may develop resistance to an endocrine drug by “loss of ER-alpha expression, ligand-independent growth factor signaling cascades that activate

kinases and ER-phosphorylation, altered expression of coactivators or coregulators that play a critical role in ER-mediated gene transcription, deregulation of the cell cycle and apoptotic machinery, and altered availability of active tamoxifen metabolites regulated by drug-metabolizing enzymes, such as CYP2D6" [84].

Antifolate resistance (hsa01523) - Mechanisms of antifolate resistance include "augmented drug export, virtue of impaired drug transport into cells, impaired activation of antifolates through polyglutamylation, increased expression and mutation of target enzymes, augmented hydrolysis of antifolate polyglutamates, and the augmentation of cellular tetrahydrofolate-cofactor pools in cells" [85].

Platinum drug resistance (hsa01524) - Platinum-based drugs cause cellular apoptosis by binding to purine DNA bases. Therefore, platinum can be resisted by "decreased binding of the drug to target (e.g., due to high intracellular pH), decreased mismatch repair, increased DNA repair, defective apoptosis, and altered oncogene expression". Other mechanisms are "increased drug efflux, decreased drug influx, intracellular detoxification by glutathione, etc." [86]

2.4 Machine learning

In the year 1959, the term machine learning was introduced by Arthur Samuel in an article published in the IBM Journal of Research and Development [87]. Machine learning refers to the ability of computer systems to solve problems without being explicitly programmed [88]. In the field of machine learning, researchers aim to study and construct algorithms for building a model. After learning from input data, the result model that can be used to make predictions on new coming data [89]. Broadly speaking, there are two main approaches for machine learning algorithms: supervised and unsupervised learning. The former starts with the goal of predicting a known output or target [90]. In contrast, in unsupervised learning, there are no outputs to predict. Instead, learning algorithms try to find naturally occurring patterns or groupings within the data [90]. Examples of supervised learning algorithms include linear regression, naive Bayes classifier, and support vector machines. In contrast, unsupervised learning algorithms include diverse clustering methods such as hierarchical clustering and k-means clustering.

2.4.1 Support Vector Machines

A support vector machine (SVM) [91] is a supervised learning model which is used for data classification and regression analysis. Like the other methods, we need to train our model first based on a suitable training set of "positive" and "negative" data points. SVM training constructs a hyperplane in order to separate training data belonging to these two classes (Figure 2.9).

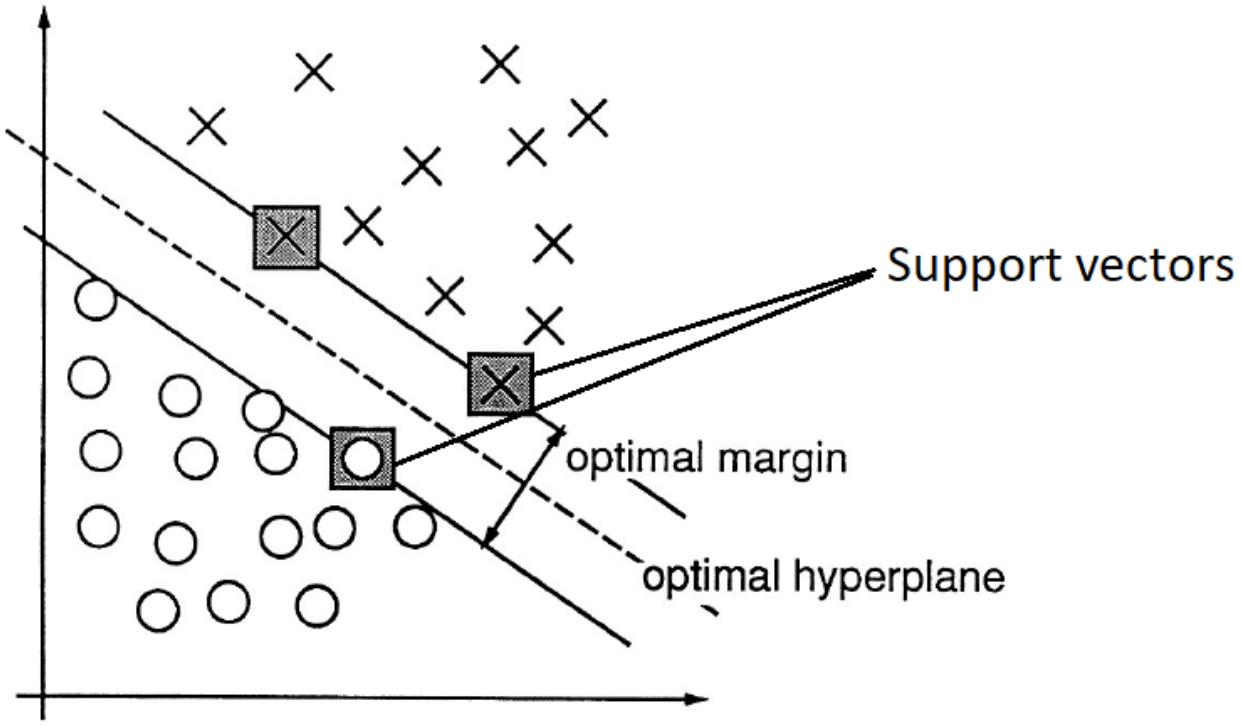

Figure 2.9 An example SVM in 2 dimensional space. Image was adapted from [91]

Let *n* points in training data be

$$(\overrightarrow{x_1}, y_1), \ldots, (\overrightarrow{x_n}, y_n)$$

where y_i indicate the class to which the point $\overrightarrow{x_i}$ belongs. Values of y_i are either -1 or 1. Each $\overrightarrow{x_i}$ is a p-dimensional vector. Our goal here is to find the "optimal hyperplane" that divides the group of points $\overrightarrow{x_i}$ for which y_i=1 from the group of points for which y_i=-1, so that the distance between the hyperplane and the nearest point $\overrightarrow{x_i}$ from either group is maximized (optimal margin).

If the training data is linearly separable, the classification function *f* is a linear function:

$$f(\vec{x}) = w^T\vec{x} + b$$

where *w* and *b* are the parameters of the classifier. The class of $\vec{x}$ is the sign of the function $f(\vec{x})$. The hyperplane can be written as the set of point $\vec{x}$ satisfying

$$f(\vec{x}) = w^T\vec{x} + b = 0$$

and the two margins as follow:

$$w^T\vec{x} + b = 1$$

$$w^T\vec{x} + b = -1$$

For every data point, we have $y_i(\omega^T\overrightarrow{x_i} + b) \geq 1$.

If the data is not linearly separable, we may allow misclassification. By adding a cost $\varepsilon_i > 0$, the optimization constraints become

$$y_i(\omega^T\overrightarrow{x_i} + b) \geq 1 - \varepsilon_i$$

If $0 < \varepsilon_i < 1$, the point $\overrightarrow{x_i}$ is correctly classified but within the margin. If $\varepsilon_i > 1$, the point is in the hyperplane or on the wrong side of it. We want to maximize the margin and minimize the cost. Another approach is using non-linear classifiers by transforming data into higher-dimensional space. This transformation is achieved using kernel functions. Examples of kernel functions include polynomial, hyperbolic tangent, and Gaussian radial basis functions.

So far, our SVM model only works with two classes (binary classifier). An approach for classifying with more than two classes is reducing the single multiclass problem into multiple binary classification problems [92]. Common methods for such reduction include: one-against-all [93], one-against-one [94], and directed acyclic graph SVM [95].

2.4.2 Model validation and evaluation

It is often useful to measure the performance of the model so that we can choose an appropriate method for a specific problem or tune the parameters of the model to improve the results. There are many metrics that can be used to measure the performance of a classifier. Performance measures are usually based on:

- Success: the class label of data point is predicted correctly
- Error: : the class label of data point is predicted incorrectly

Examples of performance metrics include:

- Error rate: proportion of incorrectly classified instances over the whole set of instances
- Accuracy: proportion of correctly classified instances over the whole set of instances

In the field of machine learning, to visualize the performance of an algorithm, people usually uses a specific table called confusion matrix (Table 2.4).

Table 2.4 Confusion matrix

	Predicted condition positive	Predicted condition negative
True condition positive	True positive (TP)	False negative (FN)
True condition negative	False positive (FP)	True negative (TN)

The following metrics can be derived from Table 2.4 :

- Accuracy (ACC) = $\frac{TP+TN}{TP+TN+FP+FN}$
- Prevalence = $\frac{TP+FN}{TP+TN+FP+FN}$
- Positive predictive value (PPV), Precision = $\frac{TP}{TP+FP}$
- False discovery rate (FDR) = $\frac{FP}{TP+FP}$
- False omission rate (FOR) = $\frac{FN}{TN+FN}$
- Negative predictive value (NPV) = $\frac{TN}{TN+FN}$
- True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{TP}{TP+FN}$
- False positive rate (FPR), Fall-out, probability of false alarm = $\frac{FP}{FP+TN}$
- Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{TN}{FP+TN}$
- False negative rate (FNR), Miss rate = $\frac{FN}{TP+FN}$
- Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$
- Negative likelihood ratio (LR−) = $\frac{FNR}{TNR}$
- Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$
- F1 score = $\frac{2TP}{2TP+FP+FN}$

In the following, three methods to estimate classifier problems will be explained. The first one is the holdout method. This method separates data into two sets, one for training (training set) and the other for testing (test set). One disadvantage of this method is that fewer labeled examples are available for training (because the test set holds some examples). Consequently, the result model may not be as good as when all the labeled examples are used for training [96]. The second method is cross-validation. In this method, data is segmented into *k* equally-sized partitions. Each iteration

uses one of the partitions for testing and the other remaining partitions for training. To use each partition for testing exactly once, this procedure is repeated *k* times. A special case of cross validation occurs when *k* is equal to the size of the data set so that each test set only contains one record. This case is called leave-one-out cross validation. The third method is bootstrap. Not like holdout or cross-validation, in which training records are sampled without replacement, in the bootstrap, a record already chosen for training is put back into the original pool of records.

2.5 Statistical hypothesis tests

"A statistical hypothesis is an assertion or conjecture concerning one or more populations" [97]. Here are some examples of statistical hypotheses:

- The mean age of cats is 10 years.
- The variable H_m, representing the height of male students, is approximately normally distributed.
- The new drug is better than penicillin.

Unless we examine the whole population, the falsity or truth of a statistical hypothesis is never known with absolute certainty. Because examining the entire population would be impossible in most real-life situations, we take a random sample from the population and use it to provide evidence that either supports or does not support the hypothesis. The hypothesis will be rejected if it is not consistent with the evidence from the selected sample. The process that leads to the decision of accepting or rejecting a statistical hypothesis is called statistical hypothesis testing.

In hypothesis testing, the term *null hypothesis* (denoted by H_0) refers to any hypothesis we want to test. We need an alternative hypothesis (H_1) in case H_0 is rejected. The alternative hypothesis is often the logical complement to null hypothesis. The three examples above now become:

- $\begin{cases} H_0: \text{The mean age of cats is 10 years} \\ H_1: \text{The mean age of cats is greater than 10 years} \end{cases}$
- $\begin{cases} H_0: \text{The variable } Hm, \text{ representing heights of male students, is approximately normally distributed} \\ H_1: \text{The variable } Hm, \text{ representing heights of male students, is not normally distributed} \end{cases}$
- $\begin{cases} H_0: \text{The new drug is the same as penicillin} \\ H_1: \text{The new drug is better than penicillin} \end{cases}$

2.5.1 Shapiro–Wilk test of normality

The Shapiro–Wilk test, published in 1965 by Samuel Sanford Shapiro and Martin Wilk [98], is a way to tell if a random sample comes from a normal distribution. The statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where

- $x_{(i)}$ is the *i*th-smallest number in the sample;
- $\bar{x} = \frac{x_1 + \cdots + x_n}{n}$ is the sample mean;
- the constants a_i are given by
    - $(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}}$
    - where $m = (m_1, \ldots, m_2)^T$
    - and $m_1, \ldots, m_2$ are the expected values of the standard normal order statistic.
    - *V* is the covariance matrix of standard normal order statistics.

The percentage point of *W* test is then computed or looked up in the table published by Shapiro and Wilk [98]. If the value is greater than the chosen alpha level, we do not have evidence to reject the null hypothesis, which means the data came from a normally distributed population. On the other hand, if this value is less than the chosen alpha level, then we have evidence that the data tested are not normally distributed, and the null hypothesis is rejected.

2.5.2 T-test

In the year 1908, William Sealy Gosset introduced the t-test in the journal Biometrika under his pen name Student [99]. The are two common types of t-tests, one-sample t-test and two-sample t-test, with one of the most important assumption that the underlying distribution which samples were taken from are normally distributed [100].

The aim of a one-sample t-test is to compare the population's mean with a specified value μ_0. The *t* statistic can be calculated as follow

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where *n* is the sample size, $\bar{x}$ is sample mean, and *s* is the standard deviation of the sample. The degree of freedom used in this test is *n-1*.

The aim of a two-sample t-test is to compare the means of two populations. If the two samples have the same variance, the *t* statistic can be calculated as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{X}_i$ is the sample mean from a sample X_1, X_2, and $s_p = \sqrt{\frac{(n_1-1)s_{X_1}^2 + (n_2-1)s_{X_2}^2}{n_1+n_2-2}}$

In two-sample t-test, the degrees of freedom for each group is $n_i - 1$, and the total number of degrees of freedom is $n_1 + n_2 - 2$.

If the two samples have unequal variances, Welch's t-test (an adaptation of Student's t-test) is applied [101]. The *t* statistic can be calculated as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$$

where $\bar{X}_i$ is the sample mean from a sample X_1, X_2, and $s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n}}$

Here s_i^2 is the unbiased estimator of the variance of sample *i*, n_i is the size of sample *i* (1 or 2). The degrees of freedom are calculated using

$$d.f. = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

After having evaluated the *t* statistic, we can compute *p*-value as explained in [100]. The null hypothesis will be rejected if the *p*-value is less than a given small alpha value.

2.5.3 Wilcoxon rank-sum test

In statistics, the Wilcoxon rank-sum test (also called Mann–Whitney U test) [102], [103] is a nonparametric test that allows two populations to be compared without making the assumption that the values are normally distributed. For this reason, the Wilcoxon rank-sum test is an alternative to the t-test. The test requires the calculation of a *U* statistic as follows

1. Merge two samples into one set, and sort this set in ascending order.
2. For each and every observation, assign a numeric rank starting with 1. We assign the same rank, which is the midpoint of unadjusted rankings, for the observations that have equal

values. E.g., the ranks of (2, 4, 4, 4, 9) are (1, 3, 3, 3, 5) (the unadjusted rank would be (1, 2, 3, 4, 5)). The sum of all the ranks is *N(N+1)/2* where *N* is the total number of observations.

3. Sum up the ranks of the observations which belong to sample 1. *U* is then given by [104]

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

where n_1 and n_2 are the number of observation in sample 1 and 2, respectively, and R_1 is the sum of the ranks in sample 1. The *U* statistic can also be calculated as

$$U' = n_2 n_1 + \frac{n_2(n_2+1)}{2} - R_2$$

where R_2 is the sum of the ranks of the observations in sample 2.

The calculated *U* or U' – whichever is larger – is compared with the two-tailed value of $U_{\alpha(2),n_1,n_2}$ of the Wilcoxon rank–sum distribution. If one of the calculated *U* values is greater than or equal to $U_{\alpha(2),n_1,n_2}$ then the null hypothesis is rejected, which means that the two populations' distributions are not the same.

2.5.4 Fisher's exact test

Fisher's exact test is a statistical test used to analyze the associations between two categorical (classification) variables [105]. The null hypothesis for the test is that there is no association between two categorical variables. Ronald Fisher said the test was motivated by Muriel Bristol, when she claimed her ability of detecting whether the milk or the tea was added first to her cup [106]. Table 2.5 shows an example result of the "lady testing tea" experiment.

Table 2.5 Result of "lady testing tea" experiment

	Actual number of cups where milk was added first	Actual number of cups where tea was added first	Row total
Predicted number of cups where milk was added first	a	b	$R_1 = a + b$
Predicted number of cups where tea was added first	c	d	$R_2 = c + d$
Column total	$C_1 = a + c$	$C_2 = b + d$	a + b + c + d (=n)

According to Fisher, the probability of obtaining any such set of values follows the hypergeometric distribution and can be computed as

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(d+b)!}{a!\,b!\,c!\,d!\,n!}$$

where $\binom{n}{k}$ is the binomial coefficient and the symbol "!" indicates the factorial operator.

Next, we find all possible matrices that have nonnegative integers consistent with the row and column sums R_i and C_j. Then we calculate the associated conditional probability for each matrix. To compute the *p*-value of the test, probabilities of the tables (matrices) that represent equal or greater deviation than the observed table are added together [107].

2.5.5 False discovery rate

When conducting statistical hypothesis tests, for example a t-test with null hypothesis that two populations have the same means, we calculate the *p*-value. If we had a *p*-value less than a chosen significant level alpha, for example *p*-value = 0.0234 and alpha = 0.05, we reject the null hypothesis and say that the means are significantly different. If the null hypothesis is actually true and we reject it, then we make a mistake. This mistake is called type I error, or a false positive. We usually like to keep this probability of a type I error small (under 5% for example).

When conducting multiple comparisons, the probability that at least one of the tests is rejected when it is actually true is computed as follow:

$$p = 1 - (1 - \alpha)^m$$

where *m* is the number of comparisons, and α is significant level. If $m = 10$ and $\alpha = 0.05$, the probability that at least one of the tests get the type I error is about 40%. If $m = 1000$ and $\alpha = 0.05$, there are (on average) 50 tests that were falsely rejected based on the null hypothesis. Table 2.6 defines the possible outcomes when testing multiple null hypotheses. The number of hypotheses *m* is known, *R* is observable variable, while *U, V, S,* and *T* are unobservable variables.

Table 2.6 Number of errors committed when testing m null hypotheses. Table is adapted from [108]

	Null hypothesis is true (H_0)	Alternative hypothesis is true (H_A)	Total
Test is declared significant (H_0 is rejected)	V	S	R
Test is declared non-significant (H_0 is not rejected)	U	T	m-R
Total	m_0	m-m_0	m

When we falsely reject null hypotheses, the proportion of errors can be computed by the random variable $Q = V/(V + S)$. Q is defined to be zero when $R = V + S = 0$. Because V and S are unobservable variable, Q is also an unobservable variable. False discovery rate (FDR) is defined to be the expectation of Q [108],

$$FDR = E(Q) = E\{V/(V+S)\} = E(V/R)$$

FDR-controlling procedures are designed in respond the need that we want to identify as many significant (reject null hypothesis) tests as possible while keeping a relatively low proportion of false positives (falsely rejecting null hypotheses). In the following, the Benjamini–Hochberg procedure [108] and the Benjamini–Hochberg–Yekutieli [109] procedure will be explained.

Benjamini–Hochberg procedure:

- Sort all *p*-values in ascending order.
- Assign ranks to the *p*-values, starting from 1.
- Calculate Benjamini-Hochberg critical value for each individual *p*-value , using the formula

  $P_{B-H} = \frac{i}{m} Q,$

  where i is the rank of individual *p*-value, m is the total number of tests, and Q is the false discovery rate (a percentage, chosen by user).

- Compare original *p*-value to the corresponding critical $P_{B\text{-}H}$; tests have original *p*-value smaller than the critical value are significant.

Benjamini–Hochberg–Yekutieli procedure controls the FDR under positive dependence assumptions [109]. This procedure is similar the one just described except that the critical value is computed as

$$P_{B-H-Y} = \frac{i}{m.c(m)} Q,$$

- If the tests are positively correlated or independent then *c(m)=1*.
- Under arbitrary dependence $c(m) = \sum_{i=1}^{m} \frac{1}{i}$
- If the tests are negative correlated, we can approximately compute *c(m)* by using the Euler–Mascheroni constant γ [110] as follow:

$$c(m) = \sum_{i=1}^{m} \frac{1}{i} \approx ln(m) + \gamma + \frac{1}{2m}$$

2.6 External tools used

2.6.1 GISTIC 2.0: Identifying genes recurrently affected by CNAs

Genomic Identification of Significant Targets in Cancer (GISTIC) is a method designed for analyzing somatic copy-number alterations (SCNA) in cancers [111]. GISTIC identifies those regions of the genome that are aberrant more often than would be expected by chance. These regions contain "driver" genes that affect the initiation or progress of tumors. In the year 2011, four years after GISTIC was introduced, Beroukhim *et al.* released GISTIC2.0 [112]. The new version can model complex cancer genomes that contain a mixture of SCNA types occurring at distinct background rates. GISTIC2.0 also provides a priori statistical confidence in interpreting copy-number analyses. Figure 2.10 shows the main steps of both versions of GISTIC.

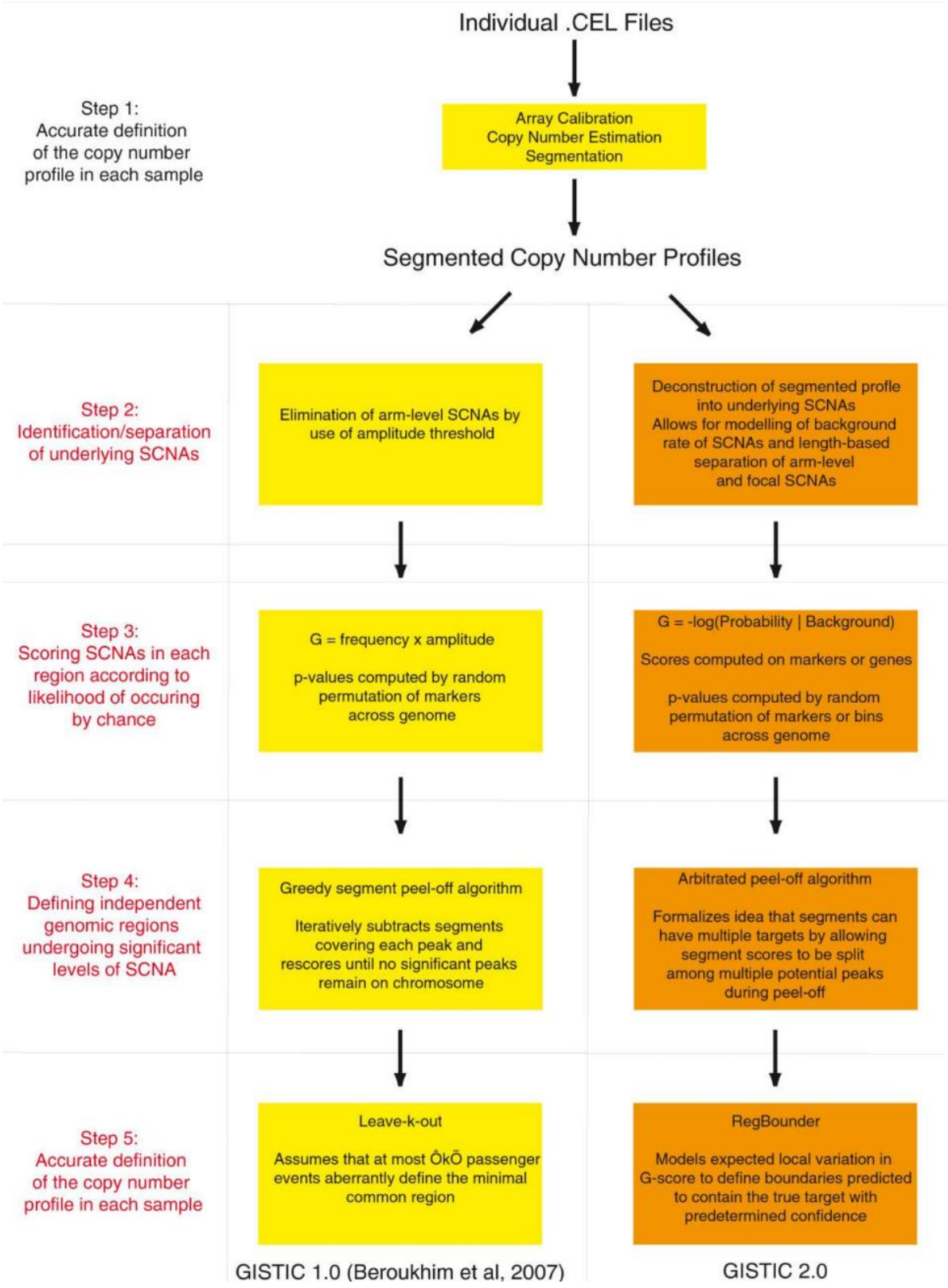

Figure 2.10 Schematic overview of GISTIC1.0 and GISTIC2.0. Image was taken from [112]

Chapter 3 Annotating the function of protein-coding genes based on Gene Ontology terms of neighboring co-expressed genes

This chapter is based on the following publication:

Tran, V.H., Barghash, A., Helms, V. Annotating the function of protein-coding genes based on Gene Ontology terms of neighboring co-expressed genes. (2018) Journal of Proteomics & Bioinformatics, V. 11, p. 868-874, doi: 10.4172/jpb.1000468.

My contribution was to design the research project and analyze the results together with the co-authors Ahmad Barghash and Volkhard Helms. I and Volkhard Helms prepared the manuscript. I collected data, implemented the machine-learning classifier and performed the calculations.

3.1 Introduction

In times of high-throughput sequencing and transcriptomics, the amount of sequencing data is quickly piling up. Yet, may proteins have still not been annotated with their cellular functions due to experimental difficulties (time-consuming and costly) involved with functional assays [113]. To address this problem, many computational methods were developed to predict the functions of proteins. The earliest methods were based on the sequence homology between proteins or on sequence motifs of proteins (e.g. PRINT-S [114], BLOCK [115], PROSITE [116], InterPro [117], transportDB [118]). As proteins exist and work as three-dimensional structures, protein structures are also a valuable indicator of similar functions between proteins [119]. Other prediction methods consider the genomic context [120]–[122] or their neighborhood in protein-protein interaction networks [123]–[125]. Recently, also some tools using natural language processing have been presented (e.g. GOstruct [126], Text-KNN [127] and PPFBM [128]).

An important yet neglected field is that of membrane proteins. According to Krogh *et al*. [9], about 21% of the *Escherichia coli* genes encode transmembrane proteins. The corresponding numbers are 21% in *Saccharomyces cerevisiae*, 30% in *Caenorhabditis elegans* and 20% in *Arabidopsis thaliana*. Transmembrane proteins play important roles, especially in mediating the interaction between cells and their surroundings. Thus, membrane proteins are important targets for drugs (about 60% of all modern medical drugs [16]). Of particular interest for the prediction of protein function is the subgroup of membrane transporters because they comprise the second largest protein family in *Homo sapiens*, next to G-protein coupled receptors. However, it is experimentally hard to identify their substrate specificities [10].

Previously, substrate specificities of membrane transporters have been predicted, for example, based on sequence homology [11] and amino acid composition [12]–[14]. Meta-methods that combine different features for functional annotation often gave improved performance compared to single-feature methods. For example, Yayun Hu *et al.* used four sequence features including amino acid composition, composition, transition and distribution properties, position-specific scoring matrices, and biochemical properties to annotate the substrate specificity of ABC transporters [15]. They reported an accuracy of 88% to distinguish between four classes of ABC transporters. Still, it is worthwhile to characterize the benefits of individual features before combining them with others.

In this study, we combined genomic context-based methods with Gene Ontology (GO annotations) [53] and gene expression data. One motivation behind considering the co-location and co-expression of neighboring genes is the principle of operons in bacterial genomes. Genes in an operon are controlled as a single unit by a single promoter [129] and thus are either expressed together or not at all. They are usually related in function too [51]. Also genes in eukaryotic genomes have been reported to have a tendency to cluster when showing similar expression, and the genes in these clusters tend to have related functions [130]–[135]. Wang and colleagues, as well as Barkai and colleagues showed that if two eukaryotic genes have the same expression levels in different conditions, they are likely to be members of the same protein complex or to participate in the same biological pathways [136], [137]. Also, Lee and Sonnhammer reported that genes involved in the same biochemical pathways tend to gather in various eukaryotic genomes [132]. These relationships between gene co-expression, neighborhood and functions have been frequently exploited in functional genomics studies, e.g. to predict protein interaction partners [138], [139], to identify and analyze gene position clusters [140] and by the STRING database [141]. A quasi-standard for functional annotation is the controlled vocabulary compiled by the Gene Ontology Consortium [53]. The Gene Ontology (GO) annotations can be used in functional profiling, functional categorizing and to predict gene function [142]. Here we combined these techniques and tested how well this method works in prokaryotes and eukaryotes.

To predict the functions of a protein, we first retrieve the neighboring genes of the respective protein-coding gene and then compute the co-expression correlation between this central gene and its neighbors. The GO term lists of the central gene and of the neighboring genes that exhibit the highest correlation to the central gene are used to create input data for a support vector machine (SVM) classifier. SVM models are then used for classifying the function of so far uncharacterized genes.

3.2 Material and methods

3.2.1 Dataset

For training and testing of the classifiers, we selected the well-studied model organisms *Escherichia coli* and *Saccharomyces cerevisiae* for which high confidence datasets are available. Later we used a *Homo sapiens* dataset to test the method. For each organism, transporter proteins and metabolic enzymes were selected. These proteins are called central proteins (and the genes encoding these are called central genes thereafter) to distinguish them from their neighboring genes.

3.2.1.1 Transporter proteins

From the Transporter Classification Database (TCDB) [143] we retrieved two sets of membrane transporters that facilitate the transport of either amino acids or sugar molecules across the membrane. Table 3.1 lists the number of proteins for the three organisms.

