Knowledge Management Approaches for predicting Biomarker and Assessing its Impact on Clinical Trials

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Avisek Deyati

aus

West Bengal, INDIA

Bonn 2016

Angefertigt mit Genehmigung

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinichen Friedrich-Wilhelms-Universität Bonn

1. **Gutachter: Prof. Dr. Martin Hofmann-Apitius**
2. **Gutachter: Prof. Dr. Sven Burgdorf**

Tag der Promotion: 13.09.2016
Erscheinungsjahr: 2016

Abstract

The recent success of companion diagnostics along with the increasing regulatory pressure for better identification of the target population has created an unprecedented incentive for the drug discovery companies to invest into novel strategies for stratified biomarker discovery. Catching with this trend, trials with stratified biomarker in drug development have quadrupled in the last decade but represent a small part of all Interventional trials reflecting multiple co-developmental challenges of therapeutic compounds and companion diagnostics. To overcome the challenge, varied knowledge management and system biology approaches are adopted in the clinics to analyze/interpret an ever increasing collection of OMICS data. By semi-automatic screening of more than 150,000 trials, we filtered trials with stratified biomarker to analyse their therapeutic focus, major drivers and elucidated the impact of stratified biomarker programs on trial duration and completion. The analysis clearly shows that cancer is the major focus for trials with stratified biomarker. But targeted therapies in cancer require more accurate stratification of patient population. This can be augmented by a fresh approach of selecting a new class of biomolecules i.e. miRNA as candidate stratification biomarker. miRNA plays an important role in tumorgenesis in regulating expression of oncogenes and tumor suppressors; thus affecting cell proliferation, differentiation, apoptosis, invasion, angiogenesis. miRNAs are potential biomarkers in different cancer. However, the relationship between response of cancer patients towards targeted therapy and resulting modifications of the miRNA transcriptome in pathway regulation is poorly understood. With ever-increasing pathways and miRNA-mRNA interaction databases, freely available mRNA and miRNA expression data in multiple cancer therapy have created an unprecedented opportunity to decipher the role of miRNAs in early prediction of therapeutic efficacy in diseases. We present a novel SMARTmiR algorithm to predict the role of miRNA as therapeutic biomarker for an anti-EGFR monoclonal antibody i.e. cetuximab treatment in colorectal cancer. The application of an optimised and fully automated version of the algorithm has the potential to be used as clinical decision support tool. Moreover this research will also provide a comprehensive and valuable knowledge map demonstrating functional bimolecular interactions in colorectal cancer to scientific community. This research also detected seven miRNA i.e. hsa-miR-145, has-miR-27a, has-

miR-155, hsa-miR-182, hsa-miR-15a, hsa-miR-96 and hsa-miR-106a as top stratified biomarker candidate for cetuximab therapy in CRC which were not reported previously. Finally a prospective plan on future scenario of biomarker research in cancer drug development has been drawn focusing to reduce the risk of most expensive phase III drug failures.

List of Abbreviations

ANOVA	Analysis of Variance
BIND	Biomolecular interaction database Network
BioPAX	Biological Pathways Exchange
CDx	Companion Diagnostics
CellML	Cellular Modelling Markup Language
CFS	Correlation based feature selection
CIN	Chromosomal instability
CRC	Colorectal Cancer
CTG	ClinicalTrials.gov
DAE	Differential Algebraic Equations
DNA	Deoxyribonucleic Acid
FAP	Familial Adenomatous Polyposis
FU	Fluorouracil
GEO	Gene Expression Omnibus
GO	Gene Ontology
HNPCC	Hereditary non-polyposis colorectal cancer
HPRD	Human Protein Reference Database
IPA	Ingenuity Pathway Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
LDA	Linear discriminant analysis
LIMMA	Linear Models for Microarray Data
MAS5	Microarray suite 5
mCRC	Metastatic Colorectal Cancer
MGED	Microarray Gene Expression Data
MI	Molecular Interaction
MIAME	Minimum Information about Microarray Experiments
MIASE	Minimum Information About a Simulation Experiment
miR	microRNA
miRNA	microRNA
MSI	Microsatellite instability
MTIs	MiRNA Target Interactions
ODEs	Ordinary Differential Equations
PAM	Partition Around Medoids
PCA	Principal Component Analysis
PPI	Protein-Protein Interaction
PSI	Proteomic Standard Initiative
RMA	Robust Multichip Average
RNA	Ribonucleic Acid
SAM	Significance Analysis of Microarray
SBGN	Systems Biology Graphical Notation
SBML	Systems Biology Markup Language
SMARTmiR	Scoring-based MARking of Therapeutic MIcroRna

SVD	Singular Value decomposition
TCGA	The Cancer Genome Atlas
UTR	Untranslated Region
VPH	Virtual Physiological Human
XML	Extensible Markup Language

Acknowledgement

I would like to express my sincere gratitude to Prof. Dr. Martin Hofmann-Apitius for accepting and guiding me throughout the projects. His excellent supervision had helped me to mature as a researcher. I will be forever indebted to Dr. Natalia Novac for selecting me and allowing me the opportunity to carry out the research in the industrial atmosphere of Merck KgaA. Her mentoring and support at professional and personal level helped me to evolve my scientific, communication, collaboration and resource management skills. I am greatly obliged to Prof. Dr. med. Joachim L. Schultze for his role in evaluating my PhD thesis as a co-supervisor. I am sincerely thankful to Dr Holger Frohlich for all the discussions and valuable suggestions. I am also grateful to my colleagues; Dr. Philipp Senger, Dr. Erfan Younesi and Ms. Shweta Bagewadi for their scientific inputs, critical analysis which helped to shape the projects.

I would like to acknowledge Merck KgaA for awarding the scholarship and providing this great research opportunity. I would like to thank the Merck HR team for all their assistance and prompt responses regarding any queries. I am thankful to the Indian community in Darmstadt, Germany, which formed a strong support system during the PhD life abroad.

I will always be indebted to my father (Mr. Sourendranath Deyati) and mother (Mrs. Aradhana Deyati) for motivating me to have a mission in life and always supported me to chase my dreams. Heartfelt thanks to my wife Baishakhi for inspiring me throughout this crucial phase and for being the pillar of my life. Above all I would like to thank my grandfather Late Mr. Dronocharan Ghosh for all his blessings, inspirational words and inculcating self-esteem within me. I am dedicating this thesis to his loving memory.

Table of Contents

List of Figures

List of Tables

Introduction

1. Cancer

Cancer is not a single disease but a combination of many diseases. Cumulatively, we call them cancer as they share a fundamental commonality: abnormal, uncontrolled growth of cells spreading throughout the body. Different forms of cancer are highly heterogeneous in terms of histology and clinical outcome as well as at the molecular level (Majewski & Bernards 2011).

1.1. Modern theories of cancer

During the last century tremendous amount of research was undertaken in the field of cancer. These researches demonstrated that cancer is the outcome of several genetic alterations occurring and accumulating inside the cell. These alterations disrupt the balance between cell proliferation and programmed cell death. Multiple factors are associated with the oncogenic processes i.e. carcinogenic exposure, infectious agents, genomic alterations and pathway modification impacting normal cellular processes.

1.1.1. Genomics

In 1962, the discovery of DNA double helix by Watson and Crick spurred a series of discoveries on gene function and the malfunction leading to mutations. This new understanding showed the disease cause dynamic changes in the genome (Hanahan & Robert A Weinberg 2011). Two types of mutations play a crucial role in initiating cancer; one that induces oncogenes from proto-oncogene with dominant gain of function while the other in tumor suppressor genes resulting in recessive loss of function.

1.1.1.1. Proto-oncogene

Proto-oncogene is a normal cellular gene that encodes a protein belonging to functional categories of growth factor, growth factor receptor, intracellular transducers, intracellular receptors and transcription factors. It can be mutated into a cancer promoting oncogene, either by changing the protein coding segment or by altering its expression (Harvey Lodish et al, 2000). Transformation of proto-oncogene to oncogene is summarized in Figure 1.

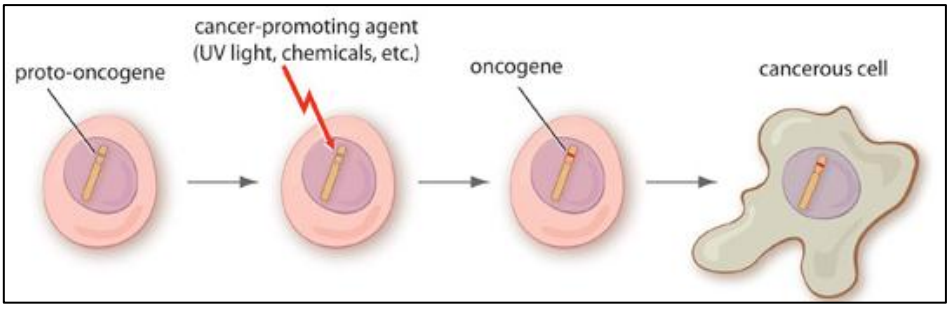


Figure 1: Transformation of proto-oncogene to oncogene and cellular fate (Harvey Lodish et al, 2000)

Normal cell can lead to be cancerous cell due to the activation of oncogenes in the presence of a carcinogen.

There are three genomic events that can trigger such a change

- Point mutation in proto-oncogene.
- Localized amplification of a chromosomal part that incorporates a proto-oncogene resulting in over expression.
- Chromosomal translocation that drives the integration of another promoter with proto-oncogene resulting in over expression.

1.1.1.2. Tumor suppressor genes

Tumor suppressor genes, also called anti-oncogene, generally encode a protein which inhibits cell proliferation. The categories of tumor suppressor genes also include cell cycle control proteins, DNA repair proteins and anti-apoptotic proteins. Losses of function mutations in these genes inhibit the cell to pause the uncontrolled growth. As only one copy of tumor suppressor genes have got the ability to control cell proliferation, both alleles of tumor suppressor genes must be lost or inactivated to promote tumor (Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore 2000).

1.1.2. Pathway centric understanding of cancer

Successive research in the last two decades discovered that the formation of cancer in human is a multistep process involving multiple proteins for sustainable proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis and activating invasion or metastasis (Figure 2). These hallmarks of cancer were proposed by Hanahan and Weinberg in 2000. Later it was again updated in 2011 by the same group. In the next section, an overview of the six hallmark of cancer is presented and discussed in the lights of recent discoveries.

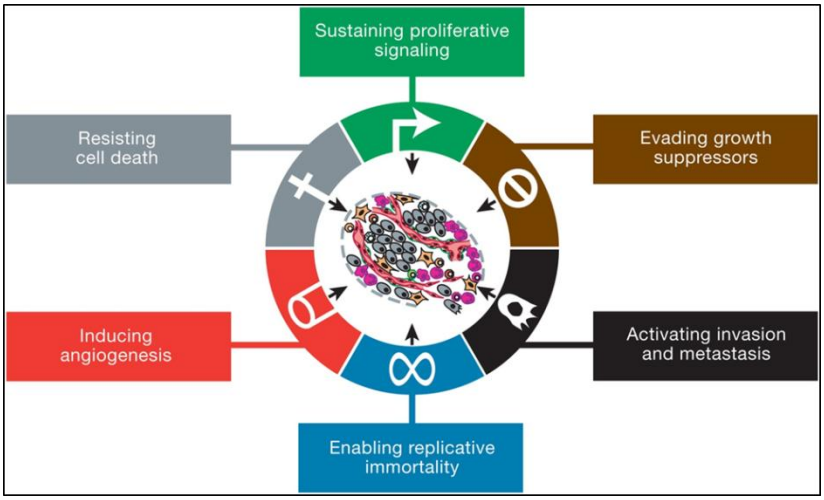


Figure 2: Hallmark of cancer (Hanahan & Robert A. Weinberg 2011)

The diagram shows six major mechanisms for the induction of cancer from normal cells, its uncontrolled growth and spread to other organs.

1.1.2.1. Sustaining proliferative signalling

The regulatory process of producing and releasing growth promoting signals are highly regulated in normal tissues. Due to this, normal tissue can maintain normal architecture and function by homeostasis of cells. Cancer cells deregulate these signals by the following means;

a. Over expression of growth factors, that binds cell surface receptors with intracellular tyrosine kinase domains.
b. By increasing surface receptors displayed at the cancer cell surface, the cell becomes hyper-responsive to the limited amount of available ligands.
c. Point mutation in the receptors results in structural alteration facilitating ligand independent firing.
d. Somatic mutations activating alternative downstream pathways lead to cell proliferation in normal cells triggered by activated growth factor receptors. The mutation in the catalytic domain of PIP3 results in hyperactive PIP3 kinase signalling including activation of Akt/PKB signal transducer (the hallmark of cancers, next generation).
e. Somatic mutations to enhance cellular proliferation by disrupting a negative feedback loop e.g. oncogenic mutations in Ras protein inhibit its GTPase activity inhibiting negative feedback loop for cellular proliferation (Hanahan & Robert A Weinberg 2011; Evan & Vousden 2001)

1.1.2.2. Evading growth suppressors

In addition to maintaining growth stimulation signal, cancer cell must also nullify the powerful negative regulation of cell proliferation by the action of tumor suppressor genes. Among a large network of molecules, two most prominent tumor suppressors are retinoblastoma-associated (RB) and TP53 proteins. They govern the decisions of cells to proliferate or activate apoptosis.

a. RB proteins integrate signals both from intracellular and extracellular sources and dictate whether a cell should proceed through its growth and division cycle. The defects in the RB pathway promote persistent cell proliferation in cancer cells. Loss of RB function is akin to the removal of a gatekeeper of cell cycle progression, hence triggering uncontrolled cellular proliferation.
b. TP53 acts within the intracellular environment to assess the degree of genome damage, the levels of nucleotide pools, growth promoting signals, level of oxygen and glucose. It permits the cell cycle to progress only if those parameters are at normal level. On the contrary, if there is irreparable damage to those subcellular components, TP53 can trigger apoptosis. A loss of function mutation in TP53 can trigger uncontrolled growth as well (Hanahan & Robert A Weinberg 2011)

1.1.2.3. Resisting cell death

Programmed cell death by apoptosis is a natural cellular mechanism to destroy the cancerous growth. Several upstream regulators and downstream effectors are the building blocks of apoptotic mechanisms. Some of the most important mechanisms of apoptosis and its blocking are listed below.

a. The regulators are divided into two network modules leading to the activation of effector proteases (caspase 8 and caspase 9). One network module processes extracellular death inducing signals i.e. Fas ligand/Fas receptors signalling. Tumor necrosis factor/TNF receptor signalling. The other one sense and integrate intracellular signals. Next the resulting activated caspases execute disassembly of cell followed by phagocytosis.
b. The Bcl-2 family of proteins has got both pro or anti apoptotic function. Bcl-2 inhibits apoptosis by inhibiting Bax and Bak. This inhibition blocks mitochondrial membrane disruption and release of cytochrome c in the cellular environment thus the cascade of proteolytic caspase is inhibited.

c. DNA damage sensors which act through TP53 tumor suppressors are the most notable one. Sensing the genomic damage, TP53 induces the apoptosis by up regulating expression of Noxa and Puma.
d. Hyperactive signalling by MYC also triggers apoptosis.
e. Insufficient IL3 signalling for the survival of lymphocytes also starts apoptosis.
f. Different cancer cell evolved mechanisms to resists apoptosis by loss of function mutation of TP53, over expression of Bcl-2 (anti-apoptotic regulator), IL-3 (survival factor) and under expression of Bax, Bak (pro-apoptotic regulators) (Hanahan & Robert A Weinberg 2011; Evan & Vousden 2001).

1.1.2.4. Enabling replicative immortality

The scientific perception was that the cancer cells must have unlimited replicative potential to generate microscopic tumors. Successive research deciphers that the length of telomere of a chromosome determines the number of cell division. Normal cell can pass through a limited number of cycles for cell division and growth. As telomeres progressively shorten with each cycle of cell division/growth resulting end to end chromosomal fusion which triggers apoptosis. However cancer cell enable replicative immortality by continuously regenerating the telomeric region of chromosome with over expression of telomerase (Hanahan & Robert A Weinberg 2011; Harley 2008)

1.1.2.5. Inducing angiogenesis

Tumor grows rapidly due to uncontrolled cell division and proliferation of cancer cell. The growth of tumors has to be supported with continuous flow of nutrients, oxygen and the resulting metabolic waste has to be excreted out. Angiogenesis synthesizes neo-vascular structure in and around the tumor to meet that needs. The cellular process of angiogenesis is controlled by the fine counter balance between inducing and inhibiting factors of angiogenesis. Some of the well-known angiogenesis inducers and inhibitors are vascular endothelial growth factor-A (VEGF-A) and thrombospondin-1(TSP-1).

a. The binding of VEGF-A to three receptor tyrosine kinase (VEGFR 1-3) trigger VEGF signalling which orchestrates the formation of new blood vessels.
b. Chronic up regulation of FGF (Fibroblast Growth Factor) enables cells to sustain angiogenesis.
c. TSP-1 binds to the trans-membrane receptors of endothelial cells and inhibits the angiogenesis.

d. Angiostatin and endostatin endogenously inhibit angiogenesis.

Tumors continuously induce angiogenesis by sustaining VEGF-VEGFR signalling by up regulation of VEGF in response to hypoxia or RAS and Myc signalling. Increased expression of pro-angiogenic molecules i.e. FGF also induce angiogenesis. Macrophages, neutrophils and mast cells can also activate angiogenesis by infiltrating pre-malignant/malignant lesions (Hanahan & Robert A Weinberg 2011; Carmeliet & Jain 2011).

1.1.2.6. Activating invasion or metastasis

Metastasis or invasion means the spread of cancer from one part of body to another. During tumor development the shape of its cells, the adhesion to other cells and to extracellular matrix degrades. The resulting tumor cells reach other parts of the body through the blood stream or lymph system. Cancer cells develop those properties by following mechanisms:

a. E-cadherin is a cell adhesive protein, helping to maintain epithelial cell sheets and quiescence of the cells to it. Down regulation or loss of function mutation in E-cadherin is a mechanism frequently used for the tumor tissues to metastasize to other body parts.
b. N-cadherin is adhesive protein helping in cell migration. The up-regulation of N-cadherin also drives the tumor cells to invade other body parts.
c. Epithelial to mesenchymal transition (EMT) is one prominent mechanism to transform the epithelial cells to invade and to resist apoptosis. A set of transcription factors namely Snail, Slug, Twist and Zeb1/2 are the key driver of EMT pathway. Some of these transcription factors down regulate E-cadherin expression to help in metastasis.
d. The interactions between cancer cells and nearby stromal cells play a crucial role for invasive growth and metastasis. Mesenchymal stem cells receive the signal from the cancer cells and in response produce CCL5. CCL5 stimulates cancer cells to invade other organs.
e. In order to facilitate invasion the tumor matrix has to be degraded. Cancer cells secret IL4 which activate macrophages. Activated macrophages disrupt the tumor matrix to support invasion (Hanahan & Robert A Weinberg 2011; Tracey A. Martin, Lin Ye, Andrew J. Sanders, Jane Lane 2013).

Other than these 6 hallmarks of cancer there are other emerging hallmarks of the disease i.e. deregulating cellular energetics and avoiding immune destruction. The hallmarks of cancer and their interconnectivities are illustrated in Figure 3.

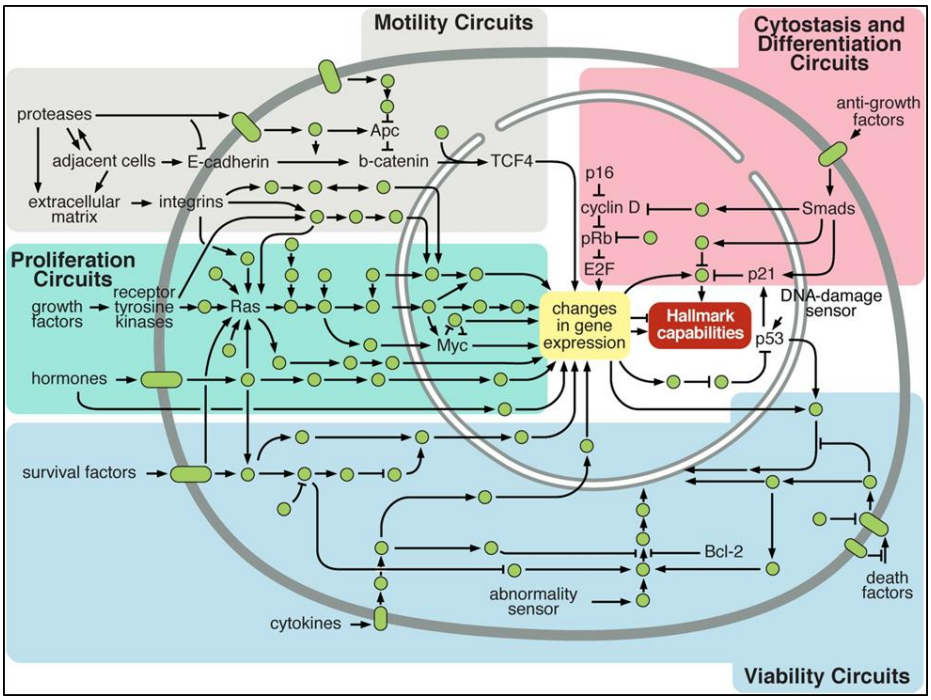


Figure 3: Intracellular signaling networks regulate the operations of cancer cells (Hanahan & Robert A. Weinberg 2011).

The figure demonstrates an elaborate integrated circuit operates within normal cells and is reprogrammed to regulate hallmark capabilities within cancer cells. Separate sub-circuits, depicted here in differently colored fields, are specialized to orchestrate the various capabilities. There is considerable crosstalk between such sub-circuits and major molecular players of each sub circuits are also visible. Each cancer cell is exposed to a complex mixture of signals from its microenvironment; each of these sub-circuits is connected with signals originating from other cells in the tumor microenvironment.

Unlike other diseases, cancer has represented one of the unvanquishable challenges to human ingenuity, resilience and perseverance. However, thanks to enormous amount of research in oncology, today millions of cancer patients extend their life span with early identification of the disease followed by treatment. Nevertheless complete cure of cancers remain elusive. In the next section, an evaluation of the most advanced cancer treatment i.e. targeted therapies is presented.

1.2. Targeted cancer treatment

Over the last three decades the scientific community has witnessed a remarkable shift in understanding the mechanisms of cancer pathogenesis. Our current understanding on that

matter has paved the way for mechanism based targeted therapy in cancer. The growing number of current targeted therapies in cancer can be categorized by their effects on one or more hallmarks of cancer as illustrated in Figure 4. The observed efficacy of each of these drugs represents a validation, whether a particular hallmark is important for tumor biology. A minute observation on the mode of action of current targeted therapy in cancer demonstrates that most of them are directed towards specific molecular targets. These targets directly or indirectly enable particular capabilities. This type of mode of action is rewarding as it represents inhibitory activity against a target while having less nonspecific toxicity (Hanahan & Robert A. Weinberg 2011).

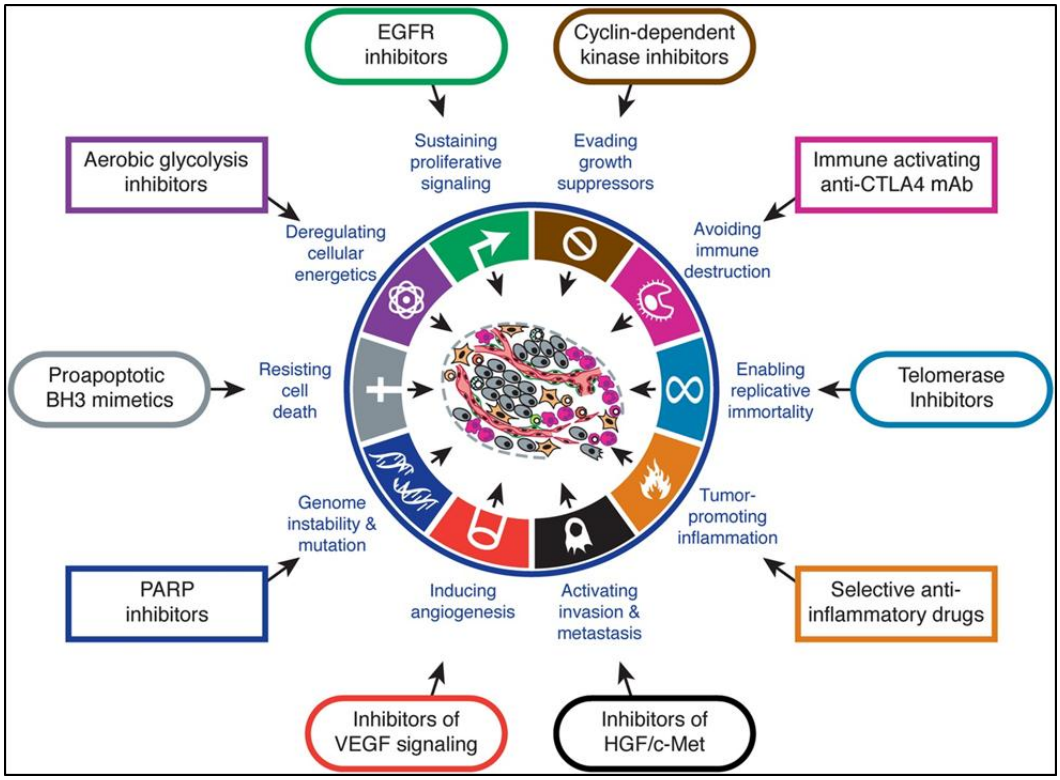


Figure 4: Recent targeted therapies of cancer and their mode of action (Hanahan & Robert A. Weinberg 2011).

Emerging or existing cancer treatments those shows promise to control with each of the acquired capabilities necessary for tumor growth and progression have been developed and are in clinical trials or in some cases approved for clinical use in treating certain forms of human cancer. Additionally, the investigational drugs are being developed to target each of the enabling characteristics and emerging hallmarks which also hold promise as cancer therapeutics.

However, in the last decade all drug companies were affected by the crisis fuelled by increased expenditure, augmented pipeline attrition rate and patent expiry of major blockbusters. The success rate of late stage clinical trials fell by 10% for phase II studies in recent years. Additionally, the number of phase III terminations doubled in the last five years (Arrowsmith 2011a; Elias 2006). Most of the phase III failed drugs target cancer. Many of these failures were supposed to be related to the clinical explorations of life extension

strategies, particularly in cancer where compounds are successful in one tumor type produces a poor outcome in another tumor type (Arrowsmith 2011b; Khanna 2012). Heterogeneity of cancer is one of the root causes of cancer drug failures. To eliminate less promising candidate drugs early in the clinical trials, selection of biomarker indicating any "off target" effects in preclinical screens will be crucial. In addressing most alarming cause of drug failures i.e. efficacy, clinical study design should consider patient stratification strategies, biomarkers, scoring systems and computational models (Khanna 2012).

So the drug discovery and development processes are in search for new models that would reduce the time taken by a drug to reach the market and increase the clinical success rate, thus satisfying regulatory authorities and patients' needs.

1.3. Recent trend in drug discovery

One of the major causes of expensive drug failure is the marginal disease improvement compared to the current standard of care when analysed in a large population of late stage trials (Frank & Hargreaves 2003a). This has propelled one of the paradigm shifts of pharmaceutical drug discovery from blockbusters to niche busters i.e. therapies targeted towards specific target molecules of specific patient populations termed as targeted therapy. The success of targeted therapy depends on identifying stratified biomarker i.e. molecular signature that will stratify the patients prior to treatment (Trusheim et al. 2007a). Companion diagnostics (CDx) (diagnosis for identifying patients benefit from a therapy based on stratified biomarker) holds great promise to improve the predictability of therapeutics interventions (Jørgensen 2013).

Currently about 10% of drug labels approved by the FDA contain pharmacogenomic information reflecting a clear trend towards targeted therapy (Frueh et al. 2008). Biomarker driven approach is being actively pursued by pharmaceutical companies as one of the next major reinventions in the field. The impact is evident through the FDA release of "Pharmacogenomic Biomarkers in Drug Labels" summarizing 136 approved drugs with 155 associated biomarkers as of 01.03.2014 (Ptolemy & Rifai 2010) (US Food and Drug Administration n.d.).

1.4. Biomarker

Biomarker is a characteristic measured and evaluated as an indicator of normal biological processes or pharmacologic responses to a therapeutic intervention (Frank & Hargreaves

2003a). The biomarker is either produced by the diseased organ (tumor) or by the body in response to a disease or therapeutic intervention. It can be applied along the whole spectrum of disease management. It can be used for risk assessment of the disease as well. During early diagnosis, biomarker can be applied for staging, grading and selection of initial therapy while in the late stages, they can be used for therapeutic dosage, monitoring, selection of additional therapy and diseases recurrence (Biomarkers Definitions Working Group, 2001). FDA in its critical path initiative emphasized on applying biomarkers as an essential tool to combat the current situation of pharmaceutical industries suffering from late stage failures and lack of successful pipeline portfolios. Recent publications suggested that a more efficient model of pharmaceutical pipeline can be designed by applying biomarkers in all stages of drug discovery and development i.e. emphasizing biomarker usage from target identification to drug marketing (Bakhtiar n.d.; Colburn 2003; Deyati et al. 2013). Figure 5 summarizes a model pipeline for the drug discovery and development applying biomarker in all stages. Based on the application in drug discovery and development, there can be five types of biomarker. The definitions of each of them are listed below:

1.7.1. Disease biomarker:
a biomarker that depicts prodromal signs to enable earlier diagnosis or allow for the outcome of interest to be determined at a more initial stage of a disease.

1.7.2. Efficacy biomarker:
a biomarker used to assess whether a drug will have clinically significant positive outcome before the treatment.

1.7.3. Safety biomarker:
a biomarker which can determine dose response for toxicity.

1.7.4. Pharmacodynamic biomarker:
a biomarker that indicates selection of an optimal dose of a drug for a patient.

1.7.5. Patient Stratification biomarker:
a biomarker that can predict the probable drug response on a selected subpopulation of patients.

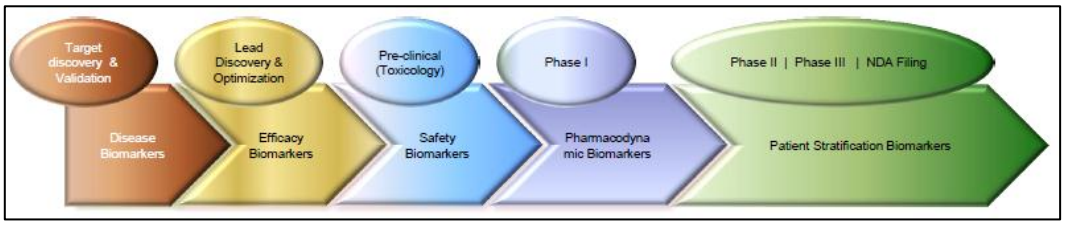


Figure 5: Model pipeline of biomarker driven drug discovery and development.

The emerging trend of applying biomarker in all five major stages of drug discovery (i.e. identifying the targets, discovery of lead molecule) and development (i.e. preclinical testing to filing the drug candidate to regulatory authorities). Each stage of drug discovery and development with its prospective biomarker are colour coded.

However, till now the use of patient stratification biomarker in late stage clinical trials stands out among other biomarker applications to cope with the most alarming issue of expensive late stage failures.

1.5. Biomarker in current clinical practice: focus on oncology (Deyati et al. 2013)

Cancer being a highly heterogeneous disease is among the first indications in moving towards targeted therapies. Different tyrosine kinase inhibitors (gefitinib, erlotinib, lapatinib, crizotinib, dasatinib, vemurafenib) and monoclonal antibodies (cetuximab, panitumumab, trastuzumab, pertuzumab, tositumumab) have been at the forefront of targeted therapies in cancer (Deyati et al. 2013; Majewski & Bernards 2011).

The success stories of targeted therapy in cancer started with commercialization of trastuzumab, cetuximab, imatinib, gefitinib. The trend has been on the rising with recent examples of Zelboraf (Vemurafenib) approved with the companion genetic test for BRAF mutation for late stage melanoma and Xalkori (Crizotinib) approved in combination with the companion genetic test for the ALK gene for late stage lung cancer (Dean & Lorigan 2012; Sai-Hong Ignatius Ou et al. 2012; Ruzzo et al. 2010a; Trusheim et al. 2007a). EGFR, Her2/neu, ALK, BRAF, Bcr-Abl, PIK3CA, JAK2, MEK, Kit, and PML-RARα are targets of recently approved targeted therapies in cancer. These target molecules and their downstream effectors are often subject to various changes on genomic, transcriptomic, proteomics and epigenetic levels. Therefore, status of those molecules (biomarker) underlies diversified patient-specific clinical responses to targeted therapies (Majewski & Bernards 2011).

As cancer is one of the prime areas of targeted therapies, we present here Table 1 summarizing stratification biomarkers currently in clinical practice in oncology along with their approved treatment. As shown in the Table 1, with few exceptions such as KRAS, most of the biomarkers are direct drug targets of the respective therapies. The majority of stratification biomarkers have been approved after the therapy went to the market (i.e. derived from the retrospective analysis of late-stage clinical trials or post-marketing surveys). It is obvious that the biology of the target and its changes under pathological conditions was not apparent during clinical development. Even in the stratified patient population, therapeutic response do not meet with equal success (Sawyers 2004; Paez et al. 2004; Kreitman 2006) suggesting that clear understanding of downstream pathways and molecular interaction networks are necessary for biomarker-driven stratified medicine. Efforts to elucidate such global downstream changes lead to host of technologies for biomarker identification reviewed in the next section.

Functional Class	Biomarker	Therapy
Kinase	EGFR	Cetuximab, Erlotinib, Gefitinib, Panitumumab
Kinase	Her2/neu	Lapatinib, Trastuzumab, Pertuzumab
Kinase	PDGFR	Imatinib
Kinase	Estrogen receptor	Fulvestrant, Exemestane
Kinase	ALK	Crizotinib
Kinase	KRAS	Cetuximab, Panitumumab
Kinase	BRAF	Vemurafenib

Immune cell surface receptor	CD20	Tositumumab
Immune cell surface receptor	CD25	Denileukin difitox
Immune cell surface receptor	CD30	Brentuximab vedotin
Immune cell surface receptor	C-Kit	Imatinib
CNV	PML-RARα	Arsenic trioxide
CNV	BCR-ABL	Dasatinib

Table 1: FDA-approved Stratification Biomarkers for targeted therapy in oncology

1.6. Biomarker discovery using high throughput technology

The rapid evolution of high throughput technologies designed for screening of biomedical samples by whole genome sequencing and microRNA (miRNA) profiling gave birth to several biological disciplines devoted to the generation and study of those multiple OMICS data. Figure 6 summarizes the latest technologies and diversification of the biomarker types based on bio molecular properties and underlying data types depending on changes detected by the respective technology.

1.6.1. Genetic biomarkers

Genetic biomarkers are biomarkers derived from technologies assessing genomic changes such as exome and whole genome sequencing, polymerase chain reaction (PCR) and Fluorescence in situ hybridization (FISH). They can accurately identify single nucleotide polymorphisms (SNPs), copy number variations (CNVs) and structural variations in the genome and delineate their functional significance in the pathophysiology of a defined phenotype. These technologies have helped to find stratification biomarkers in oncology and some of them are already in clinical practice. For example, KRAS sequencing and PCR were used to discover predictive and prognostic role of KRAS mutation in colorectal cancer and

lung cancer for anti EGFR-therapy resistance (Mauro Moroni et al. 2005; Lièvre et al. 2006; Eberhard et al. 2005; Amado et al. 2008). PCR/FISH analyses were used to show that translocation of BCR-ABL and PML-RARα may serve as predictive biomarkers conferring sensitivity to Imatinib mesylate and resistance to arsenic oxide in leukemia (Druker et al. 2001; Niu et al. 1999). The same technologies were used to identify mutation/amplification and translocation of ALK gene as biomarkers predicting the efficacy of Crizotinib treatment in late stage lung cancer (Kwak et al., 2010).

1.6.2. Expression biomarkers

Differing from the traditional single biochemical and histopathological measurements, expression biomarkers (transcriptomics biomarker) represent a fingerprint containing multiple biomarkers, which collectively indicate particular pathophysiology (Bhattacharya & Mariani 2009). Well established high throughput technologies like microarray expression profiling can identify differential expression of an entire genome at any specific sample at a given time point. There are several reports on using these technologies to identify biomarkers in specific cancer subtypes. Two expression biomarker tests are clinically approved for patient stratification in breast cancer. MammaPrint, a unique 70 gene expression profile, is a prognostic biomarker for distant recurrence of the disease following surgery in breast cancer patients (van ’t Veer et al., 2002). Oncotype DX is another gene expression signature-based biomarker test containing 16 cancer-related genes and 5 reference genes that predict the recurrence of breast cancer in Tamoxifen-treated patients with node negative, estrogen positive tumors (Paik et al., 2004).

1.6.3. Protein biomarkers

Human plasma holds the largest source of the proteome hence technologies that can measure changes in the protein profile are invaluable to identify protein biomarkers in blood. For example, mass spectrometry can capture minor changes of the protein levels and immunohistochemistry can accurately identify a specific protein in the living system. Application of proteomics in discovery of oncology biomarkers can be exhibited by immunohistochemistry-derived Her2 which is prognostic, predictive biomarkers for the sensitivity to Trastuzumab therapy in breast cancer (Lewis Phillips et al. 2008). EGFR is another pharmacodynamic biomarker discovered by immunohistochemical assay in colorectal and lung cancer samples conferring sensitivity to Cetuximab, Panitumumab and Gefitinib treatment (Saltz et al. 2004; Vanhoefer et al. 2004; Lynch et al. 2004; Freeman et al. 2009a).

Clinical acceptance of novel proteomics biomarkers suffers an anemic rate due to the lack of PCR-like amplification techniques for the vast number of analytes present in extremely small quantities in dynamic plasma (Ptolemy & Rifai 2010). As a potential solution, a novel method of immune PCR using conjugations of specific antibodies and nucleic acids is suggested which leads to 100-10000 fold signal amplification thus increasing sensitivity of protein biomarker detection (McDermed et al., 2012; Niemeyer et al., 2005).

1.6.4. Metabolic biomarkers

Ever since Otto Warburg hypothesized that altered metabolism (converting glucose carbon to lactate in oxygen rich condition) is specific to cancer cells due to mitochondrial defects, metabolic biomarkers have drawn the attention of researchers to be an effective biomarker for early cancer diagnosis and prognosis (Ward & Thompson 2012). Since then, numerous efforts have been dedicated to identify metabolic biomarkers in oncology. NMR spectroscopy, HPLC, radioimmunoassay, LC-MS, GC-MS, and enzyme immunoassay help to analyze the metabolite levels in response to pathophysiological change or treatment. Until now, only two metabolic biomarkers have made it to clinical practice. Metanephrine and Normetaephrine are two metabolites that are used to predict disease state associated to pheochromocytoma (see: http://www.accessdata.fda.gov/cdrh_docs/reviews/K032199.pdf). Despite rapid technological advancement in metabolomics, it is still impossible to differentiate metabolites derived from different sub-cellular compartments. Current fractionation methods often lead to metabolite leakage between different layers making it even more difficult for metabolite identification (Ward & Thompson 2012).

1.6.5. microRNA biomarkers

The involvement of microRNAs (miRNAs) in key cellular processes such as proliferation, cell death and negative regulation of numerous oncoproteins makes them a prime candidate as cancer biomarkers. It has been reported that cancer-specific miRNAs are detected in the blood from the earlier stages of tumor development and concentration increases as the tumor progresses over time, making them an indicator of tumor growth (Krutovskikh & Herceg 2010). Unlike other types of biomarkers, miRNAs are remarkably stable in the circulation and formalin-fixed paraffin embedded tissue, making them potentially robust oncology biomarkers. Functional miRNA species have mostly been validated in vitro using luciferase reporter activity (Krutovskikh & Herceg 2010). Microarray profiling is a powerful, high throughput technology capable of monitoring the expression of thousands of small non-

coding RNAs at a specific context. Mirage (SAGE), Stem-loop qRT-PCR for mature miRNAs, qRT-PCR for precursor miRNAs and bead-based technologies are also frequently used for microRNA profiling (Liu et al. 2008). However, no such biomarker exists in cancer clinical practice yet. It is noteworthy that genetic biomarkers witnessed a remarkable rise in clinical acceptance after the human genome project characterizing all the genes. Similar effort is needed to discover and characterize all the miRNAs in human cells to transform the potential of miRNAs as cancer biomarker into clinical success. Further understanding on how miRNAs compete with proteins to bind and control the expression of mRNA, as well as the functional interaction networks through which miRNAs act are needed for future clinical translation (Krutovskikh & Herceg 2010).

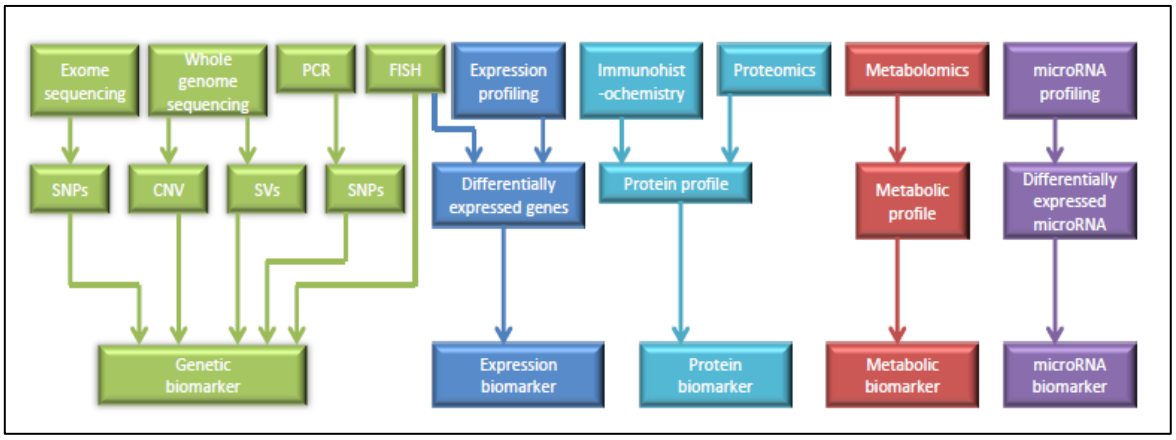


Figure 6: Data types and technologies for biomarker discovery

The figure illustrates current technologies and data types used for biomarker discovery in preclinical and clinical research. Abbreviations: CNV, copy number variations; FISH, fluorescence in situ hybridization; GCMS, gas chromatography mass spectrometry; HPLC, high-performance liquid chromatography; LCMS, liquid chromatography–mass spectrometry; NMR, nuclear magnetic resonance; PCR, polymerase chain reaction; SNPs, single nucleotide polymorphisms; SVs, structural variations.

Noting this enormous amplification of data points obtained from biomedical samples, the question arises whether these technological advances along with ever-increasing availability of the screening platforms can lead to clinical breakthroughs? To get a better understanding of the technologies contributing to the identification of currently approved stratification biomarkers in oncology; we present an OMICS wise overview on data generation platforms and types of data resulting from these platforms in Table 2.

OMICS	Technology	Biomarker	Associated no of drugs	References
GenOMICS	Fluorescence in situ hybridization or Polymerase chain reaction	ALK	1	(Janoueix-Lerosey et al. 2008)
GenOMICS	Fluorescence in situ hybridization	Her2/neu	1	(Bekaii-Saab et al. 2009)
GenOMICS	Polymerase chain reaction	BRAF	1	(Chapman et al. 2011)
GenOMICS	Polymerase chain reaction	CD20	1	(Kaminski et al. 2005)
GenOMICS	Polymerase chain reaction	PML-RARα	1	(Niu et al. 1999)
GenOMICS	Polymerase chain reaction	KRAS	1	(Heinrich et al. 2003)
GenOMICS	Sequencing	EGFR	1	(M. Moroni et al. 2005)
GenOMICS	Sequencing	KRAS	1	(Lynch et al. 2004; Freeman et al. 2009b)
GenOMICS	Sequencing	C-Kit	1	(Heinrich et al. 2003)
GenOMICS	Sequencing	BCR-ABL	1	(Takei et al. 2008)
GenOMICS	Sequencing	PDGFR	1	(Takei et al.

				2008)
ProteOMICS	Immunohistochemistry	EGFR	3	(Lewis Phillips et al. 2008; Tsao et al. 2005; Addo et al. 2002; Hochhaus et al. 2008)
ProteOMICS	Immunohistochemistry	CD25	1	(Dang et al. 2007)
ProteOMICS	Immunohistochemistry	Her2/neu	1	(Slamon et al. 2001)

Table 2: OMICS technologies for stratification biomarker discovery in oncology

As evident from Table 2, few stratification biomarkers derived from each technology are currently approved and are in clinical use for cancer. This reflects the hard and long process of developing sensitive, specific and highly predictive biomarkers for clinical decision making from high throughput data. Nevertheless, OMICS technologies have tremendous potential to discover future biomarkers and the expectations have been augmented in last 20 years supporting the huge investments in the development of these technologies (Deyati et al. 2013).

To understand the future potential of these high throughput technologies, we compared the number of published candidate biomarkers, i.e. those reported in the scientific literature, clinical trials registries or scientific conferences with the number of approved biomarkers for each technology described above. For this purpose, we retrieved all cancer-related candidate biomarkers (including disease, stratification, prognostic and diagnostic biomarkers) from GVK Bio Online Biomarker Database (GOBIOM). GOBIOM is independent manually curated biomarker related knowledge base that uses the information derived from clinical reports, annual meetings and journal articles (Jagarlapudi & Kishan 2009). During the time of writing, GOBIOM contained information on 15,732 biomarkers covering 16 therapeutic areas supported by 36,681 unique references.

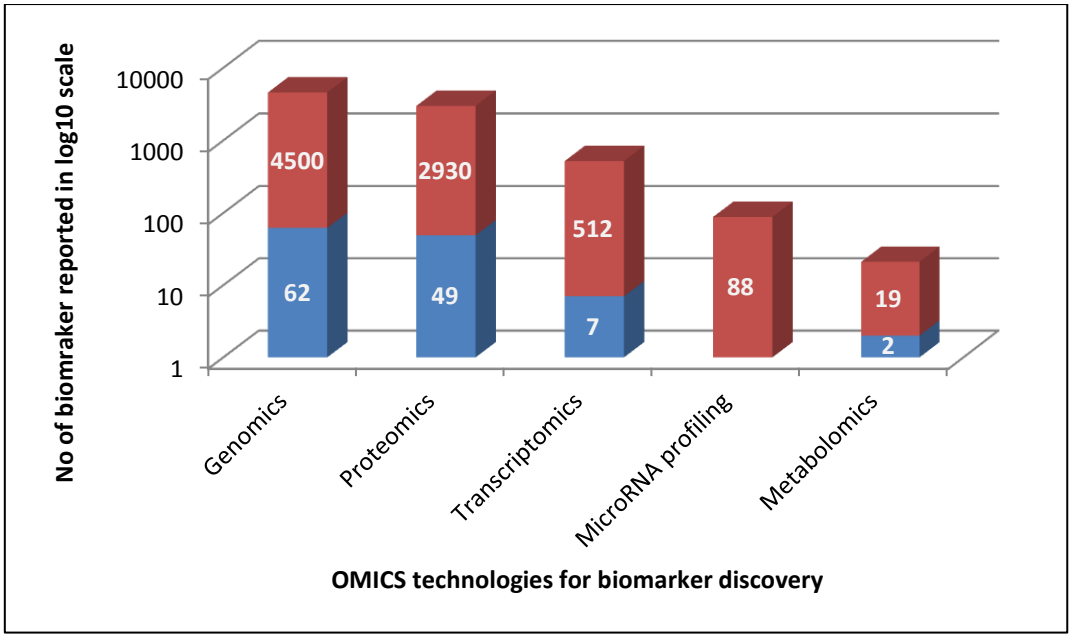


Figure 7: Comparative performance of different OMICS technologies in oncology biomarker discovery collected through the Gobiom database.

Red: Total number of candidate biomarkers reported in the public domain. Blue: Number of FDA approved biomarkers. Current contribution of OMICs technologies in oncology biomarker discovery extracted from the Gobiom database. In red: the total number of candidate biomarkers reported in the public domain. In blue: the number of FDA approved biomarkers in current clinical practice for oncology.

As evident in Figure 7, although transcriptomics technology is one of the oldest and widely used high throughput technologies, most of the candidate biomarkers are reported to be coming from genomic research followed by proteomics. Stability of the signal coming from genomic analysis as well as higher stability of the protein versus mRNA might be the reason for those biomarkers overweighting the transcriptomics derived biomarkers (Deyati et al. 2013). Never the less transcription and its regulation is one of the most important mechanisms to implement and manifest the genetic information stored in DNA. Transcriptomics technologies are also economically less expensive than proteomics technologies. State of the art analysis of transcriptomics data as well as accurate functional interpretation can potentially improve the situation (Khatri et al. 2012). In the next chapter the applied transcriptomic technologies in this thesis i.e. microarray technology along with state of the art statistical methods for data analysis are described in detail.

2. Computational Methods to Analyse Microarray Data for Cancer Biomarker Discovery

Over the years substantial research has been undertaken to identify differentially expressed disease biomarker by analyzing microarray data from patient samples. These biomarkers serve as diagnostic or prognostic indicators that dictate preferred therapeutics. This chapter is focusing on microarray technology and its data analysis in cancer. The focus will be also on quality assessment, data combination, feature selection and prediction.

Microarray technology for measuring expression profiles of genes is an invaluable tool to infer the physiological condition of the cell, tissue or an organism. Compared to measuring the expression level of proteins, measuring mRNA expression profile is easier and less expensive. Thus scientific community often applies gene expression level and its post transcription control as a reasonable substitute. A set of probes are attached to the solid surface of microarray or chip. The key principle behind the microarray technology is a process called hybridization in which a nucleotide strand binds to its unique complementary strand. Fluorescence tagged nucleotide sequences are allowed to complementary bind with its probes and based on the intensity of the fluorescence, the relative abundance of the nucleotide sequence (gene) is determined. The microarray can be categorized into two major categories i.e. single channel and two-channel. In two-channel microarray two labeled samples are hybridized on the same surface of the chip to determine the relative expression of the genes between the two samples. As the name indicates in one-color microarray gene expression profile of only one sample can be measured (Lockhart et al. 1996; DeRisi et al. 1997).

Over the years, the biomedical community has witnessed an exponential growth in publication analyzing gene expression profiles from clinical samples which are largely deposited to the Gene Expression Omnibus (GEO) (Edgar et al. 2002). Clinical gene expression data that has been analysed in this dissertation are extracted from biopsies of tumor samples and healthy tissues taken from GEO and The Cancer Genome Atlas (TCGA). In order to control bias, the gene expression profiles are typically measured on the same platforms across samples of interest. The data analysis was also followed by common pre-processing steps including background correction, imputation, and normalization.

2.1. Analysis of microarray data

The data analysis of a microarray experiment is a multi-step process. It starts with sample preparation and hybridization in microarray, which produce image files containing probe signal intensities. This probe signal intensities require image analysis, signal adjustment and data normalization to eliminate all the non-biological variability (or noise) inherent to the system. These procedures are often referred to as "data normalization". Due to the multiple microarray platforms with different target preparation and hybridization methods, numerous normalization methods have been developed. After pre-processing, the normalized gene expression data can be analysed using statistical tools and exploratory methods to extract genes or patterns with biological significance. The analysis of one-color and two color microarray data is more or less similar, the main difference between these methods lies in the normalization of the data.

2.1.1. Image Analysis

Microarray image analysis starts with placing a computer-aligned grid over the hybridized surface area. Next, software is used to measure the intensity of each spot representing each probe on the hybridized array. Each spot on the chip surface is investigated for any hybridization biases or poorly hybridized probe to evaluate the hybridization quality. The software eliminates these "bad spots" and other spot intensities are stored in a file. The signal adjustment is a crucial step to correct the background noise and processing effect, to adjust for cross hybridization caused by the binding of non-specific target (DNA or RNA) and finally to adjust for expression estimates so that proper scale dimensionality is ensured. The intensities of the pixels surrounding the spot are measured to compute the background adjustment for spotted arrays. Due to the highly dense nature of Affymetrix GeneChips®, only the probe intensities must be used to determine any adjustment. A specialized statistical algorithm called Ideal Mismatch procedure is used to calculate the adjusted signal from Affymetrix GeneChips® (Freeman et al. 2009a).

2.1.2. Data Normalization

Data normalization is a method that strives to eliminate experimental variation due to differential amount of extracted RNA, dye effects, scanner difference etc. The scientific rationale for data normalization is to achieve uniform level of gene expression for most genes and make sure the gene expression is following a normal distribution when comparing two or

more samples on a genome-wide level. Hence the expected mean intensity ratio between two channels (two-color data) or chips (one color data) should be one, otherwise the data is numerically processed to adjust this ratio to one. Normalizations include the following steps: data transformation to stabilize intensity variance across the datasets and removal of non-biological variation within a single array or between different arrays. In some normalization processes expression of house-keeping genes or spike-in controls (external control sequences from other organism) are applied. Replicates from biological and technical point of view are also very useful to minimize the effect of outliers and in enhancing data quality (Quackenbush 2001). Mean or median probe intensity values of multiple probes (one-color arrays) or replicate spots (two-color arrays) are often applied to increase the robustness of the data. The choice of normalization method is often based upon prior knowledge of the dataset. To maintain normal distribution of log intensity ratios for each slide of two-color microarrays, log centering and scaling are applied. However, LOWESS (locally weighted linear regression) or dye-swap smoothing are often preferred normalization methods where dye biases can depend on spot overall intensity or co-ordinate within the array (Ekstrøm et al. 2004; Bandyopadhyaya et al. 2002). There are several preferred normalization methods for Affymetrix GeneChip®, to name just a few Microarray Suite (MAS) 5.0 statistical algorithm from Affymetrix (Gold et al. 2005), robust multichip average (RMA) (Wang et al. 2004), model-based expression index (MBEI) (Wu et al. 2005). These normalization methods considerably differ from one other leading to different results in data analysis. Popular MAS 5.0 normalization applies a linear regression algorithm. Adjusted PM values are log transformed and a robust mean is calculated on the resulting values. Then the data is scaled using a trimmed mean after obtaining a signal for each probe set as the antilog of the resulting value (Gold et al. 2005). MBEI normalization is done with median intensity selected as a baseline array for normalization. The algorithm uses an invariant set method where numerous probes are selected *ad-hoc* as references for comparison of two samples and a non-parametric curve (running median) is fitted through the data points (Wu et al. 2005). RMA and GCRMA are a modification of MBEI normalization method. They use quantile normalization instead of median method used in MBEI. Nevertheless, apply same empirical distribution of probe intensities for each array of an experiment and the maximum background corrected; log-transformed PM intensity on each chip is evaluated. Next the original intensity values are replaced with average values and the process repeated for all intensities in descending order. After that, expression value for each probe on each GeneChip is measured by fitting an additive linear model to the normalized data. In addition to RMA

method described above, GCRMA takes into account of sequence information to address nonspecific background variation (Quackenbush 2001). Quality and comparability of the data analysis of microarray experiments has become a major challenge due to the wide range of methods for data production and normalization. As a solution Microarray Gene Expression Data Society (MGED) created a guideline called Minimum Information about Microarray Experiments (MIAME) for microarray data reporting standards. Nowadays many journals regarded the MIAME standards to publish microarray data (Brazma et al. 2001; Verducci et al. 2006).

2.1.3. Statistical analysis of microarray data

The main idea behind statistical analysis of high dimensional microarray data is to characterize the structure of the data, reduce the dimensionality and extract statistically significant pattern in it. Fold change was among the first methods used to evaluate whether genes were differentially expressed. However, nowadays it is considered an inadequate test statistic as it does not account for the variance and offers no associated level of confidence. Parametric tests are more predominantly used for the analysis of microarray data due to its normalized nature. Typical parametric tests used in microarray analysis are student's t-test or analysis of variance (ANOVA). These methods assume the data is normally distributed and try to estimate whether the variance in the data comes from the normal distribution. Often microarray experiments have a large number of observations but fewer samples which lead to test multiple hypotheses. In a biological experiment, the observed differences are expected to happen by chance and as well as due to biological variability. The required correction of the statistical tests are often achieved by the Bonferroni method and the false discovery rate (FDR) suggested by Benjamin and Hochberg (Benjamini Yoav 1995). There are ample commercial and non-commercial statistical analytical tools available for advanced data analysis and visualization. Treeview (Eisen et al. 1998), GeneCluster (Tamayo et al. 1999), SAM (Tusher et al. 2001a), dCHIP (Li & Wong 2001), Gene data expressionist, GeneSpring (Agilent Technologies) and numerous other R and Bioconductor packages provides great resources to analyse microarray data. The main purpose of these packages vary from normalization, removal of insignificant genes, statistical analysis to identify differentially expressed genes or classify genes based on different phenotypes.

One of the big challenges of analyzing microarray data with traditional classical statistics is the presence of large number of unchanged genes adding high level of noise and uncertainty. The power and reliability of the statistical test can be improved by eliminating those genes; here arises the application of gene filtering to reduce the dimensionality of the data and differentiation approach to screen significant differentially expressed genes. In the next section, some of the popular data mining methods to identify differentially expressed genes are summarized.

2.1.3.1. Data mining Algorithms to screen differentially expressed genes and feature selection

Significance analysis of Microarray (SAM): SAM identifies a set of statistically significant genes in expression analysis by assimilating a set of gene specific t tests. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. Genes with scores greater than a threshold are deemed potentially significant. The algorithm applies permutation of repeated measurements to estimate the percentage of genes estimated by chance i.e. the false discovery rate. The threshold can be adjusted smaller or larger set of genes. The algorithm incorporates the gene specific fluctuation of expression with a statistics based on the ratio of change in gene expression to standard deviation in that data for that gene. The "relative distance" d(i) in gene expression is calculated as follows.

$$d(i) = \frac{x_I(i) - x_U(i)}{s(i) + s_0}$$

$x_I(i)$ and $x_U(i)$ are defined as the average levels of expression for gene (i) within states I and U. The "gene-specific scatter" s(i) is the standard deviation of repeated expression measurements:

$$s(i) = \sqrt{a\{\sum_m [x_m(i) - x_I(i)]^2 + \sum_n [x_n(i) - x_U(i)]^2\}}$$

$a = (1/n_1 + 1/n_2)/(n_1 + n_2 - 2)$. n_1 and n_2 are the number of measurements in states I and U. The statistical distribution d(i) should be independent of the level of gene expression so that d(i) across all genes are comparable. When gene expression level is low the variance in d(i) can be high because of small values of s(i). To ensure that the variance of d(i) is independent of

gene expression by a small positive constant S_0 added to the denominator to calculate d(i) (Tusher et al. 2001a).

Limma: Limma is a package to identify differentially expressed genes in a microarray experiment by applying a linear model to the expression data for each gene. To stabilize the experimental analysis with a smaller number of array, shrinkage method like empirical bayes is used for an expression overview across all the genes in the array. The linear model with Limma requires two matrices. The first one is the designed matrix which provides a representation of the hybridized RNA targets in the array. The second one is the contrast matrix which allows the coefficients defined by the design matrix to be combined into contrasts of interest. Limma operates on a matrix of expression values in which each row represents a gene or some other relevant genomic feature and each column represent a RNA sample. The algorithm fits a linear model to each row of data and it facilitates to handle complex experimental design and to test extremely flexible hypothesis. Genomic data is highly parallel in nature. So when the sample number is low the model can borrow information between gene-wise models allowing different levels of variability between genes and between samples thus the statistical inferences get robust. All the features of the statistical models can be accessed not just for gene wise expression analysis but also for gene expression signature. Mathematically a linear model $E[y_j] = X\alpha_j$ is assumed where y_j symbolizes the expression data the gene j, X is the designed matrix and α_j is the vector of coefficient. The contrasts of interest are given by $\beta_j = C^T \alpha_j$ where C is the contrasts matrix (Ritchie et al. 2015).

Correlation based feature selection (CFS): CFS algorithm uses a correlation based heuristic to evaluate the importance of features. The heuristic takes into account the usefulness of individual features for predicting the class label along with the level of correlation among them. The underlying hypothesis of CFS is "Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other". The formalism of heuristics is as follows:

$$\text{Merit}_s = \frac{kr_{cf}}{\sqrt{k+k(k-1)r_{ff}}}$$

Where Merit_s is the heuristic "merit" of a feature subset S containing k features, r_{cf} the average feature class correlation and r_{ff} the average feature-feature intercorrelation. Equation

n can also be viewed as Pearson's correlation where all the variables have been standardized. The numerator indicates how predictive a group of features are whereas the denominator implicate how much redundancy there is among them (Hall 2000).

Mann-Whitney-U test: Mann Whitney U test is an unpaired, univariate and non-parametric test i.e. two independent random samples A and B can be compared with Mann-Whitney-U test. The Mann-Whitney-U statistics is defined as follows:

$$U = n_1 n_2 + n_2 (n_2 + 1)/2 - \sum_{i=n_1+1}^{n_2} Ri$$

When A and B are randomly collected from population, the samples are independent from each other and the measurement scale is ordinal. Sample A of size n_1 and B of size n_2 was pooled and R_i is the rank.

The aim of the test is to find out if the centrality of a test variable when it differs significantly between two groups of interest. The p-value serves as the evaluation measure of discrimination. Mann Whitney u test is especially appropriate for two class biological problems (Lehmann, E.L, Romano 2006)

2.1.3.2. Clustering Algorithm

Clustering is an unsupervised learning algorithm to separate a group of data points into a number of clusters. Clustering can be defined as the process of organizing objects into groups whose members are similar in some way. Unsupervised clustering can further be split to hierarchical clustering method and non-hierarchical clustering methods such as self-organizing maps (SOM) or K-means clustering. One of the major goals of clustering is to determine the intrinsic grouping within an unlabelled data. Clustering is one of the most popular methods as the first step in gene expression data analysis. In order to reduce the high dimensional gene expression data, clustering algorithm like PCA is applied. Strategically microarray data analysis can be applied either at the gene level to find out similarities or dissimilarities between different genes across the samples to find out the correlation between two genes or comparing the samples to find out differentially expressed genes in different experimental conditions (Korol, A.B., 2003).

Principal component analysis: Principal component analysis (PCA) is an unsupervised projection approach which reduces the dimensionality of biological data by identifying directions called principal component along which the variance of the data is maximum. So

PCA ignores those dimensions in which the data can be much variabe. Given a high dimensional data, PCA calculates a new system of coordinates. The directions of this new coordinate system calculated by PCA are the eigenvectors of the covariance matrix. In mathematical terms, the Eigen vector with the highest Eigen value computed from the covariance matrix is the first principal component. So intuitively covariance matrix reflects the shape of data points. With the help of Eigen vectors PCA captures the main axes of the shape formed by the data in n-dimensional space. Given a large dataset like microarray with multiple variables, by applying PCA we can identify variables with the highest correlation with the dependent variable (Ringnér 2008).

Singular Value decomposition: SVD is a dimensionality reduction algorithm. Mathematically SVD is matrix decomposition technique which is also the mathematical framework of PCA. Microarray data can be represented as real valued m-by-n matrix (say X). The same matrix can also be represented as the product of three matrices (Liu & Zhao 2006).

$$A = QDR^T$$

Q: Eigenvectors of AA^T which will be m-by-m matrix

D: Eigenvectors of A^TA which will be n-by-n matrix

R: Diagonal matrix with square root of the Eigen values of AA^T and A^TA (Korol, A.B., 2003)

Unsupervised Clustering: Clustering without prior knowledge about the data is termed as unsupervised clustering. Algorithms in this category treat all inputs of a set of n numbers or an n-dimensional vector. Unsupervised clustering can be done on genes, samples, time points in a time series experiment. Unsupervised clustering is based on the measure of similarity. The similarity measure between objects is also referred as a distance matrix. There are several mathematical algorithms to calculate similarities which are described below (Korol, A.B., 2003).

Suppose we are having two n-dimensional vectors $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$

a. Euclidean distance

$$d_E = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - x_{n+1})}$$

The Euclidean distance counts both direction and magnitude of the vector (Korol, A.B., 2003).

b. Manhattan distance is calculated as follows:

$$d_m(x, y) = |x_1\text{-}y_1| + |x_2\text{-}y_2| + \ldots. + |x_n\text{-}x_{n+1}|$$

Manhattan distance represents the absolute value of the difference between x_i and y_i. However, cluster using distance calculated by applying Euclidean matrix is much more compact than applying Manhattan distance (Korol, A.B., 2003).

c. Chebychev distance is calculated as follows:

$$d_{max}(x, y) = \max |x_i - y_i|$$

As the formula indicates the Chebychev distance will simply pick the largest distance between two corresponding entities. The distance matrix is very robust in avoiding the noises as long as the values do not exceed the maximum distance (Korol, A.B., 2003).

d. Correlation distance is calculated as follows:

$$d_r(x, y) = 1 - r_{xy}$$

r_{xy} is the Pearson correlation coefficient calculated as $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - x_m)(y_i - y_m)}{\sqrt{\sum_{i=1}^{n}(x_i - x_m)^2}\sqrt{\sum_{i=1}^{n}(y_1 - y_m)^2}}$

x_m and y_m are the mean values of x and y variable. The correlation distance can vary between 0 and 2 as the correlation coefficient r_{xy} takes values between -1 and 1. This distance matrix does not incorporate the magnitude of the coordinates (Korol, A.B., 2003).

e. Mahalanobis distance is calculated as follows: $d_{ml}(x, y) = \sqrt{(x - y)^T S^{-1}(x - y)}$

S is any n*n positive definite matrix and $(x\text{-}y)^T$ is the transposition of (x-y). The role of the matrix S is to distort the space as desired. If S is an identity matrix then Mahalanobis distance reduces to the classical Euclidean distance (Korol, A.B., 2003).

Clustering algorithms have been extensively used in phylogenetic analysis. The algorithm has also been adopted by the gene expression community for its accuracy and usefulness. Some of the extensively used clustering algorithms are discussed in detail in the next section.

Non-hierarchical clustering

K-Means: At first the algorithm estimates a number of clusters then calculates the centroid of each cluster and finally find out the closest cluster for each data point by minimizing the objective function (J).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left| x_i^j - c_j \right|^2$$

Where $|x_i^J - c_j|^2$ is the calculated distance between a data point x_i^j and cluster centre c_j. The distance is calculated for n data point over k cluster. After all the objects have been assigned into cluster the k centroids are recalculated followed by calculating the objective function. This process continues until the centroid no longer moves. K means clustering is one of the fastest and simplest clustering algorithms (Ma et al. 2005).

SOM: Self-organizing maps or SOM is an application of self-organizing neural network to cluster multi-dimensional data to recognise and classify features. The algorithm has a set of nodes with simple topology (two dimensional grid) and a distance function $d(N_1, N_2)$ on the nodes. Nodes are iteratively mapped into k-dimensional "gene expression" space. The position of node N at iteration i is denoted $f_i(N)$. The initial mapping f_0 is random. After subsequent iterations, a data point P is selected and the node N_P that maps nearest to P is identified. The mapping of node then adjusted by moving points toward P by the following formula:

$$f_{i+1}(N) = f_i(N) + \gamma(d(N, N_P), i)(P - f_i(N))$$

The learning rate γ decreases with distance of N from N_P with iteration number i. The point P used at each iteration is determined by random ordering of the n data points generated once and recycled as needed. The function γ is defined by $\gamma(x, i) = 0.02T/(T + 100\ i)$ for $x = \alpha(i)$ and $\gamma(x, i) = 0$. Otherwise where the radius $\alpha(i)$ decreases linearly $i(\gamma(0) = 3)$ and eventually become zero and T is the maximum number of iterations (Korol, A.B., 2003, Tamayo et al. 1999).

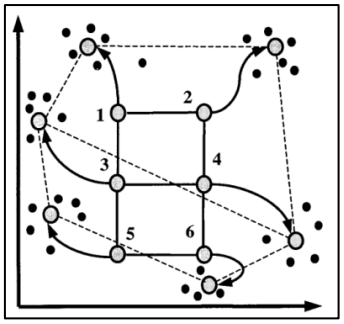


Figure 8: Principle of SOMs (Tamayo et al. 1999).

Initial geometry of nodes in 3*2 rectangular grids is indicated by solid lines connecting the nodes. Hypothetical trajectories of nodes as they migrate to fit data during successive iterations of SOM algorithm are shown. Data points are represented by black dots, six nodes of SOM by large circles, and trajectories by arrows.

Hierarchical clustering: As the name indicates hierarchical clustering uses a progressive combination of elements that are most similar and clustered them into higher order as dendrograms. Different experimental conditions can be clustered based on genes expression and different genes also can be clustered based on their expression values in different experimental conditions. Then different cluster can also be clustered by an inter-cluster distance to make a higher level cluster. So unlike k-means clustering, one can deduce the relationship between different clusters based on their distance from the root, i.e. closer the clusters are from the roots more similar they are. The hierarchical clustering are built based on two approaches top-down and bottom-up. In bottom-up method first the distance between all data points are calculated based on the different distance calculation algorithms described before to cluster the data points into the initial cluster. Next the distances between different clusters are calculated and the processes continue to group most similar clusters into higher level clusters. The formalism of top-down algorithm is quite reverse that bottom-up method. First all the data points are considered to be a part of a super cluster in the top-down approach. Next cluster is divided into two clusters by applying k-means clustering (k=2). This process is repeated until reaching cluster contains only one data point (Obulkasim & van de Wiel 2015)

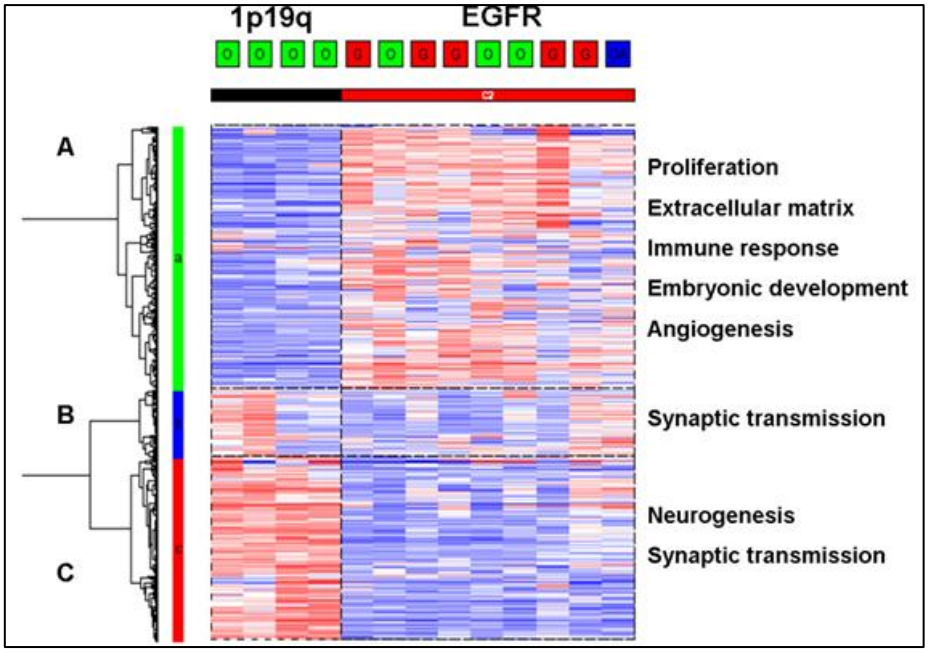


Figure 9: Hierarchical clustering based on gene expression of microarray data (Ducray et al. 2008).

Figure 9 demonstrates A typical example of application of hierarchical clustering using microarray data in cancer research. The analysis was completed on 4 oligodendrogliomas with 1p19q codeletion and 9 gliomas with *EGFR* amplification. Unsupervised hierarchical clustering was performed using the 1366 probe sets whose expression varied the most across the 13 samples (probe sets with a robust coefficient of variation superior to the 97.5th percentile). 1p19q = 1p19q codeletion, *EGFR* = *EGFR*amplification. The gliomas were classified into 2 groups according to their genomic profile. Gliomas with *EGFR* amplification were classified into one cluster irrespective of their histology (red = glioblastoma, green = grade III oligodendroglioma, blue = grade III oligoastrocytoma). Gene cluster A was enriched in genes involved in proliferation, extracellular matrix, immune response, embryonic development and angiogenesis. Gene cluster B was enriched in genes involved in synaptic transmission. Gene cluster C was enriched in genes involved in neurogenesis and synaptic transmission.

2.1.3.3. Classification algorithms

The supervised clustering or classification algorithms are developed to assign objects to predetermined classes. Supervised methods generally involve the use of a training data set and an independent validation data set. The aim is to obtain a function or rule that uses expression data to predict its class. In cases where the dataset is too small to be effectively split, a cross-validation method such as leave-one-out or class permutation procedure is often used. Classification algorithms are also widely applied to analyse gene expression data with the aim either to discover new categories within the dataset or to assign classes to a given category. Genes or samples are classified into specific groups based on the values of a set of computed variables by unsupervised clustering. Typically genes are grouped into classes based on the expression profiles in different biological conditions. In the next section state of art classification algorithms which are predominantly used for the analysis of microarray experiment are described in detail.

Linear discriminant analysis (LDA): Classical LDA projects the data onto a lower-dimensional vector space such that the ratio between classes distances is maximized, thus achieving maximum discrimination between classes. The objective is to classify unknown samples into one of the k classes n_k training samples per class, k = 1, 2, 3, …, K with m genes in each microarray. For each training sample, we observe class membership Y. and expression profile X. For simplicity, the classes has been represented by the numbers 1, 2,..., K. Note that each expression profile is a vector of length m. We assume that expression profiles from class k are distributed as N (μ_k, $\sum$), the multivariate normal distribution with mean vector μ_k and covariance matrix $\sum$. Call L(μ_k, $\sum$) the corresponding probability density function. Finally, we agree upon prior probabilities π_k that an unknown sample comes from class k, k = 1, 2,..., K (Dabney 2005).

Bayes' theorem states that the probability of a sample comes from class k, given that sample's expression profile, is proportional to the product of the class density and prior probability (Dabney 2005):

$$P_r(Y = k|X=x) \propto L(x; \mu_k, \sum) * \pi_k$$

We call Equation (n) the posterior probability that array x comes from sample k. LDA assigns the sample to the class with the largest posterior probability (Dabney 2005):

$$\hat{Y} = \arg\max_k \{L(x; \mu_k, \sum) * \pi_k\}$$

Prediction analysis of Microarray (PAM): PAM is an application of LDA based classifier for microarray analysis. PAM applies nearest shrunken centroids for class prediction. Let a_{ij} be the expression of genes i = 1, 2,…, K and let C_k be indices of the n_k samples in class k. The i^{th} component of the centroid for class K is $\bar{a}_{ik} = \sum_{j \in C_k} a_{ij}/n_k$. Then the i^{th} component of the overall centroid is $\bar{a}_i = \sum_{j=1}^{n} a_{ij}/n$ Hence the class centroids shrink towards the overall centroids after standardizing by the within-class standard deviation for each gene. This standardization has the effect of giving higher weight to genes whose expression is stable within samples of the same class (Tibshirani et al. 2002) .

$$d_{ik} = \frac{\bar{a}_{ik} - a_i}{m_k(s_i + s_0)}$$

Where s_i is the pooled within-class standard deviation for gene i:

$$s_i^2 = \frac{1}{n-K}\sum_k \sum_{j \in C_k} (a_{ij} - \bar{a}_{ik})^2$$

and $m_k = \sqrt{\frac{1}{n_k} + \frac{1}{n}}$ makes the $m_k.s_i$ equal to the estimated standard error of the numerator in dik. In the denominator, the value s_0 is a positive constant (with the same value for all genes), included to guard against the possibility of large d_{ik} values arising by chance from genes with low expression levels. s_0 is set to equal the median value of the s_i over the set of genes. Thus d_{ik} is a t statistic for gene i, comparing class k to the overall centroid.

$$\bar{a}_{ik} = \bar{a} + m_k(s_i + s_0)d_{ik}$$

The PAM method shrinks each d_{ik} toward zero, giving d'_{ik} and yielding shrunken centroids or prototypes.

$$\bar{a}_{ik}{}' = \bar{a}_i + m_k(s_i + s_0)\, d_{ik}{}'$$

Support vector machine (SVM): SVM is supervised machine learning analysis as the method uses prior knowledge to assign the class of a test set. The key idea behind the SVM is the concept of decision plane which defines decision boundaries to separate a set of objects into a different class. In doing that the algorithm often constructs hyperplanes in multi-dimensional space in separating multiple classes. Often mathematical functions or kernels are applied to transform a non-linear decision tree into a linear one. The mathematical features of SVM i.e. flexibility in choosing a similarity function, sparseness of solution when dealing with large data sets, the ability to handle large feature space have made extremely popular method for analysing microarray data (Rojas et al. 2009).

Decision Tree: Decision tree is essentially a classifier that classifies the data by posing a series of question about the features associated with the data. Each question is designated to a node and every internal node points to a child node for each possible answer to its question. The questions represented as hierarchy and encoded as a tree and that is why the classifier is called “decision tree”. An item is sorted into a class by following the path from the topmost node i.e. the root to a node without children called a leaf. The item is assigned the class of the leaf it reaches. In certain variations, each leaf contains a probability distribution over the classes that estimate the conditional probability that an item reaching the leaf belongs to a given class. Decision trees are grown by adding question nodes incrementally by using labelled training examples to guide the choice of questions. A good question will split a

collection of items with heterogeneous labels, stratifying the data in a way that there is little variance in each stratum. The most common measures to evaluate the impurity within a decision tree are entropy and Gini index. Let assume that we are building a classifier based on the decision tree using a set of E training items into m classes. Suppose pi (i=1, …., m) be the fraction of items of E that belong to class i. The entropy of the probability distribution $(pi)^{m}_{i=m}$ gives a reasonable measure of the impurity of set E. The entropy, $-\sum_{i=1}^{m} pi\, logpi$ is the lowest when a single pi equals 1 and all others are 0. But it is maximized when all pi are equal. The Gini index is computed as 1- $\sum_{i=1}^{m} pi^2$. The Gini index will be 0 when the set E contains items from only one class. Given a measure of impurity I, we choose a question that minimizes the weighted average of the impurity of the resulting children nodes. That is, if a question with k possible answers divides E into subsets $E_1, \ldots, E_k$; we choose a question to minimize $\sum_{j=1}^{k}(|E_j|/|E|)I(E_j)$ (Kingsford & Salzberg 2008).

Random Forest: Random forest method is an application of the Decision Tree model based on the generation and comparison of an ensemble of trees. RF works as a large collection of uncorrelated decision trees. It applies bagging to construct a collection of decision trees aiming at identifying a complete set of significant features. After many decision trees are generated the class assignment goes with the most predominant assigned class by those trees.