Table 3.1 Number of transporters belonging to different groups and organisms according to TCDB

		Organism		
		Escherichia coli	*Saccharomyces cerevisiae*	*Homo sapiens*
Group	Amino Acid Transporters	47	24	37
	Sugar Transporters	39	17	13

3.2.1.2 Enzymes in metabolic pathways

Beside transporter proteins, we also used enzymes of metabolic pathways in *Escherichia coli* to test our method. Four groups of metabolic pathways involved in carbohydrate, lipid, amino acid, and nucleotide metabolisms were collected. The lists of enzymes for each group were downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps, under the tag "metabolism" and the four respective sub-tags, e.g. carbohydrate metabolism, lipid metabolism, nucleotide metabolism, and amino acid metabolism [144]. The gene identifiers of the four groups are listed in Supplement table 1. The groups contain 187 genes (amino acid metabolism), 253 (carbohydrate metabolism), 45 (lipid metabolism), and 99 genes (nucleotide metabolism), respectively.

3.2.1.3 Data used for functional annotation

Neighboring genes: From the BioCyc database, we downloaded information about all genes of *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens* [145], [146]. We then rearranged the list of genes according to increasing genomic positions. Sorting these files helps in finding neighboring genes more easily. We use the term *neighboring genes* for genes on the same chromosome that have close genomic positions.

GO terms: We retrieved tab-delimited files with gene symbols and GO terms from the Gene Ontology Consortium [53].

Microarrays data: We used Pearson correlation to measure the co-expression of genes. For *Escherichia coli* we used preprocessed and normalized microarray expression data from Dataset Record GSE1121 [147] whereas for *Saccharomyces cerevisiae* we used respective data from Dataset Record GDS91 [148]. For *Homo sapiens*, we used data for colon adenocarcinoma patients from TCGA, but only selected data files from normal samples. After finding neighboring genes, the co-expression correlation between a gene and its neighbors was computed as:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

x_i is expression value of gene x in *i*th sample

y_i is expression value of gene y in *i*th sample

n is the number of samples

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; and analogously for $\bar{y}$

3.2.2 Methods

Figure 3.1 shows the basic steps in this project. To retrieve the neighboring genes, we look for them both upstream and downstream of the current gene. The number of neighbors upstream, is denoted as *n*, is equal to the number of neighbors downstream. The number of selected neighbors that possess highest co-expression correlation with current gene is denotes as *N*. A pair of number of neighbors and number of selected neighbors are written as (*n*, *N*) which we refer to as window size. In the result section, we show the results for three different window sizes (5, 3), (10, 3), and (20, 5).

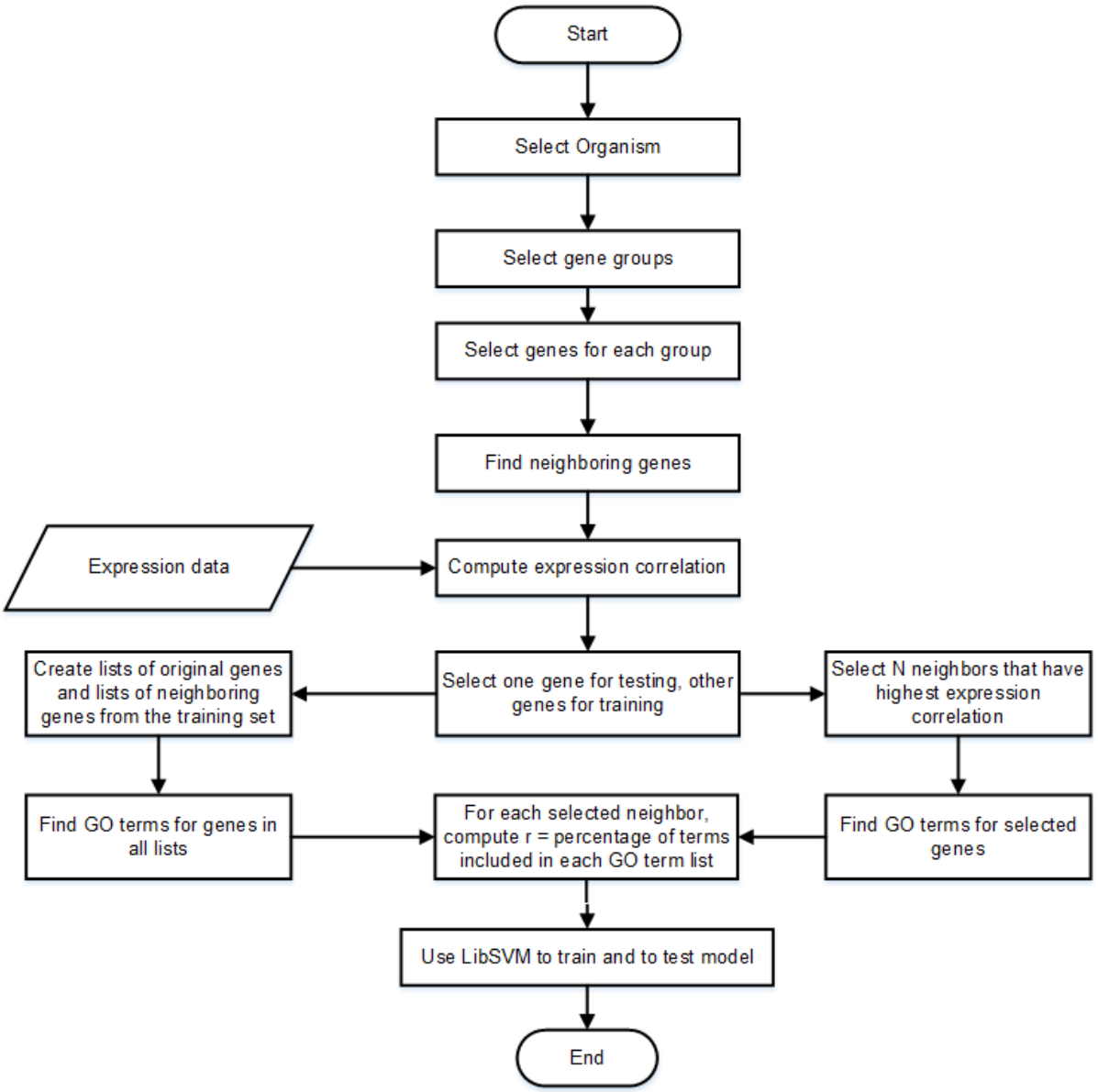

Figure 3.1 The workflow of basic steps in this project

3.2.2.1 Training and testing data for SVM

The dataset of each organism was split into two subsets, the training set and the test set. In this project, we used one record for testing and all other records for training. Then for each protein group in the training set, we created two lists. One list contains the selected genes and the other list contains all the neighbors of the selected (central) genes. After that, we retrieved the GO terms for every gene in these lists. From then on, we only worked with these lists of GO terms. For example, if we have two groups of transporter proteins (amino acid transporters and sugar transporters), then we have four lists of GO terms (the first list contains all GO terms of all amino acid transporters in the training set, the second list contains all GO terms of all neighboring genes of these amino acid

transporters, the third list contains all GO terms of all sugar transporters in the training set, and the fourth list contains all GO terms of all neighboring genes of sugar transporters).

For each central gene in the training set, we selected maximum *N* neighbors that have the highest co-expression correlation with the central gene. Then we identified the GO terms for each selected neighbor. After that, we computed the percentage of GO terms that are contained in each GO list. If this percentage was greater than or equal to a pre-selected threshold (*r*) then we assigned the value 1, otherwise we assigned the value 0. As a test, we also used real-valued functional similarities obtained from GOSemSim [149]. Yet, this strategy gave results of lower quality than the binary-valued approach. Using binary-value has a disadvantage, because a higher threshold (*r*) yields more 0 values. For some cases we did not obtain a value of 1 at all, and a vector with all 0 values is not usable for SVM. Supplement table 2, Supplement table 3 and Supplement table 4 summarize the number of genes that we found suitable to use to build the models. For gene ArtQ of *Escherichia coli* (see Figure 3.2) in the training set, for example, we selected the neighbors that had the highest co-expression levels (ArtM, ArtI and ArtP). If neighbor ArtI is selected, we compute what percentage of its GO terms are contained in each of the four lists of GO terms. If this percentage is greater than or equal to a pre-selected threshold (*r*) then we assigned the value 1, otherwise we assigned the value 0. Since we have four GO term lists, this gives four values. If we select three neighbors that have the highest co-expression correlation then we have 3x4=12 values of 0 or 1. We used these twelve features together with the group's names, that were converted to positive integer values, as class label to train the classifier. These steps were repeated for all genes in the testing set.

3.2.2.2 Support vector machine for classification

Support Vector Machine classification [91] of substrate specificity or of participation in metabolic pathways was done with the software LIBSVM [150]. LIBSVM can efficiently classify samples into multiple classes, it automatically selects a model, which can generate contours of the cross validation accuracy, and it makes cross-validation for model selection and treats unbalanced data by using a weighted SVM. In this project, we used leave-one-out cross validation. LIBSVM also provides various kernel functions and different SVM formulations. We tested our method with three kernel functions (linear, radial basis function (RBF), and sigmoid). In most cases with different threshold *r*, number of neighboring genes or organisms, RBF gave the best results. Then we proceeded using RBF and tested for different values of the *cost* parameter (0.1, 0.5, 1. 1.5, 5 and 10). The default cost parameter of 1 gave the best results. A lower value of 0.1 gave the worst accuracies. The reliability increases substantially when *cost* changes from 0.1 to 0.5. The accuracies

of *Saccharomyces cerevisiae* changed by 15%, accuracies of *Escherichia coli* by 5.6 % and of *Homo sapiens* by 11.6% at most, respectively. With *cost* parameter greater than or equal to one, the accuracies did not show remarkable changes. We also tested four different values of the *gamma* parameter (1.0, 0.8, 0.5, 0.3 and default value of *gamma*). The default value of *gamma* gave better accuracies than other values in most of the cases. For this reason, we kept the default values of all the parameters.

3.2.2.3 Model validation and evaluation

We used leave-one-out cross validation to evaluate the prediction ability of our model. In the leave-one-out cross validation, one record was used for testing, all others were used for training. The process of training and testing was repeated until all records had been used for testing once. Accuracy (ACC) was evaluated in the usual way as:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP, FP, TN and FN are true positive, false positive, true negative, and false negative, respectively.

3.3 Results

3.3.1 Transporter proteins

For illustration, Figure 3.2 shows that the *Escherichia coli* gene ArtQ has large co-expression levels with several neighbors (ArtM, ArtI and ArtP) for the selected microarray dataset. As suggested by the very similar gene names, all these genes transport amino acids. Thus we predict that ArtQ also transports amino acid.

Figure 3.2 Co-expression levels of central gene ArtQ and its neighboring genes

First, we set the number of upstream and downstream neighbors to 10 each and selected the 3 neighbors with highest co-expression correlation. Figure 3.3 shows the results for three different thresholds *r*.

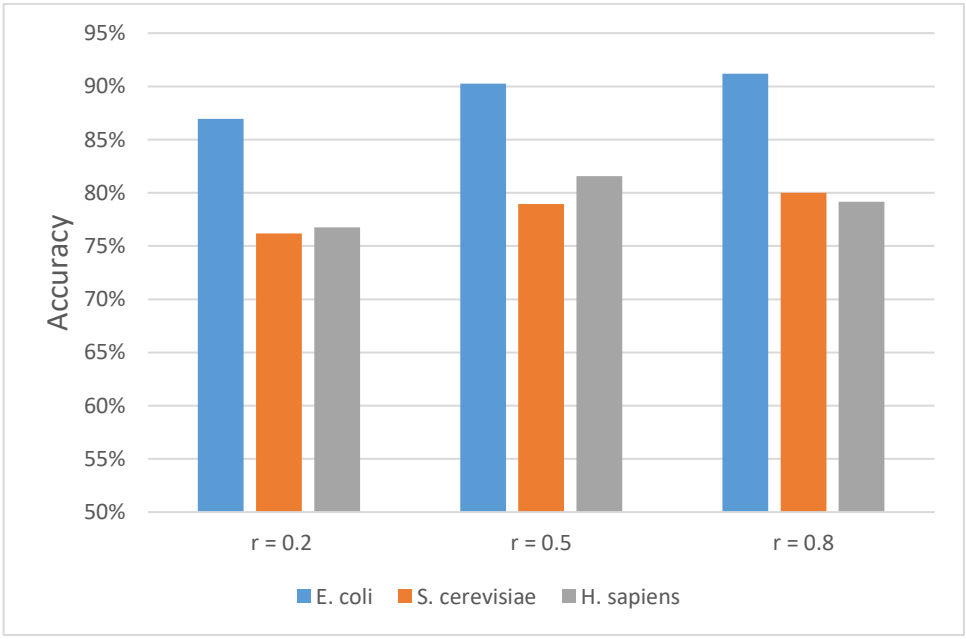

Figure 3.3 Effects of the similarity threshold r of GO terms on the accuracy of transporter substrate classification

When the threshold *r* was increased from 0.2 to 0.5, all accuracies increased likewise (*Escherichia coli*: from 87% to 90%, *Saccharomyces cerevisiae*: from 76% to 78%, *Homo sapiens*: from 77% to 82%). When the threshold was increased further from 0.5 to 0.8, the accuracies of

Escherichia coli and of *Saccharomyces cerevisiae* increased further whereas that of *Homo sapiens* decreased slightly. For *Homo sapiens*, more sugar transporters were incorrectly classified than amino acid transporters, although the number of amino acid transporters is much larger than the number of sugar transporters (Table 3.2).

Table 3.2 Number of genes that were correctly and in-correctly classified

Organism	Transporter substrate	r = 0.2		r = 0.5		r = 0.8	
		Correctly classified	Not correctly classified	Correctly classified	Not correctly classified	Correctly classified	Not correctly classified
Escherichia	Sugar	18	3	15	3	14	2
coli	Amino acid	22	3	22	1	17	1
Saccharomyces	Sugar	5	2	5	2	5	1
cerevisiae	Amino acid	11	3	10	2	7	2
Homo sapiens	Sugar	3	7	3	6	4	3
	Amino acid	30	3	28	1	15	2

Next, we varied the number of neighbors while keeping the threshold *r* at 0.5. Figure 3.4 shows the results for three cases where the windows sizes were (5, 3), (10, 3) and (20, 5), respectively. (10, 3) gave the best result for all three organisms.

For comparison, we compared our tool against two webservers that predict substrate specificities of membrane transporters from the protein sequence: (1) TrSSP (http://bioinfo.noble.org/TrSSP/) ([151]) using the options "AAindex + PSSM based (Swissprot)" and (2) TransportTP (http://bioinfo3.noble.org/transporter/) ([152]) using an E-value threshold = 0.1. The results obtained with these methods are listed in Table 3.3. Our method gave superior results (90% accuracy and higher) than TrSSP (64% in the best case) and TransportTP (54% in the best case) for *Escherichia coli* sequences. TransportTP did not provide useful results for *Saccharomyces cerevisiae* and human sequences. The results of TrSPP for human sequences were of comparable accuracy to those of our tool. *For Saccharomyces cerevisiae* sequences, TrSPP provided better results than our tool. In addition, it should be noted that our method was not able to make predictions for for transporters that have non-zero features (see methods; paragraph "Training and testing data for SVM").

Table 3.3 Comparison against alternative methods for predicting substrate specificities

Organism	Group	Number of sequences	TrSSP		TransportTP	
			Correct	Accuracy	Correct	Accuracy
Escherichia coli	aa	47	23	48.94%	10	21.28%
	sugar	39	25	64.10%	21	53.85%
Saccharomyces cerevisiae	aa	24	20	83,33%	0	0.00%
	sugar	17	16	94,12%	0	0.00%
Homo sapiens	aa	37	31	83,78%	0	0.00%
	sugar	13	10	76,92%		

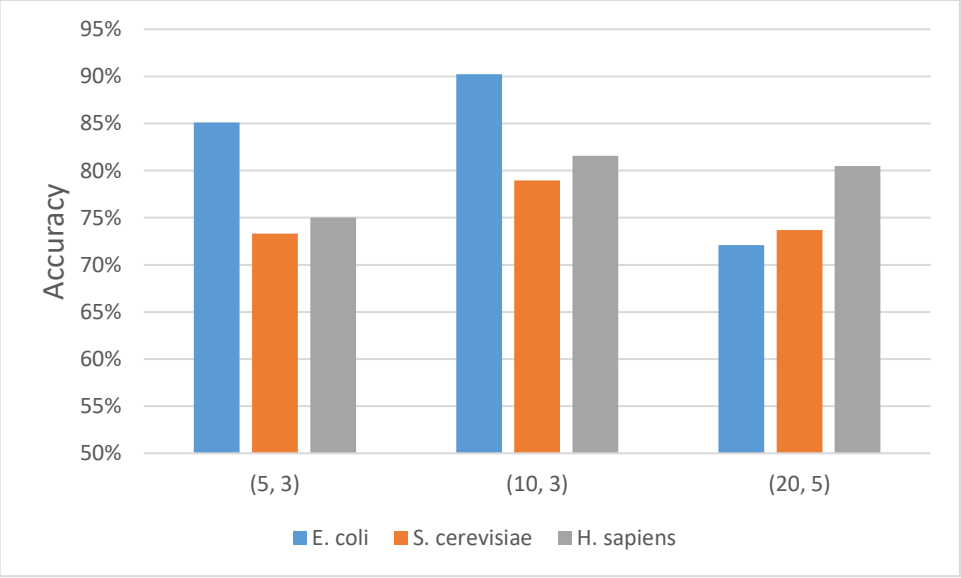

Figure 3.4 Prediction accuracy for different window sizes

3.3.2 Metabolic pathway enzymes

Next we tested the same approach for the genes coding for enzymes belonging to different groups of metabolic pathways of *Escherichia coli*. Supplement table 3 shows that, when the number of neighbors was extended, the number of genes that can be used by SVM decreased. In consequence, the accuracies decreased when we considered more neighbors (Figure 3.5). This characteristic was not found for the transporter proteins.

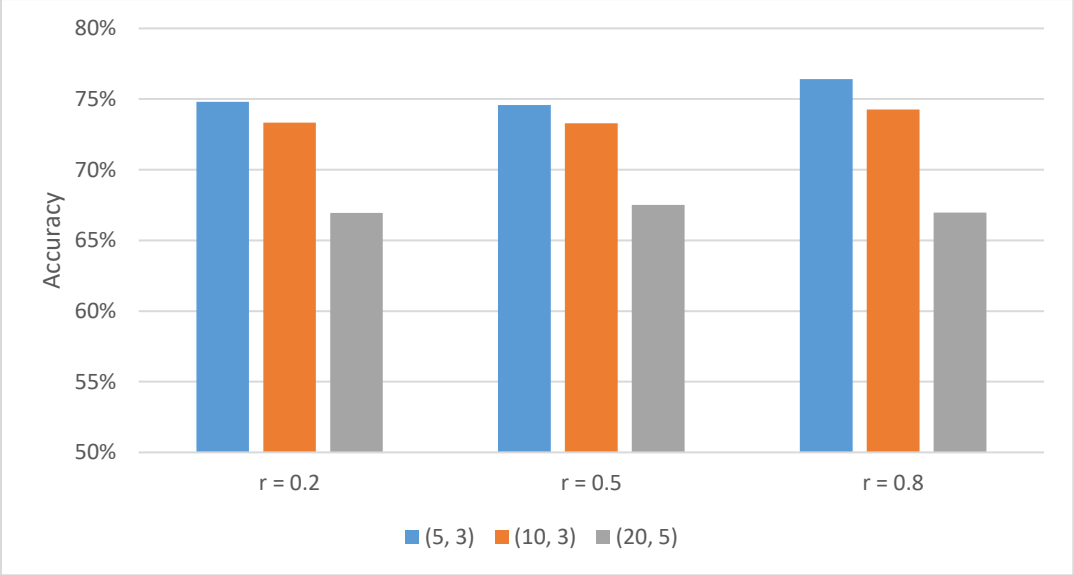

Figure 3.5 Accuracies of different thresholds r and number of neighbors when testing with enzymes of the sugar and amino acid metabolism

After testing with two groups, we tested the method with the four groups of genes involved in sugar, amino acid, lipid, and nucleotide pathways, respectively. Figure 3.6 shows that the accuracies relative to the random prediction (25%) are increased compared to the previous test. Secondly, the threshold *r* had only a small effect when we extended the number of neighbors to (20, 5).

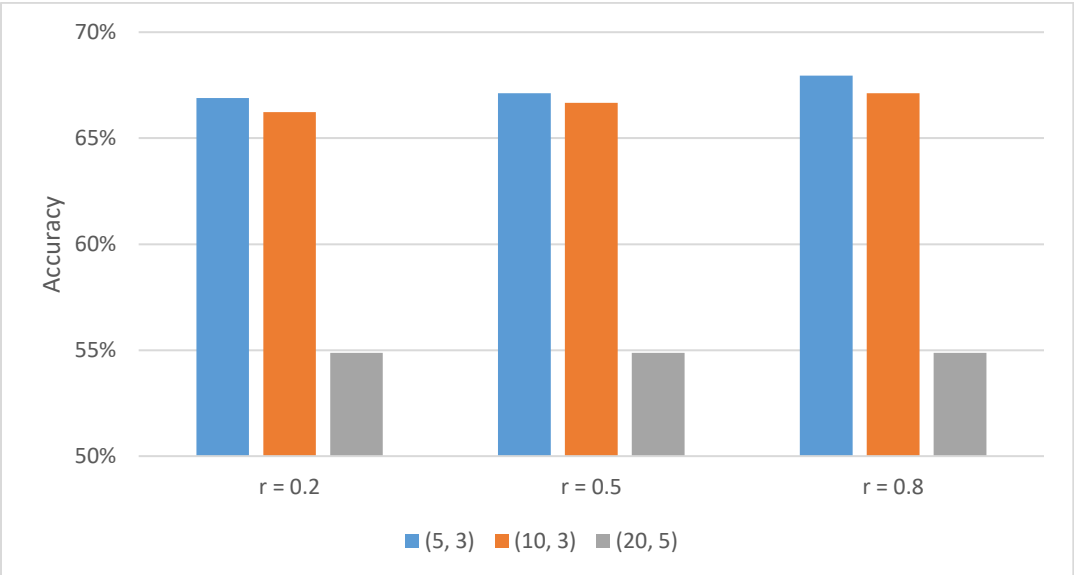

Figure 3.6 Accuracies of 4-class prediction for different thresholds and number of neighbors when testing with enzymes belonging to the sugar, amino acid, lipid and nucleotide metabolic pathways

3.4 Discussion

The main findings of our study are:

a) The function of membrane transporters and of metabolic enzymes is best associated with that of its co-expressed neighbor genes for *Escherichia coli*, followed by *Saccharomyces cerevisiae*, and by *Homo sapiens*.

b) The substrate-specificities of membrane transporters can be classified better than the membership of enzymes to four major metabolic pathway classes.

The first finding had to be expected. Operons exist in bacteria and rarely in eukaryotes (*Saccharomyces cerevisiae* and *Homo sapiens*). Junier and Rivoire recently reported that the 2034 genes of *Escherichia coli* are arranged in 740 synteny segments [153]. They found that co-expression occurs at high levels within synteny segments and low levels outside. However, it was also suggested that functionally related genes are grouped together in bacteria outside of operons in the form of so-called "uber-operons" [154].

In yeast, the most highly co-expressed pairs of neighbor genes tend to be similar in function [133], [155] . Adjacent genes are frequently (more than 25%) transcribed in the same phase(s) of the cell cycle[130].

For *Homo sapiens*, Wang and colleagues recently compared the expression profiles of bulk tissue of glioblastoma patients to expression profiles at single-cell level [136]. Interestingly, they found that co-expression in bulk samples was stronger associated with similar gene function than that in single cell samples. In the latter case, co-expressed genes showed a stronger tendency to physically interact with each other. Nevertheless, our results show that the biological functions of co-expressed neighbor genes are in all three investigated species associated with the function of the central gene.

When compared to results obtained the alternative method TrSSP, our method gave superior results for *Escherichia coli* transporters, results of comparable quality for human transporters, and results of slightly lower accuracy for *Saccharomyces cerevisiae* transporters. Since both methods take quasi-orthogonal approaches, it appears worthwhile to combine both methodologies in the future.

Now we turn to the question why function prediction gave better results for the membrane transporters than for metabolic enzymes. To us, this came as a surprise. In *Arabidopsis thaliana* (which was not studied here), Ren and colleagues reported that co-functionality was in most cases a poor predictor of co-expression, also for neighboring genes [156]. When turned around, this

suggests that co-expressed and gene neighborhood cannot be taken as guarantee for co-functionality, at least not in eukaryotic genome.

Cui and colleagues recently analyzed correlations of the expression levels of neighboring genes in *Homo sapiens* [157]. Interestingly, they distinguished between four types of genes: housekeeping genes, specific and selective genes that are either preferentially or exclusively expressed in response to physiological stimuli, and repressed genes. Importantly, they found that the direction of transcription of gene pairs (parallel or antiparallel) has at most a weak effect on the level of co-expression. This supports the approach taken in our study where we have ignored directionality of genes. Compared to randomly selected gene pairs, preferentially expressed and repressed genes showed a substantially higher co-functionality. Interestingly, this was not the case for neighboring housekeeping genes and exclusively expressed gene pairs that showed an even lower co-functionality than randomly selected gene pairs.

These results show that functional associations may be quite case-specific.

3.5 Conclusion

In this work, we focused on the classification of integral membrane transporters from three organisms (*Escherichia coli, Saccharomyces cerevisiae* and *Homo sapiens*) according to their transported substrates. The idea was to identify among the close neighbors of a query gene with unknown function those genes that show high co-expression with this gene. Then, we identified frequent GO terms among these co-expressed neighbors and used a support vector machine classifier to annotate the substrate specificity of the query gene. Training of the method was performed on groups of known amino acid and sugar transporters. For transporter proteins, the average accuracies of *Escherichia coli, Saccharomyces cerevisiae* and *Homo sapiens* were 89%, 78%, and 79%, respectively. When tested on the genes belonging to different metabolic pathways of *Escherichia coli*, the average accuracy was 75% (two classes) and 67% (four classes). In future works, this approach may be used in combination with other features such as sequence motifs, sequence similarity, and further characteristics of the protein sequence such as its amino acid composition.

Chapter 4 Copy number alterations in tumor genomes deleting antineoplastic drug targets partially compensated by complementary amplifications

This chapter is based on the following publication:

Tran, V.H., Kiemer, A., Helms, V. Copy number alterations in tumor genomes deleting antineoplastic drug targets partially compensated by complementary amplifications. (2018) Cancer Genomics & Proteomics, V. 15, p. 365-378,doi: 10.21873/cgp.20095.

My contribution was to design the research project and analyze the results together with the co-author Volkhard Helms. I, Alexandra K. Kiemer, and Volkhard Helms prepared the manuscript. I collected data and performed the calculations.

4.1 Introduction

Tumor cells differ phenotypically from normal cells, for example, by showing increased levels of proliferation and evading apoptosis [62]. At the genomic level, one common variation of tumor cells are DNA copy number changes that include both gene amplifications and deletions [158]. When these changes occur in germline cells, they are referred to as DNA copy number variations (CNV). When they occur in somatic cells, they are termed copy number alterations (CNA) [32]. It is believed that CNAs in genome sequences of cancer patients [159] may play important roles in oncogenesis and cancer therapy [160].

An important reference data set on CNAs in patients suffering from more than 30 different tumors was compiled by The Cancer Genome Atlas (TCGA) project. A pan-cancer study of these data analyzed the effect of CNAs on known oncogenic drivers and tumor suppressor genes (TSG) and identified potential new cancer drivers, TSGs and biomarkers [161]. This study also analyzed the length and the distribution of somatic CNAs along the chromosomes, identified regions that recurred significantly often and compared the number of genes in amplified and deleted regions [161]. Subsequent studies [162], [163] of CNA data from TCGA focused either on specific genes (e.g. PD-L1, CD247, IRS4, IGF2) or on the relationship between copy number events and gene expression [162], [164]. From the 33 tumor types available at TCGA today, we processed the data from 31 tumors in this study (glioblastoma multiforme, renal clear cell carcinoma, brain lower grade

glioma, lung squamous cell carcinoma, liver hepatocellular carcinoma, renal papillary cell carcinoma, kidney chromophobe carcinoma, breast invasive carcinoma, ovarian serous cystadenocarcinoma, uterine carcinosarcoma, head and neck squamous cell carcinoma, thyroid carcinoma, prostate adenocarcinoma, colon adenocarcinoma, stomach adenocarcinoma, bladder urothelial carcinoma, cervical squamous cell carcinoma and endocervical adenocarcinoma, sarcoma, acute myeloid leukemia, esophageal carcinoma, pheochromocytoma and paraganglioma, rectum adenocarcinoma, adrenocortical carcinoma, cholangiocarcinoma, lymphoid neoplasm diffuse large B-cell lymphoma, uveal melanoma, mesothelioma, thymoma, testicular germ cell tumors, uterine corpus endometrial carcinoma, pancreatic adenocarcinoma). The original publications on the datasets collected for these thirty-one tumors focused on the rate of copy number alterations, identification of recurrently amplified/deleted CNAs, the distribution of CNAs along the chromosomes, identification of oncogenes and TSGs, and clustered the tumors into subtypes. Several follow-up studies have analyzed CNA data from TGCA and analyzed copy number changes [165]–[167], recurrent copy number variations/alterations [168], [169]–[171], the effect of CNAs on specific genes [172]–[176], identified putative new druggable cancer driver genes [177], tried to predict cancer relapse [178], and studied how cancer patients may be grouped into subtypes [167], [173].