Given an ensemble of classifiers $h_1(x)$, $h_2(x)$, . . . , $h_K(x)$, and with the training set drawn at random from the distribution of the random vector Y, X, define the margin function as:

$$mg(X,Y) = ask\ I(h_k(X) = Y) - max\ av_k\ I(h_k(X) = j)$$

Where I(·) is the indicator function. The margin measures the extent to which the average number of votes at X, Y for the right class exceeds the average vote for any other class. The larger the margin, the greater is the confidence in the classification. The generalization error is given by: $PE* = P_{X,Y}\ (mg(X,Y) < 0)$

Where the subscripts X, Y indicates that the probability is over the X, Y space. (Breiman 2001, Enot et al. 2006)

2.1.4. Challenges of biomarker discovery

The described data mining algorithms have been instrumental in identifying biomarkers in different forms of cancer. However, comparing the number of approved biomarkers to those mentioned in the public domain reveals that majority of candidate biomarkers either failed or

did not reach the clinic yet. Even in the case of strong signal derived from analysis of high throughput data by data mining algorithms, its conversion to clinical practice meets with challenges of accurate functional interpretation. The interpretation of the high throughput data in the context of molecular pathophysiology of an underlying disease and specific treatment is the current rate-limiting step in the biomarker identification and validation. If properly identified, extracted and interpreted; OMICS datasets can provide valuable biological insights. Functional analysis of OMICS data requires knowledge of the molecular interactions and pathways underlying pathophysiology of diseases as well as the treatment's mode of action. Accumulated biological knowledge across different system's levels therefore needs to be collected, annotated, transformed into computer-readable form and stored in a semantically enhanced knowledge base. Such knowledge bases can then be used for knowledge-based analysis of OMICS datasets through integrative approaches that aim at finding key biological processes, pathways, interaction modules or causative network signatures that could be used as candidate biomarkers (Deyati et al. 2013). In the next section, the development and analytic view of existing knowledge bases have been presented. The focus will be on the interpretation of functional role of promising candidate biomarkers after statistical analysis of gene expression data. The emphasis will also be on biomarker identification in the scientific literature and the state-of-art knowledge representation techniques for modelling and mining approaches.

3. Knowledge Management for Biomarker Discovery

The mission of knowledge bases is to collect and systematize biomedical information through manual information extraction from primary publications in so-called curation process. The curation process organises knowledge via mapping of extracted information to an underlying ontology. Such knowledge bases provide a number of features for analysis of expression data by overlying the data onto the known pathways. Identification of the key affected pathways within the expression data followed by network analysis can potentially identify key regulatory molecules behind the respective gene signatures. The functional interpretation of microarray starts with the identification of genes with characteristic biological function. National Centre for Biotechnology Information (NCBI) developed and maintained Entrez Gene database which provides unique identifiers for gene and associated biological function. The database also provides links to external databases dedicated for the biological function and pathway association of the gene i.e. Gene Ontology (GO), KEGG, Reactome.

3.1. Gene ontology

In 2000, The Gene Ontology Consortium (http://geneontology.org/) was established with the mission to provide a controlled vocabulary for annotating homologous genes and proteins across organisms. Genes and its products are classified into three hierarchical structures to describe associated biological processes, cellular components and molecular functions (Ashburner et al. 2000)]. Quite a few bioinformatics analytical tools i.e. MAPPFinder (Doniger et al. 2003), GoMiner (Zeeberg et al. 2003)], Onto-Express (Draghici et al. 2003)] have been developed to functionally interpret high throughput data with the knowledge from Gene Ontology (GO).

However the complex biological relationship of genes, proteins and their interactions can be oversimplified by annotating them into categories of GO. Protein-protein interaction databases and Pathway databases are more potent representation of the biological systems in its entirety. So interpreting microarray data in terms of over or under expressed biological pathways leading to certain phenotypes, can better reflect systems behaviour in different experimental conditions. Nevertheless, GO contains a systematic hierarchical nomenclature of genes and proteins which has played a vital role for batch processing and future design of pathway databases (Tsui et al. 2007).

3.2. Proteins-protein interaction (PPI) databases

Ever increasing number of publications in biology provides a great source to screen molecular interactions in human and in other model organisms. However, the actual challenge lies in retrieving those interactions from the literature and deciphering the dynamics of the interaction network. In recent years, great amount of effort has been undertaken to construct the several proteins-protein interaction (PPI) database described in Table 3.

Database	Properties
Intact (Orchard et al. 2014)	• Manually curated or user submitted PPIs. • Knowledge of experimental details, such as the experimental technology used, cellular context, protein modifications, expression systems, and a confidence value is added to each interaction. • Interactions can be downloaded in MI-Tab, MI-XML, BioPAX, XGMML, RDF format for visualization in different software. • MINT database has been merged with Intact. • Currently contains 87,006 interactors having 529,495 interactions.
HPRD (Keshava Prasad et al. 2009)	• Manually curated PPI database based on experimental evidence. • Knowledge on post translational modification, tissue expression, subcellular location, domain architecture, disease association is added to each interaction. • Interactions can be downloaded as XML or tab delimited format. • Currently contains 30, 047 proteins having 41,327 interactions.
DIP (Xenarios et al. 2000)	• Manually curated as well as automatically extracted experimentally validated PPIs. • Knowledge on protein-protein relationship, properties of interacting networks, evaluation of PPIs have been added. • The interactions can be downloaded as native XML based XIN format, tab delimited format and Molecular Interaction Format (MIF)

	• Currently the database contains 27,701 proteins having 79,339 interactions.
BioGrid (Stark et al. 2006)	• A manually curated protein and genetic interaction database. • Knowledge on organism specificity, proteins function and interaction detection technology has been added. • Data can be downloaded in tab delimited and PSI-MI XML format. • Currently the database holds 557,106 unique interactions from 56,907 proteins.
MINT (Licata et al. 2012)	• A manually curated and experimentally validated molecular interaction database. • MINT was one of the first interaction databases to score each interaction reflecting its reliability. • Interaction data can be downloaded as PSI-MI XML format. • As of 2011, 235,000 binary interactions were stored in the database.
STRING (Szklarczyk et al. 2014)	• Curated as wells as predicted PPI database. • The database contains both physical and functional interactions from four sources i.e. genomic context, high-throughput experiments, co-expression and manual curation of published articles. • Each interaction has been scored and organism specific interactions can be screened. The proteins have been annotated with its structure and function. • Currently STRING database covers 9,643,763 proteins from 2,031 organisms.

BIND (Bader et al. 2003)	• Biomolecular interaction database Network or BIND archives PPIs screened through yeast two hybrid, mass spectrometry, genetic interactions and phage display. • The BIND data now has been incorporated within Biomolecular Object Network Databank (BOND). • The PPIs and interaction network with BOND can be visualized within Cytoscape.

Table 3: Protein-Protein interaction data bases

3.3. Pathway databases

In the past decades of research has led to the understanding of the protein-protein interactions (PPIs) within cells in different physiological conditions. But individual PPI cannot elucidate how cell collectively responds to cues in their internal or external environment. In addition, the discovery of the connections between each of these components promoted the reconstruction of the chain of reactions, which subsequently give rise to a signaling pathway. Ultimately, our ability to interpret the function and regulation of cell signaling pathways is crucial for understanding the ways in which cells respond to external cues and how they communicate with each other (Bauer-Mehren et al. 2009)]. So building pathways and its databases is an important milestone in the quest of functional interpretation of expression data. In the next section (Table 4) the most crucial open source pathway databases are discussed.

Pathway Databases	Description
KEGG	• Established in 1995 Kyoto Encyclopedia of Genes and Genomes (KEGG) contains manually assembled pathway maps based on the duration of published literature. • KEGG pathways can be grouped together into five categories namely metabolic pathways, Cellular processes, Genetic information processing pathways, Environmental processing pathways and pathway for human diseases. • Most of the cancer associated pathways fall into the category of environmental information processing group. This section is further grouped together into membrane transport, signal

	transduction and signaling pathways. • Other than human pathways, KEGG also contain pathways for other model organisms like rat, mouse, chimpanzee, cow and pig. • The granularity, hierarchy and graphical representation as well as annotation of the nodes of all the KEGG pathways are stored in KEGGML (KEGG Markup Language) format. (Kanehisa & Goto 2000)
Reactome	• Reactome is a manually curated peer reviewed pathway database based on published literature. It was established in 2003 in collaboration with European bioinformatics institute, Ontario Institute of Cancer Research and New York University of Medical Center. • Human pathways are hierarchically structured into following categories: cell cycle, cell-cell communication, cellular responses to stress, chromatin organization, circadian clock, developmental biology, disease, DNA repair, DNA replication, extracellular matrix organization, Hemostasis, Immune System, metabolism, metabolism of proteins, muscle contraction, expression, neuronal system, organelle biogenesis and maintenance, programmed cell death, reproduction, signal transduction, trans membrane transport of small molecules, vesicle-mediated transport. • Other than human pathways, Reactome contain pathways from 18 different organisms. • Reactome is open-source, open-data and have continuously supported the major open-data standards in the domain; including BioPAX levels 2 and 3, PSI MITAB, SBML-ML (25) and SBGN export format (Croft et al. 2014).
BIOCARTA Pathways	• Biocarta pathway is an open source, community fed pathway database maintained within Biocarta. Human and mouse Pathways are manually curated from the scientific literature. • Pathways are clustered into developmental biology, expression, hematopoiesis, immunology, metabolism and neuroscience. (Diez et al. 2010)

Pathway Interaction Database	• Pathway interaction database is a collection of human signaling and regulatory pathways curated from peer-reviewed literature with a focus on cancer pathways. Since its inception in 2006, the database has been created in collaboration between US national cancer institute and Nature Publishing group. • The database was designed to deal with two issues affecting the pathway representation i.e. the arbitrariness of pathway boundaries and the need to capture knowledge at different levels of details. • The database is open-source, open-data and has continuously supported BioPAX levels 2 and XML format (Schaefer et al. 2009).

Table 4: Open source pathway data bases

As the protein-protein interaction and pathway databases started to grow, a surge for software systems providing statistical analysis of the gene expression data followed by functional interpretation was imminent. Capturing the trend several public and commercial knowledge bases have been introduced that offer an integrated environment consisting of an annotated knowledge base and analytical tools to analyse gene expression data. The aim is to perform a full-fledged statistical analysis with GUI following functional interpretation. Overviews of knowledge bases summarizing their main features as well as published examples of their application in biomarker discovery are discussed in next section.

3.4. Integrated software systems for analysis and interpretation of expression data

3.4.1. Metacore

Metacore is an integrated commercial knowledge base from Thomson Reuters (previously GeneGo) which can support functional analysis (pathways, networks and maps) of OMICS data including microarray, sequence based gene expression, SNPs and CGH arrays, proteomics and metabolomics. It can rank the affected pathways and networks from the experimental date based on proprietary algorithms. The tool has also got filters based on disease, tissue, sub cellular localization and functional processes to capture specific network. The toxicology application of Metacore is specifically designed to discover safety, efficacy and toxicity biomarker to a chemical compound. (see:

http://www.genego.com/metacore.php). Brentnall *et al.* in collaboration with Institute of Systems Biology completed a quantitative proteomic analysis to investigate differentially expressed proteins associated with ulcerative colitis (UC) neoplastic progression. Functional analyses of differentially expressed proteins with Metacore software identified Sp1 and c-MYC as biomarkers of early and late stage of UC tumorigenesis. The same collaborative group made an ICAT-based quantitative proteomics research to analyze protein expression in chronic pancreatitis in comparison with a normal pancreas. Metacore assisted pathway analysis revealed that c-MYC as a prominent regulator in the networks of differentially expressed proteins common in pancreatic cancer and chronic pancreatitis. Another collaborative group with Bayer Schering Pharma discovered the functional link between the KRAS mutation and Erlotinib resistance in non-small cell lung carcinoma (NSCLC). The functional analysis of RNA expression data with Metacore indicated a possible correlation between differential expressions of cell adhesion proteins to NSCLC (Brentnall et al. 2009, Chen et al. 2007, Fichtner et al. 2008).

3.4.2. IPA

IPA is a manually curated commercial knowledge base from Ingenuity systems (now a part of Qiagen). Its biomarker filter is specialized to prioritize the molecular biomarker based on species specific connection to diseases, detection in body fluid, expression in specific cell type, cell line, clinical samples and also in stratification biomarker discovery based on disease state or drug response. The tool also can produce functional annotation of the biomarker including pathway association (see: http://www.ingenuity.com/science/knowledge_base.html). Using Ingenuity pathway analysis Merck & Co predicted and then experimentally validated that phospho-PRAS40 (Thr246) positively correlates with PI3K pathway activation and AKT inhibitor sensitivity in PTEN deficient mouse prostate tumor model and triple-negative breast tumor tissues. Bristol-Myers Squibb has analyzed gene expression signature of responders and non-responders to neoadjuvant ixabepilone therapy in breast cancer. Functional analysis of the data with IPA has indicated that significant deregulation of certain proliferation and cell cycle control genes can potentially predict treatment sensitivity. Cleveland clinic reported a functional analysis with IPA of the genes carrying non synonymous SNPs that may be associated with the severity of sunitinib-induced toxicity in metastatic clear cell renal cell carcinoma. As per the functional analysis those genes clustered around biological processes like interferon gamma,

TNF alpha, TGF beta 1 and amino acid metabolism molecular pathways (Andersen et al., 2010; H. Chang, C. E. Horak, P. Mukhopadhyay, C. Lowery, 2011; P. W. Faber et al 2008).

3.4.3. Pathway Studio

Pathway Studio is commercial software from Ariadne Genomics (now Elsevier) for pathway analysis as well as analysis of high throughput OMICS data. It is based on proprietary databases from Ariadne like ResNet, DiseaseFx, ChemEffect, Mamalian and Plant database to build relationships between biomolecules and design pathways. The databases are built based on proprietary NLP based relationship extraction from scientific literature. The software suite also provide state of the art network algorithm to indicate important nodes from the network perspective. The researcher can also put weightage on each relationship in the pathways based on the number of literature evidence (see: http://www.pathwaystudio.com/). A group from Harvard Medical School published functional connection of 117 highly differentially expressed genes to endometrial cancer. Pathway Studio assisted analysis of the data predicted that many of these genes are correlated to angiogenesis, cell proliferation and chromosomal instability. Furthermore, they also reported 10 key differentially regulated genes to be associated with tumor progression. Xiao *et al.* published functional analysis of EGFR regulated phosphorproteome in nasopharyngeal carcinoma (NPC) to shed light on EGFR downstream signaling. They first identified 33 unique phospho proteins by 2D-DIGE and mass spectrometry. Based on the proteomic data the group built EGFR signaling in NPC by using Pathway Studio and also validated GSTP1 as one of the key EGFR-regulated proteins which is involved in chemo-resistance in NPC cells (Wong et al. 2007; Ruan et al. 2011).

3.4.4. Oncomine

Oncomine was originally developed under Compendia bioscience (now a part of Thermo Fisher Scientific). At the moment, the service is limited to breast and colon cancer (see: http://www.compendiabio.com/). Using Oncomine a group from the University of Michigan predicted that decreased protein expression of Raf kinase inhibitor protein (RKIP) is a prognostic biomarker in prostate cancer. Another group from the same University predicted that high expression of EZH2 and ECAD was statistically significantly associated with prostate cancer recurrence after radical prostatectomy (Fu et al. 2006; Rhodes et al. 2003).

3.4.5. NextBio

NextBio (now a part of Illumina) has got two major components namely NextBio clinical and NextBio research described as follows. NextBio Clinical involves semantic based integration of the proprietary OMICS data with public knowledge to get better insight further leading to discovery of drug targets and biomarkers. NextBio Research can specially identify crucial pathways leading to a disease phenotype supported by cross studies and multiple data points. The tool also elucidates the identification of disease biomarker and analysis of pharmacokinetic profiles or toxicity indications. NextBio uses proprietary algorithms to rank the search outcomes based on the statistical significance of the correlation supported by biological data points (see: http://www.nextbio.com/b/nextbioCorp.nb). Using the NextBio platform Walia *et al.* reported that loss of breast epithelial marker hCLCA2 (chloride channel accessory protein) promotes higher risk of metastasis (Walia et al. 2012).

3.4.6. BEL and Reverse Causal Reasoning

Selventa introduced a proprietary technology called Reverse Causal Reasoning (RCR), a proprietary technology introduced by Selventa. It is a computational methodology for the interpretation of large-scale biological datasets (microarray, RNA-Seq, proteomic/phosphoproteomic and metabolomic) designed to compare different biological states (e.g., treated versus control, diseased versus normal or time course and dose analyses). RCR is an automated reasoning technique to generate novel hypothesis by processing networks of causal relationships. The generated hypotheses are also evaluated by the tool against the available data sets of differential measurements. Next, each hypothesis links a biological entity to measurable quantities that it can influence. RCR analysis attempts to answer the question "What signaling differences could lead to the observed differences in measured quantities?" As a reference for reasoning, RCR uses a network data structure called a Knowledge Assembly Model (KAM). KAM is a directed network of experimentally validated causal interactions between biological entities (e.g., mRNAs, protein activities, chemicals, processes). The causal relationships within a KAM are encoded as BEL (Biological Expression Language) statements. BEL was developed with the aim to support biological knowledge curation by providing qualitative causal inference using large data sets. Within a KAM causal edge from A to B represents prior scientific knowledge, asserting that a change in A was demonstrated to cause a change in B in one or more controlled experiments supported by a specific citation (Selventa 2011). The platform has enabled the discovery of

predictive response biomarkers by reverse engineering disease mechanisms *a priori* from molecular patient data (OMICS data). It identifies disease- and tissue-specific biomarker content that can match targeted therapies to sub-population of patients. Reverse Causal Reasoning (RCR) algorithm is used for identification of master regulators. Very recently, Selventa has introduced its open BEL framework for biomarker discovery based on mechanistic causal reasoning and demonstrated its application in stratifying responders to ulcerative colitis drug, infliximab, from non-responders based on identification of IL6 as the biomarker for alternative disease mechanisms in non-responders (Mal et al. 2012).

3.4.7. transMART

A knowledge management platform enabling integration of the OMICS data with published literature, clinical trial outcome and established knowledge from Metacore, Ingenuity IPA and NLM resources. The applications of this platform include making novel hypothesis, validating them, disease association of certain pathways, genes, SNPs and biomarker discovery. Analysis of transcriptomic data from melanoma patients using k-means clustering facility in tranSMART showed that the expression levels of *cyclin D1* increase from benign to malignant whereas in metastatic melanomas the expression level decreases, clearly delineating multiple subgroups of samples in the presumably homogenous metastatic melanoma cohort (Szalma et al. 2010).

3.4.8. KeggArray

A microarray gene expression and metabolomics data analysis tool from KEGG. It is able to map OMICS data to KEGG Pathways, Brite and genome maps (see: http://www.kegg.jp/kegg/download/kegtools.html). KeggArray was used to investigate metabolic pathways associated with the marker metabolites that were detected by two-dimensional gas chromatography mass spectrometry in tissues from 31 patients with colorectal cancer. The results led to the identification of chemically diverse marker metabolites and metabolic pathway mapping suggested deregulation of various biochemical processes (Mal et al. 2012).

Although all these databases contain manually curated knowledge, their differences in the coverage and granularity of the information reflects underlying differences in methodology of information retrieval, variability of the resources used for knowledge extraction as well as the difference in interpretation of the experimental results by the annotators. Shmelkov *et al.*

have recently carried out a comparative analysis on quality and completeness of human regulatory pathways among ten public and commercial pathway knowledge bases and found that surprisingly there is little overlap in the knowledge content of these databases (Shmelkov et al. 2011a). The authors reported that the only exception was the MetaCore pathway database whose content was validated in 84% of the cases with experimental results, compared to the low overlap of 24% obtained from KEGG database.

Beside the issue of coverage and quality, the lack of consistent standard scheme for biomarker classification and biomarker knowledge representation hampered literature searches about biomarkers. The fact that qualification of translational biomarkers requires a wide range of information on the level of sensitivity, specificity, the mechanisms of action, toxicity, and clinical performance, emphasizes the need for standardization of biomarker vocabularies and classification. Recently, a prototypical process has been suggested to ensure qualification of biomarkers based on seven types of scientific evidence (Altar et al. 2008). Similarly, the Pistoia Alliance, established by information experts from several pharmaceutical companies, has launched a project focused on developing ontological and data standards for integrating biomarker assay data and handling different endpoints [see: http://www.pistoiaalliance.org/]. Although in their nascent stages, such developments can form the basis for future biomarker standardization efforts. Therefore, next-generation knowledge bases should address above challenges by introducing efficient information retrieval/extraction tools as well as biomarker data standards (Deyati et al. 2013). Taken all together, there are both advantages and disadvantages associated with existing knowledge bases, which are summarized as in Table 5.

Advantages	Disadvantages
Evidence-supported data content	Poor annotation of metadata
Structured data representation	Lack of standard representation model
Enhanced retrieval and retention of information	Lack of flexible filtering criteria
Focused semantic context	Divergent in content and subject focus

Table 5: Summary of cons and pros of biomarker-related knowledge bases (Deyati et al. 2013).

The resolution and quality of knowledge bases are largely dependent on the granularity of the underlying ontology, quality of data retrieval and experience of annotators. Creation and maintenance of manually curated knowledge bases is becoming a tremendous task in times of ever accelerating speed of publication growth different from the slow steady process of manual curation. For example, a recent report shows that assembling a compendium of potential biomarkers for pancreatic cancer, which was carried out by systematic manual curation of the literature, took over 7,000 person hours (Harsha et al. 2009). In the absence of automated methods for retrieval of biomarker information, the slow pace of manual curation cannot guarantee that the current content of knowledge bases is comprehensive and sufficient for functional interpretation of OMICS data. Novel high throughput text-mining approaches are essential for automated biomarker knowledge processing. In the next section we describe automated biomarker information retrieval methods that can be used in support of systematic update of knowledge bases and acceleration of the biomarker-related information extraction from the unstructured text (Deyati et al. 2013).

3.5. Text-mining for identifying biomarker related information

To accelerate the speed of curation process, emerging state-of-the-art information retrieval and extraction technologies are under active development. Such tools are being powered by text-mining algorithms that automatically recognize potential biomarkers such as genes and proteins in text by a process called ‘named entity recognition’ or NER (Pennings et al. 2009). However, existing NER approaches are not sufficiently selective for the retrieval of biomarker-related content information (such as its association with drug or disease) from the

literature. Consequently, studies on biomarker relation extraction from text are considered on the basis of semantic relations between named entities such as the relation between diseases and genes or proteins (Bundschus et al. 2008). Some efforts have been recently dedicated to mining and extraction of such relationships using semantically enhanced methods (Jessen et al. 2012). One limitation of these approaches is that they do not consider additional properties of candiadate biomarker, such as measurement evidence and technique besides disease and gene names. In an attempt to overcome this limitation, Ongenaert and Dehaspe (2010) have employed different keyword lists containing terms that specify methylation biomarkers in cancer and used them in conjunction with gene names from GeneCards to generate the methylation database in cancer, PubMeth (Ongenaert et al. 2008; Deyati et al. 2013).

As a step in this direction, we have recently developed a dedicated biomarker terminology organized in six proposed classes and used it for information retrieval and extraction of biomarker knowledge embedded in the literature (Younesi et al. 2012). It was demonstrated that the application of this dedicated biomarker terminology could enhance the retrieval performance significantly through combined search for cancer-related genes and selected classes of the biomarker retrieval terminology. Further evaluation of this terminology in an independent disease area, namely Alzheimer's disease, showed that not only well-known biomarkers were retrieved successfully but also new biomarker candidates could be identified. Integration of such terminologies into search tools supporting semantic and ontological search can reduce the high number of unspecific search results and improve the retrieval rate of informative documents (Deyati et al. 2013).

Ultimately, context-sensitive biomarker information extracted from literature can be used for automated enrichment of knowledge bases and/or combined with OMICS data may generate a basis for integrated models of disease or and drugs mode of action with the aim of prospective prediction of candidate biomarkers (Butcher et al., 2004).

3.6. Knowledge representation

Currently, research within biology rapidly generates new knowledge on how genes, proteins and other substances interact. A complete description of the protein interaction network underlying cell physiology is seen as one of the major goals for proteomics by the Human Proteome Organization. The US National Human Genome Research Institute recognizes the understanding of genetic networks and protein pathways as crucial parts for two out of three important areas outlined for future genomics research. In particular, the understanding of how

pathways contribute to the function of the cells and organisms, and the development of therapeutic approaches to diseases based on this knowledge are stated as two of the grand challenges for future research. They also recognize the development for reusable software modules, new ontologies and improved technologies for database and knowledge management as means for finding solutions to these challenges in the future. To fulfill this vision a format for representation of molecular pathways that allow for exchange, integration and easy creation of software tools are needed. Evaluations have shown that XML is an interesting and easy-to-use format for information representation and recent XML-based exchange formats for pathway information, e.g. SBML, PSI MI and BioPAX (BioPAX working group, 2004, http://www.biopax.org), have been proposed (Strömbäck & Lambrix 2005).

3.6.1. BioPAX

Biopax or Biological Pathways Exchange is a standard language with the aim to integrate, exchange, visualize and analyze biological pathway data at the cellular or molecular level. BioPAX is an open collaborative effort that supports data exchange between pathway data groups and thus reduces the complexity of interchange of pathways by standardizing pathway data format. BioPAX is defined in OWL DL and represented in RDF/XML format. The progress of BioPAX is defined as levels and currently three levels are available. Level 1 focus on standardizing the metabolic network, level 2 adds molecular interaction network and the most recent level 3 deals with genetic interaction and signal transduction. Prote'ge' is widely used for viewing and editing of BioPAX ontology. In BioPAX all objects are described in a class hierarchy with Entity as the most general class. Figure 10 shows the BioPAX hierarchy when loaded in Prote'ge'. Entity has three subclasses PhysEntity, representing the interacting objects; Interaction, representing the interactions and Pathway, representing a set of interactions that together form a pathway model. PhysEntity has five subclasses, complex, protein, DNA, RNA and small molecule, describing different kinds of objects that may interact. There is a large number of subclass for interaction. To unify concepts and entities between data sources containing the same or similar information about a biological phenomenon, it is possible to provide cross reference in BioPAX format (Demir et al. 2010).

Figure 10: The hierarchical structure of BioPAX data format (Strömbäck & Lambrix 2005).

Demonstrating the , BioPAX hierarchy of "physical entities", their "interaction" within a living biological systems.

3.6.2. PSI MI

Proteomic Standard Initiative (PSI) was developed in 2002 at the HUPO meeting aiming two key areas of proteomics field i.e. mass spectrometry and protein-protein interaction data. The scope of PSI MI (Proteomic Standard Initiative for Molecular Interactions) is to exchange protein-protein interaction data. The root element of any PSI MI XML is an entry set. One or more protein-protein interactions are grouped together as an entry set based on a common reason. Each entry is a self-contained unit with six descriptions namely source, availability list, experiment list, interactor list, interaction list and attribute list. There are two forms for a PSI MI format i.e. a compact and expanded form. In the compact form, all interactors (proteins), experiments and availability statements are described once in the respective list elements, and then only referred to by references from the individual interactions in the interaction list. The expanded form contains all proteins, experiments and availability statements which are described directly in the interaction elements. In PSI MI the source of the data can be either a database or a publication. Each interaction can be specified with a type of interaction i.e. aggregation, binding, phosphorylation etc. It is also possible to set a confidence level for detecting this protein in the experiment, the role of the protein and whether the protein was tagged or over-expressed in the experiment (Strömbäck & Lambrix 2005). An example of PSI MI format is provided in Figure 11.

```
<entry>
<interactorList>
  <proteinInteractor id="Succinate>
     <names>
       <shortLabel>Succinate</shortLabel>
       <fullName>Succinate</fullName>
     </names>
  </proteinInteractor>
  ....
</interactorList>
<interactionList>
  <interaction>
     <names>
       <shortLabel> Succinate dehydrogenas catalysis </shortLabel>
       <fullName>Interaction between ....</fullName>
     </names>
     <participantList>
       <proteinParticipant>
          <proteinInteractorRef ref="Succinate"/> <role>neutral</role>
       </proteinParticipant>
       <proteinParticipant>
          <proteinInteractorRef ref="Fumarate"/><role>neutral</role>
       </proteinParticipant>
       <proteinParticipant>
          <proteinInteractorRef ref="Succdeh"/><role>neutral</role>
       </proteinParticipant>
     </participantList>
  </interaction>
</interactionList>
```

Figure 11: Example of PSI MI data format (Strömbäck & Lambrix 2005).

The BioPAX and PSI MI are designed for data exchange to and from databases, pathways and network data integration. Dynamic and quantitative aspects of biological processes, including temporal aspects of the feedback loops are not supported with those two data formats. Data format like SBML, CellML are created to support kinetic behaviour of biological systems and SBGN represents pathway diagrams.

3.6.3. SBML

Systems Biology Markup Language or SBML is an open interchange format for computer models of biological processes such as metabolism, cell signalling, gene regulatory network and infectious diseases. Currently there are three levels of SBML; level 1 was dedicated to model the metabolic network, level 2 was developed to model/simulate biomolecular network. Level 3 added a number of features like model composition, description of molecule complexes, display and layout information and spatial characteristics of models. Every SBML model contains a number of compartments where the reaction occurs. The entities involved in the reaction are called species. A range of biological objects ranges from proton, atom, complex molecule like glucose, RNA or proteins can be defined as species within a SBML model. Species can also be defined with its spatial size and charge. The interactions between molecules are represented as reactions and there can be several types of reactions i.e. transformation, transport, binding etc. Reactants, products and modifiers for reactions are specified by giving references to the relevant species. The initial concentration or change in

concentration of any reactant over time can be saved inside SBML format. The reaction kinetics, mathematical and kinetic laws of the reaction can be encoded with SBML format. In addition to reactions, SBML also contains events, defined as discrete changes in the model. It is also possible to specify what triggers the event, its time constraints and the result of the event (Strömbäck & Lambrix 2005). In systems biology community SBML largely serve the purpose to encode the semantic. As a result, it is particularly well suited to encoding models of processes such as biological reactions. The mathematics may then be derived from these process descriptions as a series of ordinary differential equations (ODEs) by simulation software when desired (Smith et al. 2014). An example of an SBML representation is provided in Figure 12.

```
<?xml version="1.0" encoding="UTF-8"?>
<sbml xmlns="http://www.sbml.org/sbml/level1"
      level="1" version="2">
  <model name="gene_network_model">
    <listOfUnitDefinitions>
       ...
    </listOfUnitDefinitions>
    <listOfCompartments>
       ...
    </listOfCompartments>
    <listOfSpecies>
       ...
    </listOfSpecies>
    <listOfParameters>
       ...
    </listOfParameters>
    <listOfRules>
       ...
    </listOfRules>
    <listOfReactions>
       ...
    </listOfReactions>
  </model>
</sbml>
```

Figure 12: Example of SBML data format (Strömbäck & Lambrix 2005).

3.6.4. CellML

CellML is an extensible markup language by the international Union of Physiological Sciences, Physiome and European Virtual Physiological Human (VPH) projects. It aims to encode mathematical models of biological processes based on systems of ordinary differential equations (ODEs) and differential algebraic equations (DAEs). The scope of CellML is to support sharing of biological models by having a unified model structure (the inter-relatedness of each part of the model); mathematical equations describing biological processes; other important metadata of the model. CellML is built on Mathematical Markup Language (MathML) for encoding the mathematical part of the model and Dublin Core for

bibliographic information. A format of CellML is presented in Figure 13. Current version of CellML is complementary to SBML (Beard et al. 2009).

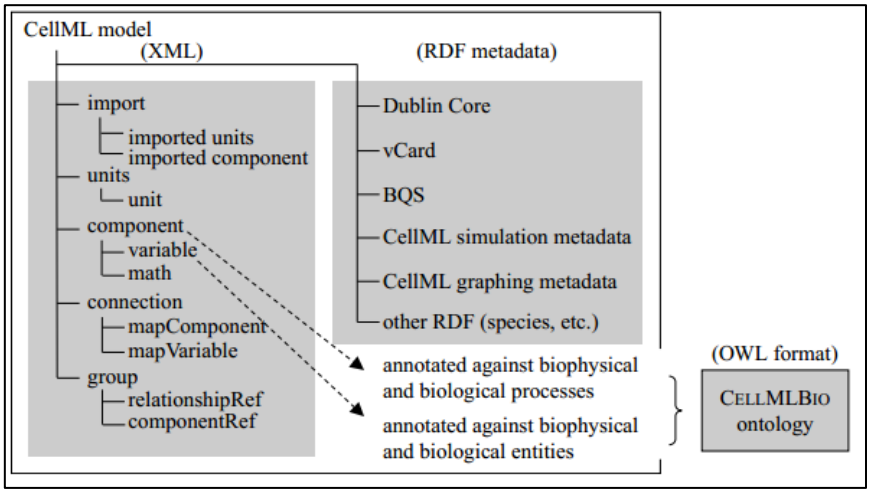


Figure 13: Entities in a CellML model (Beard et al. 2009).

The CellML model file on the left part depicting the base model with its imports, units, components, connections and groups described in XML format. The corresponding metadata is shown in RDF format. The annotation of CellML variables with biological and biophysical meaning is handled via cmeta:id links to terms stored in RDF format in a separate OWL file.

Traditionally, biochemical and cellular pathways have been graphically represented in the text books. The trend also followed in the first databases for pathways and metabolic reactions i.e. within EMP, EcoCyc and KEGG. More notations have been defined by virtue of their implementation in specialized pathway visualization software tools such as NetBuilder, Patika, JDesigner, CellDesigner. But the graphical notations used by these software tools were not standardized and their understanding relied mainly upon relating examples with one's pre-existing knowledge of biochemical processes. This ambiguity in presenting a graph was first addressed by Kurt Kohn in his Molecular Interaction Map (MIM). Nevertheless, none of these notations could become the community standard. The SBML group tried to bridge the gap with the development of SBGN (Systems Biology Graphical Notation).

3.6.5. SBGN

SBGN was formulated with the aim to specify the connectivity of the graphs and the types of the nodes and edges, but not the precise layout of the graphs. Semantics of the SBGN diagrams do not depend on the relative position of the symbols, colours, patterns, shades, shapes and thickness of the edges. There are three types of languages within SBGN.

a. SBGN process diagram: It represents biochemical reactions that change location and state of physical entities. To enable such representations different states of the

physical entities are represented separately. Process diagram cannot properly represent reaction with combinatorial explosions of states and processes. The process diagram is particularly useful to represent metabolic pathways with unambiguous transcription into biochemical events and mechanistic description of the processes.

b. SBGN entity relationship diagram: It represents the interaction between entities and rules all the controlling factors of the interaction. Unlike process diagram, physical entities are represented only once. Mechanistically it describes the relationships of the entities. The reactions with creation, destruction and translocations are not easily represented by an entity relationship diagram. Entity relationship diagram is most suitable to represent signaling pathways involving multi-state entities.

c. SBGN active flow diagram: Active flow diagram is uniquely suited to represent the influence of biological activities on each other. As it represents different activities of physical entities differently. Active flow diagram is not suitable to represent reactions with association, dissociation and multi state entities. SBGN active flow diagram is most suitable to represent functional genomics reactions and signaling pathways with simple activities (Le Novère et al. 2009). The components of SBGN diagram are represented in Figure 14.

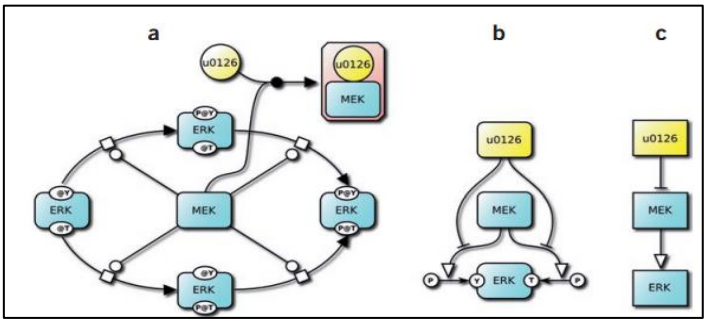


Figure 14: A SBGN representation of protein phosphorylation reaction catalyzed by an enzyme and modulated by an inhibitor (Le Novère et al. 2009).

(a) SBGN Process Diagram: Four states of ERK i.e. phosphorylated and non-phosphorylated on the tyrosine and threonine residues as well as the processes of phosphorylation by MEK and the inhibition of MEK by complexation with u0126. Note that the relationship between MEK and u0126 is not represented here. (b) Entity relationship diagram: It also shows ERK phosphorylation. At the same time relationship between MEK and u0126 is also clear. The phosphorylation sites are represented by variables, which in this example are labeled simply as 'Y' and 'T'. Unlike process diagram ERK is represented only once without the description of its different states, in entity relationship diagram. (c) Activity flow diagram: representing the activation of ERK by MEK and the inhibition of MEK by u0126. This is a simplistic representation of the activities of u0126, MEK and ERK with the abstract representations of the influences of activities on each other. But the biochemical details are missing (Le Novère et al. 2009).

BioPAX and SBGN communities have collaborated to ensure that SBGN can be used to visualize pathways in BioPAX format. The relationship among popular standard format for pathway related data is provided in Figure 15. The purpose of BioPAX and PSI-MI is to exchange the data to and from databases, pathways and biomolecular networks. SBML and CellML are designed to support mathematical simulations of biological systems. The SBGN represents a uniform notation for pathway diagrams. For example controlled vocabularies developed by PSI-MI and BioPAX can be used to annotate SBML and CellML models.

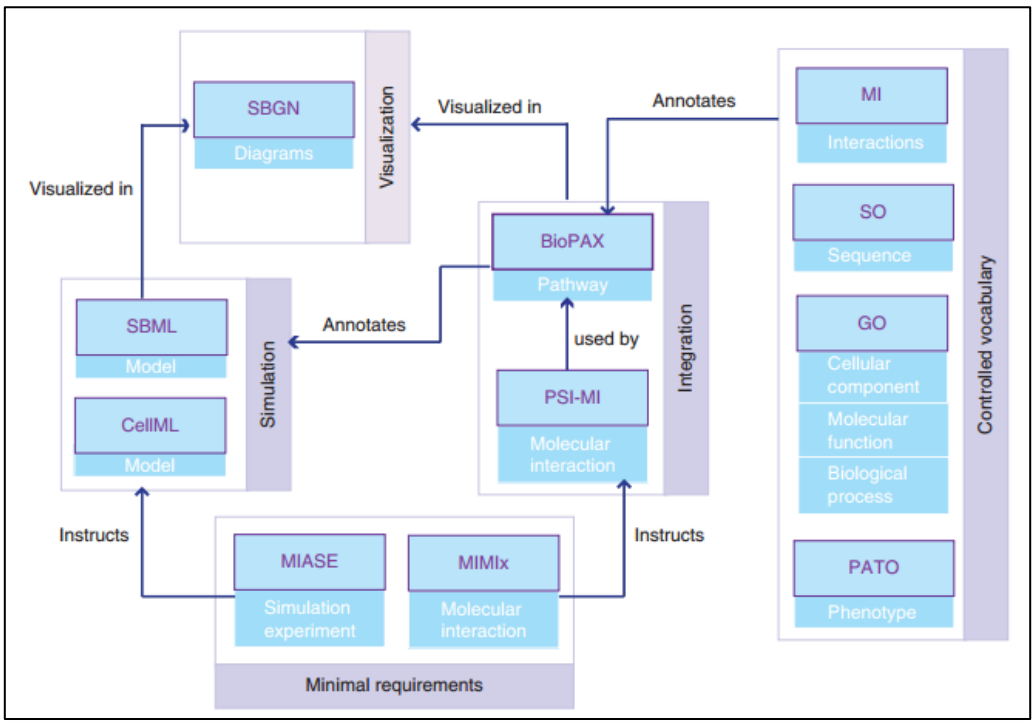


Figure 15: Inter-relationship of popular pathway data format and standard Knowledge Management Tools to represent/analyse pathway data and its downstream phenotype (Demir et al. 2010).

MIASE: Minimum Information About a Simulation Experiment (MIASE) describes the minimum set of information that must be provided to make the description of a simulation experiment available for others to use. It includes the list of models to use with their modifications, all the simulation procedures to apply according to order, the processing of the raw numerical results, and finally the description of the final output to ensure the reproducibility of simulation experiment [(Waltemath et al. 2011)].

MIMIx: Minimum Information required for reporting a Molecular Interaction Experiment (MIMIx) represents the depth of information required to describe all relevant aspects of an interaction experiment. The purpose is to ensure that the bench scientist has a checklist of the information to be supplied when describing experimental molecular interaction data in a journal article (Orchard et al. 2007).

MI: Molecular Interaction (MI) is an outcome of HUPO ProteOMICS Standards Initiative. The aim is to improve the annotation and representation of published molecular interaction data (Deutsch et al. 2015).

SO: Sequence Ontology (SO) is an ontology project for the definition of sequence features used in biological sequence annotation. SO was initially developed by the Gene Ontology Consortium (Ashburner et al. 2000).

GO: In 2000, The Gene Ontology Consortium (GO) was established with the mission to provide a controlled vocabulary for annotating homologous genes and proteins across organisms. Genes and its products are classified into three hierarchical structures to describe associated biological processes, cellular components and molecular functions (Ashburner et al. 2000).

PATO: Phenotypic Attribute Trait Ontology (PATO) designated for defining composite phenotypes and phenotype annotation (Knowlton et al. 2008).

3.7. Knowledge Visualization

Computational models of biological networks are a cornerstone of systems biology research. Over the years, quite a few modelling software tools have been developed to simulate biochemical reactions, gene transcription kinetics, cellular physiology and metabolic network. Such models promise to transform biological research by providing a podium for the following research objectives:

1. To systematically interrogate and experimentally verify knowledge of a pathway,
2. Efficiently manage the immense complexity of hundreds or potentially thousands of cellular components and interactions.
3. Reveal emergent properties and unanticipated consequences of different pathway configuration.

The major objectives of building systems biology models are to understand cellular processes, disease mechanism or drug mode of action. The snap shot of these models are built based on literature knowledge. But the dynamic properties or emergent behavior of these models are achieved with differential and/or stochastic equations based on the available contextual OMICS data. The scientific communities are looking for software tools to process, analyse and visualize these OMICS data. Several tools i.e. Pajek, Graphlet, daVinci has been published to view molecular interaction network in two dimensional space. Software like Osprey and PIMrider has got the added feature to import PPI from BIND and DIP databases. In the same way several software platforms i.e. GeneCluster, TreeView and GeneSpring have been developed for gene expression profiles. There was a need to merge molecular interaction and pathways with OMICS data in a common framework and bridging them with several other model building parameters. This need has been addressed with the development of popular platforms like Cytoscape, Celldesigner (Shannon, Markiel, Ozier, Nitin S. Baliga, et al. 2003).

3.7.1. Cytoscape

Cytoscape is a general purpose modelling environment for integrating bimolecular interaction network and states. The bimolecular interactions are represented as a network graph with nodes representing the biomolecules and edges representing the relationships between two biomolecules. Its core software component provides basic functionality for integrating arbitrary data on the graph, a visual representation of the graph, selection/filtering tools and an interface to external analytics implemented by plug-ins. The main feature of Cytoscape is described below.

Data Integration: Each data point is integrated within the network graph using Attributes. Attributes are the (name, value) pairs that map node or edge to a specific data value. The name of the node can be gene id, gene symbol, Uniprot accession no or any other ids/names. The attribute values may assume any type (e.g. text string, discrete or continuous numbers, URLs, lists).

Transfer of Annotations: The annotation feature within Cytoscape enable assigning hierarchical classification i.e. ontology of progressively more specific description of groups of nodes or edges. Annotation typically corresponds to an existing repository of knowledge that is large, complex and relatively statics such as gene ontology.

Graph Layout: This feature enable Cytoscape to visualize complex network of nodes and edges into two dimensional space. A variety of automated network layouts algorithms i.e. spring-embedded layout, hierarchical layout and circular layout are supported within Cytoscape. The spring embedded is the most widely used method for arranging general two-dimensional graphs.

Attribute-to-Visual Mapping: Attribute-to-Visual mapping is one of the most powerful visualization capabilities of Cytoscape. It controls the appearance of their associated nodes and edges based on the data attribute. Cytoscape supports a wide variety of visual properties, such as node color, shape, size, thickness as well as edge colour, thickness, shape. The data attributes are mapped to a visual property using either a lookup table or interpolation, depending on the continuous or discrete nature of the attribute.

Graph Selection and Filtering: This feature allows Cytoscape to selectively display subsets of nodes and edges according to a wide variety of criteria including by name, list of names or on the basis of attributes. More complex network selection queries are supported by a

filtering toolbox that includes a Minimum Neighbours filter, which selects nodes having a minimum number of neighbours within a specified distance in the network; a Local Distance filter, which selects nodes within a specified distance of a group of preselected nodes; a Differential Expression filter, which selects nodes according to their associated expression data; and a Combination filter, which selects nodes by arbitrary and/or combinations of other filters. The typical components of Cytoscape are represented in Figure 16.

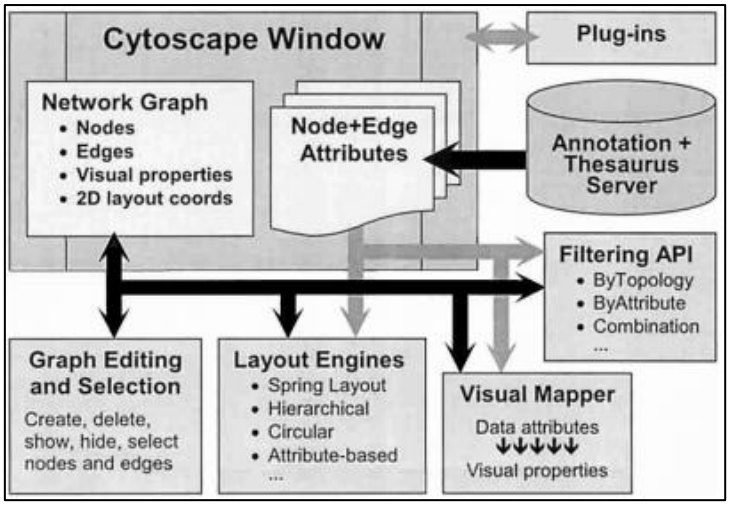


Figure 16: The components of Cytoscape (Shannon et al. 2003).

The figure shows the schematic overview of the Cytoscape core architecture. The Cytoscape window is the primary visual and programmatic interface to the software which contains the network graph and attributes data structures. Core methods that operate on these structures are graph editing, graph layout, attribute-to-visual mapping, and graph filtering. Annotations are available through a separate server.

3.7.2. CellDesigner

CellDesigner is another very popular stand-alone application for modelling and simulating biochemical networks. CellDesigner is unique software which supports the following:

a. Diverse biological objects and interactions are well defined and can be uniquely represented.
b. It is semantically and visually unambiguous.
c. It able to incorporate varied notation.
d. It can convert a graphically represented model into mathematical formulas for analysis and simulation.
e. It ensures community can freely use a notation schema.

CellDesigner supports two classes of vertexes and edges. One class of vertex, called 'state node' (SN), represents biomolecules within a biological process such as proteins, small

molecules, ions, genes and RNA. The other class is 'transition node' (TN), represents modulations imposed on the reaction, such as catalysis, inhibition, association and dissociation. In a process diagram, different states of one molecular species are represented by different SNs. SNs that represent complexes are called complex SNs (CSNs), and there are two or more SNs as components of the node. There are two types of edges: edges from a state node to a transition node (ST-Edge) and edges from a transition node to a state node (TS-Edge). There are two types of TS-edges; one that represents state changes in the molecular species (represented by a closed arrow), and one that represents translocation of the molecule (represented by an open arrow). A reaction is represented as two or more state nodes connected by edges that are connected through a transition node. Each SN may have a hierarchical internal structure defined as N-tree to represent members of a complex that are also SNs. Connectivity of internal nodes is defined by the connectivity matrix, which defines bindings among proteins that constitute a complex, as well as domains that constitute a protein. Each SN may have features that represent the modification state of residues as well as allosteric configurations (Kitano et al. 2005). Graphical notation of process diagrams are represented in Figure 17.

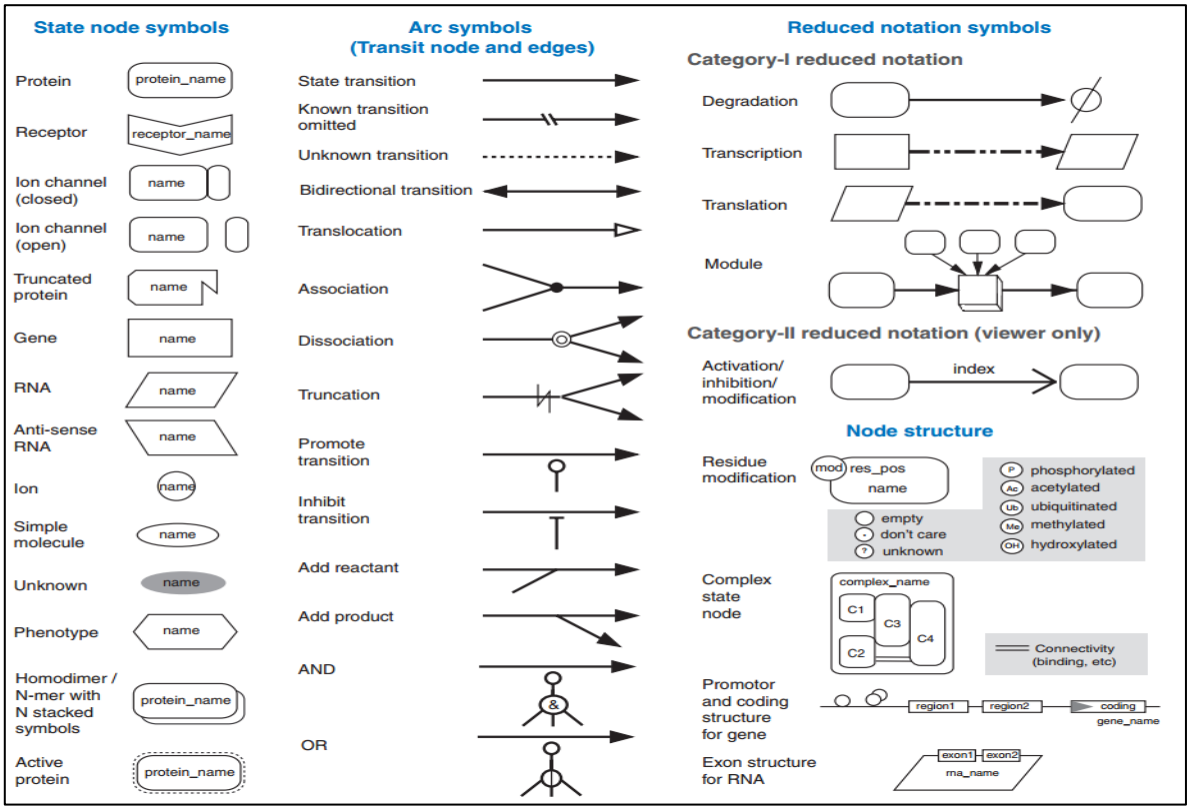


Figure 17: CellDesigner process diagrams (Kitano et al. 2005).

The figure demonstrates a set of symbols for representing biological networks with process diagrams within CellDesigner. Symbols in the process diagrams consist of visual icons for state nodes and arcs. Each arc consists of a transit node and edges. Currently, there are four reduced notations that display simplified diagrammatic symbols. The category-I reduced notation can be used during editing of the network. The category-II reduced notation is limited to viewer software, and is not permitted during the editing process because of potential confusion that could arise from the implicit nature of state transition description.

Currently pathway analytics and knowledge bases represent a useful tool for the interpretation of OMICS data, identification of upstream mechanistic drivers as well as visualisation of OMICS data assisting scientific understanding of the underlying biological processes. Though certain limitations of conventional pathway analytics hinder the use of knowledge bases as predictive tools for biomarker discovery and were recently reviewed by A. Butte (Khatri et al. 2012). Most pathways accumulated in the knowledge bases represent a mixture of findings described in different healthy and pathological conditions in various biological systems and tissues. Creation of the tissue, treatment or condition specific pathways is a challenge and is currently in focus of many commercial knowledge bases providers. Since today's knowledge bases transform multiple transcripts and SNPs to Entrez Gene id in pathway representation, granularity of the pathways should be further improved for the analysis of RNA and DNA-sequencing-derived OMICS data. Finally, existing knowledge bases contain only static information that represents "snapshots" of the system behaviour for particular condition under which the data has been obtained. Pathway interdependencies reflecting the sequence of events in pathological processes is not captured thereby limiting their use for modelling and prediction (Deyati et al. 2013).

Despite these success stories of companion diagnostics, a significant gap exists between the R&D expenditure, number of biomarker related research grants and available clinically validated biomarkers (Ptolemy & Rifai 2010). The technological advancements in the fields of genomics, transcriptomics, proteomics, metabolomics lead to a deluge of publicly available biomedical data. On the other hand, there is a vast amount of biomedical knowledge accumulated in the textual body of scientific literature and patents that could be of tremendous value for translational efforts. The lack of standardized translational algorithms allowing the use of OMICS data along with knowledge derived from scientific literature, is one major reason behind the scarcity of biomarkers currently used in the clinic. To fill the gap of OMICS data interpretation, a number of system biology approaches are suggested by the scientific community, however there is no proof of concept methodology, which can lead to the successful biomarker prediction, and it is not clear how the success of OMICS

technologies can be translated from research discovery to clinical biomarker (Deyati et al. 2013).

4. Summary of the thesis

Currently major effort in drug discovery is focused on heterogeneous disorders like cancer. The success of targeted drug development in cancer depends on how efficiently the patient will be stratified before treatment. The first scope of this thesis is to understand which OMICS technologies are currently being used for the identification of oncology biomarkers in the clinical trials and the existing methodologies for recovering biomarker-related information from the text. It also aims to draw a perspective in the integration of data and knowledge for the identification of biomarker in oncology.

Secondly the thesis strives to decipher the impact of stratified biomarker on clinical development, a retrospective analysis of clinical trials with stratified biomarker registered in ClinicalTrials.gov was achieved. The following questions were analysed:

a. What is the current frequency of stratified molecular biomarkers in clinical development?

b. What are the key technologies for biomarker identification in clinical trials?

c. What are the major funding organizations in the clinical biomarker field?

d. Which are the major disease indications targeted by stratified biomarker program?

e. At which phase stratified molecular biomarkers are mainly explored in clinical development?

f. What is the impact of biomarker program on clinical trial duration and chance of completion?

Next aim was to research the need and prospect of a novel class of biomarker i.e. miRNA to stratify the cancer patients to benefit from targeted therapy. But lack of translational algorithm which can integrate OMICS data and knowledge to predict the causal relationship between candidate miRNA and clinical outcome of a treatment in a disease condition potentially hamper the discovery of miRNA as stratified biomarker. In this direction a novel integrative algorithm i.e. SMARTmiR has been developed by combining literature knowledge and available OMICS data to identify specific miRNA as therapeutic biomarker. Finally the thesis aims to design a prospective plan on future scenario of biomarker research during cancer drug development to reduce the risk of most expensive phase III drug failures.

5. Clinical Trials and Previous Text Mining Efforts on Trials Data

Clinical trials are studies performed on number of volunteers with the aim to evaluate safety and efficacy profile of a new treatment. The concept of randomized clinical trials was first implemented in 1931 by Amberson *et al.* to study a pulmonary tuberculosis therapy. The randomizations implemented the concept of distributing patients randomly in the case control groups with sole purpose of avoiding the bias. The concept of blinded trial was also first implemented in this trial. So the patients in this trial were not aware whether they took the treatment or placebo. The volunteers need to also go through a set of stringent selection criteria termed as "Eligibility Criteria". A clinical trial is conducted through three to four phases measuring dosage, safety, and efficacy profile of the treatment (Friedman, Lawrence M., Furberg, Curt D., DeMets 2010).

5.2. ClinicalTrials.gov database

As of 1st August, 2015; ClinicalTrials.gov lists 195,624 trials covering 190 countries worldwide. All the trials are downloadable from the database as XML files. The clinical information like <<drugs>>, <<diseases>> are structured MeSH (Medical Subject Headings) terms. But the XML entries have got unstructured data as well, like trial description, eligibility criteria and other descriptions. Over the years there has been extensive focus on text mining of MEDLINE articles and abstracts but lesser text mining has been achieved has been done on clinical trials entries of ClinicalTrials.gov database. So resources of ClinicalTrials.gov can be used for potential new discoveries by text mining. Additionally trial results are often accessible much before the trials are officially completed and published making them an attractive biomedical resource to decipher trends in clinical trial domain [see: Clinicaltrials.gov].

There has been previous text mining efforts on ClinicalTrials.gov entries by using Part of Speech tagging and dictionary based approaches to simply tag an entry of interest in the trial description. The focus of this initiatives ranges from identification of gene-drug-disease relationships or development of tools for better management of clinical trials. In the next section these initiatives are reviewed in details.

5.3. Previous text mining on ClinicalTrials.gov

In 2008, Cao *et al.* completed a text mining analysis of 645 clinical trials from ClinicalTrials.gov on cancer vaccine and presented number of clinical vaccine trials per cancer type, over time, by phase, by lead sponsors, as well as trial activity relative to cancer type and survival data. The group also found out the neglected cancer areas such as bladder, liver, pancreatic, stomach, esophageal cancer in those trials. The text mining system consists of a back-end XML database, a front-end visualization interface, and the analysis component (Cao et al. 2008)

In October, 2011 Korkontzelos *et al.* presented ASCOT (Assisting Search and Creation Of Clinical Trials), an efficient search application customized for clinical trials. ASCOT uses text mining and data mining methods to enrich clinical trials with metadata to narrow down search. In addition, ASCOT integrates a feature for clinical trial practitioners in recommending eligibility criteria to volunteers based on a set of selected protocols. ASCOT employs state-of-the-art text mining technologies, clustering and term extraction algorithms applied on large clinical trial collections (Korkontzelos et al. 2012).

In 2012, Tasneem *et al*. developed a database called AACT (Aggregate Analysis of ClinicalTrials.gov) that contains data from 96,346 trials as of 27th September, 2010. The group has formulated the project with two major purposes: A) to extend the usability of ClinicalTrials.gov for research purposes. B) to develop and validate a methodology for annotating studies by clinical specialty with the help of a custom taxonomy by applying an NLM algorithm which uses Medical Subject Heading (MeSH) terms. Key design features of AACT include 1) the capacity to extend the dataset by parsing existing data; 2) linking to additional data resources, such as the Medical Subject Headings (MeSH) thesaurus; and 3) integrated metadata (Tasneem et al. 2012) .

In 2012, Li *et al*. developed a systematic approach to automatically identify pharmacogenomics (PGx) relationships between genes, drugs and diseases from trial records in ClinicalTrials.gov database. The group found out that the extracted relationships overlap significantly with the curated factual knowledge through the literature in a PGx database i.e. PharmGKB. The most relationships also appear on average 5 years earlier in clinical trials than in their corresponding publications. This suggest that clinical trials may be valuable for

both validating known and capturing new PGx related information in a more timely manner. The group collected 93,661 clinical trials as of August 2010. The group first processed these records and identified sections of interest. Second, a dictionary-based method was applied to identify PGx concepts (i.e., diseases, drugs and genes) from the pre-processed trial records. The reported gene–drug–disease relationship extraction is based on their co-occurrence in one trial record (Li & Lu 2012).

In November 2014, He *et al*. designed and built COMPACT (Commonalities in Target Populations of Clinical Trials) database to store structured eligibility criteria and trial metadata in a computable format indexed by disease topics. The COMPACT database is highly useful for clinical trial practitioner to identify common eligibility features for clinical research participant selection for a given disease indication. The group has collected 159,891 trials from ClinicalTrials.gov as of 27th January 2014 and parsed, indexed eligibility criteria text, extracted common eligibility features and developed an example analytic module called CONECT, which enables a user to mine contextual common eligibility features for trials on a certain disease from COMPACT (He et al. 2014).

In 2014, Bell *et al*. published a comprehensive characterization of clinical trials on rare diseases compared to trials on non-rare diseases registered in ClinicalTrials.gov. The group downloaded 133,128 trials as of 27th September, 2012 and by annotating medical subject heading (MeSH) descriptors to condition terms they could identify rare and non-rare disease trials. A total of 24,088 Interventional trials registered after January 1, 2006, conducted in the United States, Canada and/or the European Union were categorized as rare or non-rare. Then the group made a comparison of trials on rare and non-rare disease indication based on number of participants, number of single arm trials, number of non-randomized trials, open label, number of terminated trials, actively pursuing or waiting to commence or enrolling trials (Bell & Tudur Smith 2014)].

But none of these analyses of ClinicalTrials.gov focuses on trials with biomarker across different disease indication areas and to figure out the impact on biomarker program on clinical trials. The increasing literature evidence shows the future treatment on heterogeneous disease indication will be based on biomarker. So an analysis of trials with biomarker will provide us future landscape of stratified medicine as trials appear on an average 5 years earlier in ClinicalTrials.gov than in their corresponding publications (Friedman, Lawrence

M., Furberg, Curt D., DeMets 2010). This really motivates us to analyze Clinicaltrials.gov focusing on trials with biomarker. In the next section the analysis is described in details.

6. Impact of biomarker on drug discovery and development (Deyati, A; 2014)

Despite the advantages of the stratified medicine approach for patients and payers, such as prevention of overtreatment and an early decision for an alternative therapy (Frank & Hargreaves 2003b), the business incentive for pharmaceutical companies to invest into the co-development of the stratified molecular biomarker early on is less clear (Davis et al. 2009). Also financially it can be seen as a burden for the pharmaceutical companies as they have to bear the added cost of companion diagnostic development and yet have to cope with the reduced market size due to patient stratification (Davis et al. 2009). However, recent observations show that stratified biomarker-aided drug discovery is commercially more viable for pharmaceutical companies than post-approval research for diagnostic testing. For example clinical development of trastuzumab and imatinib with stratified biomarker have enhanced clinical, commercial success but post approval application of KRAS mutation test for cetuximab and panitumumab might not be commercially very rewarding (Loupakis et al. 2008). In order to understand the degree to which biomarker programs are implemented as well as the current trends and risks of inclusion of stratified biomarkers in clinical trials, we decided to analyse more than 150,000 clinical trials entries accumulated in ClinicalTrials.gov database.

Since the inception of ClinicalTrials.gov database in 2000 (in response to US congress law obliging NIH to publish private and federally-sponsored trials) it has rapidly become the most favoured publicly available search engine in clinical trials registry covering trials from all the geographical locations in the world. Although the database was launched in the year 2000, clinical trials with start date as early as 1970 (NCT00005125) has been incorporated in it, thus representing 43 years of clinical research. Despite the known issues with the updates, consistency and completeness of the data (Wadman 2006; Innocenzi et al. 1984), it is the oldest and largest clinical trial registry containing the information on more than 150,000 trials (August, 2013). Choosing between the clinical trials registries, we reckoned that the analysis of the largest database will give us a representative picture of the historic and current trends on the use of stratified molecular biomarkers in clinical trials.

The major question we have investigated here are as follows:

Q1. Frequency of stratified molecular biomarkers in clinical development

Q2. Key technologies for biomarker identification in clinical trials

Q3. Major funding organizations in the clinical biomarker field

Q4. Major disease indications targeted by stratified biomarker program

Q5. Phase wise distribution of stratified molecular biomarker trials

Q6. Impact of biomarker program on clinical trial duration and chance of completion

Semi-automated curation of ClinicalTrials.gov

We downloaded the entire database as XML files on 02 August, 2013 for the analysis. On that date the total number of clinical trials registered in the database was 150,504. The bigger half of the studies (121,922) was interventional i.e. analysing clinical outcome after intervention. The remaining 27,886 studies were observational. The analysis started by selecting three groups.

Interventional trials (Group 1): Focusing on 121,922 interventional trials we further filtered trials by excluding those with unknown "Overall status" to remove uncertainty regarding updates of the database. To avoid time confounders, interventional trials with "start date" between 1991 to 2013 were programmatically filtered as first trial with stratified biomarker was registered in 1991 (NCT00001271). In many trials "intervention name" and targeted disease ("condition") field can be empty, so we programmatically checked "intervention name" and "condition" field of each trial to select only those 60,629 interventional trials with a drug/biologics term as "intervention name" and a disease term as "condition".

Interventional trials with Biomarker as outcome measure (Group 2): In selecting Group 2 the "outcome measures" field of interventional trials was checked and only those with "biomarker" in it, were selected. Next, similar to Group 1 "start date", "intervention name" and "condition" fields were programmatically checked. Finally we filtered 4745 trials with biomarker as outcome measure. To check the quality of the screening method for the selection of Group 2 trials, we randomly checked 10% of it and could not find out any false positive.

Interventional trials with stratified molecular biomarker (Group 3): The prime focus of our analysis was trials with stratified molecular biomarker hence we developed a set of keywords derived from the manual annotation of 80 studies (when searched

ClinicalTrials.gov with “Cetuximab AND KRAS”). The aim of this manual annotation was to look for intuitive words which can filter a trial with stratified biomarker and a search keyword was developed (please look at Figure 18). 22,273 trials were filtered by searching the ClinicalTrials.gov with the developed keyword. Focusing on the trials that are using molecular biomarkers, we further filtered the resulting list by excluding trials without any gene or protein names by automated tagging and attained the list of 5,420 interventional studies. All the 22,273 trials with outcome of the above mentioned tagging and presence of “intervention name” are presented in Supplementary Table 1 (See: https://drive.google.com/file/d/0Bw_MQVhSKAMXSmR6TzhDcEE4eDQ/view). Automated tagging was done by running an internally developed script containing a list of all gene and its synonyms obtained from NCBI database (http://www.ncbi.nlm.nih.gov/gene/). Using the script, we screened the important XML fields of each clinical trial i.e. “title”, “purpose”, “official title”, “primary outcome measures”, “secondary outcome measures”, “detailed description” and “keywords” to filter out the trials with gene names. Further focusing on the trials that use molecular biomarkers for patient stratification prior to treatment; we manually curated those 5,420 trials. Finally 1,701 trials with stratified molecular biomarkers were filtered, and these trials were further analysed and became the basis of this paper. In ensuring the authenticity of this screening, we randomly checked 10% of the trials which did not appear in the result set (i.e. 5420-1701) for the presence of any gene names. We found out that none of the trials were false positive either. The examples of the trials filtered out by manual curation as trials with stratified molecular biomarker, please refer to Supplementary Table 2 (See: https://drive.google.com/file/d/0Bw_MQVhSKAMXZ24yLWxiak5qcjQ/view). Trials in Group 1, Group 2 and Group 3 are not overlapping.

Identification of disease terms and segmentation into therapeutic categories: One of the major aims of this analysis was to elucidate therapeutic focus of Group 3 trials. To achieve it, at first “condition” i.e. targeted disease of 1701 trials belonging to Group 3 was collected. Next, all those targeted diseases were manually curated and segmented into 15 major therapeutic areas based on MeSH disease tree. Next important focus of the analysis was to investigate the impact of stratified biomarker on trial duration of “completed” trials and trial’s status across different therapeutic categories curated in the earlier step. In order to have a balanced comparison of trial duration between target group (Group 3) and control group (Group 2), the duration period were calculated in months. The assessment of trials status was determined as follows: a trial was considered successful when “Overall status” was “completed” whereas, a trial was considered unsuccessful if “Overall status” was “terminated” (definition of

terminology: http://clinicaltrials.gov/ct2/about-studies/glossary). The comparison has been done between Group 3 and Group 1 i.e. all other interventional trials excluding Group 3 (Figure 18). For the quantitative analysis, formatting has been completed with Perl programming language based on the Document Type Definition (DTD) (http://clinicaltrials.gov/ct2/html/images/info/public.dtd) of ClinicalTrials.gov. To map diseases related synonyms to a unique identifier, MeSH disease dictionary was applied. Figure 18 shows a flow chart describing the logic and steps of our meta-analysis. All the manually curated individual disease indications belonging into 15 major therapeutic categories are listed in Supplementary Table 3 (See: https://drive.google.com/file/d/0Bw_MQVhSKAMXWHAwaFJ0b2hrUWM/view).

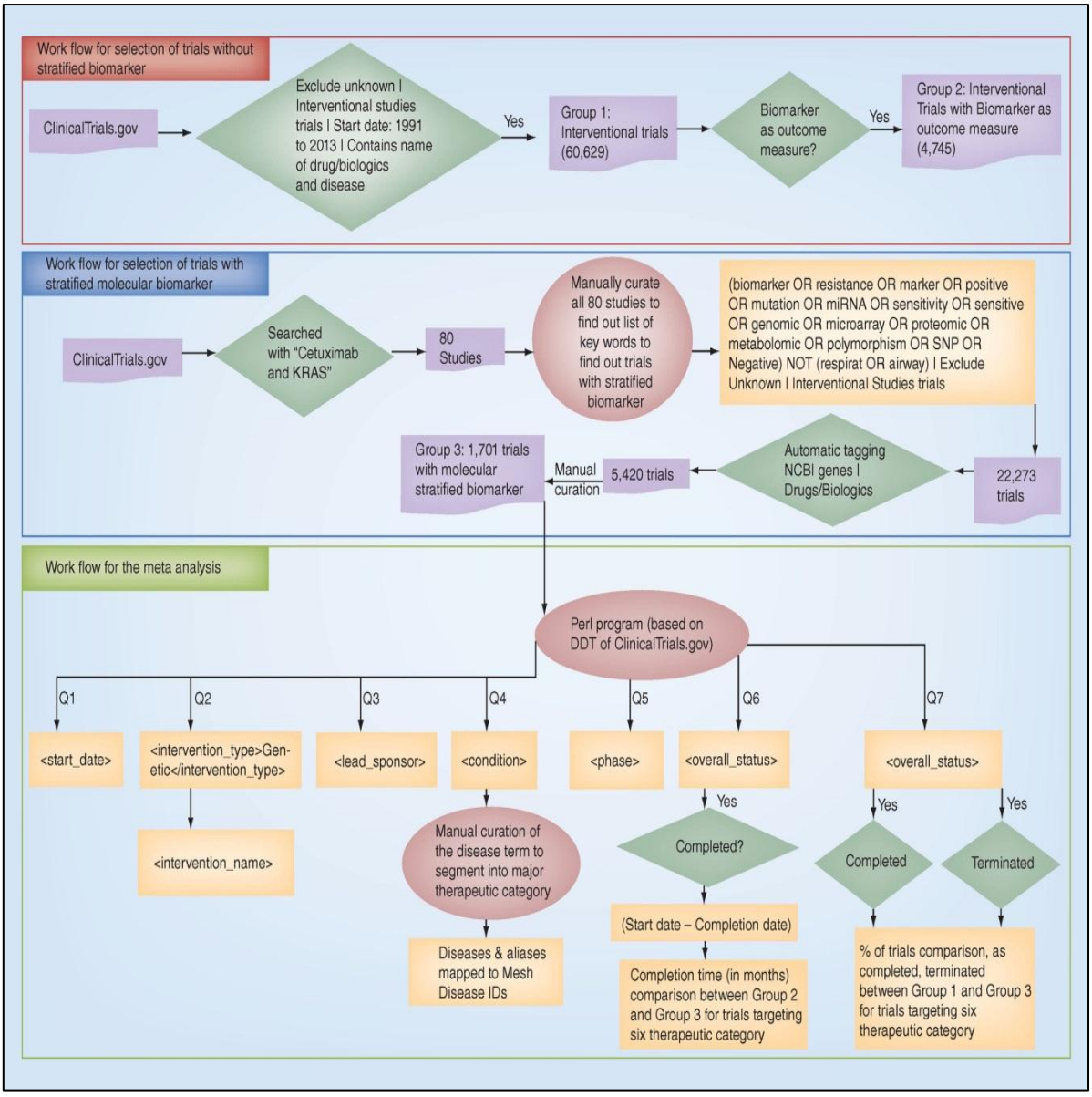


Figure 18: Workflow for the selection of trials with, without stratified molecular biomarker and meta-analysis of the ClinicalTrials.gov

Q1, Q2, Q3, Q4, Q5, Q6, Q7: major investigated questions as mentioned in page 73. The detailed description of the meta analysis consists of three steps described below.

Step1: The step1 can be separated in selection of two groups:

Group1 (Interventional trials): We used "Advanced Search" option of ClinicalTrials.gov database with two following criteria.

Criteria1: In the "Study Type" field selected "interventional studies".

Criteria2: In order to remove uncertainty regarding updates of the database, so to remove all interventional trials with unknown "overall status" from our analysis by marking "Exclude Unknown Status" in "Advanced Search".

Then we downloaded all the trials as XML file.

In avoiding time confounders interventional trials with "start date" between 1991 to 2013 were programmatically (PERL) filtered as first trial with stratified biomarker was registered in 1991 (NCT00001271). We also checked "intervention name" and "condition" field of each trial with PERL program to select only those interventional trials with a drug/biologics term as "intervention" and a disease term as "condition". As in many trials those fields can be empty. This filtering resulted in 60,629 intervention trials (Group 1).

Group2 (Interventional trials with Biomarker as outcome measure): We used "Advanced Search" option of ClinicalTrials.gov database and applied Criteria1 and Criteria2 (see step1). Further as third criteria in the "Outcome Measures" field typed "biomarker". Next, similar to Group1 programmatically checked "start date", "intervention name" and "condition" fields. Finally we filtered 4745 trials with biomarker as outcome measure.

Step2: The prime focus of our analysis was to find trials with stratified biomarker. In achieving this we first downloaded a set of 80 trials by searching ClinicalTrials.gov with keyword "Cetuximab AND KRAS". Then manually curated those 80 trials for intuitive words which can possibly filter a trial with stratified biomarker and came up with following search

Keyword: "(biomarker OR resistance OR marker OR positive OR mutation OR miRNA OR sensitivity OR Sensitive OR genomic OR microarray OR proteomic OR metabolomic OR polymorphism OR SNP OR Negative) NOT (respirat OR airway)"

Finally applied the above search keyword in "Advanced Search" option of ClinicalTrials.gov database followed by applied Criteria1 and Criteria2 (see step1). The above criteria could filter 22,273 trials. Next to find out genes and proteins, automatic tagging has been applied on 22,273 trials as we wanted to screen trials with stratified molecular biomarker (genes/proteins). We filtered out 5420 trials with tagging criteria. Finally to confirm the molecular stratification program of each of those 5420 trials, we manually curated all of them and it turn out 1701 (Group3) trials having molecular stratified biomarker program before the treatment. These 1,701 trials were further analysed and became a basis of this paper.

Step3: There are seven major questions we are investigating in this paper. In answering those questions we collected certain fields of each xml files of each clinical trial downloaded from ClinicalTrials.gov database. Below in Table 6 we listed the XML fields which were programmatically extracted for further analysis.

Question no. in Introduction	XML field
Q1	"start_date"
Q2	"intervention_name" when "intervention_type" is "Genetic"
Q3	"lead_sponsor"
Q4	"condition"
Q5	"clinical_phase"
Q6	"overall_status", filtered when value is "Completed". Duration of trials filtered through first criteria was calculated based on ("trial_start _date" – "trial_completion_date")
Q7	"overall_status"

Table 6: The analysed XML fields to answer questions rose in the introduction

Any overlap between Group 1 and Group 2 has been removed from Group 1. Similarly any overlap between Group 1 and Group 3 has been removed from Group 1. Overlap between Group 2 and Group 3 has been removed from Group 2.

6.2. Frequency of stratified molecular biomarkers in clinical development

Our primary goal was to understand whether the stratified medicine trend as a consequence of the genetic revolution, widely discussed in biomedical scientific forums; is actually translating in the use of molecular biomarkers in clinical trials. At very first step, we collected "start date" of each trial and calculated the percentage of clinical trials that are using molecular biomarkers for patients' stratification. Comparing them to the total number of clinical trials registered in the database we found 1.39% of all interventional clinical trials

(i.e. 1,701 trials out of total 121,922 interventional trials) belong to this category. In order to figure out what is the trend in implementation of biomarkers in clinical research, we analyzed the historical record of trials with stratified molecular biomarker (TSMB). The first trial using stratified molecular biomarker was registered in 1991 when CD22(+) B cell lymphoma patients were selected for treatment with IgG-RFB4-SMPT-dgA antibodies (NCT00001271). Since then, the number of such clinical trials is steadily growing and attending the peak in 2011 with 214 trials registered that year (Figure 19 a). We further calculated year wise proportion of trials with stratified molecular biomarker compare to total number of interventional trials with 95% confidence interval (CI) starting from 2000 to August, 2013. In Figure 19b, we have plotted year wise lower and upper limits of the proportion with 95% CI. Based on Figure 19b with 95% CI we can see that less than 5% of all interventional trials are using molecular biomarker for patient stratification. All the trials with stratified molecular biomarker and its start year can be found in Supplementary Table 4 (See: https://drive.google.com/open?id=0Bw_MQVhSKAMXWE1mNmFSUDRQX0U).

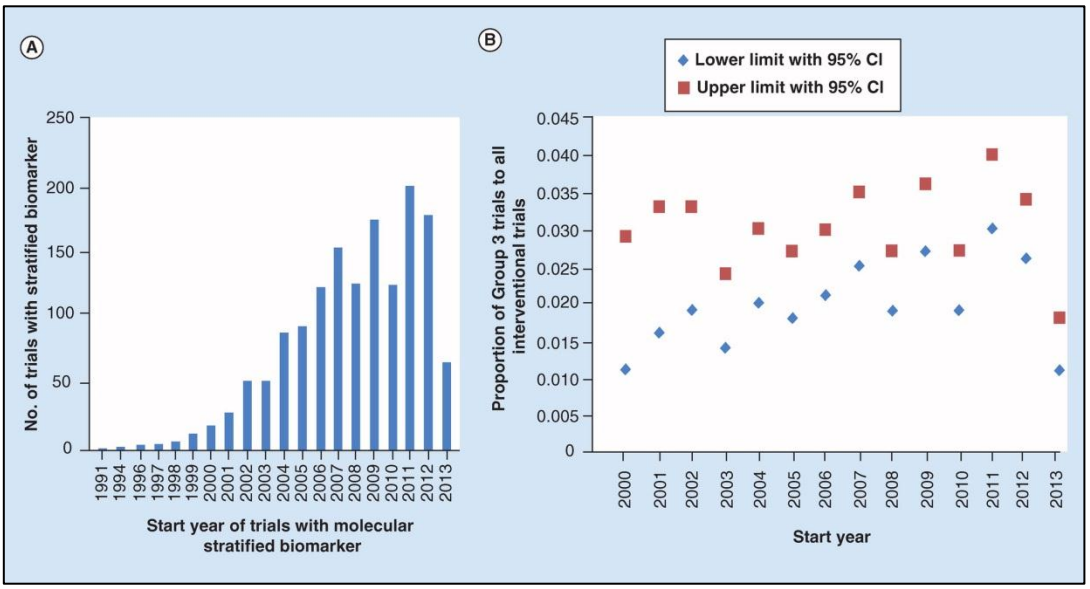


Figure 19 A: Growth of trials with stratified biomarker from 1991 to 2013. B: Year wise proportion of trials with stratified molecular biomarker compared to trials from 2000 to 2013

In figure 19 A, the steady growth rate of trials with stratified biomarker is evident. Year wise proportion of trials from 2000 to August, 2013 with stratified molecular biomarker compare to total number of interventional trials with 95% confidence interval (CI) was calculated. In Figure 19b, we have plotted year wise lower and upper limits of the proportion with 95% CI.