Tumor therapy often involves chemotherapy [179]. The current release of Drugbank (version 5.0.11, downloaded on January 12, 2018) lists 477 drugs as antineoplastic (AN) agents that are annotated to bind to 220 different protein targets. Mapping the targets of AN agents to the KEGG database of cellular pathways using the tool KEGG mapper [180] shows that 53 target proteins from this list belong to the PI3K-Akt signaling pathway, 39 to metabolic pathways, 32 to the Rap1 signaling pathway, 30 to Th17 cell differentiation, 32 to the Ras signaling pathway, and 38 to the MAPK signaling pathway. The complete list of these pathways is included as Supplement table 5.

The aim of this project was to analyze how protein targets of AN agents are affected by CNAs. To our best knowledge, no prior study addressed a related question so far. The only related work we are aware of is a study by Graham *et al*. who recently reported that recurrent patterns of DNA copy number alterations in tumors reflect metabolic selection pressures such as coordinated alteration of genes involved in glycolytic metabolism [181]. For 31 tumor types from the TCGA dataset (see list above), we compared how recurrent CNAs affected the set of protein targets of chemotherapeutic drugs in comparison with a set of housekeeping genes and a set of cancer hallmark genes

4.2 Materials and methods

Figure 4.1 summarizes the main steps of our analysis.

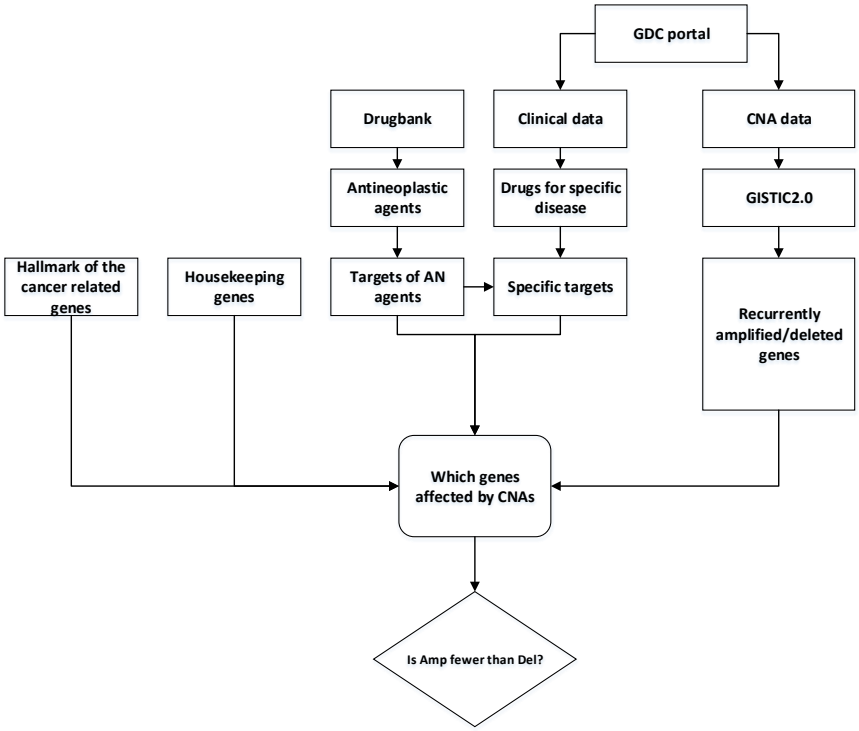

Figure 4.1 Main steps of analysis workflow

4.2.1 Data on copy number alterations

As mentioned, we analyzed genomic data from the TCGA project on CNAs observed in patients suffering from 31 different forms of tumors (listed in the introduction section). Missing from this list are the data for lung adenocarcinoma and skin cutaneous melanoma as these could not be processed with the GISTIC2.0 tool (see below). The CNA data of these patients (start and end position, chromosome, and segment mean of CNA) were downloaded from the Genomic Data Commons Portal (GDC portal) on September 29, 2017 [182].

4.2.2 Clinical data

From the clinical data provided at GDC, we extracted information on which drug treatment was given to specific patients. Thereby, the presence of CNAs in individual patient genomes was associated with the drug treatment applied to these patients. In our work, only data from patients that had both CNA and clinical data available were used.

4.2.3 Antineoplastic agents and their targets

A list of 477 ANs together with their target proteins was extracted from Drugbank [183] (version 5.0.11, downloaded on January 12, 2018). We considered only those protein targets for which

pharmacological action of the respective drug molecule is reported as "yes" in Drugbank. These 477 AN agents are reported to bind to 220 different protein targets (labeled here by their Uniprot accessions numbers). After converting Uniprot accession numbers to gene symbols, we were left with 218 genes. As "tumor-specific" drugs, we considered those drugs that were applied to the patients of a particular tumor entity according to the TCGA data files. As shown in Supplement table 6 and Supplement table 7 for drugs against lung cancer or breast cancer, these sets comprise a representative subset of the FDA-approved drug treatments for these tumors types (8 out of 16 and 23 out of 31), see https://www.cancer.gov/about-cancer/treatment/drugs/cancer-type. The sets for lung squamous cell carcinoma and breast cancer also included eight further drugs each that are not FDA-approved, but applied to TCGA patients possibly during ongoing clinical trials. Here, such drugs are labeled as "experimental drugs".

4.2.4 Gene sets

Beside the set of protein targets of AN agents, we also considered a set of 3804 housekeeping genes [184] (i.e. at least one variant of these genes is expressed in all tissues uniformly; downloaded from https://www.tau.ac.il/~elieis/HKG/ on January 13, 2018) and a set of 2338 "hallmark genes" of cancer. The latter set contains all human genes that are annotated in the Gene Ontology [53] to at least one of 37 Gene Ontology terms that were described as hallmarks of cancer [63] (downloaded from http://geneontology.org/page/download-annotations on January 13, 2018). After converting Uniprot accession numbers to symbols, this gave 2321 gene symbols in the hallmarks of cancer gene set. Figure 4.2 shows the overlap of the three gene sets.

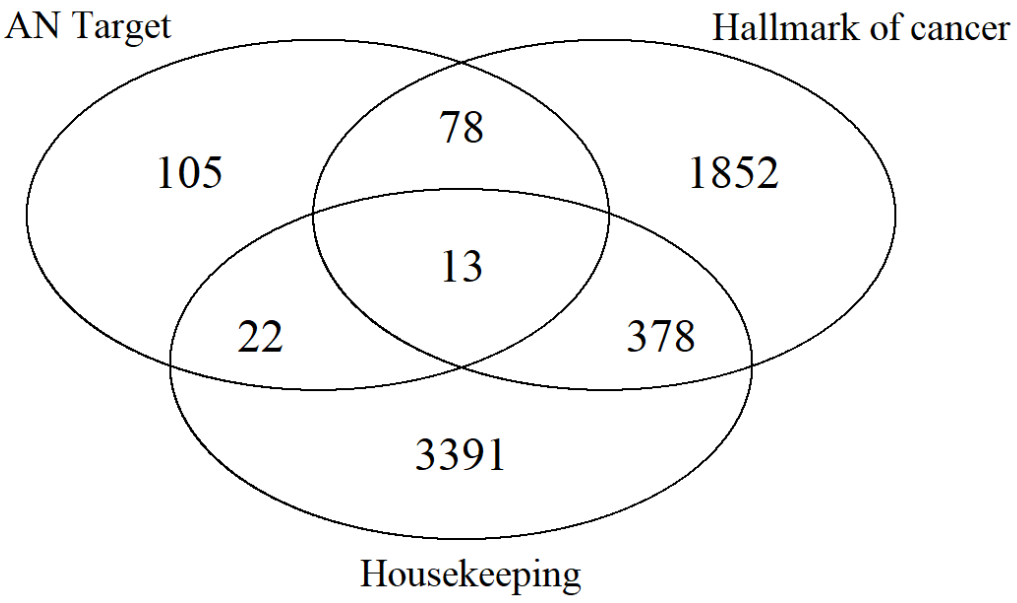

Figure 4.2 Overlap between the three gene sets

4.2.5 Genes affected by copy number alterations

Genes that are recurrently affected by CNAs were identified with the GISTIC2.0 tool version 2.0.22 [112] using segmentation files and marker files created from the CNA data of the tumor samples. Following Laddha *et al.* [185], we used 0.2 and -0.2 as thresholds for GISTIC2.0 to identify recurrent amplification and deletion peaks and the genes contained in those peaks. Uniprot accession numbers used by Drugbank were converted to gene symbols used by GISTIC2.0 by making use of data from the HUGO Gene Nomenclature Committee (HGNC database) [186] that were downloaded in January 2017. Information on genes (chromosome, start position, and end position) was based on data from Ensembl (data downloaded from http://rest.ensembl.org on January 16, 2018).

4.3 Results

4.3.1 General statistics

The aims of this work were (1) to test the hypothesis that genomic CNAs observed in tumors affect the protein targets of AN agents significantly more often than expected by chance, (2) to test whether either amplifications or deletions are more common, and (3) to study the potential relevance for chemoresistance. In principle, one can expect that eventually all genes except for the essential genes will be affected by CNAs in some patients. Hence, to get more meaningful results, our analysis was focused on the set of recurrently occurring CNAs that appear statistically more often in each individual tumor entity than expected by chance. This strategy is similar to that used by Graham *et al.* [181].

Table 4.1 lists the number of recurrently amplified and deleted genes obtained by processing the raw CNA data for the 31 considered tumors with the GISTIC2.0 program. Specified is also how many of these amplifications/deletions affect hallmark genes, housekeeping genes, and protein targets of AN drugs. Note that, in this initial analysis, protein targets of all 477 considered AN drugs were considered irrespective of whether these drugs are actually being used to treat the particular subtype of cancer. In acute myeloid leukemia, 38 of 105 cases (26.57%) received treatment prior to the time when the CNA data ware taken. For glioblastoma (22 of 590 cases) and renal clear cell carcinoma (18 of 530 cases), the number of such cases was around 4%. In all other tumors, the fraction of pre-treated patients was below 3 %. Hence, in all tumors except for acute myeloid leukemia, the detected amplifications and deletions are unlikely to reflect resistance phenomena occurring in response to treatment (Supplement table 8). As shown in Table 4.1, in twenty-nine out

of thirty-one studied tumors (the exceptions are thyroid carcinoma and kidney chromophobe), the number of recurrently deleted genes exceeded the number of recurrently amplified genes. However, this difference between the lower number of amplifications and the higher number of deletions was equally significant for the sets of all genes, antineoplastic targets, hallmark genes, and housekeeping genes (*p*-values 8.501e-09, 1.721e-08, 9.196e-09 and 8.367e-09, Wilcoxon test) and, hence, does not reflect a peculiar property of AN target genes. Supplement table 9 shows that a similar behavior is observed for genes annotated to specific cancer hallmarks.

Table 4.1 Number of genes affected by CNAs in TCGA data for the 31 considered types of tumors

Disease	Number of cases considered	Number of cases without pre-treatment	Recurrently amplified genes	Recurrently deleted genes	Amplified AN targets	Deleted AN targets	Amplified Hallmark genes	Deleted Hallmark genes	Amplified Housekeeping genes	Deleted Housekeeping genes
Breast Invasive Carcinoma	1094	1079	841	4084	5	34	123	605	76	304
Glioblastoma Multiforme	590	568	231	2176	4	12	20	286	20	190
Ovarian Serous Cystadenocarcinoma	570	569	470	3144	3	30	102	463	34	246
Uterine Corpus Endometrial Carcinoma	540	538	456	8377	3	68	84	1266	33	774
Renal Clear Cell Carcinoma	530	512	3072	5053	33	37	471	771	267	451
Head and Neck Squamous Cell Carcinoma	517	508	715	3166	8	31	121	455	82	238
Brain Lower Grade Glioma	514	511	628	5092	9	45	118	801	61	451
Thyroid Carcinoma	505	500	10	4	0	0	1	1	0	1
Lung Squamous Cell Carcinoma	503	496	1154	3866	14	43	155	577	120	305
Prostate Adenocarcinoma	497	495	497	2600	2	26	70	429	30	232
Colon Adenocarcinoma	450	447	403	2364	4	23	84	317	35	193
Stomach Adenocarcinoma	442	442	1081	4124	9	41	169	641	90	407
Bladder Urothelial Carcinoma	412	402	1248	3049	12	31	232	458	134	266
Liver Hepatocellular Carcinoma	375	374	644	2818	4	28	101	388	58	223
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	295	295	1506	3829	13	34	231	540	114	350
Renal Papillary Cell Carcinoma	290	290	299	6132	5	59	52	922	17	501
Sarcoma	260	259	2602	8101	28	82	407	1201	232	759
Acute Myeloid Leukemia	143	105	3	3714	0	29	1	593	0	344
Esophageal Carcinoma	184	184	801	6773	6	61	130	1010	77	576
Pancreatic Adenocarcinoma	184	183	597	7190	6	59	87	1072	56	595
Pheochromocytoma and Paraganglioma	178	177	56	5840	1	52	9	911	5	513
Rectum Adenocarcinoma	164	163	1116	5663	9	40	190	853	92	508
Testicular Germ Cell Tumors	134	134	2142	2811	21	31	312	443	222	260
Thymoma	124	122	0	2038	0	20	0	352	0	174

Kidney	66	66	38	1	1	0	8	1	6	0
Chromophobe Adrenocortical Carcinoma	90	89	693	5243	4	50	93	778	66	448
Mesothelioma	87	86	0	4357	0	43	0	681	0	464
Uveal Melanoma	80	80	564	3050	5	33	90	465	56	342
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	48	47	110	6580	0	68	15	1041	10	663
Uterine Carcinosarcoma	56	56	917	4615	14	42	159	689	95	439
Cholangiocarcinoma	36	35	19	2801	2	21	3	456	0	274
Max	1094	1079	3072	8377	33	82	471	1266	267	774
Min	36	35	0	1	0	0	0	1	0	0
Average	321.23	316.52	739.13	4150.16	7.26	37.84	117.35	627.94	67.35	370.68

4.3.2 Disease specific statistics

For each disease, we then extracted from the GDC clinical data files the names of the drugs that were prescribed to the respective patients. The analysis was repeated with the same numbers of cases considered as in Table 4.1, but focused on the combined set of cancer-specific targets of these drugs, see Table 4.2. This set of target proteins was termed “specific drug targets” meaning that these are targets of the drugs that are given to patients with this specific tumor entity. By way of construction, the resulting numbers of affected genes were now far smaller. In 18 tumors, no CNA-amplifications affected the specific drug targets. In contrast, sarcoma behaved as an outlier to the other extreme with eight amplified targets. In the 12 remaining tumors, only one or two cases were observed. In contrast, in 23 tumors, CNA-deletions affected the specific drug targets of these tumor types. Among the three tumors (brain lower grade glioma, sarcoma, and mesothelioma) showing the largest number of CNA-deleted targets (10, 14, 11) only mesothelioma showed significantly more deletions than amplifications (adjusted *p*-value of 0.001, Fisher’s exact test). When taking all tumor data together, the difference between specific amplified/deleted targets for the 31 tumors was significant (*p*-values of 0.00016, Wilcoxon rank-sum test).

Table 4.2 Specific drugs and drug targets of the specified disease and the number of observed CNA-amplifications or CNA-deletions affecting the specific drug targets

Disease	Number of Drugs	Number of targets proteins	CNA-amplified targets	CNA-deleted targets
Breast Invasive Carcinoma	38	32	2	4
Glioblastoma Multiforme	37	52	2	2
Ovarian Serous Cystadenocarcinoma	31	19	1	3
Uterine Corpus Endometrial Carcinoma	16	15	0	5

Renal Clear Cell Carcinoma	17	29	2	6
Head and Neck Squamous Cell Carcinoma	18	20	1	3
Brain Lower Grade Glioma	24	37	2	10
Thyroid Carcinoma	1	1	0	0
Lung Squamous Cell Carcinoma	16	16	2	2
Prostate Adenocarcinoma	11	10	0	3
Colon Adenocarcinoma	15	27	1	4
Stomach Adenocarcinoma	22	16	0	2
Bladder Urothelial Carcinoma	20	24	2	3
Liver Hepatocellular Carcinoma	12	26	0	5
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	12	12	0	4
Renal Papillary Cell Carcinoma	14	24	1	7
Sarcoma	23	34	8	14
Acute Myeloid Leukemia	0	0	0	0
Esophageal Carcinoma	11	9	0	3
Pancreatic Adenocarcinoma	15	14	1	1
Pheochromocytoma and Paraganglioma	6	3	0	2
Rectum Adenocarcinoma	12	9	2	0
Testicular Germ Cell Tumors	5	1	0	0
Thymoma	8	11	0	0
Kidney Chromophobe	5	18	0	0
Adrenocortical Carcinoma	10	16	0	2
Mesothelioma	16	30	0	11
Uveal Melanoma	1	0	0	0
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	23	15	0	4
Uterine Carcinosarcoma	10	8	0	4
Cholangiocarcinoma	3	2	0	0

Following up on Table 4.2, Supplement table 10 lists the number of patient genomes where tumor-specific AN targets were affected by CNA mutations. This data shows that, although the absolute number of CNA-affected AN target proteins is quite small (Supplement table 10), the proportion of patients harboring these CNAs is in fact rather high. Respective target amplifications and deletions occur recurrently in a sizeable fraction (0 to 90%) of all patients.

To get more insight into the molecular mechanisms at place, Table 4.3 and Table 4.4 list the gene symbols of the tumor-specific AN targets that were affected by CNA amplifications and deletions (Table 4.2) and the respective drugs that were applied to patients of these tumors. Experimental drugs were marked by label ^{EXP}, e.g. docetaxel^{EXP}. For acute myeloid leukemia that contains a sizeable fraction of pre-treated patients (26.57 %) no information about the applied drugs

is provided in the TCGA clinical data files so that we could not identify recurrent CNA amplifications or deletions of cancer-specific drug targets in this case.

Table 4.3 Gene names and corresponding drugs of specific AN targets that were recurrently amplified by CNAs. The drugs that bind to the respective AN target proteins are given in brackets. Tumors having no amplified AN targets and that are not listed in Table 4.5 are not shown.

Disease	Target gene (Drug name)
Breast Invasive Carcinoma	TOP2A (Mitoxantrone^{EXP}, Doxorubicin), EGFR (Lapatinib)
Glioblastoma Multiforme	KDR (Cabozantinib^{EXP}, Sorafenib^{EXP}), EGFR (Erlotinib^{EXP}, Gefitinib^{EXP})
Ovarian Serous Cystadenocarcinoma	VEGFA (Bevacizumab)
Uterine Corpus Endometrial Carcinoma	-
Renal Clear Cell Carcinoma	FLT4 (Sunitinib, Sorafenib, Axitinib, Pazopanib), BRAF (Sorafenib)
Head and Neck Squamous Cell Carcinoma	TYMS (Capecitabine^{EXP}, Pemetrexed^{EXP}, Fluorouracil^{EXP})
Brain Lower Grade Glioma	KIT (Imatinib^{EXP}, Sorafenib^{EXP}), EGFR (Erlotinib^{EXP}, Afatinib^{EXP})
Thyroid Carcinoma	-
Lung Squamous Cell Carcinoma	TYMS (Pemetrexed), EGFR (Erlotinib, Gefitinib)
Prostate Adenocarcinoma	-
Colon Adenocarcinoma	VEGFA (Aflibercept, Bevacizumab)
Stomach Adenocarcinoma	-
Bladder Urothelial Carcinoma	EGFR (Erlotinib^{EXP}), VEGFA (Bevacizumab^{EXP})
Liver Hepatocellular Carcinoma	-
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	-
Renal Papillary Cell Carcinoma	FLT4 (Sunitinib, Sorafenib, Pazopanib)
Sarcoma	PDGFRA (Pazopanib), HDAC2 (Vorinostat^{EXP}), FLT4 (Sorafenib^{EXP}, Pazopanib), TUBB1 (Docetaxel^{EXP}), KIT (Imatinib, Sorafenib^{EXP}, Pazopanib), KDR (Sorafenib^{EXP}, Pazopanib), PTGS2 (Sulindac^{EXP}), FGFR1 (Sorafenib^{EXP})
Pancreatic Adenocarcinoma	TYMS (Capecitabine^{EXP}, Fluorouracil)
Rectum Adenocarcinoma	TOP2A (Etoposide), VEGFA (Aflibercept, Bevacizumab)

Table 4.4 Names of genes that were recurrently deleted by CNAs. The drugs that bind to the respective AN target proteins are given in brackets. Tumors having no amplified AN targets and that are not listed in Table 4.3 are not shown.

Disease	Target gene (Drug name)
Breast Invasive Carcinoma	TUBA1A (Vinblastine), TUBB3 (Ixabepilone), PGR (Megestrol acetate), ESR2 (Tamoxifen)
Glioblastoma Multiforme	FLT1 (Sorafenib^{EXP}), FLT3 (Sorafenib^{EXP})
Ovarian Serous Cystadenocarcinoma	RRM1 (Gemcitabine), PSMB1 (Bortezomib^{EXP}), ESR2 (Tamoxifen^{EXP})
Uterine Corpus Endometrial Carcinoma	RRM1 (Gemcitabine^{EXP}), PGR (Megestrol acetate), ESR1 (Tamoxifen^{EXP}, Fulvestrant^{EXP}), ESR2 (Tamoxifen^{EXP}), VEGFA (Bevacizumab^{EXP})
Renal Clear Cell Carcinoma	FLT1 (Sunitinib, Sorafenib, Axitinib, Pazopanib), CRBN (Thalidomide^{EXP}), FLT3 (Sunitinib, Sorafenib), NR1I2 (Erlotinib^{EXP}), RAF1 (Sorafenib), FGFR2 (Thalidomide^{EXP})
Head and Neck Squamous Cell Carcinoma	RRM1 (Gemcitabine^{EXP}), BCL2 (Paclitaxel^{EXP}), MTOR (Everolimus^{EXP})
Brain Lower Grade Glioma	TUBA1A (Vinblastine^{EXP}), TOP1MT (Irinotecan^{EXP}), FLT4 (Sorafenib^{EXP}), NR1I2 (Erlotinib^{EXP}), GSR (Carmustine), PDCD1 (Pembrolizumab^{EXP}), TYMS (Capecitabine^{EXP}), FGFR2 (Thalidomide^{EXP}), ESR2 (Tamoxifen^{EXP}), FGFR1 (Sorafenib^{EXP})
Lung Squamous Cell Carcinoma	TUBB (Vincristine^{EXP}, Vinorelbine), NR1I2 (Erlotinib)
Prostate Adenocarcinoma	LHCGR (Goserelin), MAPT (Docetaxel), CYP17A1 (Abiraterone)
Colon Adenocarcinoma	MAPK11 (Regorafenib), TYMS (Raltitrexed^{EXP}, Capecitabine, Fluorouracil, Floxuridine^{EXP}), PGF (Aflibercept), FGFR2 (Regorafenib)
Stomach Adenocarcinoma	MAP2 (Docetaxel, Paclitaxel^{EXP}), NR3C1 (Dexamethasone^{EXP}, Methylprednisolone^{EXP})
Bladder Urothelial Carcinoma	HDAC2 (Vorinostat^{EXP}), BCL2 (Paclitaxel^{EXP}), TUBE1 (Vinblastine^{EXP})
Liver Hepatocellular Carcinoma	TOP2A (Doxorubicin^{EXP}), MAPK11 (Regorafenib), BRAF (Regorafenib, Sorafenib), FGFR2 (Regorafenib), FRK (Regorafenib)
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	MAP2(Paclitaxel^{EXP}), TUBB1(Paclitaxel^{EXP}), TOP1(Topotecan), ALPPL2(Amifostine^{EXP})
Renal Papillary Cell Carcinoma	IFNAR2 (Interferon Alfa-2b, Recombinant^{EXP}), RRM1 (Gemcitabine^{EXP}), FLT1 (Sunitinib, Sorafenib, Pazopanib), FLT3 (Sunitinib, Sorafenib), MS4A1 (Rituximab^{EXP}), GART (Pemetrexed^{EXP}), IFNAR1 (Interferon Alfa-2b, Recombinant^{EXP})
Sarcoma	RET (Sorafenib^{EXP}), PDGFRB (Sorafenib^{EXP}, Pazopanib), FLT1 (Sorafenib^{EXP}, Pazopanib), HDAC3 (Vorinostat^{EXP}), FLT4 (Sorafenib^{EXP}, Pazopanib), BRAF (Sorafenib^{EXP}), TYMS (Pemetrexed^{EXP}), PTGS2 (Sulindac^{EXP}), CYP19A1 (Letrozole^{EXP}), ATIC (Pemetrexed^{EXP}), MAP2 (Docetaxel^{EXP}), PGR (Megestrol acetate^{EXP}), MAP4 (Docetaxel^{EXP}), FGFR1 (Sorafenib^{EXP})
Acute Myeloid Leukemia	

Esophageal Carcinoma	RRM1 (Gemcitabine^{EXP}), TUBB1 (Docetaxel, Paclitaxel^{EXP}), BCL2 (Paclitaxel^{EXP})
Pancreatic Adenocarcinoma	RRM1 (Gemcitabine)
Pheochromocytoma and Paraganglioma	RRM1 (Gemcitabine^{EXP}), TUBB (Vincristine^{EXP})
Adrenocortical Carcinoma	RET (Sorafenib^{EXP}), BRAF (Sorafenib^{EXP})
Mesothelioma	PDGFRB (Sunitinib^{EXP}), CSF1R (Sunitinib^{EXP}), HDAC2 (Vorinostat^{EXP}), FLT1 (Sunitinib^{EXP}), ATIC (Pemetrexed), HDAC3 (Vorinostat^{EXP}), FLT3 (Sunitinib^{EXP}), FLT4 (Sunitinib^{EXP}), TUBB (Vinorelbine^{EXP}), MTOR (Temsirolimus^{EXP}, Everolimus^{EXP}), VEGFA (Bevacizumab^{EXP})
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DHFR (Methotrexate), TUBA1A (Vinblastine), TNFSF11 (Lenalidomide), NR3C1 (Dexamethasone, Prednisone, Prednisolone^{EXP})
Uterine Carcinosarcoma	TUBB1 (Docetaxel, Paclitaxel), BCL2 (Paclitaxel), MAP4 (Docetaxel, Paclitaxel), MAPT (Docetaxel, Paclitaxel)

Comparison of Table 4.3 and Table 4.4 reveals that for some tumors, there exist targets of the same drugs that were both recurrently deleted and amplified in patients of the same tumor type. Table 4.5 lists all such pairs.

Table 4.5 Drugs that bind to amplified and deleted AN targets in a single tumor type. Names of target genes are given in brackets

Disease	Drug name (Amplified target genes)	Drug name (Deleted target genes)
Glioblastoma Multiforme	Sorafenib^{EXP} (KDR)	Sorafenib^{EXP} (FLT1, FLT3)
Renal Clear Cell Carcinoma	Sunitinib (FLT4), Sorafenib (FLT4, BRAF), Axitinib (FLT4), Pazopanib (FLT4),	Sunitinib (FLT1, FLT3), Sorafenib (FLT1, FLT3, RAF1), Axitinib (FLT1), Pazopanib (FLT1)
Brain Lower Grade Glioma	Erlotinib^{EXP} (EGFR), Sorafenib^{EXP} (KIT)	Erlotinib^{EXP} (NR1I2), Sorafenib^{EXP} (FLT4, FGFR1)
Lung Squamous Cell Carcinoma	Erlotinib (EGFR)	Erlotinib (NR1I2)
Colon Adenocarcinoma	Aflibercept (VEGFA)	Aflibercept (PGF)
Renal Papillary Cell Carcinoma	Sunitinib (FLT4), Sorafenib (FLT4), Pazopanib (FLT4)	Sunitinib (FLT1, FLT3) Sorafenib (FLT1, FLT3), Pazopanib (FLT1)
Sarcoma	Sorafenib^{EXP} (FLT4, KIT, KDR, FGFR1), Sulindac^{EXP} (PTGS2), Vorinostat^{EXP} (HDAC2), Docetaxel^{EXP} (TUBB1), Pazopanib (PDGFRA, FLT4, KIT, KDR)	Sorafenib^{EXP} (RET, PDGFRB, FLT1, FLT4, BRAF, FGFR1), Sulindac^{EXP} (PTGS2), Vorinostat^{EXP} (HDAC3), Docetaxel^{EXP} (MAP2, MAP4), Pazopanib (PDGFRB, FLT1, FLT4)

4.4 Discussion

In this project, CNA and clinical data for 31 types of tumors from the TCGA project were combined with information on AN drugs from Drugbank. As shown in Table 4.1, in 29 studied tumors, the number of recurrently deleted genes exceeded the number of recurrently amplified genes. This finding is generally concordant with the results of the TCGA consortium who reported in their pan-cancer study that the 70 peak amplification regions contained a median of 3 genes each, whereas 70 peak regions of CNA deletions contained a median of 4 genes [161]. Earlier studies [161], [164] reported that CNAs promote carcinogenesis and/or tumor progression by deleting tumor suppressor genes (TSGs). In agreement with this, in the dataset studied here the patient genomes of 29 tumors contained at least one of 71 known TSGs [168] in their list of genes recurrently deleted by CNAs. In the case of uterine corpus endometrial carcinoma and lymphoid neoplasm diffuse large B-cell lymphoma, even 22 of the 71 known TSGs were recurrently affected by CNA deletions (Table 4.6).

Table 4.6 Number of tumor suppressor genes affected by CNAs in different tumors.