6.3. Key technologies for biomarker identification in clinical trials

In order to understand the technologies being applied in clinic for the detection of stratified biomarkers and their frequency of application, the "intervention name" from each TSMB was selected when "intervention type" is "Genetic". Unfortunately less than 10% of trials specify this information in the registry. TSMB trials with specified technologies are listed in Supplementary Table 5 (See: https://drive.google.com/open?id=0Bw_MQVhSKAMXeHlYN3NoVjVrQVE). After programmatically retrieving the technologies, they were manually curated and then segmented into three different OMICS technologies i.e. genomics, transcriptomics and proteomics. Genomic technologies including detection of mutations and polymorphisms by PCR, gene sequencing and cytogenetic analyses were used in 50% of trials. Supplementary Figure1 (See: https://drive.google.com/file/d/0Bw_MQVhSKAMXaHJIeUZ2bnhaNUU/view). Genomic technologies were followed by transcriptomics analysis combining various gene expression arrays and used in about 40% of trials reporting biomarker detection techniques**.** Different to the traditional single biochemical and histopathological measurements, expression profiles represent a fingerprint containing multiple biomarkers, which collectively indicate a particular pathophysiology (Bhattacharya & Mariani 2009). The rest 10% of the trials were using proteomics technologies for detection of stratified biomarker, starting from Western Blotting in early studies and ending with mass spectrometry based proteomic profiling (e.g. NCT01658566, NCT00601913). The leading role of genomic technologies in biomarker detection clearly represents the current trend in clinical biomarker discovery reflecting both stability of the genomic signal and commoditization of the genomic technologies.

6.4. Major funding organizations in the clinical biomarker field

Focusing on the 1,701 trials, we extracted the "lead sponsor" of the trials; to investigate the major players in the field investing heavily into the stratified biomarker programs. Three pharmaceutical companies such GSK, Roche and Novartis appear to be the main industrial sponsors with 3.7%, 3.1% and 2.7% (Figure 20) of all trials with stratified molecular biomarker. Therefore it is not surprising that those companies are behind most recent breakthroughs in the field of stratified medicine. GSK compound Dabrafenib with the

companion diagnostics for detection of BRAF mutation has received FDA approval for the treatment of melanoma (Ouellet D, Grossmann KF, Limentani G, Nebot N, Lan K, Knowles L, Gordon MS, Sharma S, Infante JR, Lorusso PM, Pande G, Krachey EC, Blackman SC 2013). Another GSK's MEK inhibitor trametinib has been approved with BRAF mutation as stratified biomarker (Gilmartin et al. 2011). Roche antibody-drug conjugate trastuzumab emtansine has been approved for the treatment of Her2-positive advanced breast cancer patient (Verma et al. 2012). Another Roche compound vemurafenib has been approved for the treatment of BRAF-mutated metastatic melanoma (Bollag et al. 2012). Imatinib of Novartis has got the approval for the treatment of leukemia patients with specific PDGFR and C-Kit mutations (Dupart et al. 2011). Pfizer compounds Crizotinib with the companion diagnostic (ALK5 mutation) (Sahu et al. 2013) and Maraviroc with the companion diagnostic TM test for the viral tropism have also been approved (Obermeier et al. 2012). Vast experience of these companies in conducting trials with stratified biomarker allowed for these targeted approaches and in case of Crizotinib unprecedentedly shortened the development of the drug leading to millions of savings for the company (S.-H. I. Ou et al. 2012). National Cancer Institute (12.35%) and NIAID (4.41%) seem to have the major academic drivers of trials with stratified molecular biomarker. As European agencies are not obliged to fill into the database, there was no mention of European academic player in the top 15 clinical research organizations. Other proprietary databases such TrialTrove or PharmaProjects would be better for the analysis of the European situation in the field of clinical biomarker research. Trials with stratified molecular biomarker and its lead sponsor can be found in Supplementary Table 6 (See: https://drive.google.com/open?id=0Bw_MQVhSKAMXRk1kTUxSSUVvU2M).

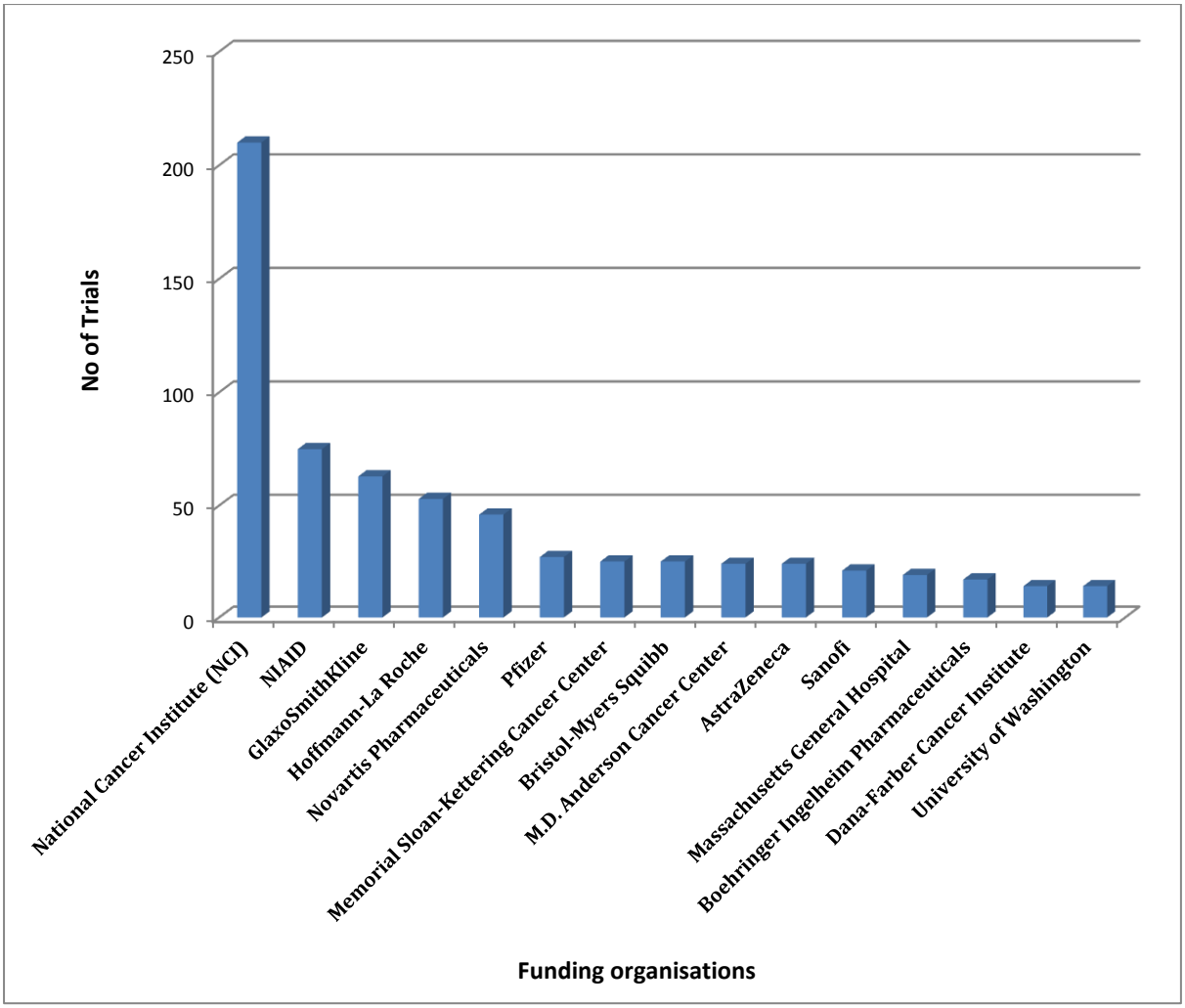


Figure 20: Major funding organizations sponsoring trials with stratified molecular biomarker.

The figure demonstrates what percentage of trials with stratified biomarker is funded by top 15 organisations driving such trials.

6.5. Major disease indications targeted by stratified biomarker program

According to our analysis (Figure 21), oncology represents more than 75% of all the trials with stratified biomarker program. Infectious disorders are next major focus of trials with stratified biomarker representing about 10% of all the analysed studies (Figure 21a). Other therapeutic areas combined, represent the rest 15% of the trials with molecular stratified biomarker and metabolic diseases leading this group (Figure 21b). Closer look at the individual indications of oncology therapeutic area reveals that most of the registered stratified molecular biomarker studies are in the field of breast cancer comprising 478 studies and constituting 28% of all trials (Figure 21c). This extensive clinical effort is translated into the marketed companion diagnostics and breast cancer patients are benefiting from it. In this

indication hormone-dependency tests, genetic susceptibility tests (such as BRCA1/2 mutations) and gene expressions analysis (such as Mammaprint and OncotypeDX) became a part of a standard clinical care (Tessari et al. 2013; Deyati et al. 2013). Lung cancer is the second most frequently targeted (20.8%) by trials with stratified molecular biomarker with 353 studies registered in ClinicalTrials.gov. Years of clinical research are translated in a number of approved biomarkers in this therapeutic area, such as Crizotinib approved in combination with the companion genetic test for the ALK5 gene for late stage lung cancer (S.-H. I. Ou et al. 2012). Leukemia and Lymphoma are the next largest oncological indications benefiting from the early discovery of stratified molecular biomarkers such as Philadelphia chromosome or other genetic translocations such as RAR-PML fusions that brought to the early discovery of the target-specific treatments, celebrating the first wins of molecular biology in clinics. Trials with stratified molecular biomarker and its targeted disease indications can be found in Supplementary Table 7 (See: https://drive.google.com/file/d/0Bw_MQVhSKAMXbHFRQTNPQ0NrZTg/view).

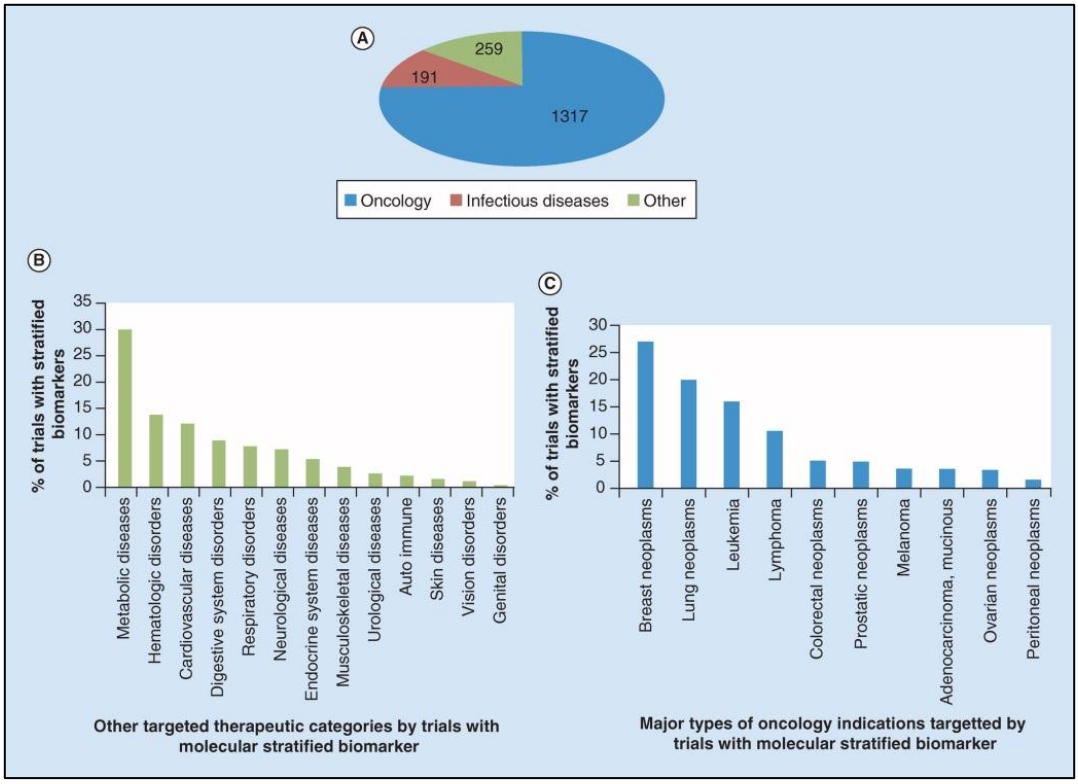


Figure 21: Major targeted therapeutic areas by trials with stratified molecular biomarker.

(A) Major indication areas targeted by trials with stratified biomarker; **(B)** Other targeted indication areas; **(C)** Major types of targeted cancer.

6.6. Phase wise distribution of stratified molecular biomarker trials

Further analysing at which stage of the clinical development the stratification biomarkers are explored, we extracted the information on the "phase" from all 1,701 TSMB (Figure 22). Around 93% of trials with stratified molecular biomarker contain the information on clinical phase. According to our analysis most of the trials, around 60%, are in phase II. This category includes the trials indicated to be in the phase I-II as well. Fifteen percent trials with stratified molecular biomarker are in phase I and III. It was interesting to see that almost 6% of all stratified molecular biomarker-associated trials are conducted at phase IV. It clearly reflects that search for stratified biomarkers are continued in the post-marketing phase as well. Trials with stratified molecular biomarker and its clinical phase can be found in Supplementary Table 8 (See: https://drive.google.com/open?id=0Bw_MQVhSKAMXQkptRTluNnRFY2s).

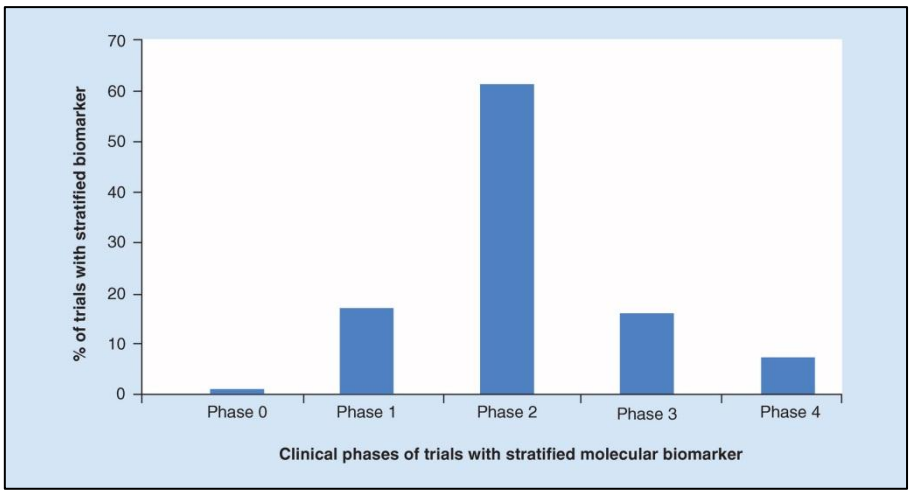


Figure 22: Phase wise distribution of trials with molecular stratified biomarker.

The figure demonstrates what percentage of trials with stratified biomarker lies in what phase.

6.7. Impact of biomarker program on clinical trial duration and chance of completion

Our next question was to analyze whether the inclusion of stratified molecular biomarker has any impact on trial duration. To answer that clinical trial duration was calculated in months as ("start date" - "completion date"). Statistical distribution of clinical trials duration, targeting disease indications of Group 2 (interventional trials with biomarker as outcome measure) and Group 3 (interventional trials with stratified biomarker) and falling into six therapeutic areas were compared in Figure 23a. The three letter abbreviations used along the x axis of Figure

23a, represents each category. The first letter of the abbreviation represents targeted therapy (e.g. O: oncology, I: infectious diseases, M: metabolic disorders, R: respiratory disorders, C: cardiovascular diseases and N: neurologic disorders) followed by two other letters representing groups (WB: Group 2, SB: Group 3). As can be seen in Figure 23a, addition of molecular stratified biomarker into clinical trials targeted to oncology, infectious diseases and neurologic disorders extend the trial duration, evident from the difference in the central tendency of data (i.e. median). We also calculated therapeutic area specific mean trial duration and difference of means with 95% confidence interval (CI) between Group2 and Group 3 trials by Welch two sample t-test and represented in Table 7. With 95% confidence interval stratification step increases trial duration by 13.3 to 2.7 months in oncology, by 27 to 7.5 months in infectious diseases and by 32.4 to 8 months in neurologic disorders. In metabolic and cardiovascular diseases the application of stratified biomarker seems to be less significant in affecting clinical trial duration. On the other hand in respiratory disorders stratified biomarkers on mean scale shorten the trial duration by 19.7 to 8 months. All the completed trials belonging to Group 2 (WB) and Group 3 (SB) and falling into above mentioned six therapeutic categories are presented in Supplementary Table 9 (See: https://drive.google.com/file/d/0Bw_MQVhSKAMXb3NocGFXRXE1TlU/view) along with its "start date", "completion date", trial duration in months.

Therapeutic area	Group 2 abbreviation, Mean trial duration in months	Group 3 abbreviation, Mean trial duration in months	Difference of means between Group 2 to Group 3 with 95% CI
Oncology	OWB, 42.2	OSB, 50.2	-13.3 to -2.7
Infectious diseases	IWB, 32	ISB, 49.3	-27 to -7.5
Metabolic disorders	MWB, 26.6	MSB, 27.5	-8 to 6.2
Respiratory disorders	RWB, 25	RSB, 11.2	8 to 19.7
Cardiovascular disorders	CWB, 27.8	CSB, 30.2	-18 to 13.3

Neurologic disorders	NWB, 21	NCB, 41.2	-32.4 to -8

Table 7: Therapeutic area specific mean trial duration and difference of means with 95% CI between Group2 and Group 3 trials

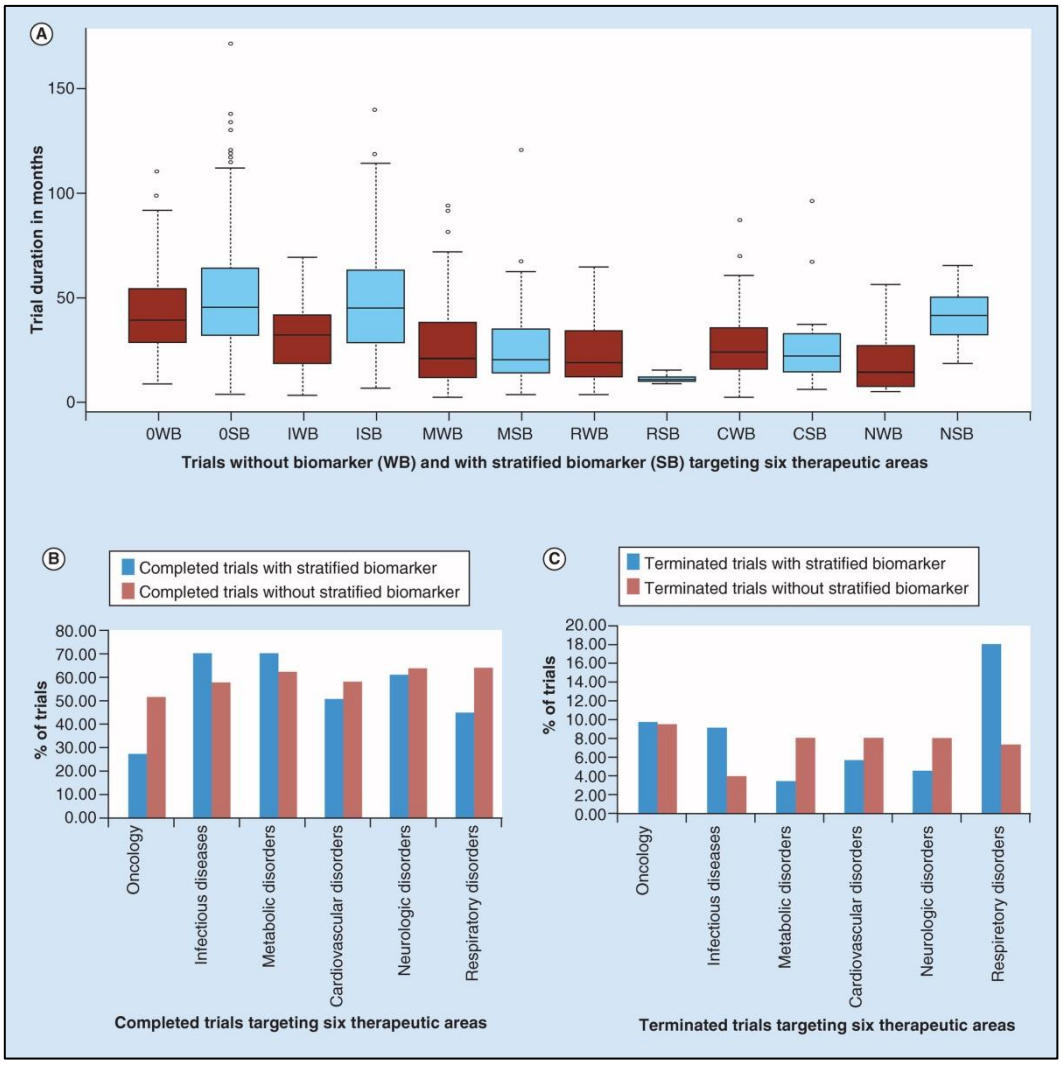


Figure 23: Impact of stratified biomarker program on trial duration and completion

(A) Comparative statistical distribution of clinical trials duration between Group 2 and Group 3. **(B)** Comparative proportion of completed trials between Group 1 and Group 3. **(C)** Comparative proportion of terminated trials between Group 1 and Group 3. Impact of stratified biomarker program on trial duration and completion.
The three-letter abbreviations used along the x axis in **(A)** represents each category. The first letter of the abbreviation represents targeted therapy (O: Oncology; I: Infectious diseases; M: Metabolic disorders; R: Respiratory disorders; C: Cardiovascular diseases; N: Neurologic disorders) followed by two other letters representing groups (WB: Group 2; SB: Group 3)

As we see, the duration of trials could be affected by the inclusion of stratified molecular biomarker; however question arises if completion itself is affected by inclusion of stratified molecular biomarkers? In addressing this question we compared the proportion of

"Completed" trials across six most targeted therapeutic areas between Group 1 (all interventional trials) and Group 3 (interventional trials with stratified biomarker) as a measure of success. To measure the rate unsuccessful trials proportion, "Terminated" trials across same six targeted therapeutic areas between Group 1 and Group 3 was compared. Proportion of "Completed" trials across six therapeutic categories of each group was calculated based on number of "Completed" trials in given therapeutic category of a group divided by total number of trials targeting the same therapeutic category of the same group. The proportion of "Terminated" trials was calculated using the same formula. Patient stratification by molecular biomarkers significantly affect the fate of the trial whether it will be successful or unsuccessful (Figure 23b, 23c). In metabolic, cardiovascular and neurologic disorder, trial termination rate is significantly higher in Group 1 (trials without a biomarker). However, fewer trials targeting respiratory and cancer disorders, are completed in Group 3 compared to the Group1 trials. If in other therapeutic areas the average percentage of terminated studies is about 8%, addition of stratified biomarkers to trials targeting respiratory disorders, doubles the chance of the trials to be terminated prematurely. In the case of oncology every third molecular biomarker trial is completed compared to every second in the non-biomarker group. According to our analysis, more than one third of Group 3 trials started after 2009 and cancer being most frequently targeted (Figure 19; 21a). Knowing that the median trial duration is ~45 months (Figure 23a), most of them were still ongoing and ~31% (411 out of 1317) of oncology trials with stratified biomarker are still in the recruiting phase. At the same time slightly higher termination of stratified biomarker trials targeting cancer, reflects the hard road of clinical development (well standardized tests, trained personnel and dedicated resources) targeting highly heterogeneous diseases like cancer (Ludwig 2012). Among all studied therapeutic areas, respiratory disorders are more challenging for stratification due to heterogeneity of the population and the fact that biomarker studies in respiratory disorders are still in their infancy (Taylor 2011; Penaloza et al. 2012; Lee et al. 2012). The result is evident in our study, which shows that more clinical trials are terminated and less completed when stratified compared to the non-stratified group. Group 3 and Group 1 trials with its "overall status" can be found in Supplementary Table 10 (See: https://drive.google.com/file/d/0Bw_MQVhSKAMXWmpid3QzUmUxT0E/view) and Supplementary Table 11 (See: https://drive.google.com/file/d/0Bw_MQVhSKAMXYjFaTWZUSG9lX0E/view).

It is evident from this analysis that percentage of the trials including patients' stratification based on molecular differentiation is still very low (less than 5%) reflecting all the challenges of biomarker development. All though a variety of OMICS technologies have been developed in recent years with the aim to identify biomarker in oncology by detailed understanding of disease pathophysiology and drug mode of action. But the search for potent biomarker is far from being over. A new class of stratified biomarker i.e. miRNA is rapidly emerging in cancer treatment with the promise of stratifying the patient population with greater confidence. But lack of translational algorithm which can integrate data and knowledge to predict the causal relationship between candidate miRNA biomarker and clinical outcome of a treatment in a disease indication potentially hamper the endeavor. In this direction a novel translational algorithm i.e. SMARTmiR, has been developed to predict the pharmacogenomic role of miRNA when treating the colorectal cancer patients with cetuximab therapy.

Next three chapters (i.e. 7, 8, 9) have been dedicated to describe each of those three concepts i.e. miRNA, colorectal cancer and cetuximab. Finally in chapter 10^{th} the published methodology of SMARTmiR has been discussed in detail.

7. microRNA

7.1. History of miRNA discovery:

The team of Victor Ambros, Rosalind Lee and Rhonda Feinbaum discovered the existence of microRNA and its regulation. They reported lin-4 gene controls the timing of *C.elegans* larval development stages. But to their surprise this gene did not code any protein instead only translated into two small RNAs 22 and 61 nucleotide long. The Ambros and Ruvkun labs also discovered that lin-4 miRNA has several anti-sense complementarities to the 3´untranslated region (UTR) region of lin14 gene. Ruvkun lab further demonstrated that after complementary binding of lin-4 RNA to lin-14, the concentration of lin-14 protein was substantially reduced without negligible changes in lin-14 mRNA. These discoveries initiated a series of other discoveries producing ample evidences of the existence and post transcriptional regulation of small RNAs, known as microRNA or miRNA (Bartel 2004). miRNA are the main regulators at post-transcriptional level. However let-7 is the first identified miRNA in human (Bartel 2004).

7.2. miRNA gene

In 2001 Lau *et al* reported that most miRNA genes transcribed from a distant part of the genome from previously annotated protein coding genomic region. This fact implies that there is independent transcriptional machinery for miRNA (Lau et al. 2001). In 2003, a group of scientists published that a minor proportion of miRNA was located in the introns of pre-miRNA. These miRNA are not transcribed from their own promoter but instead processed from the introns (Aravin et al. 2001; Aravin et al. 2003). However, in late 2004 Rodriguez *et al* proved 117 miRNAs out of 232 mammalian miRNA (~51%) are in the intronic region (Rodriguez et al. 2004). Many miRNA genes are also clustered in a certain portion of the genome. The expression of these types of miRNA falls into the category of multi-cistronic transcription (Lagos-Quintana et al. 2001). Although based on the scientific discoveries until now, majority of human miRNA genes are not clustered but isolated (Lim et al. 2003). However, these perceptions are subject to change as miRNA is comparatively new research area and new miRNA genes are constantly being reported. In 2010, a group at the Whitehead Institute sequenced 60 million small RNAs in mouse and identified 108 novel miRNA loci. The discovery opened the question how many more miRNA remain to be discovered (Chiang

et al. 2010). There are several mechanisms through which a novel miRNA can be formed which are discussed in the next section.

7.2.1. Formation of novel miRNA gene

a. As described in Figure 24 novel miRNA gene can be formed by the duplication of local or non-local miRNA gene. In case of local duplication the newly formed miRNA gene is located in the close proximity of the original and both the gene share the same transcription mechanism. miRNA gene evolved from non-local duplication are located in far apart from its original partner and has got independent transcription mechanisms (Berezikov 2011).
b. Unstructured transcripts from intron often rearrange itself into hair pin like structure which successively evolved into novel intronic miRNA (Berezikov 2011) .
c. De-novo emergences of miRNA often evolved into a transcriptional unit and produce unstructured transcripts. Later the unstructured transcripts go through the process of forming a hair-pin like structure which successively evolved into a novel miRNA (Berezikov 2011).
d. Often transposable elements can lead to the formation of novel miRNA by the formation novel miRNA like hair pin stage (Berezikov 2011).
e. tRNA, small nucleolar RNA (snoRNA) often forms miRNA through hair-pin stage.
f. Existing miRNA loci sometime undergoes antisense transcription leading to the formation of miRNA hairpins with novel mature miRNA and mRNA* (Berezikov 2011).
g. Theoretically after two rounds of genome duplication, a locus can retain a gene and miRNA in its intron or either one is lost or the entire locus can be lost (Berezikov 2011).

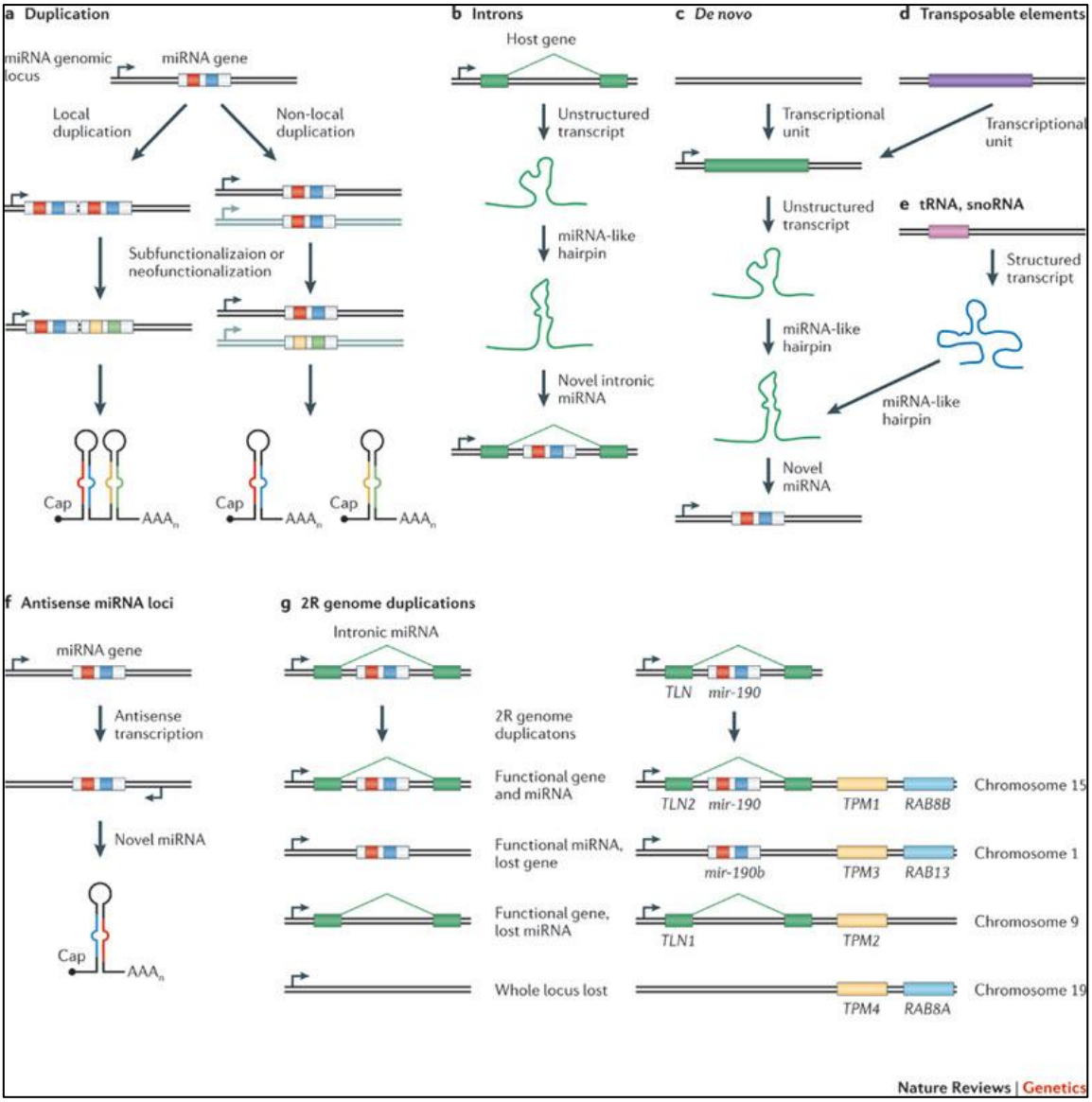


Figure 24: Genomic Sources of novel miRNA genes (Berezikov 2011).

Each biological processes (a to g) is described in detail in previous page.

7.3. miRNA biogenesis

The formation of mature miRNA is a series of processes which start with the transcription of miRNA gene followed by the nuclear and cytoplasmic processing by two RNases. The process is described in the following section.

7.3.1. Transcription of miRNA

In eukaryotes, miRNAs are transcribed by RNA polymerase II and activity of the enzyme is controlled by RNA polymerase II-associated transcription factor and epigenetic regulators. Transcription factors like p53, MYC, ZEB1, ZEB2 and MYOD1 can be positively or

negatively regulate miRNA expression. Different epigenetic events such as DNA methylation and histone modifications can also regulate the miRNA gene expression (Ha & Kim 2014). A schematic presentation of miRNA biogenesis and its control is presented in Figure 25 a and 25 b.

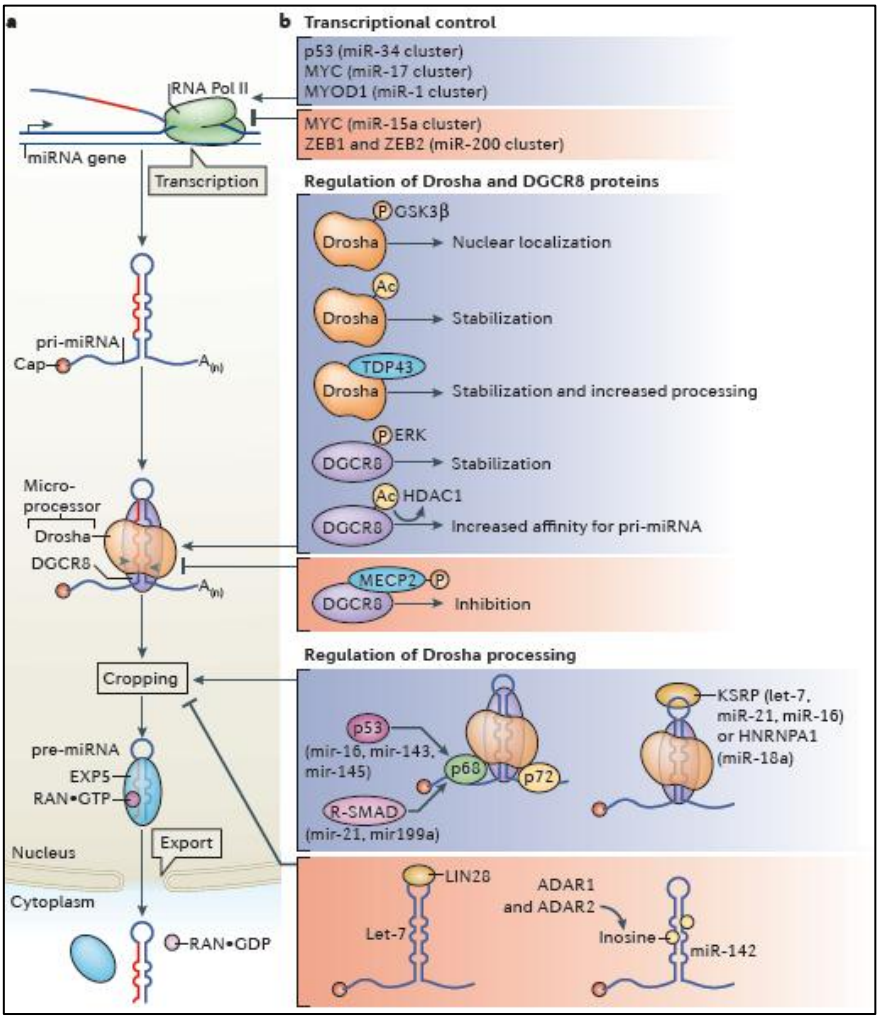


Figure 25: Nuclear event in miRNA biogenesis pathway (Ha & Kim 2014).

Schematic model of microRNA (miRNA) transcription by RNA polymerase II (Pol II), nuclear processing by the Microprocessor complex (comprising Drosha and DGCR8) and export by exportin 5 (EXP5) in complex with RAN GTP.

7.3.2. miRNA processing

After transcription miRNA goes through a series of processes to form mature miRNA. Pri-miRNA is 1 kilobases long with local stem loop structure. A typical pri-miRNA consists of 33-35 base pairs long stem, a terminal loop and single stranded RNA segments at both 3' and 5' ends. The maturation process initiates within the nucleus by Drosha, an RNase III enzyme. Drosha crop the stem loop to release small hairpin shaped RNA of ~65 nucleotide long. Unlike plants, Drosha forms a microprocessor complex with its cofactor DGCR8 for the

RNase activity in most of the animal. Before the processing of pri-miRNA, the precise recognition of pri-miRNA by the microprocessor complex is of paramount importance. The two double stranded RNA binding domain of DGCR8 precisely recognize a pri-miRNA. Drosha cleaves the hairpin structure at approximately 11 base pairs away from the basal junction between single stranded RNA and dsRNA and approximately 22 base pairs away from the apical junction linked to the terminal loop (see Figure 25). It is still quite unclear how Drosha and DGCR2 cumulatively interact with junction and the stem before cropping. The UG and CNNC motif at the basal junction and UGUG motif at the apical junction (see Figure 26) may also play a crucial role in binding other determinants for the processing of pri-miRNA. These three motifs are quite conserved and 79% of human miRNAs has got at least one of these three motifs.

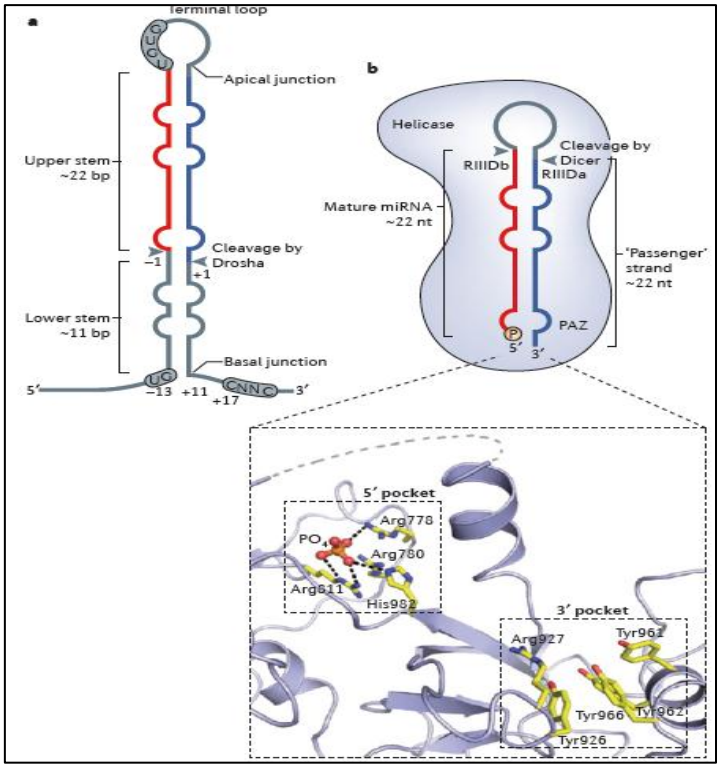


Figure 26: Recognition sites of Drosha and microprocessor complex (Ha & Kim 2014).

The Microprocessor complex (comprising Drosha and DGCR8) recognizes the single-stranded RNA tails, the stem of ~35 bp in length and a terminal loop of the primary microRNA (pri-miRNA). Microprocessor measures ~11 bp from the basal junction. ~22bp from the apical junction, and Drosha cleaves the pri-miRNA at this position | Dicer recognizes pre-miRNA. The termini of pre-miRNA are recognized by the PAZ (PIWI –Argonaute (AGO)–ZWILLE) domain of human Dicer, which contains two basic pockets: one that interacts with the 5′-phosphorylated end of the pre-miRNA and one that interacts with the 3′ end105,106. The stem of the pre-mRNA is aligned along the axis of the protein in a way that Dicer can measure a set distance from both termini (like a 'molecular ruler'), because the catalytic domains of RNase III domain a (RIIIDa) and RIIIDb are placed ~22 nucleotides (nt) away from the termini. P, phosphate.

7.3.3. Nuclear export

After the first processing of pri-miRNA by Drosha, resulting pre-miRNA is exported into the cytoplasm. In cytoplasm, the maturation process can be completed. The EXP5 protein forms a transport complex with GTP-binding nuclear protein RAN and pre-miRNA. This complex passes through the nuclear pore complex. Next GTP is hydrolysed, thus complex is disassembled and pre-miRNA is released into the cytosol.

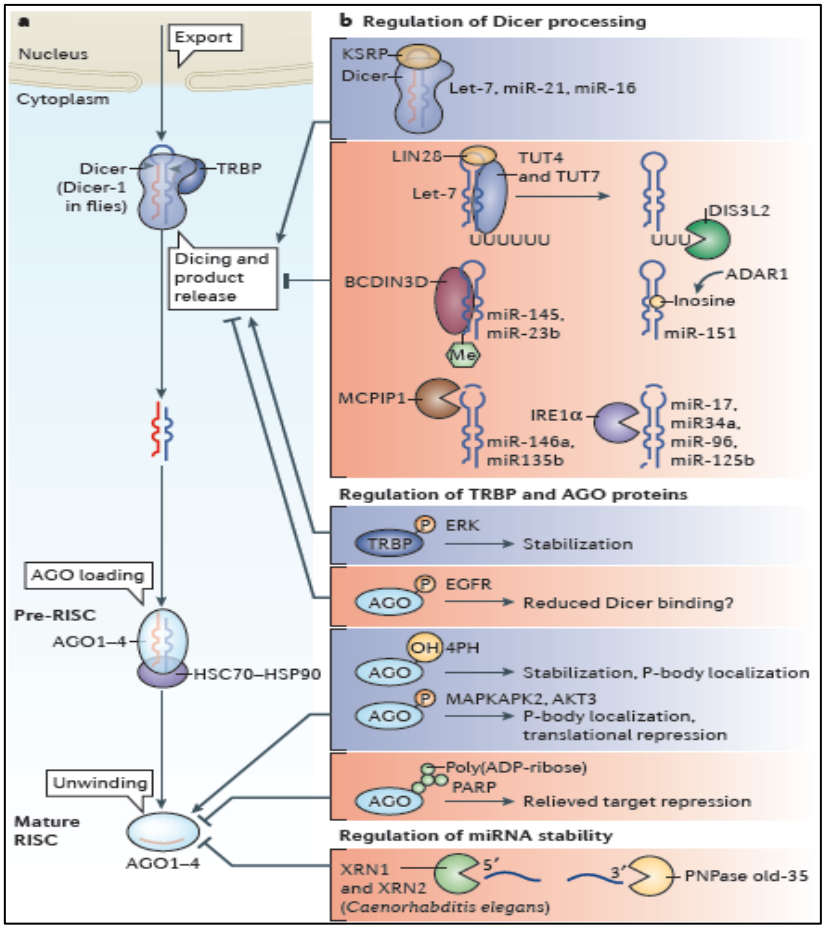


Figure 27: Cytoplasmic processing of miRNA processing (Ha & Kim 2014).

Schematic model of Dicer-mediated processing and Argonaute (AGO) loading. Dicer interacts with a double-stranded RNA-binding domain (dsRBD) protein (TAR RNA-binding protein (TRBP) in humans.

7.3.4. Cytoplasmic processing pre-miRNA

The released pre-miRNA in the cytoplasm is cleaved near its terminal loop by the Dicer complex. The Dicer interacts with double stranded RNA–binding domain (dsRBD) protein i.e. TRBP in human. The cleaving of pre-miRNA by Dicer forms a RNA duplex (see Figure 27). This RNA duplex is released and subsequently loaded onto human AGO1-4 to form RNA induced silencing complex (RISC). RISC further binds to a complex made of Heat

shock cognate 70 (HSC70) and heat shock protein 90 (HSP90). Successively the passenger strand (Blue region of the duplex in Figure 26) is discarded and remaining mature part (red part in Figure 27) remains attached to AGO1-4. Through consecutive steps mature miRNA formation and targeting of mRNA is achieved through RICS assembly. Dicer processing is also tightly regulated by post translational modification of TRBP and AGO proteins.

7.4. miRNA nomenclature

The nomenclature of miRNA is not completely consistent. The early discovered miRNA were named after their phenotypes i.e. lin-4, let-7 and lys-6. Recently discovered miRNAs are published with a sequential numbering after "mir", such that mir-2 has been discovered after miR-1. The organism specificity of a miRNA is designated by a three letter code, for example miR-1 in human will be designated as has-miR-1. If a miRNA gene encodes two sisters miRNA then they are designated as miR-125-a and miR-125-b. If identical mature sequences are expressed by distinct precursor sequences, genomic loci then numeric suffix are added at the end of the name e.g. miR-125-b-1 and miR-125-b-2. Each miRNA gene locus code two mature miRNA one from the 5' end the other from 3' end, the stand specificity is denoted as follows: miR-125-a-5p and miR-125-a-3p. However, only one of them is biologically more active and abundant called "guide" and the other "passenger" one often represented as miRNA* (Ambros et al. 2003).

7.5. miRNA expression

The miRNA has got very interesting yet very intriguing expression pattern. Several miRNA in *C.elegans* i.e. lin-4 and let-7 have developmental stage specific expression pattern. Many mammalian miRNAs express in specific tissue and organ, to name just a few; miR-1 primarily found in mammalian heart, miR-122 in liver, miR-223 in macrophages. Even the degree of expression of a miRNA which is only expressed in specific organ may have differential expression correlated to different development stages (Minami et al. 2014; Vacchi-Suzzi et al. 2013).

7.6. miRNA function

Intensive research over the years has discovered the role of miRNA in key cellular processes, i.e. development, differentiation, proliferation, cell death and metabolism. Some examples showing the relation of miRNA to those cellular processes are summarised below. The first discovered miRNA lin-4 and let-7 known to regulate the developmental stages of *C.elegans*

and without the expression of lin-4, *C.elegans* is unable to make the transition from first to second larval phase. The coordination of miR-48 and miR-84 is essential to end the larval monitoring cycle (Abbott et al. 2005). The over expression of miR-181 is the prerequisite for the terminal differentiation of the myoblasts. Expression of miR-309 cluster is responsible for the zygotic onset of *Drosophila* (Bushati & Cohen 2007). The expression of miR-214 plays an important role throughout embryogenesis of Zebrafish (Flynt et al. 2007). miR-15, miR-16, miR-17 cluster and miR-21 regulate cell death or apoptosis (Jovanovic & Hengartner 2006). miR-196, miR-302, miR-27a, miR-224, miR-16, miR-223 are well known regulators of cell proliferation and differentiation (Tsai et al. 2010; Popovic et al. 2009; Lin et al. 2008; Mertens-Talcott et al. 2007; Venugopal et al. 2010; Guo et al. 2009; Fazi et al. 2007; Felli et al. 2009). miR-122 is the first miRNA linked to metabolic control. Expressing primarily in liver miR-122, regulate hepatic cholesterol and lipid metabolism. miR-195 regulate glucose uptake by directly regulating GLUT3. miR-32 regulate a glucose transporter SLC45A3. Hexokinase catalyses the first step of glycolysis and miR-143, miR-138 regulate hexokinase. miR-375 known to regulate the lactate metabolism. miR-375, miR-124a and miR-9 known to regulate insulin metabolism (Hatziapostolou et al. 2013).

Different miRNAs are also showing the promise to be a powerful cancer biomarker (Chen et al. 2008; Bushati & Cohen 2007). These miRNAs are often referred as OncomiRs. A detailed discussion of different types of OncomiRs, their mechanism of action and involvement in different types of cancer are presented later.

7.7. Bioinformatics approaches to study miRNA regulation

The discovery of abundant miRNAs in diverse multi cellular eukaryotes raised many intriguing questions including the function of the miRNA within the cell. The key to finding out the answer lies in the way miRNA recognise its target mRNA and their cellular function. In RISC, processed mature miRNA pair with mRNA and repress the expression of target mRNA post transcription. At the beginning of the last decade, it was proposed that miRNA will specifically target a mRNA if it has sufficient complementarity to the target mRNA. Otherwise the miRNA will repress productive translation if target mRNA does not have sufficient complementarity but does have suitable constellation of miRNA complementary sites (Bartel 2009). The complementarity is mainly achieved by recognizing and binding to the 3' untranslated region of the target mRNA (Yue et al. 2009). A number of other predictive

features for miRNA target predictions have been discovered in recent times, such as dinucleotide composition of flanking sequence, strong base pairing between 3' UTR of mRNAs and miRNA seed region, thermodynamic stability of the binding site, evolutionary conservation of binding sites (particularly the seed region), secondary structure accessibility and host gene expression profile (Zheng et al. 2013). Over the years many miRNA target prediction algorithms have been proposed and in the next section an overview of the current algorithms and their features have been summarized.

7.7.1. miRNA target prediction algorithms

A list of the most popular target prediction algorithms is provided in Table 8. In the Table 8 seed matching is defined as sequence alignment between 1^{st} to 8^{th} nucleotide of the miRNA 5' end and target mRNA sequence, Conservation is defined as seed matching which is conserved across different species, Free energy refers to the minimum free energy which shows how strong the binding of miRNA with its target is (Yue et al. 2009).

Algorithm	Approach	Features	Availability
TargetScan (S)	Rule Based	1. The algorithm uses seed matching, free energy and conservation. 2. Supported Organisms: mammals, worms, files.	http://www.targetscan.org/
miRanda	Rule Based	1. Optimizes sequence complementarity based on position-specific rules and interspecies conservation. 2. The algorithm also uses free energy and conservation. 3. Supported Organisms: human, mouse, rats.	http://www.microrna.org/microrna/home.do
Pita	Rule Based	1. Investigate the role of target site accessibility, as determined by base-pairing interactions within the miRNA. 2. Also uses seed match and	http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html

		free energy. 3. Supported Organisms: Human, mice, flies and worms	
DIANA-microT	Rule Based	1. Combines conserved and non-conserved miRNA recognition elements into a final prediction score. 2. It uses a dynamic programming algorithm to have the highest scoring alignment between nine nucleotide long windows of the 3' UTR with the miRNA driver sequence. It also applies to energy. 3. Supported Organisms: Human	http://diana.cslab.ece.ntua.gr/microT
Pictar	Data Driven: Hidden Markov Model	1. It checks the seed matching, conservation of seed alignment across multiple species. It also considers the miRNA-target binding free energy. 2. Supported Organisms: vertebrates, flies and nematodes.	http://pictar.mdc-berlin.de/
RNA-hybrid	Rule based	1. It determines the most favourable hybridization between two sequences. 2. The algorithm uses seed match and free energy as a feature. 3. Supported Organism: any.	http://bibiserv2.cebitec.uni-bielefeld.de/rnahybrid
MicroInspect	Rule based	1. The software scans and	http://bioinfo.uni-

or		detects miRNA binding sites and sort the possible target sites based on free energy. 2. The algorithm uses four features seed match, free energy, binding structure and self-complementarity, 4. Supported Organism: any.	plovdiv.bg/microinspector/

Table 8: miRNA prediction algorithms (Yue et al. 2009; Zheng et al. 2013).

Current miRNA target prediction algorithm suffers from a high false positive rate (FPR). FPR is higher than 0.3 which reflect that specificity is often lower than 70%. Using conservation and functional similarities have reduced the number of false positives for miRNA target prediction but it still needs improvement. It is noteworthy that different algorithm's target prediction can vary and researchers often crosscheck or intersect predictions from different algorithms to get additional confidence on true positive (Yue et al. 2009).

7.7.2. Computational Methods to detect miRNA-mRNA regulatory relationship

Gene expression technology has emerged as important and promising evidence based resource for exploring miRNA-mRNA regulatory relationships with biological relevance. Figure 28 shows a common framework used by the state of the art computational methods to detect the regulatory relationship between mRNA-miRNA. A set of features is first selected followed by building a statistical model to predict the regulatory relationship between miRNA and its target miRNA (Figure 28). Statistical methods such as correlation, regression and Bayesian parameter learning have achieved significant results in inferring miRNA-mRNA regulatory relationships (Le et al. 2014).

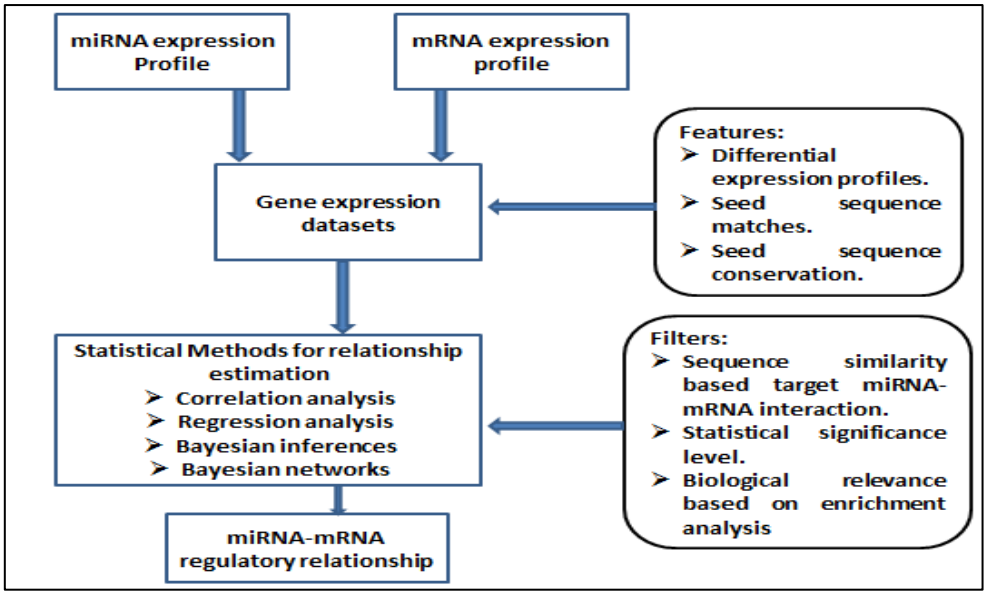


Figure 28: General workflow of existing computational methods to investigate miRNA-mRNA regulatory relationship (Le et al. 2014).

The principal hypothesis of this type of analysis assumes if a mRNA is regulated by miRNA, correlation profile at the expression level should reveal the relationship. Some of these methods also consider availability of miRNA-mRNA sequence based targeting to reduce the false discovery rate.

In 2009, Li *et al.* integrated miRNA expression data and mRNA target information to predict miRNA-mRNA interactions. The team has detected miRNA-mRNA interactions for different physiological conditions using Bayesian network structure learning with splitting-averaging strategy. The method has applied to heterogeneous data including miRNA targeting information, expression profiles of miRNAs and mRNAs, and sample categories. Variational Bayesian-Gaussian Mixture Model has been applied to integrate the score for over expression data and from sequence based prediction methods (Liu et al. 2009).

In 2011, Liang *et al.* launched mirAct web service to explore miRNA activities based on gene expression data. Given the user-uploaded gene expression data, mirAct first transforms values to ranks or Z-scores. Then, mirAct infers the regulatory effect of a miRNA via a two-step procedure. First, a sample score measuring the activity of a miRNA in a sample is obtained by comparing the expression levels of its non-targets with those of targets. In the case of rank transformation, the difference of the average ranks between a miRNA's non-targets and

targets is used. In the case of Z-score representation, the two-sample t-statistic is applied. Then, miRNA activity change across different classes of samples are investigated by examining the sample scores via Kruskal–Wallis test, (null hypothesis is that all classes have identical miRNA activity and supports the analysis of multiple-class data). In addition, Jonckheere–Terpstra trend test is implemented to examine any trend present in the miRNA activity, which might be useful for analyzing data at multiple stages (e.g. disease progression). Multiple comparisons are corrected using the Benjamini and Hochberg FDR method. It is the integration of information within a single sample and across different classes of samples that makes mirAct distinct from other tools. Furthermore, based on computationally determined miRNA sample scores, mirAct enables clustering analysis for samples and miRNAs, which facilitates the visualization and identification of miRNAs of interest (Liang et al. 2011).

In 2013, Chen *et al*. integrated different gene expression data sets from TCGA (The Cancer Genome Atlas) into a model to infer miRNA-mRNA interactions in different forms of cancer. The group developed a novel statistical method to jointly analyze expression profiles from multiple cancers to identify miRNA–gene interactions that are both common across cancers and specific to certain cancers. At first, probabilities of miRNA–mRNA interactions are calculated with the Pearson correlation and the statistical significance level is calculated by Fisher transformation. Next, an empirical Bayes method was jointly applied to infer the posterior probability of interactions across cancers. The method is suitable for analyzing multiple expression data sets (miRNA and target mRNA) of the same physiological condition i.e. cancer in this example (Chen et al. 2013).

$$r_{djk} = \frac{\sum_{i=1}^{N_d}(Y_{dij} - Y_{dj}^{m})(X_{dik} - X_{dk}^{m})}{\sqrt{\sum_{i=1}^{N_d}\left(Y_{dij} - Y_{dj}^{m}\right)^2}\left(\sqrt{\sum_{i=1}^{N_d}\left(X_{dik} - X_{dk}^{m}\right)^2}\right)}$$

$$z_{djk} = \frac{1}{2}\ln\left(\frac{1+r_{djk}}{1-r_{djk}}\right) \sim Normal\ (0,\ \frac{1}{N_d-3})$$

where r_{djk} is the Pearson correlation coefficient between mRNA j and miRNA k in disease d; Y_{dij} is the expression of gene j in individual i of disease d; X_{dik} is the expression of miRNA k in individual i of disease d; N_d is the number of individuals in disease d; z_{djk} is the z-score of each pair gained from the Fisher transformation of the Pearson correlation coefficient.

In the same year, Jacobsen *et al.* also applied regression analysis to understand the regulatory relationship of miRNA and its target mRNA across different cancer using TCGA data. The team developed a computational method and statistical score i.e. the association recurrence

(REC) score by using miRNA and mRNA expression profiles across many cancer types. In individual cancer types, pairwise miRNA-mRNA relationships are evaluated using a multivariate linear model (Figure 29), which also factors in variation (noise) in mRNA expression induced by changes in DNA copy number and promoter methylation at the mRNA gene locus. Associations are rank transformed in individual cancer types, and the method subsequently evaluates the null hypothesis that no association exists between the miRNA-mRNA pair in all cancer types to determine specificity (Jacobsen et al. 2013).

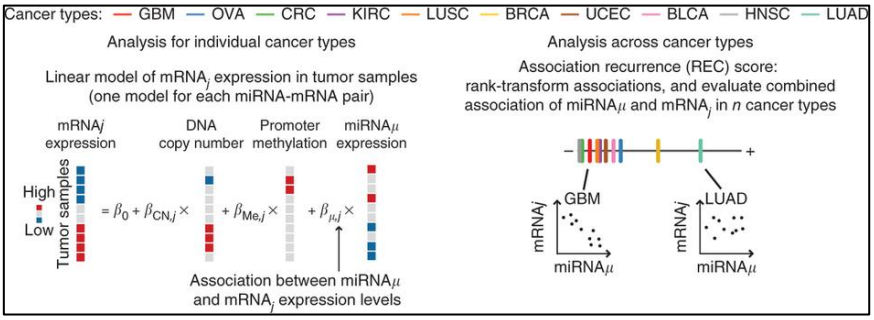


Figure 29: Overview of the statistical approach. Statistical method i.e. the linear model has been applied to evaluate recurrence of miRNA-mRNA expression association across cancer types (Jacobsen et al. 2013).

GBM: Glioblastoma multi forme, OVA: Ovarian serous cystadenocarcinoma, CRC: Colon and rectum adenocarcinoma, KIRC: Kidney renal clear-cell carcinoma, LUSC: Lung squamous-cell carcinoma, BRCA: Breast invasive carcinoma, UCEC: Uterine corpus endometrioid carcinoma, BLCA: Bladder urothelial carcinoma, HNSC: Head and neck squamous-cell carcinoma, LUAD: Lung adenocarcinoma

In 2014, Goldenberg *et al.* developed a probabilistic scoring method called TargetScore. TargetScore predicts miRNA targets as the transformed fold-change weighed by the Bayesian posteriors given target features. The team integrated 84 datasets from Gene Expression Omnibus corresponding to 77 human tissue or cells and 113 distinct transfected miRNAs. TargetScore is a novel probabilistic method for miRNA target prediction problem by integrating miRNA-overexpression data and sequence-based scores from other prediction methods. Briefly, each feature is considered an independent observed variable as input to a variable Bayesian–Gaussian mixture model (VB-GMM). Bayesian method was applied over a maximum likelihood approach to avoid over fitting. Specifically, given expression fold-change (due to miRNA transfection), a three-component VB-GMM was used to infer down regulated targets accounting for genes with little or positive fold-change (Liu et al. 2009).

In 2013, Le *et al*. designed a causality discovery-based method to uncover the causal regulatory relationship between miRNAs and mRNAs, using expression profiles of miRNAs and mRNAs. The algorithm does not incorporate any prior target information. The algorithm

basically learns the regulatory network from the expression data. Finally it simulates the intervention procedure to estimate the causal effect of miRNA on its target mRNA (Le et al. 2013).

7.8. OncomiR

In 2002 miRNA dysregulation in cancer was first discovered in chronic lymphocytic leukemia (CLL). 13q14.3 a frequently deleted region in CLL holds two miRNA cluster loci in it i.e. miR-15, miR-16 (Calin et al. 2002). Recent evidences indicated that miRNA post transcriptionally regulates many oncogenes and tumor suppressor genes. In this process, they can control all the hallmarks of cancers i.e. tumor growth, apoptosis, invasion, angiogenesis and immune evasion (Stahlhut & Slack 2013). Multiple scientific studies also suggest that miRNAs are generally down regulated in cancer and mature miRNA present at a reduced level in tumors. This phenomenon can be due to genetic loss (like in CLL), epigenetic silencing, defects in their biogenesis pathway or wide spread repression in transcription (Jansson & Lund 2012). Corroborating to this, numerous studies find out that normal miRNA biogenesis often disrupts in cancer as reduced DICER expression presenting human tumors (Karube et al. 2005; Melo et al. 2009; Pampalakis et al. 2010; Zhu et al. 2012). The oncogenic transcription factor Myc control the expression of multiple miRNA which are well known for their anti-proliferative, anti-apoptotic, anti-tumor suppressor effects e.g. let-7, miR-15a/16-1, miR-26a and miR-34a (Bui & Mendell 2010). An abridged presentation on miRNA action in cancer is provided in the next section.

7.8.1. Mechanism of action of OncomiRs

The action of OncomiRs is mostly controlled through the following mechanisms:

A. One to One miRNA-mRNA interaction: Some of the most important miRNA in cancer fall into this category i.e. they repress expression of a single target and regulation manifest as certain phenotype (Hayes et al. 2014).
B. One to many miRNA-mRNAs interactions: In some cases the expression of several miRNAs are controlled by a single miRNA resulting a sum effect determining a common phenotype (Hayes et al. 2014).
C. The functional approach: This special type of regulation where miRNA not only regulate expression of transcripts and decide the cellular fate but the targets are functionally related in a way they control expression of each other (Hayes et al. 2014).

D. The pathway approach: In pathway approach one or multiple miRNAs from same cluster i.e. regulated by the same transcription factor target several proteins' transcripts within a specific pathway. In this scenario, the pathway has several regulations at several nodal points (Hayes et al. 2014).

A schematic representation of all the four possible mechanisms is presented in Figure 30.

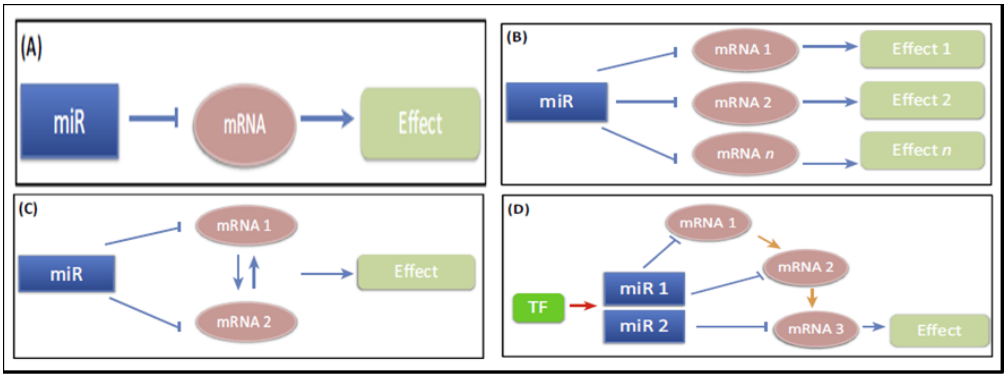


Figure 30: Mechanisms of action of OncomiRs (Hayes et al. 2014).

Approaches to studying microRNA networks. (A) The 'basic' approach: single microRNA–mRNA interaction. (B) The 'broad' approach: one microRNA to many mRNAs. (C) The 'functional' approach. (D) The pathway-based approach.

7.8.2. OncomiRs in Clinical Development

Several ongoing clinical trials aim to decipher miRNA role where specific miRNAs are being monitored for safety, pharmacokinetics, pharmacodynamics profile and for patient stratification purpose. Such clinical trials are presented in Table 9 (Hayes et al. 2014).

MicroRNA	Trial reference	Disease
miR-34a	NCT01829971	Liver cancer and liver metastases
Numerous	NCT01964508	Thyroid cancer
Circulating	NCT01722851	Breast cancer
Numerous	NCT01220427	High-risk prostate cancer
miR-10b	NCT01849952	Glioma
miR-29b	NCT02009852	Oral squamous cell carcinoma
Numerous	NCT02127073	Breast cancer
Circulating	NCT01595139	Low-grade glioma
Numerous	NCT01828918	Colorectal carcinoma
Numerous	NCT01119573	Endometrial cancer
Circulating	NCT01595126	Central nervous system cancer

miR-29 family	NCT01927354	Head and neck squamous cell carcinoma
Numerous	NCT01453465	Rhabdoid tumors
Circulating	NCT01391351	Ovarian cancer
Circulating	NCT01505699	B-cell acute lymphocytic leukemia
Numerous	NCT01957332	Breast cancer
Circulating	NCT01556178	Pediatric brain cancer
Numerous	NCT01050296	Pediatric solid tumors
Numerous	NCT00864266	Non-small-cell lung cancer

Table 9: miRNA in Clinical Trials (Hayes et al. 2014).

One of the major focuses of the thesis is to investigate the prospective role of miRNA for the stratification of colorectal cancer patient who can benefit from anti-EGFR therapy. To familiarize the reader to each of those concepts next chapter have been dedicated to describe colorectal cancer in detail.

8. Colorectal Cancer

8.2. Colorectal cancer: Definition

The colon and rectum together constitutes the terminal part of the human digesting tract, commencing at the ileocecal valve and ending at the anus. Colorectal cancer (CRC) arises in the epithelial cells lining the lumen of the colon; resulting from a multistep process. Initially neoplastic tubular adenomas originate as polypoid structures growing into the colon lumen which gradually acquire disordered villous histology (Markowitz et al. 2002). Cancerous growth is recognized when the invasive cells rupture the underlying epithelial basement membrane of the colon.

8.3. Epidemiology

Colorectal cancer is the third most common form of cancer accounting for 10% of all cancer incidence worldwide and the fourth most common cancer cause of death globally (Ferlay et al., 2012). Almost 55% of the cases occur in more developed regions of Europe, North America, Australia and New Zealand, whereas incidence is lowest in some countries of south and central Asia, parts of Africa and South America (Center et al. 2009). There is wide geographical variation in incidence across the world with incidence rates varying ten-fold in both sexes worldwide. In 2012, estimated age-standardised incidence by region for men and women ranged from 4.5 and 3.8 per 100,000 in Western Africa to 44.8 and 32.2 per 100,000 in Australia/New Zealand respectively (Ferlay et al., 2012; Jemal et al., 2011). However, in the USA and several other high income countries, incidence has stabilised or started to decrease, probably because of increased use of sigmoidoscopy and colonoscopy with polypectomy (Stock et al. 2012).

Mortality in CRC is lower (694,000 deaths, 8.5% of the total) with more deaths (52%) occurring in the less developed regions of the world. There is less variability in mortality rates worldwide (six-fold in men, four-fold in women), with the mortality rates for men and women ranging from 3.5 and 3.0 per 100,000 in Western Africa to 20.3 and 11.7 per 10,000 in Central and Eastern Europe (Center et al., 2009; Ferlay et al., 2012). Mortality has been decreasing since the 1980s in several high-income countries and countries of east Asia and eastern Europe, most probably because of improved early detection and treatment. However, the rise in mortality rates have continued in countries with poor healthcare resources

including countries in Central and South America and rural areas in China (Guo et al. 2012; Bosetti et al. 2011)

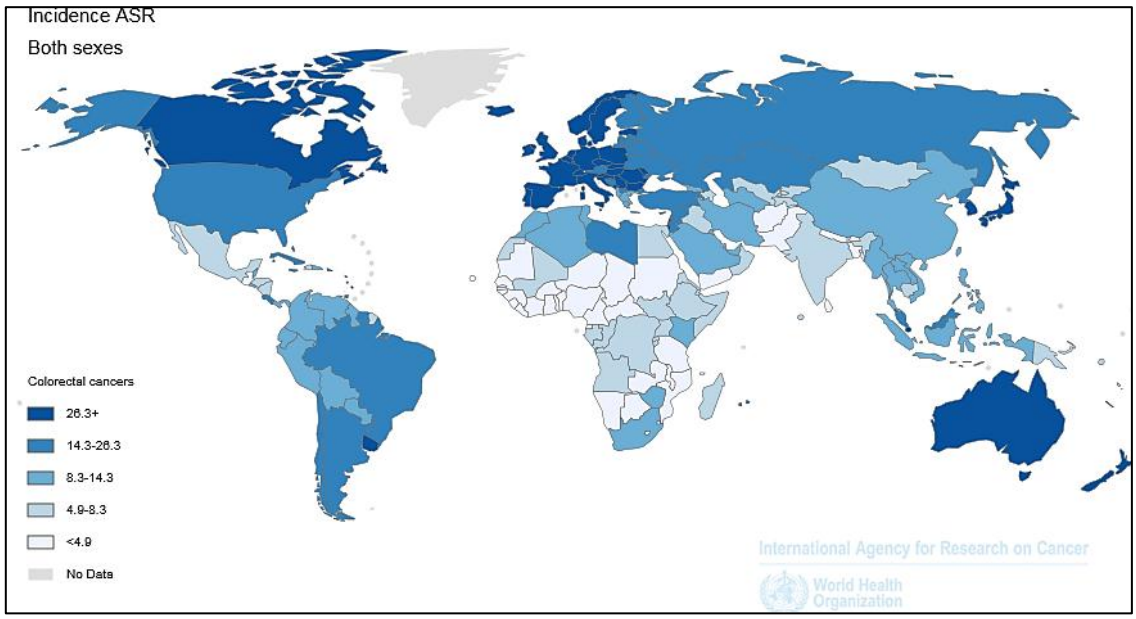


Figure 31: Colorectal cancer epidemiology

The figure shows the estimated age-standardised incidence by region for men and women in the world.

8.4. Risk factors

Several risk factors are associated with the occurrence of CRC (Brenner et al. 2014; Haggar & Boushey 2009). The factors ranging from the demographic to lifestyle and dietary factors are briefly described as following;

8.4.1. Age & Sex

The overall risk of developing CRC is equal among both gender, however, women have a higher risk for colon cancer, while men are more likely to develop rectal cancer (DeCosse et al. 1993). The chances of developing CRC increases after the age of 40, sharply inclining after the age of 50 and more than 50% CRC cases are diagnosed in individuals aged above 50. The incidence rate is more than 50 times higher in persons aged 60 to 79 years than in those younger than 40 years (Ries et al., 2008; World Cancer Research Fund, 2007)

8.4.2. Heredity

The hereditary factors contribute to substantial number of CRC cases. Studies show that 20% patients have a family history of CRC while 5-10% of the cases are a consequence of recognized hereditary conditions (Jackson-Thompson et al., 2006; Skibber et al., 2001)]. The most common inherited conditions are familial adenomatous polyposis (FAP) and hereditary

nonpolyposis colorectal cancer (HNPCC), also called Lynch syndrome (discussed in section 3.5.5.)

8.4.3. Inflammatory Bowel Disease (IBD)

Inflammatory bowel disease (IBD) is a term used to describe two diseases, ulcerative colitis and Crohn disease. The presence of IBD considerably increases the risk of CRC and the relative risk has been estimated to be between 4 - 20 fold (Janout, 2001).

8.4.4. Dietary habits

Dietary intake strongly influences the development of CRC with studies demonstrating that change in food habits might reduce the cancer burden up to 70% (Willett 2005). High red meat and processed meat consumption (Larsson & Wolk 2006; Santarelli et al. 2008) along with high fat intake (Janout, 2001) are implicated in the development of CRC, both characteristics of typical Western diet. In addition, it is considered that consuming diet low in fruits and vegetables may have a higher risk of CRC (National Institutes of Health 2006)

8.4.5. Lifestyle factors

Cigarette smoking has been found to be critical for CRC development with 12% deaths attributed to smoking (Zisman et al. 2006). The carcinogens found in tobacco expedites cancerous growth in colon and rectum with studies revealing that smoking habits could lead to an average earlier age of disease onset in men and women (Tsong et al. 2007). Alcohol consumptions is associated with early onset of CRC (Zisman et al. 2006; Tsong et al. 2007) as well as disproportionate increase in tumor in distal colon (Bazensky et al. 2007).

8.5. Stages of colorectal cancer

Staging in CRC defines the process of finding out how far the cancer has spread. The stage of a cancer is one of the most important factors in determining prognosis and treatment options. The common method for CRC classification is the TNM system adopted by American Joint Committee on Cancer (AJCC) (Edge, 2010; Greene et al., 2002). As described in Figure 32, the TNM system describes 3 key pieces of information local invasion depth (T stage), lymph node involvement (N stage) and presence of distant metastases (M stage). The stage is based on the results of the physical exam, biopsy, along with imaging tests and called clinical stage.

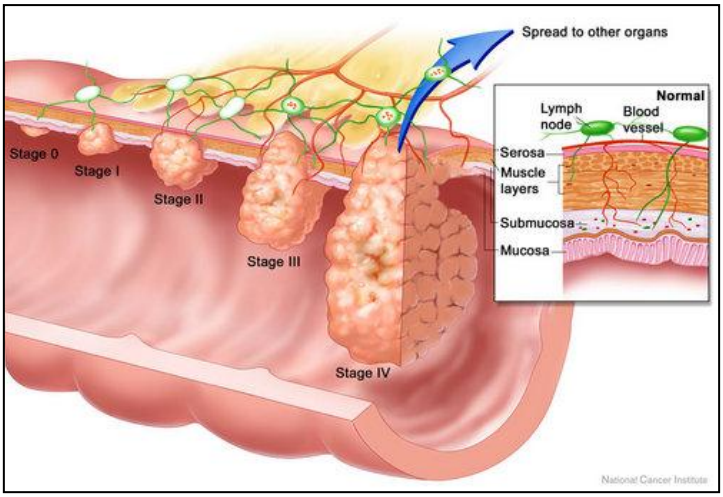


Figure 32: Stages of Colorectal Cancer

The figure demonstrates the anatomy of stage 0 to stage IV colorectal cancer. [see: https://www.bowelcanceraustralia.org/images/About_Bowel_Cancer-Bowel-Cancer-Staging_770.jpg]

Another classification system is adopted from the Union Internationale Contre le Cancer (UICC) which determines the pathologic stage after surgical removal (Sobin LH, Wittekind C 2002). The stages in combination with the clinical stages are described in Table 10.

UICC Stages	TNM	Description
Stage 0	Tis N0 M0	This stage is known as *carcinoma in situ* when the tumor is in the earliest stage has not grown beyond the inner layer of the colon or rectum
Stage I	T1-T2 N0 M0	Tumor grows through the submucosa (T1) or it may also have grown into the muscularis propria (T2).
Stage IIA	T3 N0 M0	Tumor invades the subserosa (the layer between the muscularis propria and the serosa) or the surrounding tissue of the colon or rectum.
Stage IIB	T4a N0 M0	Tumor perforates the wall of the colon or rectum through the visceral peritoneum.
Stage IIC	T4b N0 M0	The tumor invades into other nearby tissues or organs through the wall of the colon or rectum by way of the serosa.
Stage IIIA	T1-T2 N1 M0	Tumor grows through the mucosa into the submucosa (T1) and also into the muscularis propria (T2) It has spread to as many as 3 regional lymph nodes.
	T1, N2a, M0	The tumor has grown through the mucosa into the submucosa (T1). It has spread to 4 to 6 nearby lymph nodes (N2a).
Stage IIIB	T3-T4a N1 M0	Tumor has grown into the outermost layers of the colon or

		rectum (T3) or through the visceral peritoneum (T4a) and has spread to as many as 3 regional lymph nodes.
	T2-T3, N2a, M0	The tumor has grown into the muscularis propria (T2) or into the outermost layers of the colon or rectum (T3). It has spread to 4 to 6 nearby lymph nodes (N2a).
	T1-T2, N2b, M0	The tumor remains within submucosa (T1) or it may also have grown into the muscularis propria (T2). It has spread to 7 or more nearby lymph nodes (N2b).
Stage IIIC	T4a, N2a, M0	The tumor breaches through the wall of the colon or rectum (including the visceral peritoneum) but has not reached nearby organs (T4a). It has spread to 4 to 6 nearby lymph nodes (N2a).
	T3-T4a, N2b, M0	The cancer has grown into the outermost layers of the colon or rectum (T3) or through the visceral peritoneum (T4a) but has not reached nearby organs. It has spread to 7 or more nearby lymph nodes (N2b).
	T4b, N1-N2, M0	The tumor directly invades other organs or structures, including invasion of other segments of the colon or rectum by way of the serosa (T4b). It has spread to at least one nearby lymph node (N1 or N2).
Stage IVA	Any T, Any N, M1a	The tumor may or may not have grown through the wall of the colon or rectum, may or may not have spread to nearby lymph nodes. It has spread to 1 distant organ (such as the liver or lung) or set of lymph nodes (M1a).
Stage IVB	Any T, Any N, M1b	Tumor may or may not have spread through the wall of the colon or rectum and may or may not have spread to nearby lymph nodes. Cancer has spread to more than 1 organ or to the peritoneum that lines the inner wall of the abdominal cavity.

Table 10: Colorectal Cancer stages based on Union Internationale Contre le Cancer (UICC)

8.6. Molecular genetics

CRC development is a gradual process often requiring 10 or more years with the dysplastic adenomas forming the common precursor lesion (Jass 2004). Sporadic CRC has been described as a multistep model of carcinogenesis in which multiple mutations are accumulated which results in uninhibited cell proliferation and tumor development (Fearon & Vogelstein 1990; Kinzler & Vogelstein 1996).