Disease	Amplified TSG genes	Deleted TSG genes
Breast Invasive Carcinoma	0	12
Glioblastoma Multiforme	1	6
Ovarian Serous Cystadenocarcinoma	2	13
Uterine Corpus Endometrial Carcinoma	2	22
Renal Clear Cell Carcinoma	5	16
Head and Neck Squamous Cell Carcinoma	0	9
Brain Lower Grade Glioma	1	12
Thyroid Carcinoma	0	0
Lung Squamous Cell Carcinoma	2	9
Prostate Adenocarcinoma	0	9
Colon Adenocarcinoma	0	12
Stomach Adenocarcinoma	2	13
Bladder Urothelial Carcinoma	5	15
Liver Hepatocellular Carcinoma	1	9
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	0	10
Renal Papillary Cell Carcinoma	1	13
Sarcoma	6	20
Acute Myeloid Leukemia	0	13
Esophageal Carcinoma	0	18
Pancreatic Adenocarcinoma	1	19
Pheochromocytoma and Paraganglioma	0	15
Rectum Adenocarcinoma	3	18
Testicular Germ Cell Tumors	3	5
Thymoma	0	4

Kidney Chromophobe	0	0
Adrenocortical Carcinoma	0	13
Mesothelioma	0	14
Uveal Melanoma	0	10
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	0	22
Uterine Carcinosarcoma	1	14
Cholangiocarcinoma	0	10

The recurrently amplified/deleted genes of the 31 tumor types had no protein-coding gene in common. This is not unexpected as will be argued in the following. As shown in Table 4.1, recurrent CNA deletions affected on average 4150 genes, which is roughly 20% of all genes. If we assume that the 31 considered tumors are unrelated, we would expect that - by chance – an overlap of $(0.2)^{31} \times 20.000$ genes $= 4 \times 10^{-28}$ genes would be affected in all tumor groups. This number is even smaller for amplified genes. This led to the expected result that all three gene sets (AN targets, housekeeping genes, and hallmark of the cancer genes) had no gene in common that is affected by CNAs in all type of tumors.

Then, we compared how CNAs affect gene subsets comprising antineoplastic (AN) target genes, housekeeping genes, hallmark of cancer genes, or tumor-specific AN target genes. Importantly, in all these gene sets, significantly more genes were affected by deletions than by amplifications. Hence, this observation is not specific to AN target genes nor to tumor-specific AN target genes.

The tumor-specific AN target genes recurrently affected by CNA amplifications are epidermal growth factor receptor (*EGFR*), *FLT4, TYMS, TOP2A*, *KDR, VEGFA*, *BRAF, KIT, PDGFRA, HDAC2, TUBB1*, *PTGS2* and *FGFR1*. These genes belong to 13 types of tumors (Table 4.1 and Table 4.3). In the 18 remaining tumor types, no tumor-specific AN target gene was amplified. As an example, amplifications of *EGFR* gene copy numbers and overexpression of *EGFR* are known to be one of the most common alterations in non-small-cell lung carcinoma (NSCLC) cells [187]–[190] and are associated with a poor prognosis and chemoresistance. Among the histological subtypes of NSCLC, *EGFR* is most frequently expressed in squamous cells [191].

On the other hand, in 23 tumors, CNA-deletions affected specific drug targets of these tumor types. As shown in Table 4.4, CNA deletions of AN targets affected (1) the two enzymes bifunctional purine biosynthesis protein PURH (gene name *ATIC*) [192] and a subunit of ribonucleotide reductase (RRM1) that are both important for cell replication [193]; (2) the nuclear receptor NR1I2 that regulates the metabolism and efflux of xenobiotics *via* CYP3A4 and MDR1

[194]; (3) the mitochondrial and nuclear DNA topoisomerases TOP1MT and TOP2A; (4) the members of the vascular endothelial growth factor receptor family VEGFA, FLT1, FLT3, and (5) fibroblast grown factor FGFR2; (6) estrogen receptor ESR2; (7) the signaling MAP kinase MAPK11 and (8) the B-Raf Proto-Oncogen BRAF that regulates the MAP kinase/ERK signaling pathway [195]; (9) the inhibitory cell surface receptor PDCD1 that is involved in the regulation of T-cell function [196]; and finally (10) beta tubulin TUBB and the microtubule-associated protein MAP1A that is almost exclusively expressed in the brain [197], [198] (and was CNA-deleted in glioblastoma). As all of these proteins have important roles in promoting carcinogenesis, they have likely been selected as targets of antineoplastic agents. As argued above, the CNA mutations pre-existed before the onset of the therapy.

These findings of rare CNA amplifications, but frequent CNA deletions of tumor-specific drug targets have clear consequences on drug development. In future, considering CNA frequencies should certainly become a standard element of drug design efforts. These data also suggests that genomes of tumor patients may contain "compensating" mutations where one target protein of a drug is deleted and another target protein of the same drug is amplified. Unfortunately, due to space reasons we are restricted to discussing only a few of these cases in more detail.

In renal clear cell carcinoma patients that were subsequently treated with the drug molecules pazopanib, sunitinib, sorafenib, and axitinib, the target protein, FLT1, of these drugs was recurrently deleted (in 55 samples), whereas another target protein FLT4 of the same drugs was recurrently amplified (in 337 samples). Overall, 36 samples had both deleted FLT1 and amplified FLT4. FLT4 encodes a tyrosine kinase receptor of the same protein family as vascular endothelial growth factors C and D. In agreement with what is expected from the observed CNA amplification, FLT4 was previously reported to be overexpressed in renal clear cell carcinoma [199]. Besides being a recurrent target of CNA deletions here, FLT1 was also reported to be frequently silenced through promoter hypermethylation in renal clear cell carcinoma [200].

In lung squamous cell carcinoma patients subsequently treated with the drug erlotinib, one of its targets, NR1I2, was recurrently deleted (in 20 samples) and another target, epidermal growth factor receptor (EGFR), was recurrently amplified (in 186 samples). Nine samples had NR1I2 deleted and EGFR amplified at the same time. In brain lower grade glioma, NR1I2 and EGFR were also deleted and amplified, respectively. Beside these two genes, the target KIT of sorafenib was amplified while FLT4 and FGFR1 were deleted.

There exist also cases where the same target protein can be either amplified or deleted. For example, Table 4.6 shows that, FLT4 (target of sorafenib and pazopanib), and PTGS2 (target of sulindac) were observed to be either amplified or deleted in different sarcoma samples. FLT4 was amplified in 57 samples, and was deleted in 36 samples. PTGS2 was amplified in 63 samples, and was deleted in 32 samples.

4.5 Conclusion

The aim of this work was to test the hypothesis that the protein targets of AN agents in tumors are affected by genomic copy number alternations (CNAs) more strongly than expected by chance. Based on CNAs and clinical data from the TCGA repository, we found that the genome sequences of tumor patients generally contain more recurrently deleted CNAs than recurrently amplified CNAs. This is also the case for CNAs affecting target genes of the specific AN for these tumors. Interestingly, we observed certain signs of apparently compensating effects of CNAs. The data available for this study enabled us to identify CNA alterations that existed prior to therapy and that may render certain chemotherapies more or less effective. In future, it would be desirable to also collect time-series CNA data of tumor patients at time of diagnosis and at later time points. This would point to CNA alterations caused by application of certain chemotherapies and thus reflect chemoresistance.

Chapter 5 Tumor genomes frequently contain amplified resistance genes prior to treatment

My contribution was to design the research project, analyze the results, and prepare the manuscript together with the co-author Volkhard Helms. I collected data and performed the calculations.

5.1 Introduction

Chemotherapy is an important and frequently applied treatment option for tumors, besides surgery and radiation therapy. Unfortunately, the effectiveness of chemotherapy often decreases over time due to the onset of drug resistance [201]. Mansoori *et al*. reported that 90% of failures in the chemotherapy are due to the invasion and metastasis of cancers related to drug resistance [202]. Another review on breast cancer shows that 20% to 30% of HR^+ breast cancer cases resist to endocrine therapy. In case of HER-2^+ breast cancer, de novo resistance occurred in approximately 65% of patients, and about 70% of patients with disease that initially respond will ultimately develop acquired resistance [203]. The known mechanisms of drug resistance include mutations in the drug target, drug inactivation, epigenetic modifications, enhanced drug efflux, DNA damage repair, inhibition of cell death, epithelial-mesenchymal transition, aberrated activation of bypass pathways and abnormal downstream pathways [202], [204], [205].

It is well established that resistance may develop subsequent to the application of antineoplastic agents to the patient. Here, we wondered whether the genomes of untreated tumor patients are “primed” in some way to develop such forms of resistance. Such data is conveniently available at the TCGA portal where data was primarily collected prior to treatment. As a reference set of resistance genes, we considered genes belonging to four antineoplastic resistance pathways in the KEGG database [144], [206], [207]: EGFR tyrosine kinase inhibitor resistance, endocrine resistance, antifolate resistance, and platinum drug resistance. The epidermal growth factor receptor (EGFR) is a transmembrane receptor that belongs to the family of receptor tyrosine kinases [208]. The activation of EGFR may lead to cancer-cell proliferation and inhibition of apoptosis [209]. Some FDA-approved drugs belong to tyrosine kinase inhibitor category include neratinib, osimertinib and neratinib. The main mechanisms of EGFR tyrosine kinase inhibitor resistance are EGFR mutation (drug target alteration), aberrated activation of bypass pathways, abnormal downstream pathways, impairment of apoptotic pathway, and epithelial-mesenchymal transition [205].

The platinum-based drugs target DNA and induce cellular apoptosis [210]. Cisplatin, the first platinum drug, has a strong effect initially. However, the emergence of resistance is the major limitation of cisplatin-base chemotherapies. Another platinum drug, carboplatin, has a similar mode of action and leads to similar resistance patterns as cisplatin. Carboplatin was developed in order to reduce the dose-limiting toxicity of cisplatin [211]. Galluzzi *et al.* summarized four mechanisms of cisplatin resistance: pre-target resistance (e.g. increase of efflux), on-target resistance (cisplatin-resistant cells become able to tolerate unrepaired DNA lesions, or can repair adducts at an increased pace), post-target resistance (e.g. tumor cells can overcome apoptosis by defects in the signal transduction pathways), and off-target resistance [212]. To overcome resistance against cisplatin and carboplatin, oxaliplatin was developed. However, oxaliplatin resistance was reported to be accompanied by cellular influx/efflux (solute carrier superfamily of membrane transporters, copper transporter, and ABC transporters). The other resistance mechanisms (DNA adducts repair, inhibition of apoptosis) also affect the sensitivity of oxaliplatin [213].

Folate plays an important role in nucleotide biosynthesis and biological methylation [214]. Antifolates in cancer treatment interrupt the intracellular folate metabolism resulting in ineffective DNA synthesis [215], [216]. Reported mechanisms of antifolate resistance include: increased expression and mutation of target enzymes, impaired antifolate uptake, increased antifolate efflux, defective antifolate polyglutamylation, and the augmentation of cellular tetrahydrofolate-cofactor pools in cells [217], [218].

Estrogens are vital for regulating the growth and differentiation of normal, premalignant and malignant cell types through interaction with two nuclear estrogen receptors (ERalpha and ERbeta) [219], Consequently, these receptors became targets of endocrine therapies (e.g. tamoxifen) [220]. Tamoxifen is the most successful to date [221]. However, both de novo resistance and acquired resistance were observed in breast cancer patients [222]. Mechanisms of endocrine resistance include loss or modification of ER expression, epigenetics mechanisms regulating ER expression, regulation of signal transduction pathways, altered expression of coactivators or co-regulators that play a critical role in ER-mediated gene transcription, altered expression of specific microRNAs [222].

In the genomes of tumor patients, considerably more genes are affected by copy number deletions than by amplifications. This is true for the group of tumor suppressor genes, but also for general classes of genes such as housekeeping genes. In chapter 4, we analyzed how CNAs detected in the patient genomes of 31 different tumor types affect the protein targets of antineoplastic agents

[223]. We found that CNA deletions more frequently affected the targets of antineoplastic agents than CNA amplifications. In seven cancer types, we observed signs of compensatory CNAs. For example, in glioblastoma multiforme, two target genes (*FLT1, FLT3*) of the experimental drug sorafenib were recurrently deleted whereas another target (KDR) of sorafenib was recurrently amplified. In renal clear cell carcinoma, the target FLT1 of pazopanib, sunitinib, sorafenib, and axitinib was recurrently deleted, whereas FLT4 bound by the same drugs was recurrently amplified.

Here, we analyzed the same data set to identify CNAs in known resistance pathways. We found that the number of genes in all four-resistance pathways affected by CNA amplification in tumor tissues is greater than in normal tissues. In contrast, there was no significant difference between normal and tumor tissues with respect to CNA deletions.

5.2 Material and methods

5.2.1 Data on copy number alterations

As mentioned, we analyzed genomic data from the TCGA project on CNAs observed in patients suffering from 31 different forms of tumor. Missing from this list are the data for lung adenocarcinoma and skin cutaneous melanoma as these could not be processed with the GISTIC2.0 tool (see below). The CNA data of these patients (start and end position, chromosome, and segment mean of CNA) were downloaded from the Genomic Data Commons Portal (GDC portal) on September 29, 2017 [182].

5.2.2 KEGG pathways for antineoplastic resistance

From KEGG pathway (https://www.genome.jp/kegg/pathway.html), we retrieved the gene names of four antineoplastic drug resistance pathways. Table 5.1 shows the pathway ID, the name of the pathways and the number of involved genes.

Table 5.1 Antineoplastic drug resistance pathways taken from the KEGG database

Pathway ID	Pathway name	Number of Genes
hsa01521	EGFR tyrosine kinase inhibitor resistance	79
hsa01522	Endocrine resistance	96
hsa01523	Antifolate resistance	31
has01524	Platinum drug resistance	73

For comparison, we also retrieved the genes names of 28 unrelated KEGG pathways having a similar number of genes as the four resistance pathways. These pathways are listed in Supplement table 11.

5.2.3 Clinical data

From the clinical data provided at GDC, we extracted information on which drug treatment was given to specific patients. These lists were then intersected with information form https://www.cancer.gov/about-cancer/treatment/drugs/cancer-type to exclude those drugs that are not approved for specific cancer types.

5.2.4 Antineoplastic agents and their targets

A list of 477 antineoplastic (AN) agents together with their target proteins was extracted from Drugbank [183] (version 5.0.11, downloaded on January 12, 2018). We considered only those protein targets for which pharmacological action of the respective drug molecules is reported as "yes" in Drugbank. These 477 AN agents are reported to bind to 220 different protein targets (labeled here by their Uniprot accessions numbers). After converting Uniprot accession numbers to gene symbols, we were left with 218 genes. As shown in the previous section, we only considered FDA-approved drugs, and therefore we only focus on the target of these drugs.

5.2.5 Genes affected by copy number alterations

Genes that are recurrently affected by CNAs were identified with the GISTIC2.0 tool version 2.0.22 [112] using segmentation files and marker files created from the CNA data of the tumor samples. Following Laddha *et al.* [185], we used 0.2 and -0.2 as thresholds for GISTIC2.0 to identify recurrent amplification and deletion peaks and the genes contained in those peaks. Uniprot accession numbers used by Drugbank were converted to gene symbols used by GISTIC2.0 by making use of data from the HUGO Gene Nomenclature Committee (HGNC database) [186] that were downloaded in January 2018.

5.3 Results

5.3.1 Copy number alterations affect antineoplastic drug resistance pathways

Figure 5.1 summarizes the main steps of our analysis.

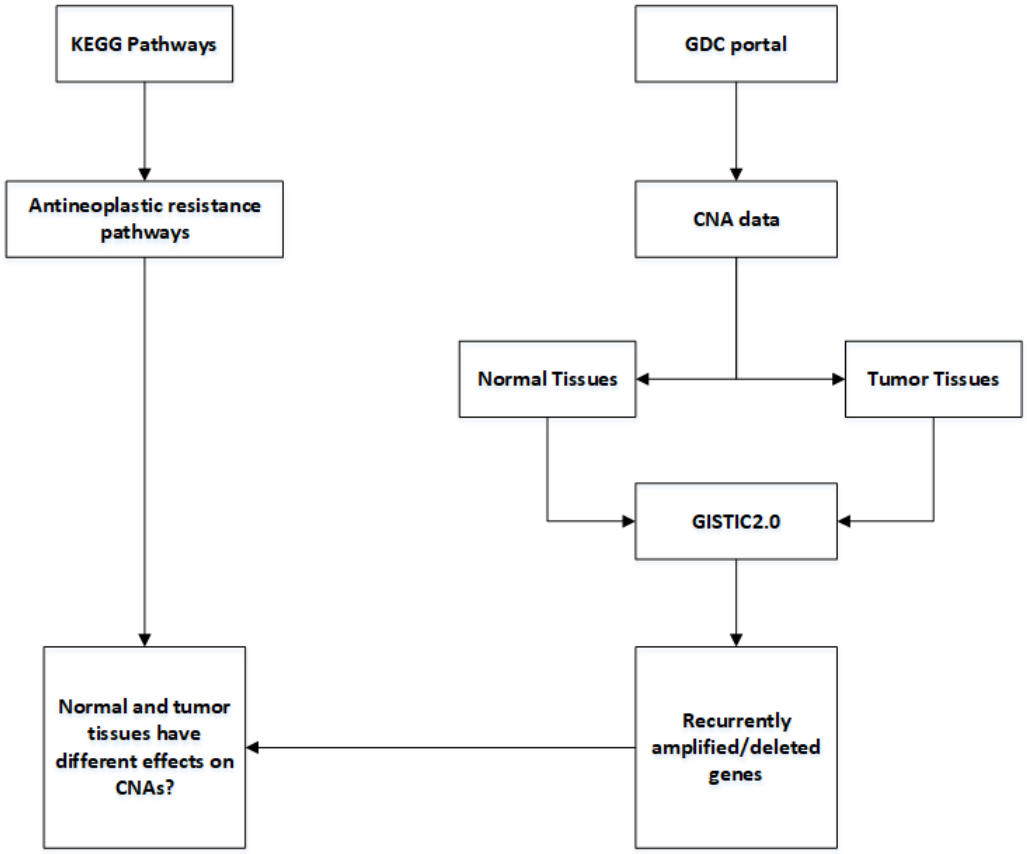

Figure 5.1 Workflow for analyzing effects of CNAs on genes in resistance pathways

Table 5.2 - Table 5.5 summarize the number of genes in each resistance pathway that are recurrently affected by CNAs in normal tissues and tumor tissues.

Table 5.2 Number of recurrently amplified resistance genes in normal tissues

Disease	hsa01521	hsa01522	hsa01523	hsa01524
Uterine Corpus Endometrial Carcinoma	1	0	0	0
Thyroid Carcinoma	4	3	2	3
Prostate Adenocarcinoma	6	3	1	1
Renal Papillary Cell Carcinoma	4	4	1	2
Cholangiocarcinoma	1	1	0	2

Table 5.3 Number of recurrently amplified resistance genes in tumor tissues

Disease	hsa01521	hsa01522	hsa01523	hsa01524
Breast Invasive Carcinoma	3	4	1	3
Glioblastoma Multiforme	6	3	0	1
Ovarian Serous Cystadenocarcinoma	4	3	0	3
Uterine Corpus Endometrial Carcinoma	5	3	0	3
Renal Clear Cell Carcinoma	14	13	0	8

Head and Neck Squamous Cell Carcinoma	5	5	1	4
Brain Lower Grade Glioma	5	6	1	2
Lung Squamous Cell Carcinoma	6	7	2	13
Prostate Adenocarcinoma	3	2	0	3
Colon Adenocarcinoma	1	1	2	1
Stomach Adenocarcinoma	7	5	1	2
Bladder Urothelial Carcinoma	10	5	2	3
Liver Hepatocellular Carcinoma	5	2	0	2
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	8	8	1	7
Renal Papillary Cell Carcinoma	2	2	0	1
Sarcoma	12	10	3	11
Esophageal Carcinoma	8	7	1	8
Pancreatic Adenocarcinoma	4	3	1	2
Rectum Adenocarcinoma	5	6	2	2
Testicular Germ Cell Tumors	8	8	2	12
Adrenocortical Carcinoma	2	4	0	2
Uveal Melanoma	1	2	1	6
Uterine Carcinosarcoma	3	4	1	3
Cholangiocarcinoma	0	1	0	0

Table 5.4 Number of recurrently deleted resistance genes in normal tissues

Disease	hsa01521	hsa01522	hsa01523	hsa01524
Breast Invasive Carcinoma	43	43	13	34
Glioblastoma Multiforme	18	22	6	26
Ovarian Serous Cystadenocarcinoma	8	11	1	11
Uterine Corpus Endometrial Carcinoma	5	4	4	6
Renal Clear Cell Carcinoma	32	34	15	30
Head and Neck Squamous Cell Carcinoma	42	52	13	46
Brain Lower Grade Glioma	35	42	11	33
Thyroid Carcinoma	14	5	6	7
Lung Squamous Cell Carcinoma	34	30	11	17
Prostate Adenocarcinoma	37	31	15	27
Colon Adenocarcinoma	29	34	12	33
Stomach Adenocarcinoma	32	25	15	30
Liver Hepatocellular Carcinoma	6	11	2	12
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	4	4	4	6
Esophageal Carcinoma	6	5	6	10
Pancreatic Adenocarcinoma	6	6	4	5
Pheochromocytoma and Paraganglioma	10	8	3	12
Thymoma	14	4	3	3
Adrenocortical Carcinoma	26	19	14	20
Mesothelioma	15	17	5	15
Cholangiocarcinoma	2	0	0	0

Table 5.5 Number of recurrently deleted resistance genes in tumor tissues

Disease	hsa01521	hsa01522	hsa01523	hsa01524
Breast Invasive Carcinoma	8	18	6	15
Glioblastoma Multiforme	5	7	1	7
Ovarian Serous Cystadenocarcinoma	9	10	4	9
Uterine Corpus Endometrial Carcinoma	21	35	11	34
Renal Clear Cell Carcinoma	19	24	5	18
Head and Neck Squamous Cell Carcinoma	8	15	4	9
Brain Lower Grade Glioma	14	23	5	7
Lung Squamous Cell Carcinoma	12	19	3	15
Prostate Adenocarcinoma	7	6	3	10
Colon Adenocarcinoma	11	8	2	13
Stomach Adenocarcinoma	11	20	7	9
Bladder Urothelial Carcinoma	8	12	2	14
Liver Hepatocellular Carcinoma	13	17	5	15
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	13	21	3	11
Renal Papillary Cell Carcinoma	18	24	7	24
Sarcoma	23	33	8	21
Acute Myeloid Leukemia	10	15	6	11
Esophageal Carcinoma	25	34	9	25
Pancreatic Adenocarcinoma	23	34	4	24
Pheochromocytoma and Paraganglioma	16	25	6	22
Rectum Adenocarcinoma	19	26	7	20
Testicular Germ Cell Tumors	8	11	1	11
Thymoma	7	13	1	5
Adrenocortical Carcinoma	17	17	5	19
Mesothelioma	16	19	8	24
Uveal Melanoma	10	10	5	15
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	23	30	7	23
Uterine Carcinosarcoma	17	26	5	12
Cholangiocarcinoma	6	11	3	14

In the diseases that are not listed in Table 5.2 - Table 5.5, none of the genes in the four considered resistance pathways was affected by CNAs.

Then, we applied the Wilcoxon test to check whether the genes belonging to the four resistance pathways are comparably often affected by CNAs in normal tissues and in tumor tissues or not, see Table 5.6. In fact, for all resistance pathways, significantly more resistance genes were subject to CNA amplifications in tumor genomes than in normal genomes. On the other hand, CNA

deletions had similar effects on resistance genes in both tissues. As previously noticed [223], the number of CNA deletions generally exceeds the number of CNA amplifications.

For comparison, we repeat the same analysis for 28 unrelated KEGG pathways of similar size. Apparently, very similar finding apply to these pathways as for resistance pathways: significantly more amplifications in tumor tissues than in normal tissues, essentially no difference for CNA deletions (Supplement table 12).

Table 5.6 Adjusted P-values of Wilcoxon test

CNA type	hsa01251	hsa01522	hsa01523	hsa01524
Amplification	$1{,}20 \times 10^{-5}$	$01{,}56 \times 10^{-6}$	$3{,}74 \times 10^{-3}$	$3{,}83 \times 10^{-6}$
Deletion	1.000	0.105	1.000	0.499

5.3.2 Copy number alterations affect antineoplastic targets

Next, we investigated the effect of CNA amplifications on the protein targets of the four drug categories. The workflow for this analysis is shown in Figure 5.2.

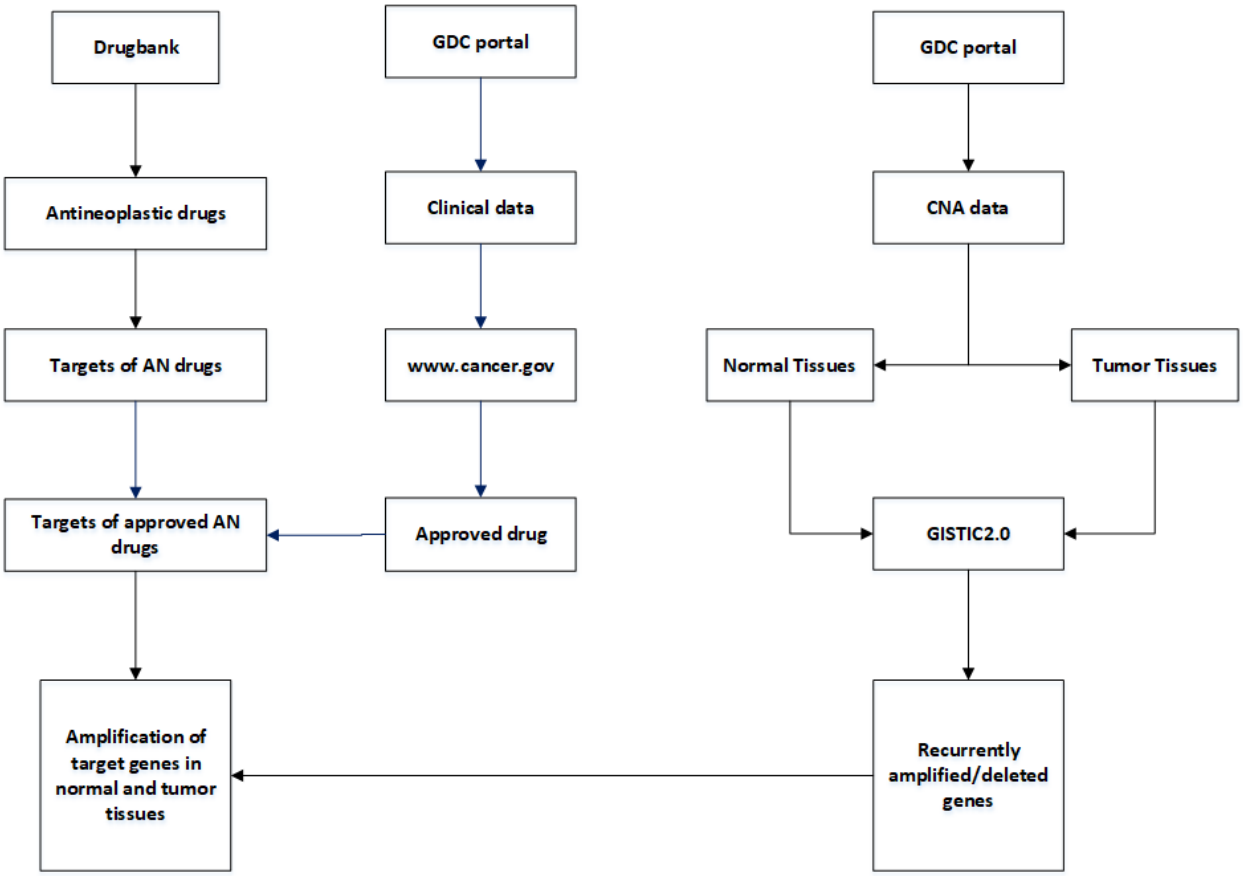

Figure 5.2 Workflow for analyzing effects of CNAs on targets of antineoplastic

Table 5.7 lists the number of drugs that were given to patients in each cancer type, the number of FDA approved drugs, and the subset of approved drugs belonging to the four resistance categories. The relatively small size of these subsets reflects the broad spectrum of drug targets. Table 5.8 shows the number of protein targets of each drug-resistance category. Table 5.9 and Table 5.10 show the number of targets of FDA-approved drugs affected by CNA amplifications in specific cancer types.