8.6.1. Adenoma carcinoma sequence

The main conclusions drawn regarding sporadic CRC are as follows;

- CRC is a result of mutational activation of oncogenes and the inactivation of tumor suppressor genes
- The accumulation of multiple genetic mutations, rather than the order in which they occur, determines the biologic behavior of the tumor
- Somatic mutations in at least 4-5 genes are required for malignant transformation. Recent genome-wide sequence studies have enumerated about 80 mutated genes per colorectal cancer, however less than 15 mutations were considered to drive the tumor development (Wood et al. 2007)

The distinct pathways which have been associated with CRC are as follows (Bogaert & Prenen 2014b);

8.6.2. Chromosomal instability (CIN):

This is the most common pathway (70%), incorporating numerical (aneuploidy) or structural chromosomal abnormalities resulting cell-to cell variability (Lengauer et al. 1997), characterized by frequent loss-of-heterozygosity (LOH) at tumor suppressor gene loci and chromosomal gene arrangements. CIN is an efficient mechanism for causing the physical loss of wild-type copy in specific oncogenes and tumor suppressor genes.

8.6.3. Mutational inactivation of tumor-suppressor genes

APC: Mutations in adenomatous polyposis coli (APC) gene are important in early cell transformation, with approximately 70–80% of sporadic colorectal adenomas and carcinomas have somatic mutations that inactivate APC (Kinzler & Vogelstein 1998; Segditsas & Tomlinson 2006). In fact, APC mutations are suggested to be the rate-limiting event in most adenoma development. The reason can be found in presence of somatic APC mutations even in earliest lesions within microscopic adenomas (Kinzler & Vogelstein 1996) as well the APC mutation frequency being same for minute adenomas and advanced carcinomas (Aoki & Taketo 2007; Polakis 2007). APC gene function is to modulate the levels of β-catenin protein by proteolysis as a component of the β-catenin degradation complex. Mutations in APC gene product leads to loss of function of oncoprotein β-catenin, which binds to its nuclear partner activating transcription factor which regulates cellular activation (Segditsas & Tomlinson 2006; Markowitz & Bertagnolli 2009). This entire cascade of events causes the constitutive activation of Wnt signalling, regarded as the initial event in CRC. Somatic mutations and

deletions that inactivate both copies of APC are present in most sporadic colorectal adenoma and cancers. In a small subgroup of tumors with wild-type APC, mutations of β-catenin that render the protein resistant to the β-catenin degradation complex activate Wnt signalling (Vogelstein B 2002; Korinek et al. 1997).

p53: Mutation of TP53 leads to inactivation of the p53, a key genetic step in CRC. Inactivated p53 is exhibited in sporadic CRC patients, in up to 75% cases. In most tumors, the two TP53 alleles are inactivated by a missense mutation that inactivates the transcriptional activity of p53 and a 17p chromosomal deletion that eliminates the second TP53 allele (Baker et al., 1990; DeVita et al., 2008). Wild-type p53 mediates G1 cell cycle arrest to facilitate DNA repair during replication or to induce apoptosis (Vazquez et al. 2008). Studies show that the inactivation of TP53 often coincides with the transition of large adenomas into invasive carcinomas (Baker, Preisinger, et al. 1990). Identification of p53 mutations in colorectal cancer has prognostic significance. Persons with tumor that have a p53 mutation have worse outcome and shorter survival than persons whose tumor do not have a p53 mutation (Kressner et al. 1999).

TGF-β: The mutational inactivation of TGF-β signalling pathway is a critical step in CRC. The frameshift mutation within the *TGFBR2* coding sequence leads to mismatch-repair defects within about one third of CRC cases (Grady et al. 1999; Derynck et al. 2001). TGF-β signalling is abrogated in about half of CRC cases with wild type mismatch repair by missense mutations in *TGFBR2* kinase domain, or by somatic mutations in the downstream components SMAD4, SMAD2 or SMAD3 (Sjöblom et al. 2006; Wood et al. 2007). Somatic mutations inactivating TGF-β signalling pathway mostly occur with the transition from adenoma to high-grade dysplasia or carcinoma (Grady et al. 1998).

8.6.4. Activation of oncogenic pathways

RAS & BRAF: Somatic mutations of RAS and BRAF activates the mitogen-activated protein kinase (MAPK) signalling pathway which develops in 37% and 13% of CRC cases respectively (Nosho et al. 2008; Davies et al. 2002). Mutations in RAS is considered to activate the GTPase activity that signals directly to RAF while mutations in BRAF signal BRAF serine–threonine kinase activity which in turn augments the MAPK signalling cascade (Tannapfel et al. 2003). BRAF mutations are detected even in small polyps are more common in hyperplastic polyps, serrated adenomas, and proximal colon cancers (Nosho et al. 2008).

8.7. Microsatellite instability pathway

The microsatellite instability (MSI) pathway causes the inactivation of the DNA mismatch repair (MMR) genes characterised by the accumulation of many insertion or deletion mutations at microsatellites spread along the genome (Boland & Goel 2010). Microsatellites are nucleotide repeat sequences of 1-6 base pairs in length and insertion/deletions within the microsatellites located in DNA coding region leads to frameshift mutations. The silencing of the MMR genes inactivates the DNA repair activity and leads to accumulation of mutations (Bogaert & Prenen 2014b). MSI accounts for 15% of the sporadic CRC caused by hypermethylation of the gene promotor for the MMR enzyme (usually MLH1) leading to gene silencing (Weisenberger et al. 2006). MSI is frequently reported in HNPCC patients, which is caused by a germline mutation of one of the MMR genes (Boland & Goel 2010). The heightened forms of MSI cancers are characterised by the localisation at proximal colon, mucinous cell type, presence of tumor infiltrating lymphocytes and synchronous occurrence with additional tumors (Bogaert & Prenen 2014b; Jung et al. 2012).

8.7.1. Inherited forms

Hereditary forms contribute to 3-10% of all colorectal cancer cases. Inherited colon cancer is consequence of a germline mutation with the phenotypic features depending on the specific gene that is mutated (Calvert & Frucht 2002).

8.7.1.1. Familial Adenomatous Polyposis (FAP)

FAP is an autosomal dominant syndrome inherited by a germline mutation in the APC gene. FAP affects approximately 1 in 12,000 individuals and accounts for ~0.5% of all CRCs (Fearon 2011). More than 1000 different mutations of the APC gene are described as a cause of FAP, resulting in a truncated APC protein (Zeichner et al. 2012). FAP is characterized by hundreds to thousands of adenomatous colorectal polyps that are developed by the third or fourth decade of life (Lynch & de la Chapelle 2003; Rustgi 2007). If left untreated at an early stage, there is a 100% risk of developing CRC by the age of 40, mean age being 36. Most patients have a family history of the disease, however approximately 25% emerge as ‘de novo’ gene mutations in the APC gene (Galiatsatos & Foulkes 2006).

8.7.1.2. Hereditary non-polyposis colorectal cancer (HNPCC)

HNPCC or Lynch syndrome is the most common inherited form of CRC, caused by an autosomal germline mutation in one of several mismatch repair (MMR) genes (Jasperson et al. 2010). About 2-5% of all CRC cases are attributed to HNPCC. Germ-line MSH2 and MLH1 mutations account for approximately 70% of known mutations in HNPCC patients along with PMS1, PMS2, MSH3, and MSH6 contributing to other reported cases (Abdel-Rahman & Peltomäki 2008). The combination of a germline mutation in an MMR gene with inactivation of the remaining normal allele, results in loss of MMR function and accumulation of mutations in microsatellites (Bogaert & Prenen 2014b). Individuals with HNPCC have fewer polyps than FAP, however the polyps are highly likely to progress to cancer. In fact, the rapid progression of the polyps in HNPCC is termed as accelerated tumorigenesis requiring only 3–5 years instead of the 20–40 years estimated for most sporadic CRCs (Dove-Edwin et al. 2006).

Indeed, a rich history of investigation has discovered several genes and pathways crucial for colorectal cancer initiation and progression; these include WNT, RAS-MAPK, PI3K, TGF-β, P53 and DNA mismatch repair pathways. Despite the background, we still did not have a fully integrated view of the genetic changes and system level understanding on CRC and their significance for colorectal tumorigenesis. A more integrated and comprehensive understanding of colorectal cancer achieved with worldwide collaborations on cancer research i.e. The Cancer Genome Atlas Network (TCGA). As any other form of cancer CRC is a genetic disorder and manifest itself through a system of interlinked pathways. The knowledge on mutation rate and altered pathways conceded from TCGA research on CRC has been presented in the next section.

8.8. Colorectal cancer through integrated OMICS data and computational models

High throughput technological advancement has enabled scientists to comprehensively characterize colorectal cancer tumor in genomic, epigenomic, transcriptomic, proteomic and metabolOMICS level. In 2012, TCGA published multi-dimensional analysis of human colorectal cancer. A genome scale analysis of 276 samples analysing exome sequence, DNA

copy number, promoter methylation, mRNA and miRNA expression was conducted within TCGA consortium. The consortium has published that 16% of colorectal carcinomas were found to be hypermutated: three-quarters of these had the expected high microsatellite instability, usually with hypermethylation and MLH1 silencing and one-quarter had somatic mismatch-repair gene and polymerase ε (POLE) mutations. Significant mutations were observed in 24 genes i.e. ARID1A, SOX9 and FAM123B in addition to the expected APC, TP53, SMAD4, PIK3CA and KRAS mutations. Recurrent copy number alterations have been found in ERBB2, IGF2. Frequent chromosomal translocation includes fusion of NAV2 and TCF7L1 has been reported.

8.8.1. Mutations

TCGA consortium published 32 genes with recurrent somatic mutations in hypermutated and non-hypermutated colorectal cancers. Among non-hypermutated tumors eight most frequently mutated genes were APC, TP53, KRAS, PIK3CA, FBXW7, SMAD4, TCF7L2 and NRAS as indicated in Figure 33. The mutations in KRAS and NRAS were mainly located in codon 12/13 and codon 61. Other frequently mutated genes are CTNNB1, SMAD2, FAM123B and SOX9. Disproportionate mutations also have been found out in tumor suppressor genes i.e. in ATM and ARID1A. Mutations in ACVR2A, APC, TGFBR2, MSH3, MSH6, SLC9A9 and TCF7L2 have been observed in hyper mutated tumors along with BRAF (V600E). However, two genes that were frequently mutated in the non-hypermutated cancers were significantly less frequently mutated in hypermutated tumors: TP53 (60 versus 20%, $P < 0.0001$) and APC (81% versus 51%, $P = 0.0023$; both Fisher's exact test) (Muzny et al. 2012).

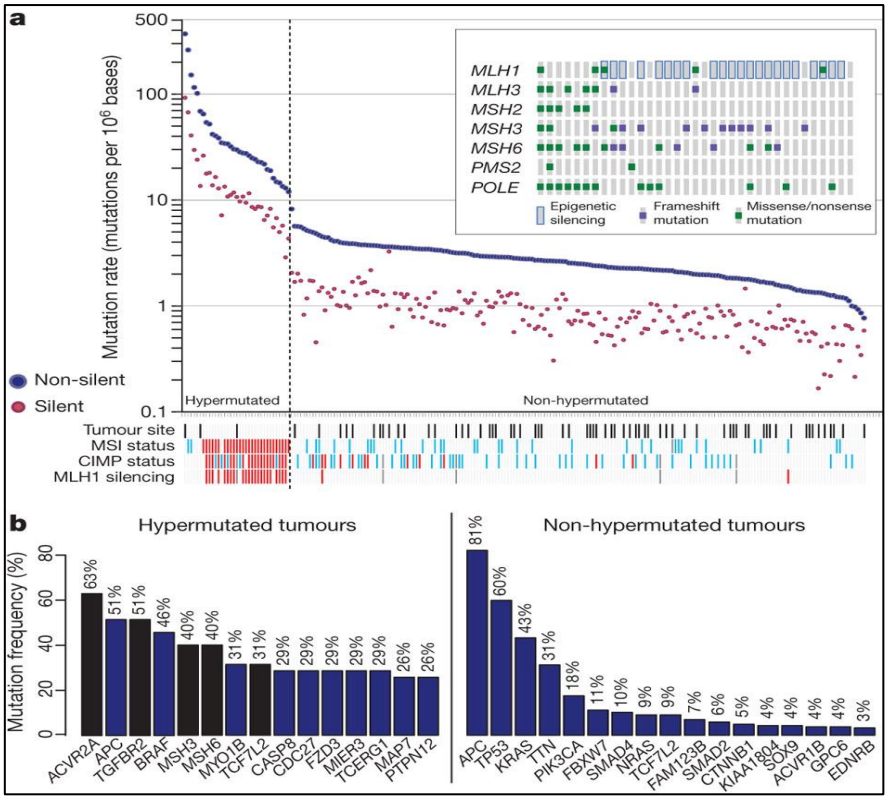


Figure 33: Mutation Frequencies in Human Colorectal cancer (Muzny et al. 2012).

a. Frequency of mutations in 224 studied colon and rectum cancer samples. Notably there is a clear separation between hypermutated and non-hypermutated samples. Red: MSI (Microsatelite Instability) high, CIMP (CpG island methylator phenotype) high. Light Blue: MSI low, CIMP low. Black: rectum, White: Colon. Grey: no data. b. Significantly mutated genes in hypermutated and non-hypermutated tumors. Blue bars represent genes identified by the MutSig algorithm and black bars represent genes identified by manual examination of sequence data.

8.8.2. Altered Pathways in CRC

Recurrent alteration in WNT, MAPK, PI3K, TGF-β and p53 pathways has been reported based on mutation, copy number and mRNA expression changes in 195 analysed tumors (Figure 34). WNT signaling pathways were altered in 93% of all tumors. Biallelic inactivations of APC or activating mutations of CTNNB1 were detected in 80% of cases. There were also mutations in SOX9 and mutations and deletions in TCF7L2 as well as the DKK family members and AXIN2, FBXW7, ARID1Aand FAM123B. A few mutations in FAM123B have previously been described in CRC and were also detected (Muzny et al. 2012). Genetic alterations in the PI3K and RAS–MAPK pathways are common in CRC. In addition to IGF2 and IRS2 overexpression, it was found that mutually exclusive mutations in PIK3R1 and PIK3CA as well as deletions in PTEN in 2%, 15% and 4% of non-hypermutated tumors, respectively. We found that 55% of non-hypermutated tumors have alterations in KRAS, NRAS or BRAF, with a significant pattern of mutual exclusivity. It has been observed co-occurrence of alterations involving the RAS and PI3K pathways in one-third of

tumors (P=0.039, Fisher's exact test). The TGF-β signalling pathway is known to be deregulated in CRC. It has been reported that genomic alterations in TGFBR1, TGFBR2, ACVR2A, ACVR1B, SMAD2, SMAD3 and SMAD4 in 27% of the non-hypermutated and 87% of the hypermutated tumors. It was also evaluated the alterations in TP53 in 59% of non-hypermutated cases and alterations in ATM, a kinase that phosphorylates and activates P53 after DNA damage, in 7% cases. Alterations in these two genes showed a trend towards mutual exclusivity (P = 0.016) (Muzny et al. 2012).

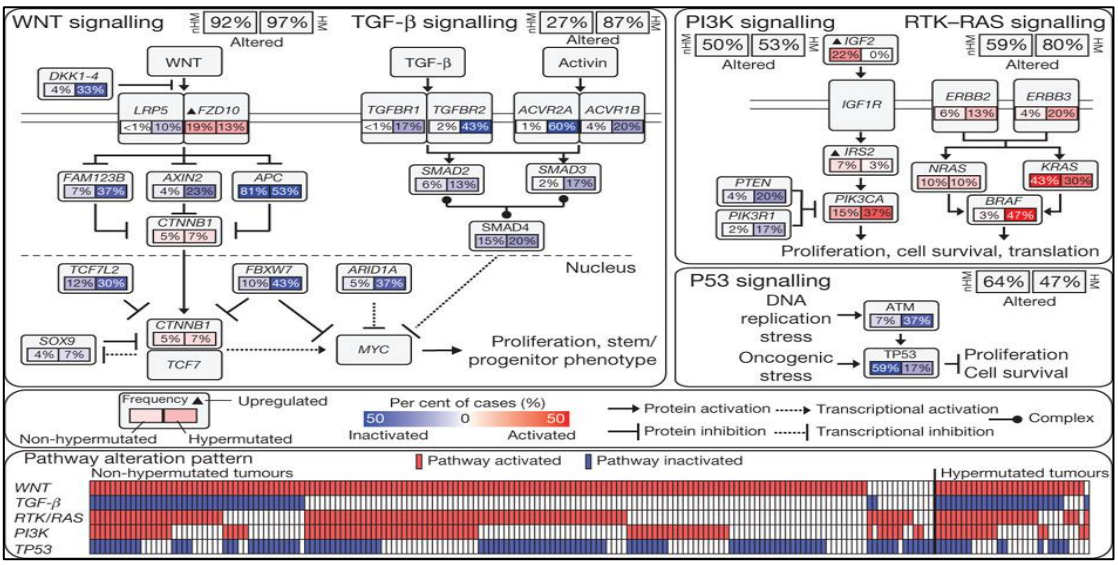


Figure 34: Diversity and frequency of genetic changes leading to dysregulation of signalling pathways in colorectal cancer (Muzny et al. 2012).

Non-hypermutated (nHM; n=165) and hypermutated (HM; n=30) samples with complete data were analysed separately. Alterations are defined by somatic mutations, homozygous deletions, high-level focal amplifications, and, in some cases, by significant up- or downregulation of gene expression (IGF2, FZD10, SMAD4). Alteration frequencies are expressed as a percentage of all cases. Red denotes activated genes and blue denotes inactivated genes. Bottom panel shows for each sample if at least one gene in each of the five pathways described in this figure is altered.

8.8.3. Computational model in colorectal cancer research:

Spurred by rapid advances in computational technology, modelling techniques, ever increasing collection of OMICS data and availability of well-defined pathways in CRC; mathematical modelling began to play a role in colorectal cancer research. In 2006, the dynamics of normal intestinal crypts has been already described with mathematical modelling by Van Leeuwen et al (Van Leeuwen et al. 2006)]. In 2011, Community Research and Development Information Service (CORDIS) under European commission has launched

SYSCOL (Systems biology of colorectal cancer) project. SYSCOL aims to systematically map out the changes and variations in the genetic code that increase individuals' risks of developing colorectal cancer, using the tools of systems biology. The knowledge on the variants will be used to identify the mechanisms of colorectal cancer growth and subsequently to develop a colorectal cancer model. The model will describe cellular pathways that contribute to tumor formation and explain in detail how the genetic disposition of an individual can activate the expression of genes that cause uncontrolled cell growth and lead to colorectal cancer. The dynamics of colorectal cancer is expected to be clear as the SYCOL project will conclude [See: http://cordis.europa.eu/result/rcn/53128_en.html]. In 2013 Nyga *et al*. reported a novel three-dimensional in-vitro biomimetic colorectal cancer model using HT29 cells and connective tissues around the cells (Nyga et al. 2013)]. In the same year, Roznovăţ *et al.* built a network based model for genetic and epigenetic events observed at different stages of colon cancer with a focus on the gene relationships and tumor pathways (Roznovăţ & Ruskin 2013)]. In March 2015, Sottoriva *et al.* reported a validated 'Big Bang' model to provide a quantitative framework to interpret tumor growth dynamics and the origins of intratumoral heterogeneity with important clinical implications (Sottoriva et al. 2015).

One of the major focuses of the thesis is to investigate the prospective role of miRNA for the stratification of colorectal cancer patient who can benefit from anti-EGFR therapy. In the previous two chapters the miRNA and colorectal cancer have been described in detail. The next chapter dedicated to describe the different treatments in colorectal cancer with major focus on cetuximab treatment thus explaining one of the key concepts of the thesis.

9. Treatment of Colorectal Cancer

The choice of treatment in colorectal cancer depends on the stage of the cancer. The treatment options are as follows: surgery, radiation therapy, chemotherapy or targeted therapy.

9.2. Stage specific treatment of colorectal cancer

9.2.1. Stage 0

In stage 0 the cancers have developed beyond the inner lining of the colon, the removal of the tumor by surgery is the most practiced treatment. The surgery can be done either by removal of the polyps also called polypectomy. In some cases colectocomy i.e. colon resection may also be required in case of bigger tumor (Society n.d.; NIH n.d.; ASCO n.d.).

9.2.2. Stage 1

In case of cancers that are a part of the polyp formation, the complete removal of the polyp without any cancer cells at the margin is the first line of treatment. If there are cancer cells at the edges of the polyps additional surgery is often recommended. Additional surgery may be needed in case the polyps cannot be removed completely or has been removed in many differential pieces. If the cancer is not in polyps then colectomy i.e. removal of the colon that carry the cancer and nearby lymph node, are generally performed (Society n.d.; NIH n.d.; ASCO n.d.).

9.2.3. Stage II

In stage II cancers may have spread into nearby tissue but not spread to the lymph nodes, the first line of treatment is the removal of the part of the colon carrying the tumor along with nearby lymph nodes. But additional chemotherapy with 5-FU and leucovorin or capecitabine are often recommended for the following condition of the tumor (Society n.d.; NIH n.d.):

- **i.** Microscopy reveals that cancer is at high grade.
- **ii.** Cancer has been grown into nearby organs.
- **iii.** Less than 12 lymph nodes has been operated and removed.
- **iv.** Cancer cells have been found at the edge of removed specimen hence some cancer cells may have been left out.

v. The colon has been obstructed by the growth of cancer.

vi. Cancer perforated the colon.

9.2.4. Stage III

In this stage cancer has been spread to the nearby lymph nodes but yet to spread other parts of the body. Partial colectomy (removal of the part of colon having the cancer along with nearby lymph nodes) along with adjuvant chemo i.e. either FOLFOX (5-FU, leucovorin and oxaliplatin) or CapeOx (Capecitabine and Oxaliplatin) are the most preferred treatment. If some cancer cells might have been left out additionally radiotherapy is performed (Society n.d.; NIH n.d.).

9.2.5. Stage IV

The choice of surgery in Stage IV CRC depends on extent of metastasis. If there are small areas of lung or liver metastasis which can be completely removed along with cancerous part of the colon then surgical removal of those parts along with nearby lymph nodes is performed. In some cases hepatic artery infusion may be used if the cancer has spread to the liver. On the contrary if metastases in liver/lung are too larger to be removed then chemotherapy is applied before and after the surgery. In some cases cancer can be too widespread then chemotherapy is the main treatment option. But in most cases patient with stage IV CRC receives chemo along with targeted therapies to control the growth of cancer tumor. The most commonly used medicines are as follows (Society n.d.; NIH n.d.):

i. Folfox: leucovorin, 5-FU, oxaliplatin

ii. Folfiri: leucovorin, 5-FU, irinotecan

iii. CapeOX: capecitabine and oxaliplatin

iv. Any other combination from the above including bevacizumab or cetuximab

v. 5-FU and leucovorin with or without bevacizumab

vi. Capecitabine with or without bevacizumab

vii. FOLFOXIRI: leucovorin, 5-FU, oxaliplatin, and irinotecan

viii. Irinotecan, with or without cetuximab

ix. Cetuximab alone

x. Panitumumab alone

xi. Regorafenib alone

Based on the focus of this thesis cetuximab treatment has been discussed in next section.

9.3. Cetuximab

9.3.1. Background of inventing cetuximab

In 1975 the team of Graham Carpenter identified a 170 KDa trans-membrane receptor which upon its ligand (EGF) binding increases ^{32}p incorporation in A431 epidermoid carcinoma cells. To describe the receptor they coined the term EGFR i.e. epidermal growth factor receptor. During 1984 Eckhart *et al* demonstrated that the phosphorylation of the tyrosine residue of EGFR might be a crucial step for tumorigenesis. Two decade long researches has identified EGFR as a receptor tyrosine kinase and further elucidate the process for EGFR activation, regulation of downstream pathways by activated EGFR as critical component for cell proliferation and survival (Toni M. Brand et al. 2011). In the year 1983 Sato *et al* identified four immunoglobulin G (IgG) and demonstrated that three of those antibodies namely M225 IgG, M528 IgG and M579 IgG blocked 95% of the EGF binding to EGFR in human A431 cells. The same group also shown that each antibody effectively blocked EGF induced phosphorylation of EGFR resulting reduced proliferative potential in the cell line model. M225 reported to have more affective anti-EFGR property (Toni M. Brand et al. 2011; Gill et al. 1984; Kawamoto et al. 1983; Sato et al. 1983; Masui et al. 1984). These series of discoveries draws the attention of pharmaceutical industry sensing a potential targeted therapy in cancer.

The phase trial of M225 was successful but all the patients produced human-anti-mouse antibodies. To nullify this immunologic reaction M225 was converted into human-murine chimera. The part of the variable region of mouse M225 which bind to EGFR was integrated IgG1 Fc isotype from human for its potential to enhance the immune contribution of C225. The structure of C225 is provided in detail in Figure 35 (Toni M. Brand et al. 2011).

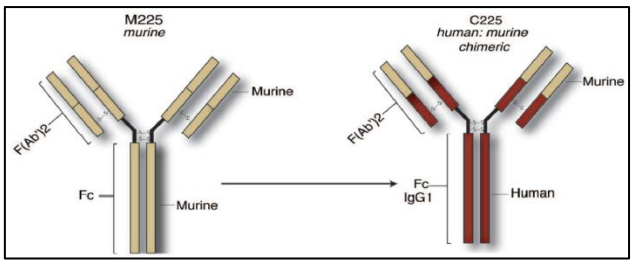


Figure 35: **The Structure of C225** (Toni M. Brand et al. 2011).

The design of cetuximab. Three murine antibodies designated M225 igG, M528 igG and M579 igG with activity against the eGFR were developed. Further testing identified M225 as being the most efficacious for anti-eGFR activity and was moved into Phase i clinical trials. Although successful, patients developed human-anti-mouse antibodies (HAMA) and therefore M225 was converted to a human:murine chimera, C225, with an igG1 FC isotype.

Subsequently, the anti-EGFR antibody C225 IgG1 entered the clinical trial. The mode of action of this biologics is controlled by binding to the extracellular domain III of EGFR i.e. by blocking EGF binding to EGFR. Later this chimeric antibody was marketed as Cetuximab (ICM-225, Erbitux^{TM}). To present a detail idea of the consequence of blocking EGF from binding on EGFR hence blocking all the downstream pathways are presented in the next section.

9.3.2. EGFR biology and downstream pathway

EGFR is a member of EGF receptor tyrosine kinase family with four other members ErbB1/HER1, Her2/neu (ErbB2), HER3 (ErbB3) and HER4 (ErbB4). Structurally these receptors are similar, each having an extracellular ligand binding domain, single membrane spanning region, a juxtamembrane nuclear localization signal (NLS), a cytoplasmic tyrosine kinase domain and C-terminal tail housing several tyrosine residues for propagating downstream signalling. Upon binding the extracellular ligand the receptor dimerizes, allowing cytoplasmic EGFR-TK to activate in a tail to head fashion as indicated in Figure 36.

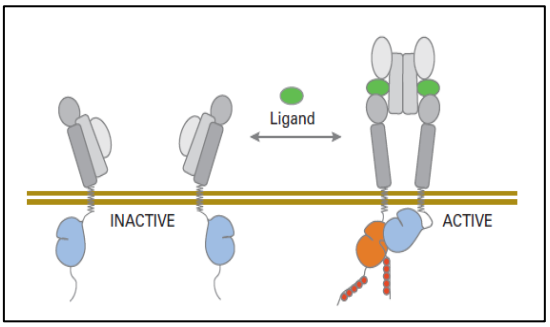


Figure 36: Modelling the effect of ligand binding to the EGFR receptors (Kumar et al. 2008).

A schematic diagram of EGFR family activation based on crystal structures. On extracellular ligand binding the receptor dimerizes allowing the cytoplasmic EGFR-TK to activate in a tail-to-head fashion.

Upon ligand binding EGFR activate three downstream pathways RAS/RAF/MEK/ERK, PI3K/AKT/mTOR and PLCγ/PKC. Other two important downstream pathways are SRC tyrosine kinases and STAT activation.

9.3.2.1. RAS/RAF/MEK/ERK pathway:

This pathway leads to cell proliferation and play a central role in many human cancers. Ligand binding to EGFR triggers its activation and subsequent phosphorylation. Activated phospho-tyrosine residues in its C-terminal, act as a binding site for Grb2, a SH2-containing protein. In the next stage Grb2 recruits the guanine nucleotide exchange factor SOS through its SH3 domain. This promotes binding of GTP to RAS and subsequently activated RAS triggers the Map kinase (MAPK) cascade. Next Ras-GTP complex binds and activates RAF kinase (MAPKKK). Activated RAF binds to and phosphorylate MEK (MAPKK). Activated MAPKK then phosphorylate ERK1/2 (MAPK). Activated ERK kinases can activate several other kinases including MNK1, MNK2, MSK1, MSK2, RSK, MAPK. In turn these kinases phosphorylate several transcription factors including Elk-1, PPARγ, STAT1, STAT3, C-myc and AP-1. These activated transcription factors increased expression of several genes involved in cellular proliferation, most notably cyclin D1 (Sebolt-Leopold & Herrera 2004).

9.3.2.2. PI3K/AKT/mTOR pathway

Activated EGFR also recruit PI3K to the cell membrane. PI3K phosphorylation of phosphatidylinositol-4,5-biphosphate yields the second messenger phosphatidylinositol-3,4,5-triphosphate (PIP3). PIP3 serves as a membrane bound docking site for AKT. Once PIP3 and AKT together stationed on the plasma membrane, AKT is phosphorylated by two kinases i.e. PDK1 and mTORC2. One of the crucial effector of AKT activation is mTORC1. AKT activate mTORC1 via TSC2. Phosphorylation of TSC2 by AKT cancels inhibition of the small G-protein RHEB. This initiates activation of mTOR. Activated mTOR phosphorylate p70-S6 kinase 1 initiate protein synthesis via S6 ribosomal subunit. Over all activation of protein synthesis is a crucial step in cancer cell growth and survival. Activated AKT also influence cell proliferation by activating ample cellular factors (Hanahan & Robert A. Weinberg 2011; Bogaert & Prenen 2014a).

9.3.2.3. PLCγ/PKC pathway:

This pathway plays a crucial role in mediating the effects of activated EGFR. Phospholipase C (PLC) interacts with phosphor-tyrosine residues of EGFR via its SH2 domain. The interaction activates PLC mediated by PIP3. Activated PLC interacts with plasma membrane and it cleaves PIP2 to inositol triphosphate (IP3) and diacylglycerol (DAG). Subsequently IP3 bind its receptors at the endoplasmic reticulum to induce the calcium influx into the cell. DAG activates PKC at the plasma membrane. Activated PKC is a potent serine/threonine kinase second messenger capable of phosphorylating a plethora of substrates leading to cell proliferation, apoptosis, cell survival and cell migration (Hanahan & Robert A. Weinberg 2011; Bogaert & Prenen 2014a).

All these above mentioned pathways their downstream transcription factors and phenotypes i.e. cell survival, proliferation, invasion, metastasis and angiogenesis, are described in Figure 37.

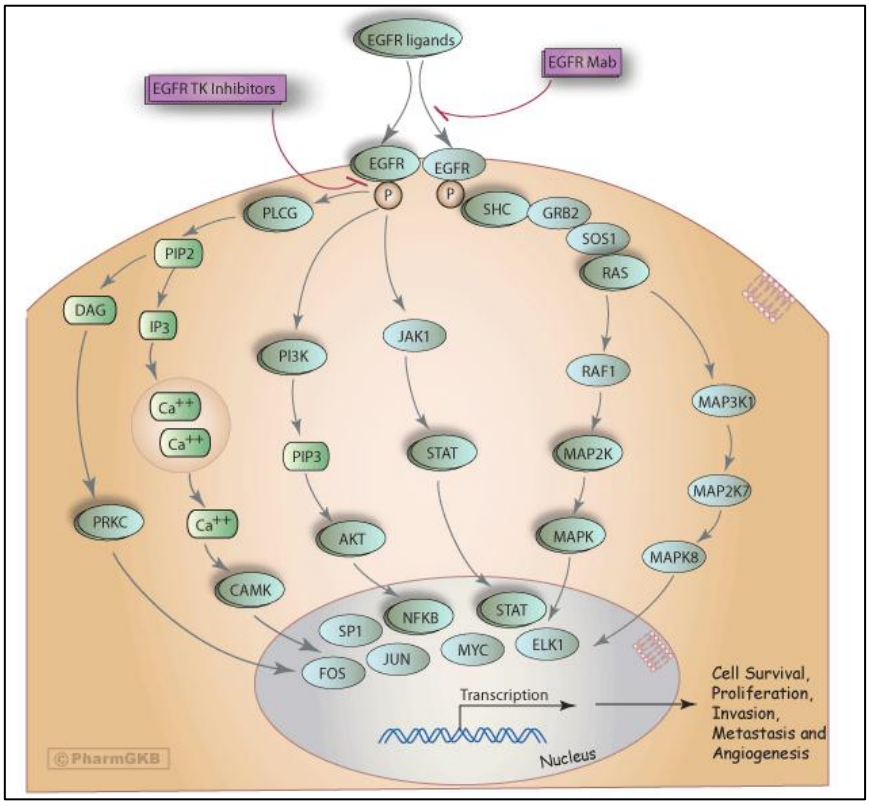


Figure 37: Downstream signaling of activated EGFR (Li Gong n.d.)

The figure shows the downstream signaling pathways which activate upon binding of EGF to EGFR leading to cell proliferation, invasion, metastasis and angiogenesis.

9.3.3. Mode of action and efficacy of cetuximab

Cetuximab effectively blocks the binding of EGF to EGFR i.e. blocks EGFR phosphorylation. It also promotes EGFR internalization and reduces cell proliferation in metastasis colorectal cancer. Detailed mechanisms of actions of cetuximab treatment are summarized in Figure 38.

a. Cetuximab has got more affinity for EGFR than two of its ligand namely TGFα, EGF i.e. cetuximab effectively block ligand induced EGFR phosphorylation. So all the three EGFR downstream pathways discussed in the previous section are blocked (Gill et al. 1984; Kawamoto et al. 1983; Sato et al. 1983; Masui et al. 1984).
b. Other reports suggested that cetuximab sterically hinder the binding of EGFR to other HER family members (Li et al. 2005).
c. Cetuximab also promotes the internalization or degradation of EGFR i.e. nullifying the chance of downstream signaling (Sunada et al. 1986).
d. Cetuximab treatment in different cancer models or human tumor xenografts has shown an increased expression of cell cycle inhibitor $p27^{kip1}$ which drives the formation of $p27^{kip1}$-Cdk2 complexes. This complex arrests the cell cycle at G1 phase (Huang et al. 1999; Wu et al. 1996).
e. Treatment with cetuximab dramatically decrease the pro-angiogenic factors i.e. inhibit angiogenesis. Cetuximab therapy may also lead to decreased invasion or metastasis spread of the tumor cells (Tortora et al. 1999; Liu et al. 2000).
f. Cetuximab treatment increases expression of apoptosis promoting Bax and decreases cell survival promoting Bcl2. By modulating the expression of those two proteins the monoclonal antibody drives the shrinkage of tumor (Wu et al. 1995).
g. Cetuximab also mediate antibody dependent cytotoxicity of the tumor cells (Kimura et al. 2007).

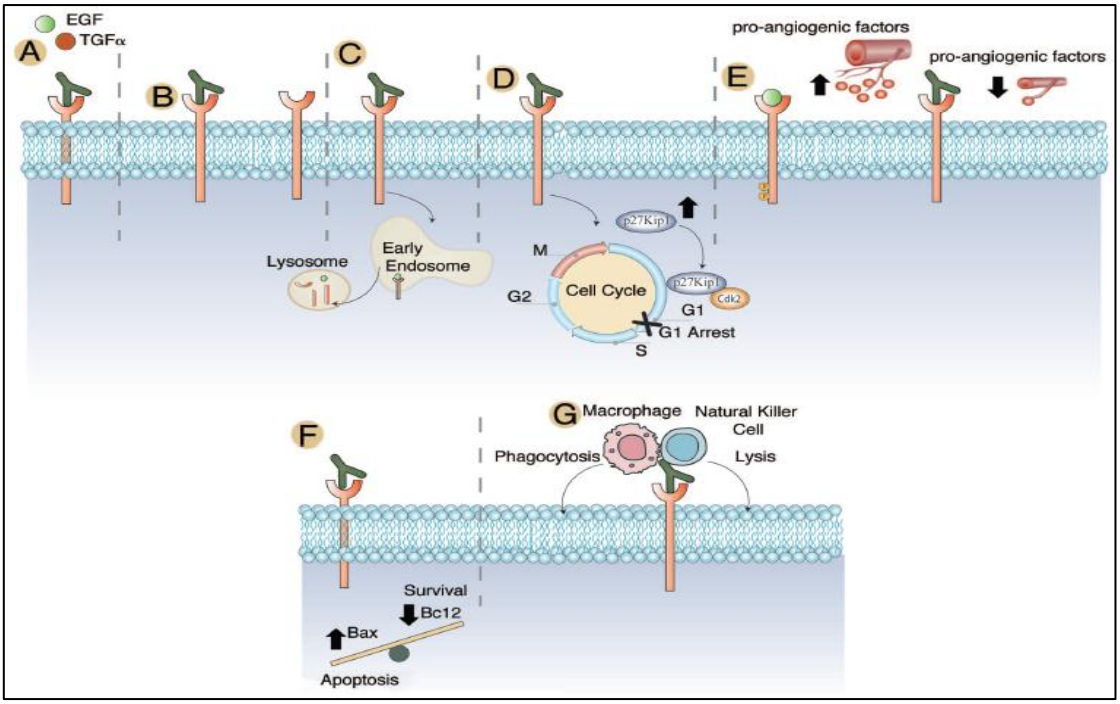


Figure 38: Mode of action of Cetuximab therapy (Toni M. Brand et al. 2011).

The mechanism of action of cetuximab. All the biological processes (A to G) are described above.

However, multiple lines of evidences suggest that only 10-20% patients with mCRC benefit from cetuximab treatment (Chung, Shia, Nancy E. Kemeny, et al. 2005; Cunningham et al. 2004a). Over the years intense research has been undertaken to discover biomarker that can stratify metastasis colorectal cancer patients which are likely to respond prior to the treatment. Some stratified biomarker has been identified and also approved by the regulatory bodies, described in the next section.

9.3.4. Stratified biomarker of cetuximab therapy

9.3.4.1. KRAS mutation

KRAS mutation is one most important indicator whether a CRC patient will be sensitive or resistant to cetuximab therapy. KRAS is a GTPase which connect EGFR to RAF/MEK/ERK pathway. After binding to GTP, KRAS activate RAF i.e. triggers the pathway leading to cellular proliferation. Mutations in codon 12 or 13 impair the intrinsic GTPase activity and show resistance to GAPs. This causes the accumulation of active mutant RAS in the cancer cells. This accumulated active form of RAS still capable to trigger RAF/MEK/ERK pathway in EGFR independent manner (Trahey & McCormick 1987; Toni M. Brand et al. 2011).

Lièvre *et al* has shown that the KRAS mutation is significantly associated with the cetuximab resistance activity in CRC (Lièvre et al. 2006).

9.3.4.2. BRAF mutation

In 2008 Nicolantonio et al shown that efficacy of anti-EGFR monoclonal antibody in CRC hindered by BRAF V600E mutation (Di Nicolantonio et al. 2008). BRAF is kinase belong to RAF family. GTP bound KRAS activate BRAF which triggers RAS/RAF/MEK/ERK pathway (Niault & Baccarini 2010). But mutant BRAF can still activate RAS/RAF/MEK/ERK pathway in EGFR independent manner (Lito et al. 2013).

9.3.4.3. EGFR gene copy number

Increased gene copy number is associated with the response of cetuximab therapy in CRC (Mauro Moroni et al. 2005). Subsequent large clinical trials of CRC patients treated with cetuximab confirmed the relationship as well (Personeni et al. 2008). But the exact mechanism for this phenomena is unknown (Toni M. Brand et al. 2011).