Table 5.7 Number of drugs for each cancer type

Disease	All Drugs	Approved Drugs	Endocrine	Folic Acid	Platinum	Tyrosine Kinase Inhibitor
Breast Invasive Carcinoma	38	23	7	1	0	1
Glioblastoma Multiforme	37	4	0	0	0	0
Ovarian Serous Cystadenocarcinoma	31	9	0	0	2	0
Lung Adenocarcinoma	16	9	0	1	1	2
Uterine Corpus Endometrial Carcinoma	16	1	0	0	0	0
Renal Clear Cell Carcinoma	17	8	0	0	0	4
Head and Neck Squamous Cell Carcinoma	18	2	0	1	0	0
Brain Lower Grade Glioma	24	4	0	0	0	0
Thyroid Carcinoma	1	1	0	0	0	0
Lung Squamous Cell Carcinoma	16	8	0	1	1	2
Prostate Adenocarcinoma	11	8	6	0	0	0
Skin Cutaneous Melanoma	32	9	0	0	0	0
Colon Adenocarcinoma	15	9	0	0	1	1
Stomach Adenocarcinoma	22	4	0	0	0	0
Bladder Urothelial Carcinoma	20	4	0	0	1	0
Liver Hepatocellular Carcinoma	12	2	0	0	0	2
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	12	3	0	0	0	0
Renal Papillary Cell Carcinoma	14	6	0	0	0	3
Sarcoma	23	4	0	0	0	1
Acute Myeloid Leukemia	0	0	0	0	0	0
Esophageal Carcinoma	11	1	0	0	0	0
Pancreatic Adenocarcinoma	15	5	0	0	0	1
Pheochromocytoma and Paraganglioma	6	0	0	0	0	0
Rectum Adenocarcinoma	12	7	0	0	1	0
Testicular Germ Cell Tumors	5	4	0	0	1	0
Thymoma	8	0	0	0	0	0
Kidney Chromophobe	5	4	0	0	0	2
Adrenocortical Carcinoma	10	0	0	0	0	0
Mesothelioma	16	1	0	1	0	0
Uveal Melanoma	1	0	0	0	0	0
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	23	13	0	1	0	0
Uterine Carcinosarcoma	10	0	0	0	0	0
Cholangiocarcinoma	3	0	0	0	0	0

Table 5.8 Number of target genes in each group of approved drugs

Disease	Targets of Endocrine	Targets of Folic Acid	Targets of Platinum	Targets of Tyrosine Kinase Inhibitor
Breast Invasive Carcinoma	5	1	0	2
Lung Adenocarcinoma	0	4	0	2
Renal Clear Cell Carcinoma	0	0	0	12
Head and Neck Squamous Cell Carcinoma	0	1	0	0
Lung Squamous Cell Carcinoma	0	4	0	2
Prostate Adenocarcinoma	4	0	0	0
Colon Adenocarcinoma	0	0	0	18
Liver Hepatocellular Carcinoma	0	0	0	19
Renal Papillary Cell Carcinoma	0	0	0	12
Sarcoma	0	0	0	6
Pancreatic Adenocarcinoma	0	0	0	2
Kidney Chromophobe	0	0	0	12
Mesothelioma	0	4	0	0
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	0	1	0	0

Table 5.9 Number of AN targets amplified by CNAs in normal tissues

Disease	Targets of Endocrine	Targets of FolicAcid	Targets of Platinum	Targets of Tyrosine Kinase
Renal Papillary Cell Carcinoma	0	0	0	3

Table 5.10 Number of AN targets amplified by CNAs in tumor tissues

Disease	Targets of Endocrine	Targets of Folic Acid	Targets of Platinum	Targets of Tyrosine Kinase Inhibitor
Breast Invasive Carcinoma	0	0	0	1
Renal Clear Cell Carcinoma	0	0	0	2
Lung Squamous Cell Carcinoma	0	1	0	1
Renal Papillary Cell Carcinoma	0	0	0	1
Sarcoma	0	0	0	4

The diseases that are not listed in Table 5.8 - Table 5.10 have no AN targets affected by CNAs.

5.4 Discussion

In this project, CNA and clinical data for 31 types of tumors from the TCGA project were combined with information on AN drugs from Drugbank. As shown in Table 5.6 the difference between normal and tumor tissues is significant for CNA amplification but not for CNA deletions. With the target genes of four drug categories, only three target genes of EGFR tyrosine kinase inhibitors were affected by CNA amplifications in normal tissues of renal papillary cell carcinoma (Table 5.9). In case of tumor tissues, the target genes of five diseases were affected by CNA amplifications (Table 5.10).

The relation between gene amplification and drug resistance was mentioned decades ago. In 1984, Robert T. Schimke reported the relation of MTX resulting from the amplification of DHFR gene [224]. In 1991, P. Borst and R. Brown published their reviews on drug resistance and gene amplification. In both reviews, the authors pointed out the role of amplification of multi-drug resistance genes in cancer [225], [226]. Many studies since then focusing on drug resistance verified that gene amplifications influence drug resistance [202], [227]–[232]. Our study again confirms that the amplification of genes belonging to known resistance pathways in tumor tissues support the ability of drug resistance.

In acute myeloid leukemia, 38 of 105 cases (26.57%) received treatment prior to the time when the CNA data were taken. For glioblastoma (22 of 590 cases) and renal clear cell carcinoma (18 of 530 cases), the number of such cases was around 4%. In all other tumors, the fraction of pre-treated patients was below 3 %. Hence, in all tumors except for acute myeloid leukemia, the detected amplifications and deletions are unlikely to reflect resistance phenomena occurring in response to treatment (Supplement table 13). However, our result showed that, there are more amplified genes in resistance pathways in tumor tissues than in normal tissues, this suggest that when tumor cells developed, they also gained the ability of drug resistance. This type of resistance is not intrinsic resistance (pre-existent) neither acquired resistance (induced by drugs) (these concepts used by Theodor H Lippert and colleges [233]).

Supplement table 14 - Supplement table 17 show number of cancer types that affect each genes in four considered resistance pathways. The first two columns of these tables show that: at most, only one cancer type in which CNA amplifications of normal tissues affect each genes. While the number of cancer types in which CNA amplifications of tumor tissues affect genes is significantly higher (columns 5 and 6). Some genes commonly affected by CNA amplifications in tumor tissues includes: IGF1R (hsa01521, hsa01522) affected in eleven cancer types; EGFR

(hsa01521, hsa01522) affected in nine cancer types; PIK3CA (hsa01521, hsa01522, hsa01524) affected in seven cancer types; KRAS (hsa01521, hsa01522) affected in seven cancer types; GAS6 (hsa01521), VEGFA (hsa01521) affected in seven cancer types; IKBKB (hsa01523) affected in six cancer types; FASLG (hsa01524), AKT3 (hsa01524), and POLH (hsa01524) affected in five cancer types; TYMS (hsa01523), GGH (hsa01523) affected in three cancer types. Insulin like growth factor 1 receptor (IGF1R) belongs to the family of transmembrane tyrosine kinase receptors [234]. IGF1R play an important role in tyrosine kinase inhibitor resistance by aberrated activation of bypass pathway [235], [236]. Another frequently affected gene is epidermal growth factor receptor (EGFR). Like IGF1R, this gene also belongs to the family of receptor tyrosine kinases [208]. Amplifications of *EGFR* gene copy numbers and overexpression of EGFR are known to be one of the most common alterations in non-small-cell lung carcinoma (NSCLC) cells [187]–[190] and are associated with a poor prognosis and chemoresistance. Belong to three of four pathways (except hsa10523), phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (PIK3CA) and PI3K-Akt signaling pathways can control key cellular processes involved in apoptosis, protein synthesis, metabolism, and cell cycle [237]–[240].The activation of this pathway also play a key role in drug resistance [241]–[244].

It is known that CNA is an mechanism for acquired resistance of chemotherapy [228], [229], however with the available data from TCGA we cannot argue about acquired resistance (CNA profile in respond to chemotherapy treatment). Integrated analysis of gene expression and CNVs/CNAs give promising results [245]–[247]. In future, we may apply this approach for better understand the mechanisms of drug resistance in cancer treatment.

5.5 Conclusion

To better understand cancer drug resistance, a big challenge in cancer treatment, in this work we provided a landscape of copy number alternations effects on four antineoplastic resistance pathways across 31 cancers. Based on CNAs data from the TCGA repository, we found that the genome sequences of tumor tissues contain more recurrently amplified CNAs of genes in antineoplastic resistance pathways than normal tissues. Not only the genes in the four resistance pathways, the targets of FDA-approved drugs also affected by tumor tissues more than normal tissues (Table 5.9 and Table 5.10). This supports an important mechanism of drug resistance: amplification of drug targets. We found out that some genes (e.g. PIK3CA, EGFR, and IGF1R that play important role in drug resistance) affected by circa 22% to 35% cancer types (Supplement table 14 - Supplement table 17). In ongoing work, we are extending our analysis by combining gene expression and CNAs data. Because the genes only function when their corresponding proteins exist, by analyzing

expression data, we may have more evidence about the effect of CNAs amplification on drug resistance.

Chapter 6 Conclusions and outlook

Presented in this thesis are three projects where we analyzed different sorts of genomic data that are related to transmembrane proteins, and genomic copy number alterations. We aimed to predict substrates which are transported by transmembrane proteins. We also investigated the effects caused by copy number alterations on the target protein of antineoplastic agents, and on the genes in antineoplastic resistance pathways in cancer patients.

In the first project, we proposed a computational method to classify membrane transporters from three organisms (*Escherichia coli*, *Saccharomyces cerevisiae* and *Homo sapiens*) according to their transported substrates. Promoted by the idea of operon, our method focuses of neighboring genes that show high co-expression with the query gene. Then, we identified frequent GO terms among these co-expressed neighbors and used a support vector machine classifier to annotate the substrate specificity of the query gene. For transporter proteins from *Escherichia coli, Saccharomyces cerevisiae* and *Homo sapiens*, the average accuracies were 89%, 78%, and 79%, respectively. When tested on the genes belonging to different metabolic pathways of *Escherichia coli*, the average accuracy was 75% (two classes) and 67% (four classes). This suggests that transfer of functional associations between co-expressed neighboring genes may be case-specific. In future works, this approach may be used in combination with other features such as sequence motifs, sequence similarity, and further characteristics of the protein sequence such as its amino acid composition.

The second project aimed at testing the hypothesis that the protein targets of AN agents in tumors are affected by genomic copy number alternations (CNAs) more strongly than expected by chance. By analyzing CNAs and clinical data of 31 tumor types from TCGA, we found that the genome sequences of tumor patients generally contain more recurrently deleted CNAs than recurrently amplified CNAs. This is also the case for CNAs affecting target genes of the specific AN for these tumors. We observed certain signs of apparently compensating effects of CNAs. For example, in glioblastoma multiforme, two target genes (*FLT1*, *FLT3*) of the experimental drug sorafenib were recurrently deleted whereas another target (*KDR*) of sorafenib was recurrently amplified. In renal clear cell carcinoma, the target *FLT1* of pazopanib, sunitinib, sorafenib, and axitinib was recurrently deleted, whereas *FLT4* bound by the same drugs was recurrently amplified. The data available for this study enabled us to identify CNA alterations that existed prior to therapy and that may render certain chemotherapies more or less effective. In future, it would be desirable to also collect time-series CNA data of tumor patients at time of diagnosis and at later time points.

This would point to CNA alterations caused by the application of certain chemotherapies and thus reflect chemoresistance.

The third project continues the idea of chemoresistance as suggested in the second one. We still used CNAs data from the TCGA repository, but not only data from tumor tissues like in the second project. In the third project, we utilized CNAs data from both normal and tumor tissues. We found that the genome sequences of tumor tissues contain more recurrently amplified CNAs of genes in antineoplastic resistance pathways than normal tissues. AN targets of FDA-approved drugs were amplified in normal tissues of only one cancer type (Table 5.9) while they were amplified in tumor tissues of five cancer types (Table 5.10). This is in support for an important mechanism of drug resistance: amplification of drug targets. We also found out that some genes (e.g. PIK3CA, EGFR, and IGF1R) play important roles in drug resistance and were affected by circa 22% to 35% cancer types. In future work, this analysis may be extended by combining gene expression data and CNA data. The genes only function when they are expressed. Hence, by analyzing expression data, we may have more evidence about the effect of CNAs amplification on drug resistance.

References

[1] R. Dahm, “Friedrich Miescher and the discovery of DNA,” *Dev. Biol.*, vol. 278, no. 2, pp. 274–288, 2005.

[2] H. H. Goldstine, “A Brief History of the Computer,” *Proc. Am. Philos. Soc.*, vol. 121, no. 5, pp. 339–345, 1977.

[3] J. D. WATSON and F. H. C. CRICK, “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid,” *Nature*, vol. 171, p. 737, Apr. 1953.

[4] F. Sanger and A. R. Coulson, “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase,” *J. Mol. Biol.*, vol. 94, no. 3, pp. 441–448, 1975.

[5] H. Garland, “Design Innovations in Personal Computers,” *Computer (Long. Beach. Calif).*, vol. 10, pp. 24–27, 1977.

[6] P. Muir *et al.*, “The real cost of sequencing: scaling computation to keep pace with data generation,” *Genome Biol.*, vol. 17, p. 53, Mar. 2016.

[7] U. I. Schwarz, M. Gulilat, and R. B. Kim, “The Role of Next-Generation Sequencing in Pharmacogenetics and Pharmacogenomics,” *Cold Spring Harb. Perspect. Med.*, p. a033027, 2018.

[8] P. Paneth and A. Dybala-Defratyka, *Kinetics and Dynamics: From Nano- to Bio-Scale*. Springer Netherlands, 2010.

[9] A. Krogh, B. Larsson, G. von Heijne, and E. L. . Sonnhammer, “Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes,” *J. Mol. Biol.*, vol. 305, pp. 567–580, 2001.

[10] A. César-Razquin *et al.*, “A Call for Systematic Research on Solute Carriers,” *Cell*, vol. 162, no. 3, pp. 478–487, Jul. 2015.

[11] A. Barghash and V. Helms, “Transferring functional annotations of membrane transporters on the basis of sequence similarity and sequence motifs,” *BMC Bioinformatics*, vol. 14, no. 1, p. 343, 2013.

[12] M. M. Gromiha and Y. Yabuki, “Functional discrimination of membrane proteins using machine learning techniques,” *BMC Bioinformatics*, vol. 9, p. 135, Mar. 2008.

[13] N. S. Schaadt, J. Christoph, and V. Helms, “Classifying Substrate Specificities of Membrane Transporters from Arabidopsis thaliana,” *J. Chem. Inf. Model.*, vol. 50, no. 10, pp. 1899–1905, Oct. 2010.

[14] N. S. Schaadt and V. Helms, “Functional classification of membrane transporters and channels based on filtered TM/non-TM amino acid composition,” *Biopolymers*, vol. 97, no. 7, pp. 558–567, 2012.

[15] Y. Hu, Y. Guo, Y. Shi, M. Li, and X. Pu, “A consensus subunit-specific model for annotation of substrate specificity for ABC transporters,” *RSC Adv.*, vol. 5, no. 52, pp. 42009–42019, 2015.

[16] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, “How many drug targets are there?,” *Nat. Rev. Drug Disc*, vol. 5, pp. 993–996, 2006.

[17] M. M. Gottesman, “Mechanisms of Cancer Drug Resistance,” *Annu. Rev. Med.*, vol. 53, no. 1, pp. 615–627, Feb. 2002.

[18] L. Pray, “Discovery of DNA structure and function: Watson and Crick,” *Nat. Educ.*, vol. 1(1):100, 2008.

[19] S. Anderson *et al.*, “Sequence and organization of the human mitochondrial genome,” *Nature*, vol. 290, p. 457, Apr. 1981.

[20] J. Cairns, “The Chromosome of Escherichia coli,” *Cold Spring Harb. Symp. Quant. Biol.* , vol. 28, pp. 43–46, Jan. 1963.

[21] D. J. Mason and D. M. Powelson, “NUCLEAR DIVISION AS OBSERVED IN LIVE BACTERIA BY A NEW TECHNIQUE,” *J. Bacteriol.*, vol. 71, no. 4, pp. 474–479, Apr. 1956.

[22] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, “Crystal structure of the nucleosome core particle at 2.8 Å resolution,” *Nature*, vol. 389, p. 251, Sep. 1997.

[23] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, “Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution††We dedicate this paper to the memory of Max Perutz who was particularly inspirational and supportive to T.J.R. in the early stages of this study.,” *J. Mol. Biol.*, vol. 319, no. 5, pp. 1097–1113, 2002.

[24] C. L. Woodcock, “A milestone in the odyssey of higher-order chromatin structure,” *Nat. Struct. &Amp; Mol. Biol.*, vol. 12, p. 639, Aug. 2005.

[25] A. P. Wolffe and D. Guschin, “Review: Chromatin Structural Features and Targets That Regulate Transcription,” *J. Struct. Biol.*, vol. 129, no. 2, pp. 102–122, 2000.

[26] A. Griswold, “Genome Packaging in Prokaryotes: the Circular Chromosome of E. coli,” *Nat. Educ.*, vol. 1(1):57, 2008.

[27] M. C. Negritto, “Repairing Double-Strand DNA Breaks,” *Nat. Educ.*, vol. 3(9):26, 2010.

[28] A. Aguilera and B. Gómez-González, “Genome instability: a mechanistic view of its causes and consequences,” *Nat. Rev. Genet.*, vol. 9, p. 204, Mar. 2008.

[29] P. Huertas, “DNA resection in eukaryotes: deciding how to fix the break,” *Nat. Struct. Mol. Biol.*, vol. 17, no. 1, pp. 11–16, Jan. 2010.

[30] P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira, “Mechanisms of change in gene copy number,” *Nat Rev Genet*, vol. 10, 2009.

[31] A. Thapar and M. Cooper, “Copy Number Variation: What Is It and What Has It Told Us About Child Psychiatric Disorders?,” *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 52, no. 8, pp. 772–774, Aug. 2013.

[32] W. Li, A. Lee, and P. K. Gregersen, “Copy-number-variation and copy-number-alteration region detection by cumulative plots,” *BMC Bioinformatics*, vol. 10, no. 1, p. S67, 2009.

[33] R. Redon *et al.*, “Global variation in copy number in the human genome,” *Nature*, vol. 444, no. 7118, pp. 444–454, Nov. 2006.

[34] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, “Copy Number Variation in Human Health, Disease, and Evolution,” *Annu. Rev. Genomics Hum. Genet.*, vol. 10, pp. 451–481, 2009.

[35] S. Volik *et al.*, “Decoding the fine-scale structure of a breast cancer genome and transcriptome,” *Genome Res.*, vol. 16, no. 3, pp. 394–404, Mar. 2006.

[36] J. M. Korn *et al.*, “Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs,” *Nat. Genet.*, vol. 40, p. 1253, Sep. 2008.

[37] V. Seshan and A. Olshen, “DNAcopy: DNA copy number data analysis.” 2018.

[38] J. Guo, “Transcription: the epicenter of gene expression,” *J. Zhejiang Univ. Sci. B*, vol. 15, no. 5, pp. 409–411, May 2014.

[39] H. Lodish, A. Berk, P. Matsudaira, and C. A. Kaiser, *Molecular Cell Biology 5th Edition, Modern Genetic Analysis 2nd Edition & Cd-rom*. Macmillan Higher Education, 2004.

[40] V. Mathura and P. Kangueane, *Bioinformatics: A Concept-Based Introduction*. Springer US, 2008.

[41] W. A. Whyte *et al.*, “Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes,” *Cell*, vol. 153, no. 2, pp. 307–319, 2013.

[42] S. Clancy, “DNA Transcription,” *Nat. Educ.*, vol. 1(1):41, 2008.

[43] S. Clancy, “Translation: DNA to mRNA to Protein,” *Nat. Educ.*, vol. 1(1):101, 2008.

[44] T. E. Dever, “A New Start for Protein Synthesis,” *Science (80-.).*, vol. 336, no. 6089, p. 1645 LP-1646, Jun. 2012.

[45] F. H. C. CRICK, L. BARNETT, S. BRENNER, and R. J. WATTS-TOBIN, “General Nature of the Genetic Code for Proteins,” *Nature*, vol. 192, p. 1227, Dec. 1961.

[46] S. L. Berg JM, Tymoczko JL, “Biochemistry. 5th edition. Chapter 3, Protein Structure and Function,” *New York: W H Freeman*, 2002. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK21177/.

[47] M. B. Miller and Y.-W. Tang, “Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology,” *Clin. Microbiol. Rev.*, vol. 22, no. 4, pp. 611–633, Oct. 2009.

[48] R. Bumgarner, “DNA microarrays: Types, Applications and their future,” *Curr. Protoc. Mol. Biol.*, vol. 0 22, p. Unit-22.1., Jan. 2013.

[49] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, “Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells,” *PLoS One*, vol. 9, no. 1, p. e78644, Jan. 2014.

[50] F. Jacob, D. Perrin, C. Sánchez, and J. Monod, “L’opéron : groupe de gènes à expression coordonnée par un opérateur [C. R. Acad. Sci. Paris 250 (1960) 1727–1729],” *C. R. Biol.*, vol. 328, no. 6, pp. 514–520, 2005.

[51] A. E. Osbourn and B. Field, “Operons,” *Cell. Mol. Life Sci. *, vol. 66, no. 23, pp. 3755–3775, Dec. 2009.

[52] A. Ralston, “Operons and Prokaryotic Gene Regulation,” *Nat. Educ.*, vol. 1(1):216, 2008.

[53] M. Ashburner *et al.*, “Gene Ontology: tool for the unification of biology,” vol. 25, no. 1. pp. 25–29, May-2000.

[54] C. A. Ball *et al.*, “Integrating functional genomic information into the Saccharomyces Genome Database,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 77–80, Jan. 2000.

[55] “The FlyBase database of the Drosophila Genome Projects and community literature. The FlyBase Consortium.,” *Nucleic Acids Res.*, vol. 27, no. 1, pp. 85–88, Jan. 1999.

[56] J. A. Blake, J. T. Eppig, J. E. Richardson, M. T. Davisson, and the Mouse Genome Database Group, “The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 108–111, Jan. 2000.

[57] M. Ringwald, J. T. Eppig, J. A. Kadin, J. E. Richardson, and the Gene Expression Database Group, “GXD: a Gene Expression Database for the laboratory mouse: current status and recent enhancements,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 115–119, Jan. 2000.

[58] B. Smith, J. Williams, and S.-K. Steffen, “The Ontology of the Gene Ontology,” *AMIA Annu. Symp. Proc.*, vol. 2003, pp. 609–613, 2003.

[59] N. Skunca, A. Altenhoff, and C. Dessimoz, “Quality of computationally inferred gene ontology annotations,” *PLoS Comput. Biol.*, vol. 8, no. 5, pp. e1002533–e1002533, May 2012.

[60] S. Carbon *et al.*, “AmiGO: online access to ontology and annotation data,” *Bioinformatics*, vol. 25, no. 2, pp. 288–289, Jan. 2009.

[61] D. Hanahan and R. A. Weinberg, “The Hallmarks of Cancer,” *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[62] D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: the next generation,” *Cell*, vol. 144, 2011.

[63] A. Suzuki *et al.*, “Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines,” *Nucleic Acids Res.*, vol. 42, no. 22, pp. 13557–13572, Dec. 2014.

[64] “Nomenclature and Symbolism for Amino Acids and Peptides,” *Eur. J. Biochem.*, vol. 138,

no. 1, pp. 9–37, Sep. 2018.

[65] L. Pauling, R. B. Corey, and H. R. Branson, “The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 37, no. 4, pp. 205–211, Apr. 1951.

[66] C. M. Venkatachalam, “Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units,” *Biopolymers*, vol. 6, no. 10, pp. 1425–1436, Sep. 2018.

[67] J. F. Leszczynski and G. D. Rose, “Loops in globular proteins: a novel category of secondary structure,” *Science (80-.).*, vol. 234, no. 4778, p. 849 LP-855, Nov. 1986.

[68] J. S. Fetrow, “Omega loops: nonregular secondary structures significant in protein function and stability.,” *FASEB J.*, vol. 9, no. 9, pp. 708–717, Jun. 1995.

[69] J. W. Pelley, “3 - Protein Structure and Function,” J. W. B. T.-E. I. B. Pelley, Ed. Philadelphia: Mosby, 2007, pp. 19–28.

[70] H. W. Schroeder and L. Cavacini, “Structure and Function of Immunoglobulins,” *J. Allergy Clin. Immunol.*, vol. 125, no. 2 0 2, pp. S41–S52, Feb. 2010.

[71] S. L. Berg JM, Tymoczko JL, “Section 8.3, Enzymes Accelerate Reactions by Facilitating the Formation of the Transition State,” in *Biochemistry. 5th edition*, New York: W H Freeman, 2002.

[72] N. Neave, Ed., “Behavioural endocrinology,” in *Hormones and Behaviour: A Psychological Approach*, Cambridge: Cambridge University Press, 2007, pp. 48–68.

[73] L.-H. Gu and P. A. Coulombe, “Keratin function in skin epithelia: a broadening palette with surprising shades,” *Curr. Opin. Cell Biol.*, vol. 19, no. 1, pp. 13–23, 2007.

[74] T. J. Silhavy, D. Kahne, and S. Walker, “The bacterial cell envelope,” *Cold Spring Harb. Perspect. Biol.*, vol. 2, no. 5, pp. a000414–a000414, May 2010.

[75] L. Burri *et al.*, “Integral membrane proteins in the mitochondrial outer membrane of Saccharomyces cerevisiae,” *FEBS J.*, vol. 273, no. 7, pp. 1507–1515, Apr. 2006.

[76] S. Shao and R. S. Hegde, “Membrane protein insertion at the endoplasmic reticulum,” *Annu. Rev. Cell Dev. Biol.*, vol. 27, pp. 25–56, 2011.

[77] J. Saier Milton H., V. S. Reddy, B. V Tsu, M. S. Ahmed, C. Li, and G. Moreno-Hagelsieb,

"The Transporter Classification Database (TCDB): recent advances," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D372–D379, Jan. 2016.

[78] S. J. Singer and G. L. Nicolson, "The fluid mosaic model of the structure of cell membranes," *Science (80-.).*, vol. 175, pp. 720–731, 1972.

[79] M. K. Ahmad, "Drug Targets for Cancer Treatment: An Overview," *Medicinal Chemistry*, vol. 5, no. 3. OMICS International., 2015.

[80] P. Ananya and B. Santanu, "Chemistry and biology of DNA-binding small molecules," *Curr. Sci.*, vol. 102, no. 2, pp. 212–231, 2012.

[81] M. Kanehisa, "KEGG Pathway Maps," *Kanehisa Laboratories*, 2011. [Online]. Available: https://www.genome.jp/kegg/kegg3a.html.

[82] M. Kanehisa, "KEGG PATHWAY Database," *Kanehisa Laboratories*, 2018. [Online]. Available: https://www.genome.jp/kegg/pathway.html.

[83] M. Kanehisa, "EGFR tyrosine kinase inhibitor resistance - Homo sapiens (human)," *Kanehisa Laboratories*, 2016. [Online]. Available: https://www.genome.jp/kegg-bin/show_pathway?map=hsa01521&show_description=show.

[84] M. Kanehisa, "Endocrine resistance - Homo sapiens (human)," *Kanehisa Laboratories*, 2017. [Online]. Available: https://www.genome.jp/kegg-bin/show_pathway?map=hsa01522&show_description=show.

[85] M. Kanehisa, "Antifolate resistance - Homo sapiens (human)," *Kanehisa Laboratories*, 2016. [Online]. Available: https://www.genome.jp/kegg-bin/show_pathway?map=hsa01523&show_description=show.

[86] M. Kanehisa, "Platinum drug resistance - Homo sapiens (human)," *Kanehisa Laboratories*, 2016. [Online]. Available: https://www.genome.jp/kegg-bin/show_pathway?map=hsa01524&show_description=show.

[87] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, 1959.

[88] J. R. Koza, F. H. I. Bennett, D. Andre, and M. A. Keane, "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming," in *Artificial Intelligence in Design '96*, Springer, Dordrecht, 1996, pp. 151–170.

[89] N. Boyko, P. Mykhailyshyn, and Y. Kryvenchuk, “Use a cluster approach to organize and analyze data inside the cloud,” *ECONTECHMOD An Int. Q. J. Econ. Technol. Model. Process.*, vol. 7, 2018.

[90] R. C. Deo, “Machine Learning in Medicine,” *Circulation*, vol. 132, no. 20, pp. 1920–1930, Nov. 2015.

[91] C. Cortes and V. Vapni, “Support-Vector Networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[92] K.-B. Duan and S. S. Keerthi, “Which Is the Best Multiclass SVM Method? An Empirical Study BT - Multiple Classifier Systems,” 2005, pp. 278–285.

[93] L. Bottou *et al.*, “Comparison of classifier methods: a case study in handwritten digit recognition,” in *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5)*, 1994, vol. 2, pp. 77–82 vol.2.

[94] S. Knerr, L. Personnaz, and G. Dreyfus, “Single-layer learning revisited: a stepwise procedure for building and training a neural network BT - Neurocomputing,” 1990, pp. 41–50.

[95] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, “Large Margin DAGs for Multiclass Classification,” in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, 1999, pp. 547–553.

[96] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*, Fifth. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA., 2005.

[97] R. E. Walpole, R. H. Myers, S. L. Myers, and K. E. Ye, *Probability and Statistics for Engineers and Scientists*. Pearson Education, 2011.

[98] S. S. SHAPIRO and M. B. WILK, “An analysis of variance test for normality (complete samples)†,” *Biometrika*, vol. 52, no. 3–4, pp. 591–611, Dec. 1965.

[99] STUDENT, “THE PROBABLE ERROR OF A MEAN,” *Biometrika*, vol. 6, no. 1, pp. 1–25, Mar. 1908.

[100] A. Ugoni and B. Walker, *THE t TEST: An Introduction*, vol. 4. 1995.

[101] B. L. WELCH, “THE GENERALIZATION OF ‘STUDENT’S’ PROBLEM WHEN

SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED," *Biometrika*, vol. 34, no. 1–2, pp. 28–35, Jan. 1947.

[102] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

[103] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *Ann. Math. Stat.*, vol. 18, no. 1, pp. 50–60, 1947.

[104] J. H. Zar, *Biostatistical Analysis (5th Edition)*. Prentice-Hall, Inc., 2007.

[105] P. Warner, "Testing association with Fisher's Exact test," *J. Fam. Plan. Reprod. Heal. Care*, vol. 39, no. 4, p. 281 LP-284, Oct. 2013.

[106] R. A. Fisher, *The design of experiments. 1935*. Edinburgh: Oliver and Boyd, 1935.

[107] E. W. Weisstein, "Fisher's Exact Test," *MathWorld--A Wolfram Web Resourc*. [Online]. Available: http://mathworld.wolfram.com/FishersExactTest.html.

[108] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995.

[109] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Stat.*, vol. 29, no. 4, pp. 1165–1188, 2001.

[110] E. W. Weisstein, "Euler-Mascheroni Constant," *MathWorld--A Wolfram Web Resource*. [Online]. Available: http://mathworld.wolfram.com/Euler-MascheroniConstant.html.

[111] R. Beroukhim *et al.*, "Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma," *Proc. Natl. Acad. Sci.*, vol. 104, no. 50, p. 20007 LP-20012, Dec. 2007.

[112] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers," *Genome Biol.*, vol. 12, no. 4, p. R41, 2011.

[113] S. Sivashankari and P. Shanmughavel, "Functional annotation of hypothetical proteins – A review," *Bioinformation*, vol. 1, no. 8, pp. 335–338, Dec. 2006.

[114] T. K. Attwood *et al.*, "PRINTS prepares for the new millennium.," *Nucleic Acids Research*, vol. 27, no. 1. pp. 220–225, Jan-1999.

[115] J. Henikoff, G. and S. Henikoff, “Blocks database and its applications,” *Methods Enzym.*, vol. 266, pp. 88–104, 1996.

[116] K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch, “The PROSITE database, its status in 1999,” *Nucleic Acids Res*, vol. 27, pp. 215–219, 1999.

[117] N. J. Mulder *et al.*, “InterPro - An integrated documentation resource for protein families, domains and functional sites,” *Brief. Bioinforma.*, vol. 3, no. 3, pp. 225–235, 2002.

[118] Q. Ren, K. H. Kang, and I. T. Paulsen, “TransportDB: a relational database of cellular membrane transport systems,” *Nucleic Acids Res.*, vol. 32, no. 2, pp. D284–D288, 2004.

[119] R. D. Sleator and P. Walsh, “An overview of in silico protein function prediction,” *Arch Microbiol*, vol. 192, pp. 151–155, 2010.

[120] P. Hu *et al.*, “Global Functional Atlas of Escherichia coli Encompassing Previously Uncharacterized Proteins,” *PLoS Biol*, vol. 7, no. 4, pp. 1–19, 2009.

[121] M. Huynen, B. Snel, W. Lathe, and P. Bork, “Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences,” *Genome Research*, vol. 10, no. 8. pp. 1204–1210, Aug-2000.

[122] T. Doerks, C. von Mering, and P. Bork, “Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes,” *Nucleic Acids Research*, vol. 32, no. 21. Oxford, UK, pp. 6321–6326, 2004.

[123] B. Schwikowski, P. Uetz, and S. Fields, “A network of protein-protein interactions in yeast,” *Nat Biotech*, vol. 18, no. 12, pp. 1257–1261, Dec. 2000.

[124] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, “Global protein function prediction from protein-protein interaction networks,” *Nat Biotech*, vol. 21, no. 6, pp. 697–700, Jun. 2003.

[125] S. Letovsky and S. Kasif, “Predicting protein function from protein/protein interaction data: a probabilistic approach,” *Bioinforma.* , vol. 19, no. suppl 1, pp. i197–i204, Jul. 2003.

[126] C. S. Funk, I. Kahanda, A. Ben-Hur, and K. M. Verspoor, “Evaluating a variety of text-mined features for automatic protein function prediction with GOstruct,” *Journal of Biomedical Semantics*, vol. 6. London, 2015.

[127] A. Wong and H. Shatkay, “Protein Function Prediction using Text-based Features extracted

from the Biomedical Literature: The CAFA Challenge," *BMC Bioinformatics*, vol. 14, no. Suppl 3. p. S14, 2013.

[128] K. Taha and P. D. Yoo, "Predicting the functions of a protein from its ability to associate with other molecules," *BMC Bioinformatics*, 2016.

[129] D. S. C. M. J. Jacob F. Perrin, "Operon: a group of genes with the expression coordinated by an operator," *C. R. Hebd. Seances Acad. Sci.*, vol. 250, pp. 1727–1729, 1960.

[130] R. J. Cho *et al.*, "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle," *Mol. Cell*, vol. 2, no. 1, pp. 65–73, Jul. 1998.

[131] B. A. Cohen, R. D. Mitra, J. D. Hughes, and G. M. Church, "A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression," *Nat Genet*, vol. 26, no. 2, pp. 183–186, Oct. 2000.

[132] J. M. Lee and E. L. L. Sonnhammer, "Genomic Gene Clustering Analysis of Pathways in Eukaryotes," *Genome Res.*, vol. 13, no. 5, pp. 875–882, May 2003.

[133] L. D. Hurst, E. J. B. Williams, and C. Pál, "Natural selection promotes the conservation of linkage of co-expressed genes," *Trends Genet.*, vol. 18, no. 12, pp. 604–606, Dec. 2002.

[134] M. J. Lercher and L. D. Hurst, "Co-expressed Yeast Genes Cluster Over a Long Range but are not Regularly Spaced," *J. Mol. Biol.*, vol. 359, no. 3, pp. 825–831, Jun. 2006.

[135] A. T. Ghanbarian and L. D. Hurst, "Neighboring genes show correlated evolution in gene expression," *Mol. Biol. Evol.* , Mar. 2015.

[136] J. Wang *et al.*, "Single-Cell Co-expression Analysis Reveals Distinct Functional Modules, Co-regulation Mechanisms and Clinical Outcomes," *PLoS Comput Biol*, vol. 12, no. 4, p. e1004892, Apr. 2016.

[137] J. Ihmels, R. Levy, and N. Barkai, "Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae," *Nat Biotech*, vol. 22, no. 1, pp. 86–92, Jan. 2004.

[138] X. Liu, B. Liu, Z. Huang, T. Shi, Y. Chen, and J. Zhang, "SPPS: A Sequence-Based Method for Predicting Probability of Protein-Protein Interaction Partners," *PLoS One*, vol. 7, no. 1, p. e30938, Jan. 2012.

[139] R. Jansen *et al.*, "A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data," *Science (80-.).*, vol. 302, no. 5644, pp. 449–453, Oct.

2003.

[140] T. Dottorini, N. Senin, G. Mazzoleni, K. Magnusson, and A. Crisanti, “Gepoclu: a software tool for identifying and analyzing gene positional clusters in large-scale gene expression analysis,” *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–15, 2011.

[141] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, “STRING: a database of predicted functional associations between proteins,” *Nucleic Acids Res.*, vol. 31, no. 1, pp. 258–261, Jan. 2003.

[142] S. Yon Rhee, V. Wood, K. Dolinski, and S. Draghici, “Use and misuse of the gene ontology annotations,” *Nat Rev Genet*, vol. 9, no. 7, pp. 509–515, Jul. 2008.

[143] M. H. Saier, C. V Tran, and R. D. Barabote, “TCDB: the Transporter Classification Database for membrane transport protein analyses and information,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D181–D186, Jan. 2006.

[144] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[145] I. M. Keseler *et al.*, “EcoCyc: fusing model organism databases with systems biology,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D605–D612, Jan. 2013.

[146] and N. L.-B. P. D. K. C. A. O. C. M.-K. L. G. P. K. D. A. S. T. N. D. Victor Kunin, “EcoCyc: fusing model organism databases with systems biology,” *Nucleic Acids Res.*, vol. 33, pp. 6083–6089, 2005.

[147] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson, “Integrating high-throughput and computational data elucidates bacterial networks,” *Nature*, vol. 429, no. 6987, pp. 92–96, May 2004.

[148] G. C. R. K. L. Brem R. B. Yvert, “Genetic dissection of transcriptional regulation in budding yeast,” *Science (80-.).*, vol. 296, pp. 752–755, 2002.

[149] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, “GOSemSim: an R package for measuring semantic similarity among GO terms and gene products,” *Bioinforma. *, vol. 26, no. 7, pp. 976–978, Apr. 2010.

[150] C. J. Chang C. C. Lin, “LIBSVM: A Library for Support Vector Machines,” *Acm Trans. Intell. Syst. Technol.*, vol. 2, 2011.

[151] N. K. Mishra, J. Chang, and P. X. Zhao, “Prediction of Membrane Transport Proteins and Their Substrate Specificities Using Primary Sequence Information,” *PLoS One*, vol. 9, no. 6, p. e100278, Jun. 2014.

[152] H. Li, V. A. Benedito, M. K. Udvardi, and P. X. Zhao, “TransportTP: A two-phase classification approach for membrane transporter prediction and characterization,” *BMC Bioinformatics*, vol. 10, no. 1, p. 418, 2009.

[153] I. Junier and O. Rivoire, “Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation,” *PLoS One*, vol. 11, no. 5, p. e0155740, May 2016.

[154] W. C. Lathe III, B. Snel, and P. Bork, “Gene context conservation of a higher order than operons,” *Trends Biochem. Sci.*, vol. 25, no. 10, pp. 474–479, Oct. 2000.

[155] J. F. Poyatos and L. D. Hurst, “The determinants of gene order conservation in yeasts,” *Genome Biol.*, vol. 8, no. 11, pp. R233–R233, Nov. 2007.

[156] X.-Y. Ren, M. W. E. J. Fiers, W. J. Stiekema, and J.-P. Nap, “Local Coexpression Domains of Two to Four Genes in the Genome of Arabidopsis,” *Plant Physiol.*, vol. 138, no. 2, pp. 923–934, Jun. 2005.

[157] X.-J. Cui, L. Cai, Y.-Q. Xing, X.-J. Zhao, and C.-X. Shi, “Influence factors on the correlations between expression levels of neighboring pattern genes,” *Biosystems*, vol. 139, pp. 23–28, Jan. 2016.

[158] J. R. Pollack *et al.*, “Genome-wide analysis of DNA copy-number changes using cDNA microarrays,” *Nat Genet*, vol. 23, no. 1, pp. 41–46, Sep. 1999.

[159] R. Beroukhim *et al.*, “The landscape of somatic copy-number alteration across human cancers,” *Nature*, vol. 463, no. 7283, pp. 899–905, Feb. 2010.

[160] T. Santarius, J. Shipley, D. Brewer, M. R. Stratton, and C. S. Cooper, “A census of amplified and overexpressed human cancer genes,” *Nat Rev Cancer*, vol. 10, no. 1, pp. 59–64, Jan. 2010.

[161] T. I. Zack *et al.*, “Pan-cancer patterns of somatic copy number alteration,” *Nat Genet*, vol. 45, no. 10, pp. 1134–1140, Oct. 2013.

[162] J. Budczies *et al.*, “Pan-cancer analysis of copy number changes in programmed death-

ligand 1 (PD-L1, CD274) – associations with gene expression, mutational load, and survival," *Genes, Chromosom. Cancer*, vol. 55, no. 8, pp. 626–639, Aug. 2016.

[163] J. Weischenfeldt *et al.*, "Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking," *Nat Genet*, vol. 49, no. 1, pp. 65–74, Jan. 2017.

[164] M. Zhao and Z. Zhao, "Concordance of copy number loss and down-regulation of tumor suppressor genes: a pan-cancer study," *BMC Genomics*, vol. 17, no. 7, p. 532, 2016.

[165] H. Chen, H. Xing, and N. R. Zhang, "Estimation of Parent Specific DNA Copy Number in Tumors using High-Density Genotyping Arrays," *PLOS Comput. Biol.*, vol. 7, no. 1, p. e1001060, Jan. 2011.

[166] R. Xi *et al.*, "Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion," *Proc. Natl. Acad. Sci.* , vol. 108, no. 46, pp. E1128–E1136, Nov. 2011.

[167] R. Jörnsten *et al.*, "Network modeling of the transcriptional effects of copy number aberrations in glioblastoma," *Mol. Syst. Biol.*, vol. 7, no. 1, Apr. 2011.

[168] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer Genome Landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, Mar. 2013.

[169] L. Zhang, Y. Yuan, K. H. Lu, and L. Zhang, "Identification of recurrent focal copy number variations and their putative targeted driver genes in ovarian cancer," *BMC Bioinformatics*, vol. 17, no. 1, p. 222, 2016.

[170] P. Freire *et al.*, "Exploratory Analysis of the Copy Number Alterations in Glioblastoma Multiforme," *PLoS One*, vol. 3, no. 12, p. e4076, Dec. 2009.

[171] Q. Zhang *et al.*, "CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data," *Bioinformatics*, vol. 26, no. 4, pp. 464–469, Feb. 2010.

[172] Y. Wang *et al.*, "Genomic DNA Copy-Number Alterations of the let-7 Family in Human Cancers," *PLoS One*, vol. 7, no. 9, p. e44399, Sep. 2012.

[173] R. G. Verhaak *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, 2010.

[174] L. Li *et al.*, “SQSTM1 Is a Pathogenic Target of 5q Copy Number Gains in Kidney Cancer,” *Cancer Cell*, vol. 24, no. 6, pp. 738–750, 2013.

[175] A. A. Hakimi *et al.*, “Adverse Outcomes in Clear Cell Renal Cell Carcinoma with Mutations of 3p21 Epigenetic Regulators BAP1 and SETD2: A Report by MSKCC and the KIRC TCGA Research Network,” *Clin. Cancer Res.*, vol. 19, no. 12, p. 3259 LP-3267, Jun. 2013.

[176] E. Kerr, E. Gaude, F. Turrell, C. Frezza, and C. P. Martins, “Mutant Kras copy number defines metabolic reprogramming and therapeutic susceptibilities,” *Nature*, vol. 531, no. 7592, pp. 110–113, Mar. 2016.

[177] Y. Chen *et al.*, “Identification of Druggable Cancer Driver Genes Amplified across TCGA Datasets,” *PLoS One*, vol. 9, no. 5, p. e98293, May 2014.

[178] H. Hieronymus *et al.*, “Copy number alteration burden predicts prostate cancer relapse,” *Proc. Natl. Acad. Sci. *, vol. 111, no. 30, pp. 11139–11144, Jul. 2014.

[179] A. Sudhakar, “History of Cancer, Ancient and Modern Treatment Methods,” *J. Cancer Sci. Ther.*, vol. 1, no. 2, pp. 1–4, Dec. 2009.

[180] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, “KEGG for integration and interpretation of large-scale molecular data sets,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D109–D114, Jan. 2012.

[181] N. A. Graham *et al.*, “Recurrent patterns of DNA copy number alterations in tumors reflect metabolic selection pressures,” *Mol. Syst. Biol.*, vol. 13, no. 2, Feb. 2017.

[182] R. L. Grossman *et al.*, “Toward a Shared Vision for Cancer Genomic Data,” *N. Engl. J. Med.*, vol. 375, no. 12, pp. 1109–1112, Sep. 2016.

[183] D. S. Wishart *et al.*, “DrugBank: a comprehensive resource for in silico drug discovery and exploration,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D668–D672, Jan. 2006.

[184] E. Eisenberg and E. Y. Levanon, “Human housekeeping genes, revisited,” *Trends Genet.*, vol. 29, no. 10, pp. 569–574, Oct. 2013.

[185] S. V Laddha, S. Ganesan, C. S. Chan, and E. White, “Mutational Landscape of the Essential Autophagy Gene BECN in Human Cancers,” *Mol. Cancer Res.*, vol. 12, no. 4, p. 485 LP-490, Apr. 2014.

[186] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford, “Genenames.org: the HGNC resources in 2015,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D1079–D1085, Jan. 2015.

[187] H. J. Lee *et al.*, “Protein overexpression and gene amplification of epidermal growth factor receptor in nonsmall cell lung carcinomas: Comparison of four commercially available antibodies by immunohistochemistry and fluorescence in situ hybridization study,” *Lung Cancer*, vol. 68, no. 3, pp. 375–382, 2010.

[188] A. V. López-Malpartida, M. D. Ludeña, G. Varela, and J. García Pichel, “Differential ErbB receptor expression and intracellular signaling activity in lung adenocarcinomas and squamous cell carcinomas,” *Lung Cancer*, vol. 65, no. 1, pp. 25–33, 2009.

[189] S. Dacic *et al.*, *Significance of EGFR Protein Expression and Gene Amplification in Non-Small Cell Lung Carcinoma*, vol. 125. 2006.

[190] F. R. Hirsch *et al.*, “Epidermal Growth Factor Receptor in Non–Small-Cell Lung Carcinomas: Correlation Between Gene Copy Number and Protein Expression and Impact on Prognosis,” *J. Clin. Oncol.*, vol. 21, no. 20, pp. 3798–3807, Oct. 2003.

[191] J. Pancewicz-Wojtkiewicz, “Epidermal growth factor receptor and notch signaling in non-small-cell lung cancer,” *Cancer Med.*, vol. 5, no. 12, pp. 3572–3578, Dec. 2016.

[192] J. L. Riley, “PD-1 signaling in primary T cells,” *Immunol. Rev.*, vol. 229, no. 1, pp. 114–125, May 2009.

[193] I. P. Foskolou *et al.*, “Ribonucleotide Reductase Requires Subunit Switching in Hypoxia to Maintain DNA Replication,” *Mol. Cell*, vol. 66, no. 2, p. 206–220.e9, Apr. 2017.

[194] B. Zhang, W. Xie, and M. D. Krasowski, “PXR: a xenobiotic receptor of diverse function implicated in pharmacogenetics,” *Pharmacogenomics*, vol. 9, no. 11, pp. 1695–1709, Nov. 2008.

[195] D. W. Hommes, M. P. Peppelenbosch, and S. J. H. van Deventer, “Mitogen activated protein (MAP) kinase signal transduction pathways and novel anti-inflammatory targets,” *Gut*, vol. 52, no. 1, pp. 144–151, Jan. 2003.

[196] H.-T. Jin, R. Ahmed, and T. Okazaki, “Role of PD-1 in Regulating T-Cell Immunity BT - Negative Co-Receptors and Ligands,” R. Ahmed and T. Honjo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 17–37.

[197] L. J. Leandro-García *et al.*, “Tumoral and tissue-specific expression of the major human β-tubulin isotypes,” *Cytoskeleton*, vol. 67, no. 4, pp. 214–223, 2010.

[198] Y. Liu, J. W. Lee, and S. L. Ackerman, “Mutations in the Microtubule-Associated Protein 1A (Map1a) Gene Cause Purkinje Cell Degeneration,” *J. Neurosci.*, vol. 35, no. 11, pp. 4587–4598, Mar. 2015.

[199] R. Beroukhim *et al.*, “Patterns of gene expression and copy-number alterations in VHL disease-associated and sporadic clear cell carcinoma of the kidney,” *Cancer Res.*, vol. 69, no. 11, pp. 4674–4681, Jun. 2009.

[200] J. Y. Kim *et al.*, “Decreased efficacy of drugs targeting the vascular endothelial growth factor pathway by the epigenetic silencing of FLT1 in renal cancer cells,” *Clin. Epigenetics*, vol. 7, p. 99, Sep. 2015.

[201] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston, “Cancer drug resistance: an evolving paradigm,” *Nat. Rev. Cancer*, vol. 13, p. 714, Sep. 2013.

[202] B. Mansoori, A. Mohammadi, S. Davudian, S. Shirjang, and B. Baradaran, “ The Different Mechanisms of Cancer Drug Resistance: A Brief Review ,” *Adv. Pharm. Bull.*, vol. 7, no. 3, pp. 339–348, Sep. 2017.

[203] Y. Tang, Y. Wang, M. F. Kiani, and B. Wang, “Classification, Treatment Strategy, and Associated Drug Resistance in Breast Cancer,” *Clin. Breast Cancer*, vol. 16, no. 5, pp. 335–343, 2016.

[204] G. Housman *et al.*, “Drug Resistance in Cancer: An Overview,” *Cancers (Basel).*, vol. 6, no. 3, pp. 1769–1792, Sep. 2014.

[205] L. Huang and L. Fu, “Mechanisms of resistance to EGFR tyrosine kinase inhibitors,” *Acta Pharm. Sin. B*, vol. 5, no. 5, pp. 390–401, 2015.

[206] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, “KEGG: new perspectives on genomes, pathways, diseases and drugs,” *Nucleic Acids Res.*, vol. 45, no. Database issue, pp. D353–D361, Jan. 2017.

[207] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “KEGG as a reference resource for gene and protein annotation,” *Nucleic Acids Res.*, vol. 44, no. Database issue, pp. D457–D462, Jan. 2016.

[208] S. R. Hubbard and W. T. Miller, “Receptor tyrosine kinases: mechanisms of activation and signaling,” *Curr. Opin. Cell Biol.*, vol. 19, no. 2, pp. 117–123, Apr. 2007.

[209] F. Ciardiello and G. Tortora, “EGFR Antagonists in Cancer Treatment,” *N. Engl. J. Med.*, vol. 358, no. 11, pp. 1160–1174, Mar. 2008.

[210] M. J. Comstock, “Platinum, Gold, and Other Metal Chemotherapeutic Agents, Copyright, ACS Symposium Series, FOREWORD,” in *Platinum, Gold, and Other Metal Chemotherapeutic Agents*, vol. 209, C. M. Joan, Ed. AMERICAN CHEMICAL SOCIETY, 1983, pp. i–vi.

[211] S. Dilruba and G. V Kalayda, “Platinum-based drugs: past, present and future,” *Cancer Chemother. Pharmacol.*, vol. 77, no. 6, pp. 1103–1124, 2016.

[212] L. Galluzzi *et al.*, “Molecular mechanisms of cisplatin resistance,” *Oncogene*, vol. 31, p. 1869, Sep. 2011.

[213] E. Martinez-Balibrea *et al.*, “Tumor-Related Molecular Mechanisms of Oxaliplatin Resistance,” *Mol. Cancer Ther.*, vol. 14, no. 8, p. 1767 LP-1776, Aug. 2015.

[214] Y.-I. Kim, “Folate and cancer: a tale of Dr. Jekyll and Mr. Hyde?,” *Am. J. Clin. Nutr.*, vol. 107, no. 2, pp. 139–142, Feb. 2018.

[215] J. Chon, P. J. Stover, and M. S. Field, “Targeting Nuclear Thymidylate Biosynthesis,” *Mol. Aspects Med.*, vol. 53, pp. 48–56, Feb. 2017.

[216] D. B. Longley, D. P. Harkin, and P. G. Johnston, “5-Fluorouracil: mechanisms of action and clinical strategies,” *Nat. Rev. Cancer*, vol. 3, p. 330, May 2003.

[217] Y. G. Assaraf, “Molecular basis of antifolate resistance,” *Cancer Metastasis Rev.*, vol. 26, no. 1, pp. 153–181, 2007.

[218] N. Gonen and Y. G. Assaraf, “Antifolates in cancer therapy: Structure, activity and mechanisms of drug resistance,” *Drug Resist. Updat.*, vol. 15, no. 4, pp. 183–210, 2012.

[219] N. Platet, A. M. Cathiard, M. Gleizes, and M. Garcia, “Estrogens and their receptors in breast cancer progression: a dual role in cancer proliferation and invasion,” *Crit. Rev. Oncol. Hematol.*, vol. 51, no. 1, pp. 55–67, 2004.

[220] E. V Jensen and V. C. Jordan, “The Estrogen Receptor,” *Clin. Cancer Res.*, vol. 9, no. 6, p. 1980 LP-1989, Jun. 2003.

[221] E. A. Musgrove and R. L. Sutherland, “Biological determinants of endocrine resistance in breast cancer,” *Nat. Rev. Cancer*, vol. 9, p. 631, Sep. 2009.

[222] R. García-Becerra, N. Santos, L. Díaz, and J. Camacho, “Mechanisms of Resistance to Endocrine Therapy in Breast Cancer: Focus on Signaling Pathways, miRNAs and Genetically Based Resistance,” *International Journal of Molecular Sciences* , vol. 14, no. 1. 2013.

[223] H. A. V. U. TRAN, A. K. KIEMER, and V. HELMS, “Copy Number Alterations in Tumor Genomes Deleting Antineoplastic Drug Targets Partially Compensated by Complementary Amplifications,” *Cancer Genomics - Proteomics* , vol. 15, no. 5, pp. 365–378, Sep. 2018.

[224] R. T. Schimke, “Gene Amplification, Drug Resistance, and Cancer,” *Cancer Res.*, vol. 44, no. 5, p. 1735 LP-1742, May 1984.

[225] P. Borst, “Genetic Mechanisms of Drug Resistance: A Review,” *Acta Oncol. (Madr).*, vol. 30:1, pp. 87–105, 1991.

[226] B. R., “Gene amplification and drug resistance,” *J. Pathol.*, vol. 163, no. 4, pp. 287–292, Jul. 1991.

[227] P. V Schoenlein, “Role of gene amplification in drug resistance BT - Anticancer Drug Resistance: Advances in Molecular and Clinical Research,” L. J. Goldstein and R. F. Ozols, Eds. Boston, MA: Springer US, 1994, pp. 167–200.

[228] K. Yasui *et al.*, “Alteration in Copy Numbers of Genes as a Mechanism for Acquired Drug Resistance,” *Cancer Res.*, vol. 64, no. 4, p. 1403 LP-1410, Feb. 2004.

[229] M. Jim Yen, I.-M. Shih, V. E. Velculescu, and T.-L. Wang, “Amplification in DNA Copy Numbers as a Mechanism of Acquired Drug Resistance BT - Cancer Drug Resistance,” B. A. Teicher, Ed. Totowa, NJ: Humana Press, 2006, pp. 531–540.

[230] S. Varma, Y. Pommier, M. Sunshine, J. N. Weinstein, and W. C. Reinhold, “High Resolution Copy Number Variation Data in the NCI-60 Cancer Cell Lines from Whole Genome Microarrays Accessible through CellMiner,” *PLoS One*, vol. 9, no. 3, p. e92047, Mar. 2014.

[231] H. Harry *et al.*, “Modulation of chemotherapeutic drug resistance in neuroblastoma SK-N-AS cells by the neural apoptosis inhibitory protein and miR-520f,” *Int. J. Cancer*, vol. 136, no. 7, pp. 1579–1588, Aug. 2014.

[232] S. Mishra and J. R. Whetstine, “Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive,” *Mol. Cell. Biol.* , vol. 36, no. 7, pp. 1050–1063, Apr. 2016.

[233] T. H. Lippert, H.-J. Ruoff, and M. Volm, “Intrinsic and Acquired Drug Resistance in Malignant Tumors,” *Arzneimittelforschung*, vol. 58, no. 06, pp. 261–264, 2008.

[234] R. Baserga, “The Insulin-like Growth Factor I Receptor: A Key to Tumor Growth?,” *Cancer Res.*, vol. 55, no. 2, p. 249 LP-252, Jan. 1995.

[235] N. Peled *et al.*, “Insulin-like growth factor-1 receptor (IGF-1R) as a biomarker for resistance to the tyrosine kinase inhibitor gefitinib in non-small cell lung cancer,” *Cell. Oncol.*, vol. 36, no. 4, pp. 277–288, 2013.

[236] M. Guix *et al.*, *Acquired resistance to EGFR tyrosine kinase inhibitors in cancer cells is mediated by loss of IGF-binding proteins*, vol. 118. 2008.

[237] G. Song, G. Ouyang, and S. Bao, “The activation of Akt/PKB signaling pathway and cell survival,” *J. Cell. Mol. Med.*, vol. 9, no. 1, pp. 59–71, May 2007.

[238] V. Duronio, “The life of a cell: apoptosis regulation by the PI3K/PKB pathway,” *Biochem. J.*, vol. 415, no. 3, p. 333 LP-344, Nov. 2008.

[239] S. Koyasu, “The role of PI3K in immune cells,” *Nat. Immunol.*, vol. 4, p. 313, Apr. 2003.

[240] J. A. Engelman, J. Luo, and L. C. Cantley, “The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism,” *Nat. Rev. Genet.*, vol. 7, p. 606, Aug. 2006.

[241] M. Falasca, “PI3K/Akt Signalling Pathway Specific Inhibitors: A Novel Strategy to Sensitize Cancer Cells to Anti-Cancer Drugs,” *Current Pharmaceutical Design*, vol. 16, no. 12. pp. 1410–1416, 2010.

[242] J. Dong, B. Zhai, W. Sun, F. Hu, H. Cheng, and J. Xu, “Activation of phosphatidylinositol 3-kinase/AKT/snail signaling pathway contributes to epithelial-mesenchymal transition-induced multi-drug resistance to sorafenib in hepatocellular carcinoma cells,” *PLoS One*, vol. 12, no. 9, p. e0185088, Sep. 2017.

[243] J. A. McCubrey *et al.*, “Roles of signaling pathways in drug resistance, cancer initiating cells and cancer progression and metastasis,” *Adv. Biol. Regul.*, vol. 57, pp. 75–101, 2015.

[244] E. Tokunaga *et al.*, “Activation of PI3K/Akt signaling and hormone resistance in breast

cancer," *Breast Cancer*, vol. 13, no. 2, pp. 137–144, 2006.

[245] K. Ohshima *et al.*, "Integrated analysis of gene expression and copy number identified potential cancer driver genes with amplification-dependent overexpression in 1,454 solid tumors," *Sci. Rep.*, vol. 7, p. 641, Apr. 2017.

[246] Y.-C. Tang and A. Amon, "Gene copy number alterations: A cost-benefit analysis," *Cell*, vol. 152, no. 3, pp. 394–405, Jan. 2013.

[247] E. R. Gamazon and B. E. Stranger, "The impact of human copy number variation on gene expression," *Brief. Funct. Genomics*, vol. 14, no. 5, pp. 352–357, Sep. 2015.

Supplementary material

Supplement table 1: Gene symbols of four metabolic pathway groups

Amino Acid	Sugar	Lipid	Nucleotide
gstA, ygfK, hcaB, hcaC, ygfM, hcaE, luxS, katG, sdaA, sdaB, mmuM, astC, astB, astE, astD, malY, patD, pheA, gdhA, lysA, ynfE, ynfF, lysC, hcaF, asd, adiA, argC, argB, tdcG, argE, argD, mhpE, argA, gor, tdcB, argG, mhpD, argF, mhpC, argI, mhpB, argH, mhpA, pepN, pepA, pepB, sufS, pepD, ilvD, hisB, ilvC, hisC, hisD, ilvE, hisF, hisG, hisH, hisI, ynjE, ilvA, hisA, trpD, trpE, trpB, trpC, sseA, tynA, xdhD, gshA, gshB, trpA, cysK, puuD, proA, proB, proC, dsdA, puuA, cysE, puuB, gss, puuC, selA, cadA, cysM, selD, dapA, dapB, putA, feaB, ansB, ansA, dapE, dapF, dapD, alaA, mnaT, nadB, tyrA, tyrB, asnA, asnB, avtA, panC, panD, ydiB, paaD, aroK, paaE, paaG, aroL, dcyD, aroG, paaA, aroF, paaB, paaC, aroH, murE, aroC, yjhH, aroB, murD, ldcC, aroE, murC, aroD, murI, paaI, paaJ, aroA, murF,	bcsA, fruA, fruB, manY, manX, aceF, aceE, ugd, sucC, manA, sucD, cmtB, scpB, rhmD, frdA, cmtA, frdB, ppsA, gapA, frdC, frdD, rhmA, glf, glk, xylA, xylB, malZ, aceB, malX, chiA, aceA, fdhF, ebgC, yeaD, ebgA, amyA, acuI, gmd, yedP, tpiA, garL, glgX, garR, glxR, glgB, glgC, garD, rbsK, glgP, ppc, ldhA, bcsZ, uxuB, appA, glgA, fruK, manZ, bglA, bglB, uxuA, ghrB, rhaB, rhaA, agaI, agaD, tdcE, agaC, tdcD, agaB, suhB, rhaD, gatY, acs, agaV, gatA, gatB, gatD, gph, nanK, fumA, fumB, talA, fumC, talB, nanA, nanE, fdoI, sdhD, fdoH, pck, sdhC, fdoG, purU, dld, sdhB, sdhA, otsB, ulaA, eno, otsA, yccX, ttdB, ulaF, ulaG, ttdA, ulaD, ulaE, ulaB, ulaC, gatZ, gcd, agp, ptsG, mtlD, mtlA, gcl, lyxK, glmM, acnA, acnB, glmU, yiaK, maeB, maeA, hchA, eutG, eutE, eutD, tktA, tktB, fbp, pfo, poxB, sgbE, treF,	fabI, fabH, fabG, plsC, fabZ, plsB, dhaL, dhaK, cdh, aslA, fabF, glpK, fabD, fabB, fabA, glpQ, yciA, pgpA, pgpB, pgpC, psd, ugpQ, glpA, glpB, glpC, glpD, ynbB, eutC, eutB, pgsA, aas, eutA, dgkA, tesB, cdsA, gpsA, tesA, pldA, pldB, fadE, fadD, plsX, plsY, clsB, clsA	dut, hiuH, cysC, dnaX, nrdB, nrdA, tmk, nrdF, nrdE, nrdD, dnaE, cdd, apt, pnp, upp, dnaQ, dnaN, guaC, cmk, guaD, xapA, gmk, rpoB, rpoA, rpoC, paoB, dgt, paoC, paoA, rdgB, umpG, nudF, nudE, apaH, guaA, guaB, rihB, thyA, hpt, ndk, ppx, psuG, psuK, holE, mazG, polA, add, ade, cpdB, gpp, adk, gpt, holA, holB, holC, holD, rpoZ, udk, udp, ushA, amn, dcd, purT, purM, purN, yjjG, spoT, deoD, cyaA, deoA, rutF, rutE, purC, purD, rutB, rutA, rutD, rutC, pyrE, allE, pyrF, purK, pyrG, purL, pyrH, purE, allB, allA, allD, pyrC, purH, allC, pyrD, tdk, xdhA, relA, gsk, xdhB, xdhC

paaK, yafJ, ggt, dadX, phnJ, phnL, phnM, phnG, mtn, phnH, alr, phnI, aspC, speE, speD, speG, tauD, speF, speA, aspA, speC, phnP, speB, thrB, thrC, dadA, thrA, ltaE, gltB, metL, gltD, patA, tnaA, dcm, astA, serA, metC, metB, serC, metA, serB, metG, metE, metK, metH, glsA, glsB, kbl, ddlB, tdh, ddlA, betA, yagE, paaZ, betB	mqo, treC, treA, treB, gntK, sgbH, cpsG, dgoK, ydjG, pgi, pgl, pgk, yadI, mak, dgoD, dgoA, prpB, prpC, prpD, prpE, fucI, chbF, fucK, gloA, gloB, fucO, ydiF, cpsB, hyi, mgsA, fucA, rpiB, kbaZ, nagE, rpiA, galE, nagK, murA, galK, galM, murB, ycjU, alsK, fbaB, fbaA, ybhJ, ycjM, pfkA, pfkB, oxc, ybhC, kbaY, nagB, nagA, sgcC, eda, edd, sgcB, sgcA, arnB, arnA, arnD, arnC, phnN, alsE, dmlA, sgbU, wecC, kduD, wecB, glpX, kduI, srlB, srlA, yihQ, gltA, srlD, yihU, uidA, malS, idnK, malQ, malP, yqhD, srlE, uxaB, lldD, uxaC, glcB, uxaA, glcD, glcE, ascF, glcF, araD, deoC, ascB, araA, araB, crr, aldB, wcaG, aldA, rpe, rspB, rspA, wcaN, kdgK, ybiW, yagH, galT, galU, murQ, nagZ, murP, pflB, pflD, gudD

Supplement table 2 Usable membrane proteins when retrieving 10 neighbors upstream, 10 neighbors downstream and selecting 3 neighbors that have highest co-expression correlations

Threshold r	Organism	Gene group	Total	No neighbors	No expression	Not usable for SVM	Usable for SVM
0.2	*Escherichia coli*	Sugar	39	0	10	8	21
		Amino acid	47	0	14	8	25
	Saccharomyces cerevisiae	Sugar	17	0	10	0	7
		Amino acid	24	0	10	0	14
	Homo sapiens	Sugar	13	1	0	2	10
		Amino acid	37	3	1	0	33
0.5	*Escherichia coli*	Sugar	39	0	10	11	18
		Amino acid	47	0	14	10	23
	Saccharomyces cerevisiae	Sugar	17	0	10	0	7
		Amino acid	24	0	10	2	12
	Homo sapiens	Sugar	13	1	0	3	9
		Amino acid	37	3	1	4	29
0.8	*Escherichia coli*	Sugar	39	0	10	13	16
		Amino acid	47	0	14	15	18
	Saccharomyces cerevisiae	Sugar	17	0	10	1	6
		Amino acid	24	0	10	5	9
	Homo sapiens	Sugar	13	1	0	5	7
		Amino acid	37	3	1	16	17

Supplement table 3 Two groups of metabolic pathway enzymes with different thresholds and number of neighbors

Threshold r	Windows size	Gene group	Total	No neighbors	No expression	Not usable for SVM	Usable for SVM
0.2	(5, 3)	Sugar	253	6	52	50	145
		Amino acid	187	2	40	44	101
	(10, 3)	Sugar	253	6	52	88	107
		Amino acid	187	2	40	72	73
	(20, 5)	Sugar	253	6	52	118	77
		Amino acid	187	2	40	104	41
0.5	(5, 3)	Sugar	253	6	52	51	144
		Amino acid	187	2	40	49	96
	(10, 3)	Sugar	253	6	52	89	106
		Amino acid	187	2	40	75	70
	(20, 5)	Sugar	253	6	52	118	77
		Amino acid	187	2	40	105	40
0.8	(5, 3)	Sugar	253	6	52	59	136
		Amino acid	187	2	40	52	93
	(10, 3)	Sugar	253	6	52	92	103
		Amino acid	187	2	40	77	68
	(20, 5)	Sugar	253	6	52	120	75
		Amino acid	187	2	40	105	40

Supplement table 4 Four groups of metabolic pathway enzymes with different thresholds and number of neighbors

Threshold r	Windows size	Gene group	Total	No neighbors	No expression	Not usable for SVM	Usable for SVM
0.2	(5, 3)	Sugar	253	6	52	65	130
		Amino acid	187	2	40	51	94
		Lipid	45	1	8	14	22
		Nucleotide	99	2	26	21	50
	(10, 3)	Sugar	253	6	52	132	63
		Amino acid	187	2	40	101	44
		Lipid	45	1	8	26	10
		Nucleotide	99	2	26	37	34
	(20, 5)	Sugar	253	6	52	157	38
		Amino acid	187	2	40	127	18
		Lipid	45	1	8	32	4
		Nucleotide	99	2	26	49	22
0.5	(5, 3)	Sugar	253	6	52	67	128
		Amino acid	187	2	40	52	93
		Lipid	45	1	8	15	21
		Nucleotide	99	2	26	21	50
	(10, 3)	Sugar	253	6	52	132	63
		Amino acid	187	2	40	101	44
		Lipid	45	1	8	26	10
		Nucleotide	99	2	26	38	33
	(20, 5)	Sugar	253	6	52	157	38
		Amino acid	187	2	40	127	18
		Lipid	46	1	8	33	4
		Nucleotide	99	2	26	49	22
0.8	(5, 3)	Sugar	253	6	52	67	128
		Amino acid	187	2	40	52	93
		Lipid	45	1	8	16	20
		Nucleotide	99	2	26	25	46
	(10, 3)	Sugar	253	6	52	132	63
		Amino acid	187	2	40	101	44
		Lipid	45	1	8	26	10
		Nucleotide	99	2	26	39	32
	(20, 5)	Sugar	253	6	52	157	38
		Amino acid	187	2	40	127	18
		Lipid	45	1	8	32	4
		Nucleotide	99	2	26	49	22

Supplement table 5 KEGG pathways that contain target proteins of antineoplastic drugs

Pathway ID	Pathway name	Number of AN targets
hsa05200	Pathways in cancer	70
hsa04151	PI3K-Akt signaling pathway	53
hsa01100	Metabolic pathways	39
hsa04010	MAPK signaling pathway	38
hsa05165	Human papillomavirus infection	36
hsa04060	Cytokine-cytokine receptor interaction	35
hsa04014	Ras signaling pathway	32
hsa04015	Rap1 signaling pathway	32
hsa04659	Th17 cell differentiation	30
hsa05166	HTLV-I infection	27
hsa05205	Proteoglycans in cancer	26
hsa04510	Focal adhesion	26
hsa04630	Jak-STAT signaling pathway	25
hsa05206	MicroRNAs in cancer	24
hsa05167	Kaposi's sarcoma-associated herpesvirus infection	24
hsa04640	Hematopoietic cell lineage	24
hsa01521	EGFR tyrosine kinase inhibitor resistance	24
hsa05203	Viral carcinogenesis	22
hsa05224	Breast cancer	21
hsa04658	Th1 and Th2 cell differentiation	21
hsa04660	T cell receptor signaling pathway	21
hsa04810	Regulation of actin cytoskeleton	21
hsa05164	Influenza A	20
hsa04380	Osteoclast differentiation	20
hsa00230	Purine metabolism	20
hsa05162	Measles	20
hsa05215	Prostate cancer	19
hsa05034	Alcoholism	19
hsa05226	Gastric cancer	18
hsa05230	Central carbon metabolism in cancer	18
hsa05223	Non-small cell lung cancer	18
hsa05161	Hepatitis B	18
hsa04210	Apoptosis	18
hsa04919	Thyroid hormone signaling pathway	18
hsa04620	Toll-like receptor signaling pathway	17
hsa04024	cAMP signaling pathway	17
hsa01522	Endocrine resistance	17
hsa04540	Gap junction	17
hsa05152	Tuberculosis	16
hsa05202	Transcriptional misregulation in cancer	16
hsa05323	Rheumatoid arthritis	16
hsa04917	Prolactin signaling pathway	15
hsa04933	AGE-RAGE signaling pathway in diabetic complications	15

hsa05218	Melanoma	15
hsa05220	Chronic myeloid leukemia	15
hsa04066	HIF-1 signaling pathway	15
hsa05212	Pancreatic cancer	15
hsa04012	ErbB signaling pathway	14
hsa04550	Signaling pathways regulating pluripotency of stem cells	14
hsa04650	Natural killer cell mediated cytotoxicity	14
hsa05142	Chagas disease (American trypanosomiasis)	14
hsa05214	Glioma	14
hsa05225	Hepatocellular carcinoma	14
hsa05221	Acute myeloid leukemia	14
hsa05418	Fluid shear stress and atherosclerosis	13
hsa05160	Hepatitis C	13
hsa04370	VEGF signaling pathway	13
hsa05219	Bladder cancer	13
hsa05169	Epstein-Barr virus infection	13
hsa05140	Leishmaniasis	13
hsa05222	Small cell lung cancer	13
hsa04921	Oxytocin signaling pathway	13
hsa04218	Cellular senescence	13
hsa04621	NOD-like receptor signaling pathway	13
hsa04926	Relaxin signaling pathway	13
hsa04722	Neurotrophin signaling pathway	13
hsa05168	Herpes simplex infection	12
hsa04072	Phospholipase D signaling pathway	12
hsa05133	Pertussis	12
hsa00240	Pyrimidine metabolism	12
hsa04064	NF-kappa B signaling pathway	12
hsa04662	B cell receptor signaling pathway	12
hsa04514	Cell adhesion molecules (CAMs)	12
hsa04217	Necroptosis	12
hsa05145	Toxoplasmosis	12
hsa04360	Axon guidance	12
hsa04912	GnRH signaling pathway	12
hsa04657	IL-17 signaling pathway	12
hsa05231	Choline metabolism in cancer	12
hsa04664	Fc epsilon RI signaling pathway	12
hsa04668	TNF signaling pathway	11
hsa05216	Thyroid cancer	11
hsa04726	Serotonergic synapse	11
hsa04071	Sphingolipid signaling pathway	11
hsa04062	Chemokine signaling pathway	11
hsa04150	mTOR signaling pathway	11
hsa04068	FoxO signaling pathway	11
hsa04145	Phagosome	11
hsa04913	Ovarian steroidogenesis	10

hsa04915	Estrogen signaling pathway	10
hsa05321	Inflammatory bowel disease (IBD)	10
hsa04144	Endocytosis	10
hsa04080	Neuroactive ligand-receptor interaction	10
hsa05130	Pathogenic Escherichia coli infection	9
hsa05340	Primary immunodeficiency	9
hsa04750	Inflammatory mediator regulation of TRP channels	9
hsa05213	Endometrial cancer	9
hsa04371	Apelin signaling pathway	9
hsa05211	Renal cell carcinoma	9
hsa04270	Vascular smooth muscle contraction	9
hsa01523	Antifolate resistance	9
hsa04520	Adherens junction	8
hsa04920	Adipocytokine signaling pathway	8
hsa04020	Calcium signaling pathway	8
hsa04910	Insulin signaling pathway	8
hsa04720	Long-term potentiation	8
hsa04672	Intestinal immune network for IgA production	8
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	8
hsa05131	Shigellosis	8
hsa03410	Base excision repair	8
hsa04916	Melanogenesis	8
hsa00140	Steroid hormone biosynthesis	8
hsa05210	Colorectal cancer	8
hsa04530	Tight junction	8
hsa04611	Platelet activation	8
hsa04666	Fc gamma R-mediated phagocytosis	8
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	8
hsa04730	Long-term depression	8
hsa04914	Progesterone-mediated oocyte maturation	7
hsa05146	Amoebiasis	7
hsa05410	Hypertrophic cardiomyopathy (HCM)	7
hsa04140	Autophagy - animal	7
hsa05132	Salmonella infection	7
hsa04022	cGMP-PKG signaling pathway	7
hsa05020	Prion diseases	7
hsa05414	Dilated cardiomyopathy (DCM)	6
hsa04725	Cholinergic synapse	6
hsa04670	Leukocyte transendothelial migration	6
hsa04114	Oocyte meiosis	6
hsa04110	Cell cycle	6
hsa05010	Alzheimer's disease	6
hsa05031	Amphetamine addiction	6
hsa03050	Proteasome	6
hsa01524	Platinum drug resistance	6
hsa05030	Cocaine addiction	5

hsa04931	Insulin resistance	5
hsa04622	RIG-I-like receptor signaling pathway	5
hsa03030	DNA replication	5
hsa05032	Morphine addiction	5
hsa04940	Type I diabetes mellitus	5
hsa05332	Graft-versus-host disease	5
hsa05330	Allograft rejection	5
hsa05143	African trypanosomiasis	5
hsa00480	Glutathione metabolism	5
hsa04728	Dopaminergic synapse	5
hsa05416	Viral myocarditis	5
hsa05134	Legionellosis	5
hsa04930	Type II diabetes mellitus	5
hsa04723	Retrograde endocannabinoid signaling	5
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	5
hsa04724	Glutamatergic synapse	5
hsa04261	Adrenergic signaling in cardiomyocytes	5
hsa04115	p53 signaling pathway	5
hsa04512	ECM-receptor interaction	5
hsa05144	Malaria	5
hsa04310	Wnt signaling pathway	4
hsa05320	Autoimmune thyroid disease	4
hsa04623	Cytosolic DNA-sensing pathway	4
hsa05014	Amyotrophic lateral sclerosis (ALS)	4
hsa00590	Arachidonic acid metabolism	4
hsa04925	Aldosterone synthesis and secretion	4
hsa05322	Systemic lupus erythematosus	4
hsa03420	Nucleotide excision repair	4
hsa00670	One carbon pool by folate	4
hsa04713	Circadian entrainment	3
hsa04960	Aldosterone-regulated sodium reabsorption	3
hsa00330	Arginine and proline metabolism	3
hsa04213	Longevity regulating pathway - multiple species	3
hsa00983	Drug metabolism - other enzymes	3
hsa05100	Bacterial invasion of epithelial cells	3
hsa03320	PPAR signaling pathway	3
hsa04350	TGF-beta signaling pathway	3
hsa05150	Staphylococcus aureus infection	3
hsa04727	GABAergic synapse	3
hsa04340	Hedgehog signaling pathway	3
hsa00350	Tyrosine metabolism	3
hsa04612	Antigen processing and presentation	2
hsa04137	Mitophagy - animal	2
hsa00260	Glycine, serine and threonine metabolism	2
hsa05310	Asthma	2
hsa00340	Histidine metabolism	2

hsa04924	Renin secretion	2
hsa05016	Huntington's disease	2
hsa00790	Folate biosynthesis	2
hsa05110	Vibrio cholerae infection	2
hsa00565	Ether lipid metabolism	2
hsa00982	Drug metabolism - cytochrome P450	2
hsa00380	Tryptophan metabolism	2
hsa04971	Gastric acid secretion	2
hsa04330	Notch signaling pathway	2
hsa04961	Endocrine and other factor-regulated calcium reabsorption	2
hsa04152	AMPK signaling pathway	2
hsa00360	Phenylalanine metabolism	2
hsa04211	Longevity regulating pathway	2
hsa00592	alpha-Linolenic acid metabolism	2
hsa00591	Linoleic acid metabolism	2
hsa00564	Glycerophospholipid metabolism	2
hsa04918	Thyroid hormone synthesis	2
hsa04966	Collecting duct acid secretion	1
hsa00190	Oxidative phosphorylation	1
hsa04390	Hippo signaling pathway	1
hsa05012	Parkinson's disease	1
hsa05204	Chemical carcinogenesis	1
hsa00860	Porphyrin and chlorophyll metabolism	1
hsa04972	Pancreatic secretion	1
hsa04911	Insulin secretion	1
hsa03010	Ribosome	1
hsa00760	Nicotinate and nicotinamide metabolism	1
hsa00730	Thiamine metabolism	1
hsa00450	Selenocompound metabolism	1
hsa04922	Glucagon signaling pathway	1
hsa05217	Basal cell carcinoma	1
hsa00270	Cysteine and methionine metabolism	1
hsa00120	Primary bile acid biosynthesis	1
hsa04215	Apoptosis - multiple species	1
hsa04976	Bile secretion	1
hsa04141	Protein processing in endoplasmic reticulum	1
hsa04070	Phosphatidylinositol signaling system	1
hsa04923	Regulation of lipolysis in adipocytes	1
hsa04136	Autophagy - other	1
hsa00900	Terpenoid backbone biosynthesis	1
hsa04970	Salivary secretion	1
hsa04610	Complement and coagulation cascades	1
hsa04721	Synaptic vesicle cycle	1

Supplement table 6 Drugs applied against Lung Carcinoma. The first column contains FDA-approved drugs against Non-Small Cell Lung Cancer (58 drugs). The second column contains a subset of the drugs from the first column after removing duplicated ones (25 drugs). The third column contains drugs that were applied to Lung Squamous Cell Carcinoma patients in TCGA (16 drugs). The drugs marked in red are found in both column (2 and 3). Eight of sixteen drugs applied to Lung Squamous Cell Carcinoma patients in the TCGA panel were FDA-approved.

Approved drugs from Cancer.org	Compact list of approved drugs	Drugs applied for patients in TCGA
Methotrexate	Methotrexate	**Erlotinib**
Methotrexate LPF (Methotrexate)		**Paclitaxel**
Mexate (Methotrexate)		Vincristine
Mexate-AQ (Methotrexate)		**Pemetrexed**
Abitrexate (Methotrexate)		Topotecan
Folex (Methotrexate)		Doxorubicin
Folex PFS (Methotrexate)		**Gemcitabine**
Paclitaxel	**Paclitaxel**	Oxaliplatin
Paclitaxel Albumin-stabilized Nanoparticle Formulation		**Docetaxel**
Abraxane (Paclitaxel Albumin-stabilized Nanoparticle Formulation)		**Gefitinib**
Taxol (Paclitaxel)		**Vinorelbine**
Afatinib Dimaleate	Afatinib Dimaleate	**Carboplatin**
Gilotrif (Afatinib Dimaleate)		Irinotecan
Everolimus	Everolimus	Cisplatin
Afinitor (Everolimus)		Temozolomide
Alectinib	Alectinib	Etoposide
Alecensa (Alectinib)		
Pemetrexed Disodium	**Pemetrexed Disodium**	
Alimta (Pemetrexed Disodium)		
Brigatinib	Brigatinib	
Alunbrig (Brigatinib)		
Atezolizumab	Atezolizumab	
Tecentriq (Atezolizumab)		
Bevacizumab	Bevacizumab	
Avastin (Bevacizumab)		
Carboplatin	**Carboplatin**	
Paraplat (Carboplatin)		
Paraplatin (Carboplatin)		
Ceritinib	Ceritinib	
Zykadia (Ceritinib)		
Crizotinib	Crizotinib	
Xalkori (Crizotinib)		
Ramucirumab	Ramucirumab	
Cyramza (Ramucirumab)		
Dabrafenib	Dabrafenib	
Tafinlar (Dabrafenib)		

Docetaxel Taxotere (Docetaxel)	**Docetaxel**
Erlotinib Hydrochloride Tarceva (Erlotinib Hydrochloride)	**Erlotinib Hydrochloride**
Gemcitabine Hydrochloride Gemzar (Gemcitabine Hydrochloride)	**Gemcitabine Hydrochloride**
Gefitinib Iressa (Gefitinib)	**Gefitinib**
Pembrolizumab Keytruda (Pembrolizumab)	Pembrolizumab
Mechlorethamine Hydrochloride Mustargen (Mechlorethamine Hydrochloride)	Mechlorethamine Hydrochloride
Trametinib Mekinist (Trametinib)	Trametinib
Vinorelbine Tartrate Navelbine (Vinorelbine Tartrate)	**Vinorelbine Tartrate**
Necitumumab Portrazza (Necitumumab)	Necitumumab
Nivolumab Opdivo (Nivolumab)	Nivolumab
Osimertinib Tagrisso (Osimertinib)	Osimertinib

According to https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/what-is-non-small-cell-lung-cancer.html, Lung Squamous Cell Carcinoma is a sub-type of Non-Small Cell Lung Cancer.

Supplement table 7 Drugs against breast cancer. The first column contains FDA-approved drugs against breast cancer (71), the second column contains the compact list after removing duplicates (31 drugs), and the third column contains drugs that were applied to patients in the TCGA panel (38). The drugs marked in red are found in both column (2 and 3). 23 out of 38 drugs applied to TCGA-patients were FDA-approved drugs.

Approved drugs from Cancer.org	Compact list of approved drugs	Drugs applied for patients in TCGA
Abemaciclib	Abemaciclib	**Tamoxifen**
Verzenio (Abemaciclib)		**Anastrozole**
Methotrexate	**Methotrexate**	**Paclitaxel**
Abitrexate (Methotrexate)		**Toremifene**
Folex (Methotrexate)		Vincristine
Folex PFS (Methotrexate)		**Fluorouracil**
Methotrexate LPF (Methotrexate)		**Capecitabine**
Mexate (Methotrexate)		**Doxorubicin**
Mexate-AQ (Methotrexate)		**Letrozole**
Paclitaxel	**Paclitaxel**	Pegfilgrastim
Paclitaxel Albumin-stabilized Nanoparticle Formulation		**Vinblastine**
Taxol (Paclitaxel)		**Trastuzumab**
Abraxane (Paclitaxel Albumin-stabilized Nanoparticle Formulation)		**Cyclophosphamide**
Everolimus	**Everolimus**	Prednisone
Afinitor (Everolimus)		**Gemcitabine**
Anastrozole	**Anastrozole**	Carboplatin
Arimidex (Anastrozole)		**Megestrol acetate**
Exemestane	**Exemestane**	Rituximab
Aromasin (Exemestane)		**Ixabepilone**
Capecitabine	**Capecitabine**	Cisplatin
Xeloda (Capecitabine)		Bevacizumab
Cyclophosphamide	**Cyclophosphamide**	Ifosfamide
Clafen (Cyclophosphamide)		Triptorelin
Cytoxan (Cyclophosphamide)		**Epirubicin**
Neosar (Cyclophosphamide)		**Exemestane**
Docetaxel	**Docetaxel**	Pemetrexed
Taxotere (Docetaxel)		**Pamidronate**
Doxorubicin Hydrochloride	**Doxorubicin Hydrochloride**	**Goserelin**
Epirubicin Hydrochloride	**Epirubicin Hydrochloride**	**Lapatinib**
Ellence (Epirubicin Hydrochloride)		**Methotrexate**
Eribulin Mesylate	Eribulin Mesylate	**Everolimus**
Halaven (Eribulin Mesylate)		Mitomycin
Pamidronate Disodium	**Pamidronate Disodium**	**Docetaxel**
Aredia (Pamidronate Disodium)		Leuprolide
Toremifene	**Toremifene**	Trabectedin
Fareston (Toremifene)		Vinorelbine
Fulvestrant	**Fulvestrant**	**Fulvestrant**

Faslodex (Fulvestrant)		Mitoxantrone
Letrozole	**Letrozole**	
Femara (Letrozole)		
Fluorouracil Injection	**Fluorouracil Injection**	
5-FU (Fluorouracil Injection)		
Gemcitabine Hydrochloride	**Gemcitabine Hydrochloride**	
Gemzar (Gemcitabine Hydrochloride)		
Goserelin Acetate	**Goserelin Acetate**	
Zoladex (Goserelin Acetate)		
Palbociclib	Palbociclib	
Ibrance (Palbociclib)		
Ixabepilone	**Ixabepilone**	
Ixempra (Ixabepilone)		
Ribociclib	Ribociclib	
Kisqali (Ribociclib)		
Lapatinib Ditosylate	**Lapatinib Ditosylate**	
Tykerb (Lapatinib Ditosylate)		
Olaparib	Olaparib	
Lynparza (Olaparib)		
Megestrol Acetate	**Megestrol Acetate**	
Neratinib Maleate	Neratinib Maleate	
Nerlynx (Neratinib Maleate)		
Tamoxifen Citrate	**Tamoxifen Citrate**	
Nolvadex (Tamoxifen Citrate)		
Pertuzumab	Pertuzumab	
Perjeta (Pertuzumab)		
Thiotepa	Thiotepa	
Trastuzumab	**Trastuzumab**	
Kadcyla (Ado-Trastuzumab Emtansine)		
Herceptin (Trastuzumab)		
Ado-Trastuzumab Emtansine		
Vinblastine Sulfate	**Vinblastine Sulfate**	
Velban (Vinblastine Sulfate)		
Velsar (Vinblastine Sulfate)		

Supplement table 8 Treatment history of TCGA patients

Disease	Nr. Of patient with treatment history	Nr. Of patient without treatment history	Nr. Of patient with treatment history information is not available
Breast Invasive Carcinoma	13	1079	2
Glioblastoma Multiforme	21	568	1
Ovarian Serous Cystadenocarcinoma	1	569	0
Uterine Corpus Endometrial Carcinoma	2	538	0
Renal Clear Cell Carcinoma	18	512	0
Head and Neck Squamous Cell Carcinoma	9	508	0
Brain Lower Grade Glioma	3	511	0
Thyroid Carcinoma	5	500	0
Lung Squamous Cell Carcinoma	7	496	0
Prostate Adenocarcinoma	2	495	0
Colon Adenocarcinoma	3	447	0
Stomach Adenocarcinoma	0	442	0
Bladder Urothelial Carcinoma	10	402	0
Liver Hepatocellular Carcinoma	1	374	0
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	0	295	0
Renal Papillary Cell Carcinoma	0	290	0
Sarcoma	0	259	1
Acute Myeloid Leukemia	38	105	0
Esophageal Carcinoma	0	184	0
Pancreatic Adenocarcinoma	1	183	0
Pheochromocytoma and Paraganglioma	1	177	0
Rectum Adenocarcinoma	1	163	0
Testicular Germ Cell Tumors	0	134	0
Thymoma	2	122	0
Kidney Chromophobe	0	66	0
Adrenocortical Carcinoma	1	89	0
Mesothelioma	1	86	0
Uveal Melanoma	0	80	0
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	1	47	0
Uterine Carcinosarcoma	0	56	0
Cholangiocarcinoma	1	35	0

Supplement table 9 Amplifications and deletions of genes annotated with various cancer hallmarks. For this, we annotated human genes following Suzuki et al [63] *according to ten hallmarks of cancer: Activating Invasion and Metastasis, Resisting Cell Death, Evading Growth Suppressors, Avoiding Immune Destruction, Inducing Angiogenesis, Deregulating Cellular Energetics, Genome Instability and Mutation, Tumor Promoting Inflammation, Enabling Replicative Immortality, Sustaining Proliferative Signaling. For each hallmark, we retrieved the genes that were annotated by GO terms of the hallmark. We then checked the number of genes that were affected by amplifications and deletions.*

Hallmark	*p*-value (Wilcoxon)	*p*-value (T-test)
Activating Invasion and Metastasis	2.41E-08	3.52E-09
Resisting Cell Death	1.09E-08	1.50E-09
Evading Growth Suppressors	5.37E-09	2.91E-10
Avoiding Immune Destruction	0.00029905	0.000314
Inducing Angiogenesis	1.72E-08	3.06E-09
Deregulating Cellular Energetics	1.73E-07	6.77E-08
Genome Instability and Mutation	2.36E-09	7.68E-11
Tumor Promoting Inflammation	6.88E-09	1.27E-10
Enabling Replicative Immortality	5.66E-08	4.99E-08
Sustaining Proliferative Signaling	4.16E-08	9.16E-09

Supplement table 10 Number of cases when the specific AN targets were affected by CNAs.

Disease	Number of all cases	Number of amplified cases	Percentage of amplified cases	Number of deleted cases	Percentage of deleted cases
Breast Invasive Carcinoma	1094	596	54.5%	662	60.5%
Glioblastoma Multiforme	590	528	89.5%	170	28.8%
Ovarian Serous Cystadenocarcinoma	570	316	55.4%	526	92.3%
Uterine Corpus Endometrial Carcinoma	540	0	0.00%	206	38.2%
Renal Clear Cell Carcinoma	530	358	67.6%	468	88.3%
Head and Neck Squamous Cell Carcinoma	517	155	30.0%	321	62.1%
Brain Lower Grade Glioma	514	147	28.6%	295	57.4%
Thyroid Carcinoma	505	0	0.0%	0	0%
Lung Squamous Cell Carcinoma	503	291	57.9%	349	69.4%
Prostate Adenocarcinoma	497	0	0.0%	109	21.9%
Colon Adenocarcinoma	450	81	18.0%	268	59.6%
Stomach Adenocarcinoma	442	0	0.0%	128	29.0%
Bladder Urothelial Carcinoma	412	196	47.6%	227	55.1%
Liver Hepatocellular Carcinoma	375	0	0.0%	273	72.8%
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	295	0	0.0%	182	61.7%
Renal Papillary Cell Carcinoma	290	41	14.1%	91	31.4%
Sarcoma	260	196	75.4%	225	86.5%
Acute Myeloid Leukemia	143	0	0.0%	0	0%
Esophageal Carcinoma	184	0	0.0%	146	79.4%
Pancreatic Adenocarcinoma	184	40	21.7%	19	10.3%
Pheochromocytoma and Paraganglioma	178	0	0.0%	80	44.9%
Rectum Adenocarcinoma	164	69	42.1%	0	0%
Testicular Germ Cell Tumors	134	0	0%	0	0%
Thymoma	124	0	0%	0	0%
Kidney Chromophobe	66	0	0%	0	0%
Adrenocortical Carcinoma	90	0	0%	60	66.7%
Mesothelioma	87	0	0%	68	78.2%
Uveal Melanoma	80	0	0%	0	0%
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	48	0	0%	17	35.4%
Uterine Carcinosarcoma	56	0	0%	52	92.9%
Cholangiocarcinoma	36	0	0%	0	0%

Supplement table 11 Pathways that have similar size with four AN resistance pathways

Pathway ID	Pathway name	Number of genes
path:hsa04260	Cardiac muscle contraction - Homo sapiens (human)	78
path:hsa05220	Chronic myeloid leukemia - Homo sapiens (human)	78
path:hsa00983	Drug metabolism - other enzymes - Homo sapiens (human)	79
path:hsa01521	***EGFR tyrosine kinase inhibitor resistance - Homo sapiens (human)***	***79***
path:hsa03018	RNA degradation - Homo sapiens (human)	79
path:hsa04610	Complement and coagulation cascades - Homo sapiens (human)	79
path:hsa01522	***Endocrine resistance - Homo sapiens (human)***	***96***
path:hsa04713	Circadian entrainment - Homo sapiens (human)	96
path:hsa04925	Aldosterone synthesis and secretion - Homo sapiens (human)	96
path:hsa04972	Pancreatic secretion - Homo sapiens (human)	96
path:hsa05146	Amoebiasis - Homo sapiens (human)	96
path:hsa00052	Galactose metabolism - Homo sapiens (human)	31
path:hsa00410	beta-Alanine metabolism - Homo sapiens (human)	31
path:hsa00512	Mucin type O-glycan biosynthesis - Homo sapiens (human)	31
path:hsa01523	***Antifolate resistance - Homo sapiens (human)***	***31***
path:hsa04710	Circadian rhythm - Homo sapiens (human)	31
path:hsa05310	Asthma - Homo sapiens (human)	31
path:hsa03320	PPAR signaling pathway - Homo sapiens (human)	72
path:hsa04520	Adherens junction - Homo sapiens (human)	72
path:hsa05210	Colorectal cancer - Homo sapiens (human)	72
path:hsa05218	Melanoma - Homo sapiens (human)	72
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC) - Homo sapiens (human)	72
path:hsa00562	Inositol phosphate metabolism - Homo sapiens (human)	73
path:hsa01524	***Platinum drug resistance - Homo sapiens (human)***	***73***
path:hsa00980	Metabolism of xenobiotics by cytochrome P450 - Homo sapiens (human)	74
path:hsa01230	Biosynthesis of amino acids - Homo sapiens (human)	74
path:hsa04918	Thyroid hormone synthesis - Homo sapiens (human)	74
path:hsa05140	Leishmaniasis - Homo sapiens (human)	74

Supplement table 12 adjusted p-values of Wilcoxon tests

Pathway	Amplification	Deletion
hsa04260	0.00003	1
hsa05220	0.00001	1
hsa00983	0.00073	1
hsa01521	0.00009	1
hsa03018	0.00014	1
hsa04610	0.00199	1
hsa01522	0.00001	0.76282
hsa04713	0.01121	1
hsa04925	0.00004	1
hsa04972	0.00003	1
hsa05146	0.00062	1
hsa00052	1.00000	1
hsa00410	0.00055	1
hsa00512	0.14636	1
hsa01523	0.02712	1
hsa04710	0.03739	1
hsa05310	0.06720	1
hsa03320	0.00114	1
hsa04520	0.00069	1
hsa05210	0.00018	1
hsa05218	0.00003	1
hsa05412	0.00565	1
hsa00562	0.00233	1
hsa01524	0.00003	1
hsa00980	0.00130	1
hsa01230	0.00428	1
hsa04918	0.00086	1
hsa05140	0.00779	1

Supplement table 13: Treatment history of patients

Disease	Nr. of patients with treatment history	Nr. of patients without treatment history	Nr. of patients where treatment history information is not available
Breast Invasive Carcinoma	13	1079	2
Glioblastoma Multiforme	21	568	1
Ovarian Serous Cystadenocarcinoma	1	569	0
Uterine Corpus Endometrial Carcinoma	2	538	0
Renal Clear Cell Carcinoma	18	512	0
Head and Neck Squamous Cell Carcinoma	9	508	0
Brain Lower Grade Glioma	3	511	0
Thyroid Carcinoma	5	500	0
Lung Squamous Cell Carcinoma	7	496	0
Prostate Adenocarcinoma	2	495	0
Colon Adenocarcinoma	3	447	0
Stomach Adenocarcinoma	0	442	0
Bladder Urothelial Carcinoma	10	402	0
Liver Hepatocellular Carcinoma	1	374	0
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	0	295	0
Renal Papillary Cell Carcinoma	0	290	0
Sarcoma	0	259	1
Acute Myeloid Leukemia	38	105	0
Esophageal Carcinoma	0	184	0
Pancreatic Adenocarcinoma	1	183	0
Pheochromocytoma and Paraganglioma	1	177	0
Rectum Adenocarcinoma	1	163	0
Testicular Germ Cell Tumors	0	134	0
Thymoma	2	122	0
Kidney Chromophobe	0	66	0
Adrenocortical Carcinoma	1	89	0
Mesothelioma	1	86	0
Uveal Melanoma	0	80	0
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	1	47	0
Uterine Carcinosarcoma	0	56	0
Cholangiocarcinoma	1	35	0

Supplement table 14 number of cancer types affect each gene in hsa10521

Normal tissues				Tumor tissues			
Amplification		Deletion		Amplification		Deletion	
Genes	Number of diseases	Genes	Number of diseases	Genes	Number of diseases	Genes	Number of diseases
PDGFRB	1	BAD	12	IGF1R	11	NRAS	19
HGF	1	PDGFD	11	EGFR	9	FGFR2	14
NRG2	1	RPS6KB2	11	PIK3CA	7	PDGFD	12
EIF4E1B	1	KDR	10	GAS6	7	FOXO3	11
SOS1	1	HRAS	10	VEGFA	7	PIK3CD	10
EIF4E2	1	PDGFRA	10	KRAS	7	AKT1	10
MET	1	FGF2	9	FGFR3	6	EIF4E1B	10
PDGFA	1	JAK1	9	AKT3	5	HRAS	9
TGFA	1	GAB1	9	PDGFA	5	JAK1	9
PIK3R1	1	EIF4E1B	9	AKT1	4	NRG1	9
EGFR	1	SOS1	9	JAK2	4	PIK3R1	9
BCL2L11	1	EIF4E2	9	PRKCG	4	MTOR	9
PLCG2	1	TGFA	9	PDGFRA	4	PTEN	8
MAPK1	1	BCL2L11	9	PRKCA	4	AKT3	8
BRAF	1	PDGFC	9	GRB2	4	EIF4E2	8
IL6	1	EIF4E	9	EIF4E1B	3	ERBB3	8
GSK3B	0	EGF	9	MET	3	BAD	8
ARAF	0	FGFR3	9	KDR	2	BRAF	8
PTEN	0	PDGFRB	8	IL6R	2	BCL2	8
PIK3CD	0	NRG2	8	GAB1	2	FGF2	7
PIK3CB	0	IL6R	7	SHC1	2	SHC4	7
FGF2	0	SHC1	7	PDGFC	2	EIF4E	7
IGF1R	0	PIK3R3	7	STAT3	2	KRAS	7
AKT2	0	PIK3R1	7	NF1	2	PDGFRB	6
AKT3	0	NRAS	7	PIK3CD	1	MAP2K2	6
KDR	0	NF1	7	PIK3CB	1	GAS6	6
AKT1	0	PIK3CD	6	FGF2	1	PIK3R3	6
JAK2	0	AKT3	6	MAP2K1	1	PDGFC	6
HRAS	0	ERBB2	6	HGF	1	EIF4EBP1	6
IL6R	0	STAT3	6	SHC2	1	PLCG2	6
JAK1	0	MTOR	6	SHC3	1	VEGFA	6
PRKCG	0	GSK3B	5	PIK3R2	1	RPS6KB2	6
PDGFRA	0	PTEN	5	FOXO3	1	IL6R	5
MAP2K1	0	PIK3CB	5	NRAS	1	PRKCG	5
MAP2K2	0	JAK2	5	BCL2L11	1	NRG2	5
PRKCB	0	NRG1	5	ERBB2	1	RAF1	5
RPS6	0	PRKCA	5	EIF4EBP1	1	SOS2	5
GAB1	0	FOXO3	5	MAPK1	1	SHC2	5
NRG1	0	ERBB3	5	BRAF	1	SHC1	5

PRKCA	0	EIF4EBP1	5	MTOR	1	EGF	5
PIK3CA	0	IGF1	5	RPS6KB2	1	GSK3B	4
GAS6	0	RPS6KB1	5	BAX	1	PIK3CB	4
RAF1	0	BCL2	5	FGFR2	1	JAK2	4
SOS2	0	GRB2	5	GSK3B	0	PIK3CA	4
SHC4	0	FGFR2	5	ARAF	0	SHC3	4
SHC2	0	HGF	4	PTEN	0	PDGFA	4
SHC3	0	PIK3CA	4	AKT2	0	PIK3R2	4
SHC1	0	RAF1	4	HRAS	0	STAT3	4
SRC	0	MET	4	JAK1	0	IGF1	4
PDGFB	0	PDGFA	4	PDGFRB	0	BAX	4
PIK3R3	0	EGFR	4	MAP2K2	0	AKT2	3
PIK3R2	0	BRAF	4	PRKCB	0	GAB1	3
FOXO3	0	VEGFA	4	RPS6	0	SOS1	3
NRAS	0	IL6	4	NRG1	0	MET	3
ERBB3	0	KRAS	4	NRG2	0	SRC	3
PDGFD	0	RPS6	3	RAF1	0	ERBB2	3
ERBB2	0	SHC3	3	SOS1	0	PLCG1	3
PDGFC	0	PDGFB	3	EIF4E2	0	AXL	3
EIF4EBP1	0	MAPK1	3	SOS2	0	GRB2	3
PLCG1	0	BCL2L1	3	SHC4	0	BCL2L1	3
EIF4E	0	IGF1R	2	SRC	0	IGF1R	2
MAPK3	0	AKT2	2	PDGFB	0	KDR	2
EGF	0	AKT1	2	TGFA	0	PDGFRA	2
BAD	0	PRKCG	2	PIK3R3	0	PRKCB	2
STAT3	0	MAP2K1	2	PIK3R1	0	RPS6	2
IGF1	0	MAP2K2	2	ERBB3	0	PDGFB	2
MTOR	0	GAS6	2	PDGFD	0	MAPK3	2
VEGFA	0	SOS2	2	PLCG2	0	MAP2K1	1
RPS6KB1	0	SHC4	2	PLCG1	0	TGFA	1
AXL	0	SHC2	2	EIF4E	0	RPS6KB1	1
RPS6KB2	0	SRC	2	MAPK3	0	NF1	1
NF1	0	PIK3R2	2	EGF	0	ARAF	0
BCL2	0	PLCG1	2	BAD	0	HGF	0
BAX	0	AXL	2	IGF1	0	PRKCA	0
GRB2	0	BAX	2	IL6	0	EGFR	0
KRAS	0	PRKCB	1	RPS6KB1	0	BCL2L11	0
FGFR3	0	PLCG2	1	AXL	0	MAPK1	0
FGFR2	0	MAPK3	1	BCL2	0	IL6	0
BCL2L1	0	ARAF	0	BCL2L1	0	FGFR3	0

Supplement table 15 number of cancer types affect each gene in hsa10522

Normal tissues				Tumor tissues			
Amplification		Deletion		Amplification		Deletion	
Genes	Number of diseases	Genes	Number of diseases	Genes	Number of diseases	Genes	Number of diseases
SOS1	1	BAD	12	IGF1R	11	CDKN2A	21
HBEGF	1	CCND1	11	EGFR	9	NRAS	19
ADCY3	1	RPS6KB2	11	MDM2	9	PRKACB	16
ADCY2	1	HRAS	10	PIK3CA	7	MAPK12	15
ADCY1	1	SOS1	9	KRAS	7	MAPK11	15
PIK3R1	1	ADCY3	9	AKT3	5	DLL4	14
ABCB11	1	ABCB11	9	AKT1	4	RB1	13
EGFR	1	MAPK10	9	MMP9	4	CDKN1B	12
MAPK9	1	PRKACB	8	NOTCH3	4	NOTCH2	12
MAPK1	1	HBEGF	8	PTK2	4	PIK3CD	10
BRAF	1	JUN	8	GRB2	4	AKT1	10
RB1	0	NOTCH2	7	NOTCH2	3	FOS	10
CDKN1A	0	SHC1	7	ADCY8	3	DLL1	10
CDKN1B	0	PIK3R3	7	MAPK9	3	MAPK8	10
ARAF	0	ADCY2	7	GNAS	3	E2F2	10
PIK3CD	0	PIK3R1	7	SHC1	2	JAG2	10
PIK3CB	0	MAPK9	7	ADCY4	2	ESR2	10
IGF1R	0	NRAS	7	JAG2	2	HRAS	9
CCND1	0	CDKN2C	7	CDKN1A	1	ADCY4	9
CYP2D6	0	PIK3CD	6	PIK3CD	1	PIK3R1	9
AKT2	0	AKT3	6	PIK3CB	1	ADCY6	9
GPER1	0	MED1	6	CCND1	1	MAPK9	9
PRKACG	0	NCOR1	6	PRKACG	1	MTOR	9
AKT3	0	ERBB2	6	PRKACB	1	AKT3	8
AKT1	0	E2F2	6	MED1	1	BAD	8
PRKACA	0	MTOR	6	MAP2K1	1	BRAF	8
HRAS	0	PIK3CB	5	FOS	1	SP1	8
PRKACB	0	TP53	5	SHC2	1	BCL2	8
MED1	0	ADCY8	5	NOTCH1	1	SHC4	7
MAP2K1	0	ADCY6	5	SHC3	1	JUN	7
MAP2K2	0	ADCY5	5	PIK3R2	1	CDKN2C	7
MMP2	0	MAPK8	5	ADCY1	1	CDK4	7
NCOA3	0	IGF1	5	ABCB11	1	MDM2	7
FOS	0	RPS6KB1	5	DLL1	1	KRAS	7
MMP9	0	CDK4	5	NRAS	1	CCND1	6
ADCY9	0	SP1	5	ERBB2	1	PRKACG	6
NCOR1	0	MDM2	5	E2F1	1	MAP2K2	6
null*	0	BCL2	5	MAPK1	1	PIK3R3	6
PIK3CA	0	GRB2	5	E2F3	1	E2F3	6

RAF1	0	CDKN1A	4	JUN	1	ESR1	6
SOS2	0	CDKN1B	4	JAG1	1	RPS6KB2	6
TP53	0	PIK3CA	4	BRAF	1	CDKN1A	5
SHC4	0	RAF1	4	MAPK14	1	RAF1	5
NOTCH2	0	ADCY1	4	MTOR	1	SOS2	5
NOTCH3	0	EGFR	4	RPS6KB2	1	HBEGF	5
SHC2	0	BRAF	4	BAX	1	SHC2	5
NOTCH1	0	MAPK14	4	RB1	0	SHC1	5
SHC3	0	PTK2	4	CDKN1B	0	NOTCH4	5
SHC1	0	MAPK13	4	ARAF	0	ADCY3	5
SRC	0	KRAS	4	CYP2D6	0	ADCY7	5
NOTCH4	0	CYP2D6	3	AKT2	0	BIK	5
ADCY4	0	PRKACG	3	GPER1	0	MAPK14	5
PIK3R3	0	NOTCH1	3	PRKACA	0	MAPK13	5
PIK3R2	0	SHC3	3	HRAS	0	MAPK10	5
ADCY8	0	NOTCH4	3	MAP2K2	0	PIK3CB	4
ADCY7	0	DLL1	3	MMP2	0	CYP2D6	4
DLL1	0	MAPK1	3	NCOA3	0	PIK3CA	4
ADCY6	0	E2F3	3	ADCY9	0	SHC3	4
DLL3	0	CDKN2A	3	NCOR1	0	PIK3R2	4
ADCY5	0	BIK	3	null	0	ADCY5	4
DLL4	0	ESR1	3	RAF1	0	IGF1	4
NRAS	0	ESR2	3	SOS1	0	PTK2	4
MAPK8	0	MAPK12	3	SOS2	0	BAX	4
ERBB2	0	MAPK11	3	TP53	0	AKT2	3
E2F1	0	RB1	2	HBEGF	0	PRKACA	3
E2F2	0	IGF1R	2	SHC4	0	MED1	3
E2F3	0	AKT2	2	SRC	0	MMP2	3
MAPK3	0	AKT1	2	NOTCH4	0	NCOA3	3
JAG2	0	PRKACA	2	ADCY3	0	MMP9	3
JUN	0	MAP2K1	2	PIK3R3	0	SOS1	3
CDKN2C	0	MAP2K2	2	ADCY2	0	NOTCH3	3
JAG1	0	NCOA3	2	PIK3R1	0	SRC	3
CDKN2A	0	FOS	2	ADCY7	0	ADCY8	3
BAD	0	MMP9	2	ADCY6	0	ABCB11	3
BIK	0	SOS2	2	DLL3	0	DLL3	3
IGF1	0	SHC4	2	ADCY5	0	ERBB2	3
MAPK14	0	NOTCH3	2	DLL4	0	E2F1	3
ESR1	0	SHC2	2	MAPK8	0	JAG1	3
ESR2	0	SRC	2	E2F2	0	CARM1	3
MTOR	0	ADCY4	2	MAPK3	0	GNAS	3
PTK2	0	PIK3R2	2	CDKN2C	0	GRB2	3
MAPK12	0	DLL3	2	CDKN2A	0	IGF1R	2
MAPK13	0	DLL4	2	BAD	0	TP53	2
MAPK10	0	E2F1	2	BIK	0	NOTCH1	2
MAPK11	0	JAG2	2	IGF1	0	MAPK3	2

RPS6KB1	0	JAG1	2	ESR1	0	MAP2K1	1
CDK4	0	CARM1	2	ESR2	0	ADCY9	1
SP1	0	GNAS	2	MAPK12	0	NCOR1	1
CARM1	0	BAX	2	MAPK13	0	ADCY2	1
RPS6KB2	0	MMP2	1	MAPK10	0	RPS6KB1	1
GNAS	0	ADCY9	1	MAPK11	0	ARAF	0
MDM2	0	ADCY7	1	RPS6KB1	0	GPER1	0
BCL2	0	MAPK3	1	CDK4	0	null	0
BAX	0	ARAF	0	SP1	0	ADCY1	0
GRB2	0	GPER1	0	CARM1	0	EGFR	0
KRAS	0	null	0	BCL2	0	MAPK1	0

* null indicates that the gene in KEGG pathway has no corresponding gene symbol

Supplement table 16 number of cancer types affect each gene in hsa10523

Normal tissues				Tumor tissues			
Amplification		Deletion		Amplification		Deletion	
Genes	Number of diseases	Genes	Number of diseases	Genes	Number of diseases	Genes	Number of diseases
ATIC	1	RELA	11	IKBKB	6	ABCC2	13
DHFR	1	FOLR3	11	ABCC5	4	CHUK	13
IL6	1	FOLR2	11	TYMS	3	MTHFR	9
IL1B	1	FOLR1	11	GGH	3	SLC46A1	7
SLC46A1	0	ATIC	9	ABCC4	2	SHMT2	7
DHFR2	0	NFKB1	9	SLC46A1	1	RELA	7
SHMT2	0	IL1B	9	SHMT1	1	ABCC4	7
SHMT1	0	ABCG2	9	MTHFR	1	NFKB1	7
ALOX12	0	SLC46A1	7	IL1B	1	DHFR	7
TYMS	0	DHFR	7	DHFR2	0	FOLR3	7
TNF	0	SHMT1	6	SHMT2	0	FOLR2	7
RELA	0	MTHFR	6	ALOX12	0	FOLR1	7
IKBKB	0	SHMT2	5	TNF	0	IKBKB	6
FPGS	0	ALOX12	5	RELA	0	SLC19A1	6
IKBKG	0	TYMS	5	ATIC	0	TNF	5
SLC19A1	0	IKBKB	5	FPGS	0	ATIC	5
ABCC3	0	ABCC3	5	IKBKG	0	GGH	5
ABCC4	0	ABCC2	5	SLC19A1	0	ABCG2	5
ABCC1	0	CHUK	5	ABCC3	0	TYMS	3
ABCC2	0	GGH	5	ABCC1	0	ABCC5	3
CHUK	0	ABCC5	4	ABCC2	0	ALOX12	2
ABCC5	0	IL6	4	CHUK	0	FPGS	2
MTHFR	0	TNF	3	IZUMO1R	0	ABCC3	1
GGH	0	FPGS	3	NFKB1	0	ABCC1	1
IZUMO1R	0	ABCC4	2	DHFR	0	GART	1
NFKB1	0	ABCC1	1	IL6	0	DHFR2	0
FOLR3	0	DHFR2	0	FOLR3	0	SHMT1	0
FOLR2	0	IKBKG	0	FOLR2	0	IKBKG	0
FOLR1	0	SLC19A1	0	FOLR1	0	IZUMO1R	0
GART	0	IZUMO1R	0	GART	0	IL6	0
ABCG2	0	GART	0	ABCG2	0	IL1B	0

Supplement table 17 number of cancer types affect each gene in hsa10524

Normal tissues				Tumor tissues			
Amplification		Deletion		Amplification		Deletion	
Genes	Number of diseases	Genes	Number of diseases	Genes	Number of diseases	Genes	Number of diseases
CASP8	1	BAD	12	MDM2	9	CDKN2A	21
MSH6	1	BIRC2	11	PIK3CA	7	GSTO2	16
MSH2	1	BIRC3	11	FASLG	5	GSTO1	16
MSH3	1	GSTP1	11	AKT3	5	ATM	16
PIK3R1	1	FADD	11	POLH	5	GSTM4	15
MAPK1	1	ATM	11	AKT1	4	GSTM3	15
BID	1	CASP8	9	BIRC5	4	GSTM2	15
CYCS	1	CASP3	9	MGST3	4	GSTM1	15
TOP2A	0	MSH6	9	BID	4	GSTM5	15
TOP2B	0	MSH2	9	BIRC2	3	ABCC2	13
CDKN1A	0	MGST2	9	BIRC3	3	PIK3CD	10
PIK3CD	0	GSTM4	8	GSTP1	3	REV3L	10
FASLG	0	GSTM3	8	GSTA5	3	AKT1	10
PIK3CB	0	GSTM2	8	GSTA4	3	CASP3	9
BRCA1	0	GSTM1	8	GSTA3	3	MGST1	9
BBC3	0	GSTM5	8	GSTA2	3	PIK3R1	9
CASP9	0	FASLG	7	GSTA1	3	CASP9	8
ATP7B	0	MSH3	7	TOP2A	2	ATP7B	8
AKT2	0	MGST3	7	CASP3	2	AKT3	8
CASP3	0	PIK3R3	7	MGST2	2	BIRC2	8
AKT3	0	PIK3R1	7	FADD	2	BIRC3	8
REV3L	0	TOP2A	6	CDKN1A	1	BAD	8
AKT1	0	PIK3CD	6	PIK3CD	1	BCL2	8
ATP7A	0	BRCA1	6	PIK3CB	1	MAP3K5	7
POLH	0	CASP9	6	CASP8	1	MSH3	7
MAP3K5	0	AKT3	6	REV3L	1	GSTA5	7
ABCC2	0	GSTO2	6	MAP3K5	1	GSTA4	7
GSTO2	0	GSTO1	6	MSH6	1	GSTA3	7
APAF1	0	ERBB2	6	MSH2	1	GSTA2	7
PDPK1	0	PIK3CB	5	MGST1	1	GSTA1	7
GSTO1	0	REV3L	5	XPA	1	MDM2	7
PIK3CA	0	ABCC2	5	GSTT2	1	FAS	7
BIRC5	0	APAF1	5	GSTT1	1	BBC3	6
TP53	0	BIRC5	5	PIK3R2	1	POLH	6
BIRC2	0	TP53	5	ERBB2	1	APAF1	6
BIRC3	0	PMAIP1	5	MAPK1	1	GSTP1	6
GSTP1	0	MDM2	5	GSTT2B	1	MGST3	6
MGST3	0	BCL2	5	GSTM4	1	PIK3R3	6
MGST1	0	FAS	5	GSTM3	1	FADD	6

XPA	0	TOP2B	4	GSTM2	1	CDKN1A	5
GSTT2	0	CDKN1A	4	GSTM1	1	FASLG	5
XIAP	0	POLH	4	ERCC1	1	PMAIP1	5
MGST2	0	PIK3CA	4	BAX	1	BAK1	5
PIK3R3	0	MGST1	4	GSTM5	1	PIK3CB	4
GSTT1	0	MLH1	4	TOP2B	0	BRCA1	4
PIK3R2	0	GSTA5	4	BRCA1	0	PIK3CA	4
ERBB2	0	GSTA4	4	BBC3	0	PIK3R2	4
PMAIP1	0	GSTA3	4	CASP9	0	BAX	4
BAK1	0	GSTA2	4	ATP7B	0	TOP2A	3
FADD	0	GSTA1	4	AKT2	0	TOP2B	3
GSTT2B	0	CYCS	4	ATP7A	0	AKT2	3
MAPK3	0	MAP3K5	3	ABCC2	0	MSH6	3
GSTM4	0	XPA	3	GSTO2	0	MSH2	3
GSTM3	0	GSTT2	3	APAF1	0	MGST2	3
GSTM2	0	GSTT1	3	PDPK1	0	ERBB2	3
GSTM1	0	MAPK1	3	GSTO1	0	ERCC1	3
CDKN2A	0	BAK1	3	MSH3	0	BCL2L1	3
SLC31A1	0	BID	3	TP53	0	CASP8	2
BAD	0	GSTT2B	3	XIAP	0	BIRC5	2
MLH1	0	CDKN2A	3	PIK3R3	0	TP53	2
GSTA5	0	SLC31A1	3	PIK3R1	0	XPA	2
GSTA4	0	BCL2L1	3	PMAIP1	0	MAPK3	2
GSTA3	0	BBC3	2	BAK1	0	MLH1	2
ERCC1	0	ATP7B	2	MAPK3	0	PDPK1	1
GSTA2	0	AKT2	2	CDKN2A	0	BID	1
GSTA1	0	AKT1	2	SLC31A1	0	ATP7A	0
MDM2	0	PDPK1	2	BAD	0	GSTT2	0
BCL2	0	PIK3R2	2	MLH1	0	XIAP	0
FAS	0	ERCC1	2	BCL2	0	GSTT1	0
BAX	0	BAX	2	FAS	0	MAPK1	0
ATM	0	MAPK3	1	CYCS	0	GSTT2B	0
GSTM5	0	ATP7A	0	ATM	0	SLC31A1	0
BCL2L1	0	XIAP	0	BCL2L1	0	CYCS	0