9.3.4.4. Over expression of EGFR ligand

In 2007 Khambata-Ford *et al* has shown the correlation between expression of two EGFR ligands i.e. EREG and AREG to cetuximab treated CRC patients. CRC patients with high expression of Epiregulin (EREG) and Apiregulin (AREG) are more like to have disease control when treated with cetuximab treatment (Khambata-Ford, Christopher R. Garrett, et al. 2007).

Discovery of these biomarkers has enabled clinicians to somewhat predict the response of cetuximab therapy in CRC patients prior to the therapy. However selection of right cohort of patients is still far from complete as cetuximab resistance mechanism remains poorly understood. To address the issue scientific community is looking for addition biomarkers predicting the efficacy of the treatment. In this regard miRNA has recently drawn the attention of scientific community for its capacity to regulate expression of multiple onco genes, tumor suppressor genes and pathways central to cancer formation. Identifying the potential of miRNA as stratified biomarker, a novel SMARTmiR algorithm has been developed to predict the role of miRNA as therapeutic biomarker for cetuximab treatment in colorectal cancer. The SMARTmiR methodology is presented in the next chapter.

10. The Prospect of miRNA as Therapeutic Biomarker in Colorectal Cancer (Deyati et al., 2015)

Colorectal cancer (CRC) is one of the most prevalent cancers, with 1.2 million new cases every year. It is the second most commonly diagnosed cancer in females and the third most common in males; the highest incidence occurs in developed countries. By the age of seventy years, one out of every two citizens in the Western world develop benign adenomas that evolve into malignant carcinomas at an estimated yearly rate of 0.1 to 0.25% (Jass 2004). In United States alone, the annual cost of CRC treatment is forecasted to reach $17.7 billion by 2020. However, using simultaneous strategies that reduce risk factors, increasing screening and treatment could avoid 101,353 deaths resulting in $33.9 billion in savings in reduced productivity loss (Bradley et al. 2011).

EGFR, a transmembrane receptor tyrosine kinase has been identified as one of the most promising targets for treating metastatic colorectal cancer (mCRC). Among the 20 molecules listed by the National Cancer Institute for the treatment of mCRC (Anon n.d.), cetuximab is one of the most successful monoclonal antibodies (Saltz et al. 2004; Chung, Shia, Nancy E Kemeny, et al. 2005). However, multiple lines of evidence suggest that only 10-20% patients with mCRC benefit from cetuximab treatment (Chung, Shia, Nancy E Kemeny, et al. 2005; Cunningham et al. 2004b). The selective efficacy, side effects and high treatment costs of cetuximab result in the need for focused research to decipher the resistance mechanisms to cetuximab. The response of the scientific community to this need is evident through the increased number of publications suggesting that the mutational status of KRAS, BRAF and PIK3CA, differential expression of PTEN, EGFR ligand (AREG, EREG) and EGFR gene copy number variation could serve as therapeutic biomarkers for anti-EGFR monoclonal antibody treatment in CRC (Frattini et al. 2007; Sartore-Bianchi, Martini, et al. 2009; Perrone et al. 2009; Mauro Moroni et al. 2005; Prenen et al. 2009; Sartore-Bianchi, Di Nicolantonio, et al. 2009). However, none of these singular molecular changes could accurately predict the response of CRC patients to cetuximab therapy. Because CRC is a system-level disorder that involves multiple molecular mechanisms to support proliferative signalling, resist cell death, induce angiogenesis and metastasis; molecules such as miRNAs that regulate signalling pathways by affecting the expression of multiple proteins might serve as more potent

therapeutic biomarkers (Hanahan & Robert A Weinberg 2011; Rukov et al. 2013). However, little is known regarding the role of miRNAs as a therapeutic biomarker for cetuximab treatment in CRC.

miRNA expression is typically dysregulated in cancer cells and this dysregulation has a high degree of tissue specificity, miRNAs could be used as diagnostic and therapy-related biomarkers. Additionally, miRNAs have an unusually high stability in formalin-fixed tissues, from which they could be extracted with minimal degradation (Tang et al. 2006). Moreover, the techniques of miRNA analysis from a single cell are established, allowing for the analysis of small amounts of miRNAs with increasing sensitivity for potential biomarker assays (Lim et al. 2005). The roles of miRNAs in CRC, EGFR signalling regulation and cetuximab treatment outcome are also evident. Multiple reports regarding the miRNA dysregulation in metastatic colorectal cancer have been published. Some groups reported that major colorectal cancer biomarkers such as EGFR and RAS are regulated by mir-7 and let-7 respectively, thus profoundly affecting downstream signalling (Webster et al. 2009; Lee et al. 2011; Xu et al. 2012; Liu et al. 2011; Li et al. 2012). Recently, Bissonnette et al. demonstrated the correlation between EGFR signalling and miR-143, mir-145 in murine colon cancer models (Zhu et al. 2011). Another group discovered that cetuximab-mediated EGFR inhibition abrogates the age-related increase of miR21, which is related to age-dependent colorectal cancer (Nautiyal et al. 2012).

These findings cumulatively suggest a potentially significant role of miRNAs in EGFR signalling in CRC. Therefore, these evidences compelled us to create a workflow to identify the most critical miRNAs important for cetuximab resistance in CRC that could be further used for potential biomarker assay development. In this study, we propose ranked miRNA candidates that might contribute to cetuximab resistance in CRC patients. The inference is based on the integration of multiple published and predicted miRNA findings into cellular pathways that lead to oncogenesis and metastasis. We prioritised the biomarker candidates based on a novel algorithm, i.e., SMARTmiR (Scoring-based MARking of Therapeutic MIcroRna), that combines the network parameters and literature-derived evidence. Finally, the significance of our prediction was strongly supported by recently published experimental data that are derived from cetuximab resistant CRC patients. This study provides an actionable insight into the novel mechanism of cetuximab resistance mediated by miRNAs that might lead to identification of miRNAs as biomarkers, thereby predicting optimum responses to the drug.

10.2. SMARTmiR workflow

Herein, we propose a predictive algorithm, i.e., SMARTmiR, that combines knowledge and data-driven approaches to identify miRNAs contributing to the therapeutic effects of cetuximab in CRC patients. The algorithm consists of the following four steps (Figure 39):

Step 1: Construction of pathway maps leading to oncogenesis and metastasis in CRC.

Step 2: Identification of miRNA candidate biomarkers via miRNAome screening.

Step 3: Ranking of miRNAs based on accumulated evidence and the effects on the cellular process of CRC patients treated with cetuximab.

Step 4: Validation of the prediction based on experimental data.

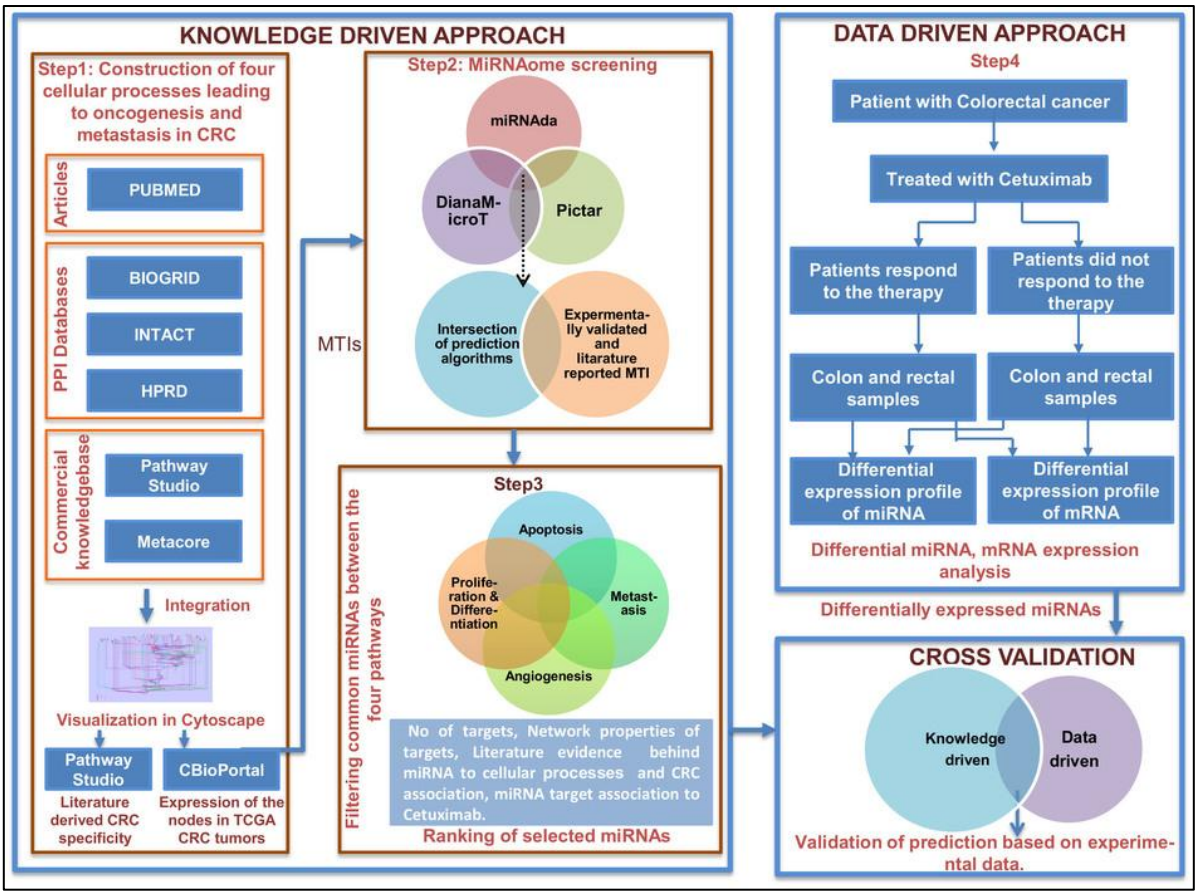


Figure 39: SMARTmiR workflow for the selection of miRNAs as candidate biomarkers conferring cetuximab resistance in colorectal cancer

10.2.1. Construction of molecular pathway maps leading to CRC oncogenesis and metastasis

Following the first step in the workflow as illustrated in Figure 39, we created a comprehensive bio molecular space for the mechanisms of action of cetuximab therapy in colorectal cancer. In doing so we assembled four pathway maps linked to four fundamental cellular processes in oncogenesis and metastasis; namely apoptosis, proliferation and differentiation, angiogenesis, and metastasis. These pathway maps integrate known pathways from Metacore (Thomson Reuters, New York, USA) that lead to the four cellular processes and protein-protein interactions from IntAct, BioGRID and HPRD (Human Protein Reference Database) Databases (Ekins et al. 2007, Shannon, Markiel, Ozier, Nitin S Baliga, et al. 2003). The pathways that were integrated to build the four pathway maps leading to the cellular processes are listed in Table 11. All of the molecules in the maps (nodes) are annotated with the Entrez gene ID, HGNC gene symbols and corresponding UniProt IDs. The distinct features of the pathways are as follows: (A) Node shapes and colours correspond to their functional category (e.g., receptor, ligand, transcription factor, kinase); (B) All of the edges have directionality; (C) The edges are differentiated by shapes and colour corresponding to the type of interaction (such as binding, catalysis, phosphorylation, transcription regulation, transformation etc.); (D) Reactions (edges) are tagged to PubMed IDs as evidence. Four pathway maps corresponding to proliferation and differentiation, apoptosis, angiogenesis, and metastasis are available in standardised SBML (Systems Biology Markup Language) (Hucka et al. 2003) for file exchange between different tools. The four pathways are provided as four Supplementary Pathway maps (i.e., Apoptosis.xml, Proliferation_Diffrentiation.xml, Angiogenesis.xml and Metastasis.xml) (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s16.zip). The Pubmed IDs supporting an edge can be found by selecting respective edge in the pathway map specific xml files after opening those in Cytoscape.

TCGA RNA Seq data in colon and rectal adenocarcinoma (2012) has been analysed to demonstrate the CRC specificity of the constructed maps (Anon 2012).

Cellular processes	Pathways assembled
Proliferation and Differentiation	EGFR signalling pathway, EGFR signalling via small GTPases, EGFR signalling via PIP3, EPO-induced MAPK pathway, ERK5 in cell proliferation, PDGF signalling via MAPK cascades, PDGF signalling via STATs and NF-kB
Apoptosis	Apoptosis and survival Anti-apoptotic TNFs/NF-kB/Bcl-2 pathway, Apoptosis and survival_Anti-apoptotic TNFs/NF-kB/IAP pathway, Apoptosis and survival p53-dependent apoptosis, EPO-induced Jak-STAT pathway, EPO-induced PI3K, AKT pathway and Ca(2+) influx, Signal transduction of AKT signalling, Signal transduction of PTEN pathway
Angiogenesis	VEGF-family signalling and activation, Angiopoietin - Tie2 signalling, VEGF signalling via VEGFR2 - generic cascades, S1P1 signalling pathway Development Role of IL-8 in angiogenesis, Thrombospondin-1 signalling
Metastasis	TGF-beta receptor signalling, TGF-beta-dependent induction of EMT via MAPK, Regulation of epithelial-to-mesenchymal transition (EMT), Immune response of Oncostatin M signalling via JAK-Stat in human cells FGF2-dependent induction of EMT, TGF-beta-induction of EMT via ROS, NOTCH-induced EMT, HGF-dependent inhibition of TGF-beta-induced EMT, TGF-beta-dependent induction of EMT via RhoA, PI3K and ILK, Immune response of Oncostatin M signalling via MAPK in human cells, HGF signalling pathway

Table 11: Detailed listing of the pathways used for assembling the proliferation and differentiation, apoptosis, angiogenesis, and metastasis processes.

10.2.2. Identification of miRNAs candidate biomarkers via miRNAome screening

In the second step, we identified miRNA-target interactions (MTIs) for all of the nodes in the four constructed pathway maps. The exhaustive search for these miRNAs included both experimentally validated miRNAs and predicted miRNAs.

10.2.2.1. Experimentally validated and literature reported miRNAs

Through screening public and commercial sources for experimentally validated MTIs, we identified two major resources that are widely used because of their extensive coverage and the quality of scientific evidence, i.e., TarBase and Pathway Studio (Elsevier, Amsterdam, Netherlands) (Nikitin et al. 2003). TarBase hosts manually curated MTIs that are experimentally validated (Vergoulis et al. 2012). It also incorporates entries from other well-known databases, such as miRecords (Xiao et al. 2009), miRTarBase (Hsu et al. 2011) and miR2Disease (Jiang et al. 2009). All of the literature-derived MTIs that were extracted for our study from the two sources described above are included in Supplementary Table 12 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s11.xls).

10.2.2.2. Predicted miRNAs

Several computer-aided algorithms are available for the identification of MTIs. We used Diana-Micro T 3.0, Pictar and TargetScanS to identify predicted MTIs for the four developed pathways, focusing in particular on their overlapping predictions. These identified miRNAs have not been encoded into the pathway maps and are used to further filter and rank potential candidate miRNAs (described below).

10.2.2.3. Ranking of miRNAs based on accumulated evidence and their effect on the system

A novel raking formula was developed to rank the miRNAs for their potential to serve as therapeutic biomarkers for cetuximab treatment in CRC patients (Equation 1). To enhance the accuracy of the ranking function, the four assembled processes were integrated in one *Integrated map*, representing a comprehensive knowledge space of functional molecular networks that lead to the cellular process those are mostly dysregulated in colorectal cancer. Each miRNA is ranked based on its topological properties, network properties of its targets and based on literature-derived evidence of miRNA regulating signalling pathways that are important for colorectal carcinogenesis, miRNA target's (i.e., nodes in the *Integrated map*) relations to cetuximab. All of the literature-derived evidence for the candidate miRNA's relationships to cellular processes, CRC and the target's relations to cetuximab are provided in Supplementary Table 13 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s12.xls), Supplementary Table 14 (See:

http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s13.xls) and Supplementary Table 15 (http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s14.xls); respectively. The network parameters (Betweenness centrality, Node degree) were calculated using the CentiScaPe plugin (Scardoni et al. 2009).

$$S_j^{mir} = \deg(mir_j) + \frac{\sum_{i=1}^{n} R_i * T_i}{\sum_{i=1}^{n} T_i} + E_j^{Pathway} + E_j^{CRC} + \sum_{i=1}^{n} E_i^{Cetuximab}$$

S_j^{mir} is the score of j[th] miRNA. Each feature is given equal weightage due its natural importance and is normalised between 0 and 1 such that the calculated score could have a maximum value of 5. The overall scoring function was implemented in the Perl programming language.

The first feature of miRNA ranking is the node degree of the j[th] miRNA, i.e., the no. of targets of the j[th] miRNA in the *Integrated map*, as defined as the deg(mir_j). The node degree corresponds to the no. of nodes adjacent to a given node v. The degree allows for an immediate evaluation of the regulatory relevance of the node. For example, in signalling networks, proteins with a very high degree are interacting with several other signalling proteins and are likely to be regulatory hubs. The second parameter was calculated based on the weighted node degree of all of the targets of j[th] miRNA, depicted as $\sum_{i=1}^{n} R_i * T_i / \sum_{i=1}^{n} T_i$, where T_i is the node degree, and R_i is rank of i[th] target based on node degree, targeted by j[th] miRNA. Additionally, betweenness is calculated considering couples of nodes (v1, v2) and by counting the no. of shortest paths linking v1 and v2 and passing through a node n. The betweenness of a node in a biological network, such as a protein-signalling network, could indicate the relevance of a protein as functionally capable of holding together communicating proteins. Betweeness centrality was applied to identify the importance of a node in each of the four pathways (see the Results section). The additional features of SMARTmiR are literature-derived evidence reflecting the direct effect of each miRNA candidate in cancer aetiology, progression, spread and miRNA specific relationship to colorectal cancer. This third feature calculates the total amount of literature evidence (PubMed IDs) linking the j[th] miRNA to the four cellular processes relevant to tumor progression (proliferation and diffrentiation, apoptosis, metastasis, and angiogenesis), depicted as $E_j^{Pathway}$. Similarly, the fourth feature calculates the total amount of literature evidence (PubMed IDs) that reflects the association of the j[th] miRNA to colorectal cancer, depicted as E_j^{CRC}.

The ultimate goal of SMARTmiR is to prioritise the role of a candidate miRNA as a therapeutic biomarker for cetuximab treatment in CRC. Therefore, the last feature, $\sum_{i=1}^{n} E_i^{Cetuximab}$, calculates the total amount of literature evidence (PubMed IDs) of the j^{th} miRNA target's (i^{th}) related to cetuximab.

Separately we have also calculated the miRNA scoring with betweeness centrality of the nodes as second feature; maintaining first, third, fourth and fifth features of SMARTmiR identical. The resulting scoring is provided in Supplementary Table 16 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s15.xls).

10.2.3. Validation of predicted miRNA biomarkers

We used a list of differentially expressed miRNAs in cetuximab sensitive and resistant CRC patients (KRAS and VRAF wild type) (Mosakhani et al. 2012) to validate the miRNAs predicted and ranked by our workflow as significant therapeutic biomarker candidates for cetuximab treatment. To further substantiate the expression of targeted mRNAs by the most significant miRNA biomarkers, we also analysed the mRNA expression data (Khambata-Ford, Christopher R Garrett, et al. 2007). The mRNA expression data analysis was performed using the appropriate R software packages. In particular, we used the MAS5 normalisation method (Irizarry et al. 2003) and applied the SAM (Significance Analysis of Microarray) package to identify differentially expressed mRNAs (Tusher et al. 2001b). All differentially expressed mRNA with its corresponding p-value can be observed in the fourth column of Table 13.

10.3. Evalution of SMARTmiR workflow

10.3.1. Construction of molecular pathway maps crucial for cetuximab mode of action in CRC

The efficacy of cetuximab treatment in responsive colorectal cancer is mostly depend on the state of four cellular processes namely apoptosis, proliferation and differentiation, angiogenesis, epithelial-to-mesenchymal transition/metastasis (Toni M Brand et al. 2011a). First a comprehensive framework for the analysis and a structured overview of the bio molecular space for the mechanisms of action of cetuximab therapy in colorectal cancer was created by assembling four pathway maps which are representative of those cellular processes in CRC.

The landscapes of the calculated network parameters for each of the four constructed pathways are presented in Figure 40. The top three nodes of each pathway (Jak2, Akt, p53 in apoptosis; SMAD, TGF-β, Shc in metastasis; VEGFR, VEGF, Src in Angiogenesis; EGFR, c-Fos, c-Raf in proliferation and differentiation) are well-known players in the corresponding processes (Hanahan & Robert A Weinberg 2011; Toni M Brand et al. 2011a; Sakurai & Kudo 2011; Wittekind & Neid 2005). The high no. of the nodes representing receptors and kinases in all four pathway maps emphasise the crucial role of these molecules in oncogenesis and metastasis. Our maps also demonstrate that the roles of some molecules, such as transcription factors, in metastasis are better understood and more well-known compared with the other three processes.

To understand the degree of the cross-talk and the overlap between the four pathway maps, we integrated all four pathway maps into a resulting *integrated map*. Table 12 presents a statistical overview of the pathway maps and the percentage of literature validation of the nodes related to the corresponding cellular processes.

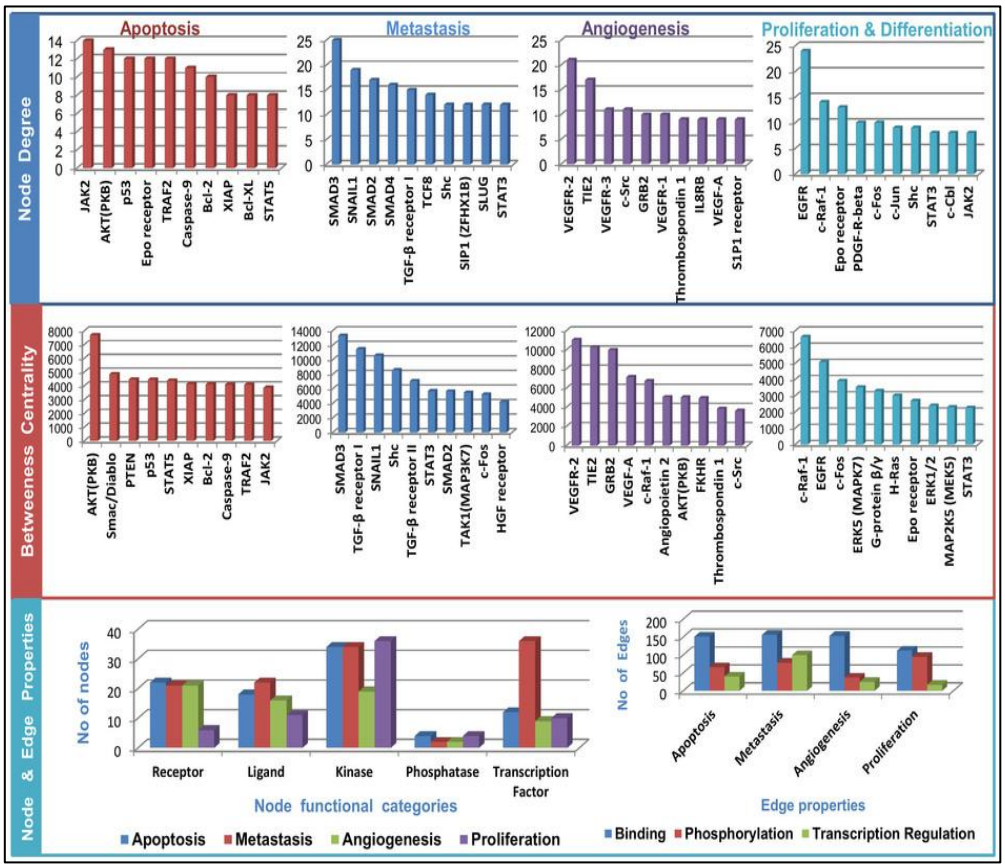


Figure 40: Landscape of the four cellular processes in terms of the node degree, betweenness centrality, functional category of nodes and edges.

Node degree: no of edges connected a node; Betweenness centrality: the no. of shortest paths from all of the vertices to all of the others that pass through that node.

A list of common nodes in all four pathways is provided in Supplementary Table 17 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s2.xls).

Cellular Processes	No. of Nodes	No. of Edges	Percentage of published relationships of nodes to the cellular processes
Apoptosis	179	305	55.3% (99/179)
Proliferation and Differentiation	151	261	45% (68/151)
Angiogenesis	201	289	36.8% (74/201)
Metastasis	245	471	46% (113/245)

Table 12: Statistical overview of the assembled pathway maps representing four cellular processes

The high percentage of literature validations (4th column of Table 12) confirms that the constructed pathways represent a good reflection of the current knowledge accumulated in the scientific literature. The metastasis pathway map is the largest in terms of the no. of nodes and edges, followed by angiogenesis, apoptosis, proliferation and differentiation. The integrated map consists of 465 nodes and 792 directed edges, demonstrating a large overlap among the molecules involved in the four processes. The literature validations of the node's relationships to the cellular processes are provided in Supplementary Table 18 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s3.xls).

To analyse the CRC specificity of the assembled pathway maps, we first explored the published TCGA (The Cancer Genome Atlas) data regarding the differential expression of the nodes' mRNAs in colon and rectum adenocarcinoma (Anon 2012). This analysis demonstrated that 96% of the nodes' mRNAs are differentially expressed (more than two standard deviations from the mean) in at least one of the 244 studied tumor samples. In addition, we examined the association of the pathway maps' nodes to colorectal cancer in the published scientific domain and determined that 45.3% of them (211/465) have been reported in the literature to have a relationship with CRC. All of those relationships are included in Supplementary Table 19 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s4.xls). The

differential expression of each mRNA in the TCGA tumor samples is provided in Supplementary Table 20 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s5.xls).

10.3.2. miRNAome screening for putative candidate biomarker

Upon construction of comprehensive pathway maps and verification of their association with the four cellular processes and in CRC, we screened the published miRNAome for miRNAs targeting at least one node of the four pathway maps, which could serve as potential biomarkers for further analysis. In our screening, we considered experimentally validated miRNA target interactions (MTIs) and computationally predicted MTIs. To increase the chance of the predicted MTIs being biologically relevant, overlapping MTIs between three different miRNA-target prediction algorithms (Krutovskikh & Herceg 2010; Witkos et al. 2011) were used. One method that can be used to compare the quality of MTI prediction is to calculate the percentage of the prediction that has already been experimentally validated. In that direction, the percentage of experimentally validated MTIs is calculated for predictions from three individual algorithms, i.e., Pictar, miRanda, DianaMicroT, and from the intersection of their predictions. Experimentally validated MTIs are obtained from TarBase and Pathway Studio. As demonstrated in Figure 41, the percentage obtained from intersection of the three prediction algorithms is higher than that obtained using any of the three prediction algorithms used individually (Figure 41). Using the intersection of three prediction algorithms, we are able to capture 17.5% of the experimentally validated MTIs (average for four processes). However, the results for each of the prediction algorithms are less impressive: Pictar (3.9%), miRanda (3.4%), DianaMicroT (1.7%). Therefore, the assumption of using the intersection of the three prediction algorithms proved to be a better approach and was used for further analysis. Total number of experimentally validated MTIs for all prediction software and their intersect for all four pathway maps are provided in Supplementary Table 21 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s6.xls).

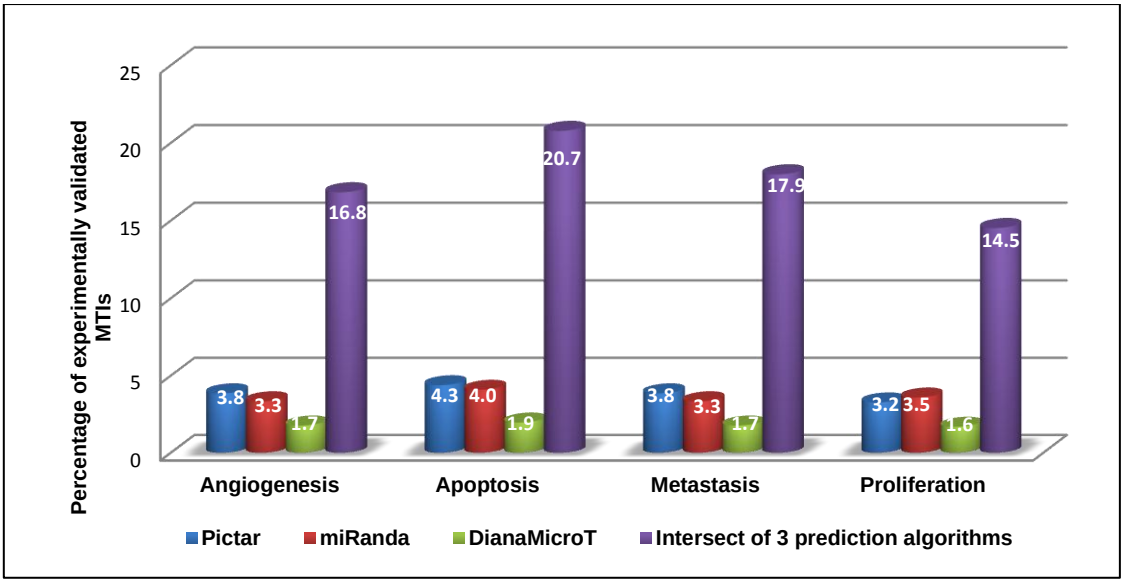


Figure 41: Comparative performance of Pictar, miRanda, DianaMicroT and the intersection of the three algorithms in capturing validated miRNA-target interactions.

The miRNAome screening revealed 335 miRNAs that target at least one node of the four assembled pathway maps. We further analysed the 335 miRNAs based on their ability to participate in all four processes, thus having a higher probability to be therapeutic biomarkers for cetuximab treatment in CRC patients. This analysis resulted in the selection of 188 miRNAs that interact with targets in all four of the pathway maps, and these selected miRNAs were used for the ranking and validation (Figure 42).

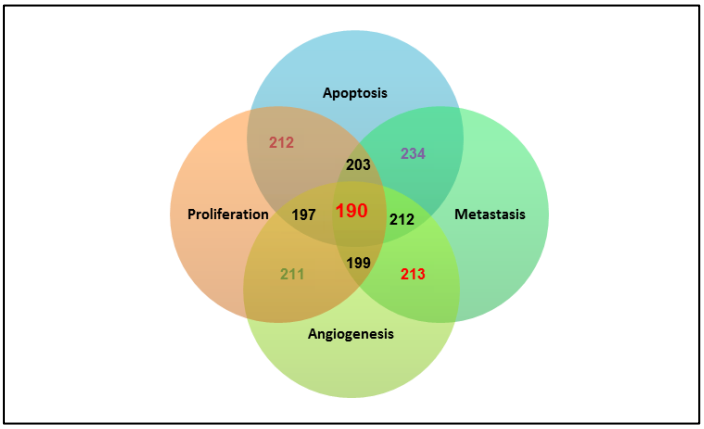
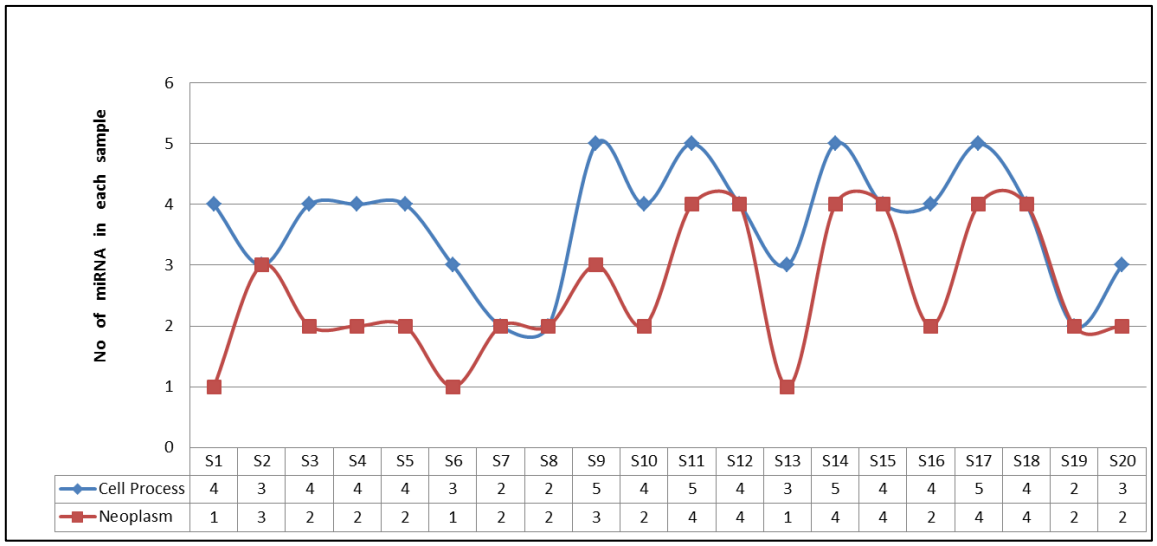


Figure 42: Quantities of miRNA species targeting each pathway and cross-sections.

To evaluate the relationship of those 188 miRNAs to the four cellular processes and to the neoplasm, we performed a random sampling approach. We randomly selected 20 non-overlapping sets containing 5 miRNAs each (S1, S2,.., S20 in Figure 43). Next the association between each miRNA (from each of those 20 samples) to four cellular processes (apoptosis, proliferation and differentiation, metastasis, angiogenesis) and to neoplasms were searched in the scientific literature. The number of miRNAs from each sample that are published to regulate at least one of those four cellular process and neoplasms are summarised in Figure 43. It is evident from Figure 43 that 74% of the miRNA candidates are known to regulate at least one of the four cellular processes and 51% of those miRNA are linked to different forms of neoplasms.



	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20
Cell Process	4	3	4	4	4	3	2	2	5	4	5	4	3	5	4	4	5	4	2	3
Neoplasm	1	3	2	2	2	1	2	2	3	2	4	4	1	4	4	2	4	4	2	2

Figure 43: The relationships of miRNAs from twenty randomly collected non-overlapping samples (five miRNAs each) to cell processes (angiogenesis, apoptosis, proliferation and differentiation, metastasis) and neoplasms.

The column S1, S2, S3, ……, S20; denote the 20 non overlapping samples having 5 miRNA in each of them. Each sample collected randomly from total 188 miRNA. The corresponding values in the shell for each sample with row starts with "Cell process" denote total no. of miRNA from that sample is linked to any of the four cellular processes and the relation is published. Similarly, the corresponding values in the shell for each sample with row starts with "Neoplasm" denote total no. of miRNA from that sample is linked to Neoplasms and the relation is published.

All 20 of the samples and the relationship between the miRNAs from each sample to cellular process and the neoplasm are provided in Supplementary Table 22 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s7.xls), Supplementary Table 23 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s8.xls) and Supplementary Table 24 (See:

http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s9.xls). Based on the node degree, the top 5 miRNAs from each of the four pathway maps are provided in Supplementary Figure 2 (See: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s1.pdf).

10.3.3. Prioritization of the selected miRNAs

Despite the potentially strong impact of the selected miRNAs on the fundamental molecular mechanisms underlying CRC, their relation to cetuximab treatment was not used in previous filtering of the 188 miRNAs, and the resulting candidate list remains too long for experimental validation. To further prioritise miRNA candidates that have the highest likelihood of acting as therapeutic biomarkers for cetuximab treatment, we ranked the miRNA candidates by applying a newly designed SMARTmiR algorithm, as described in the Materials and Methods section. The details of the 10 top-ranked miRNAs, including their scores and evidence of their interaction with cetuximab, and the role of the targets are summarised in Table 13. According to our predictions, those ten miRNAs might serve as the best candidates for therapeutic biomarkers for cetuximab treatment in CRC patients. All 188 miRNA with its SMARTmiR score and Higo gene id of the targets are provided in Supplementary Figure 3 (See the second figure: http://www.nature.com/srep/2015/150126/srep08013/extref/srep08013-s1.pdf).

miRNA	Score	No of Targets	Hugo Gene Id of differentially expressed miRNA Targets in Cetuximab sensitive to resistant CRC patients	No of TF, KNS	miRNA Fold Change
hsa-miR-21	3.57	33	NOTCH1 (0.004868), CD47 (0.031247), BRCA1 (0.013018), APAF1 (0.012285)	4, 3	8.1
hsa-miR-34a	2.21	28	YY1 (0.029778), NOTCH1 (0.004868), CD47 (0.031247), BRCA1 (0.013018), MYC (0.002497)	4, 2	4.6
hsa-	2.15	33	MKL2 (0.004586), CD47	7, 2	N/A

miR-145			(0.031247), MYC (0.002497)		
hsa-miR-27a	2.01	48	MAPK14 (0.009214), APAF1 (0.012285)	4, 9	5.8
hsa-miR-17	1.85	35	MAPK14 (0.009214), BRCA1 (0.013018), E2F1 (0.002186), EREG (3.41e-06)	3, 8	9.7
hsa-miR-155	1.73	29	CD47 (0.031247), FOXO3 (0.012217)	6, 4	N/A
hsa-miR-182	1.65	39	MKL2 (0.004586), CD47 (0.031247), FOXO3 (0.012217)	5, 2	N/A
hsa-miR-15a	1.59	25	BRCA1 (0.013018), PRKAR2A (0.021049)	1, 4	6.2
hsa-miR-96	1.53	38	PLCB4 (0.015726), FOXO3 (0.012217)	4, 5	N/A
hsa-miR-106a	1.48	26	NOTCH1 (0.004868), BRCA1 (0.013018), E2F1 (0.002186)	2, 6	N/A

Table 13: Top 10 miRNAs along with their scores, expression values, MTI, expression of MTI and miRNA in cetuximab sensitive to resistant CRC patients

1st column: top 10 miRNA candidates; 2nd column: calculated score based on developed scoring function; 3rd column: total no of miRNA targets in the "*Integrated map*"; 4th column: Hugo gene ID of miRNA targets those were differentially expressed in cetuximab sensitive compared with resistant CRC patients with significant p values; 5th column: no of transcription factors (TFs) and kinases (KNSs) among the targets; 6th column: expression value of the miRNA in cetuximab sensitive compared with resistant CRC patients (KRAS, BRAF wild type) published by Mosakhani *et al* (Mosakhani et al. 2012).

10.3.4. Validation of the prediction based on published experimental results

The optimal method of validating any systems biology prediction is through experimental results. In our case, the analysis of differentially expressed miRNAs in cetuximab-resistant CRC tumor samples would provide such validation. Recently, a group at the University of Helsinki studied differential miRNA expression patterns of 33 cetuximab-treated patients with metastatic colorectal cancer (Mosakhani et al. 2012). That group tested the association of each miRNA with the overall survival (OS) by applying a Cox proportional hazards regression model, and published a list of the 60 most differentially expressed miRNAs in patients with an extremely poor prognosis (resistant patients). According to our analysis, 85% (51 of 60 miRNAs) of the resistant patient-derived differentially expressed miRNAs are present in our list of the selected 188 miRNAs. Moreover, five of the ten top-ranked predicted miRNAs were found to be highly differentially expressed in resistant patients, exhibiting 4.6 to 9.7 fold changes (Table 13). More studies with greater no. of patients are needed to further validate our prediction; however, we consider this initial evidence to be very encouraging and to prove the applicability of our methodology.

11. Discussion

FDA in its critical path initiative emphasized on applying biomarkers as an essential tool to combat current situation of late stage drug failures. Number of recent publications suggested that more efficient drug discovery model can be designed by applying biomarkers in all stages of drug discovery and development i.e. emphasizing biomarker usage from target identification to drug marketing (Colburn 2003; Bakhtiar n.d.). However until now the usage of patient stratification biomarker in late stage clinical trials stands out among other biomarker applications to cope with the most alarming issue of expensive late stage failures. According to the simulation-based analysis performed by FDA and MIT consortium, early biomarker program is predicted to be an invaluable model for drug development (Trusheim et al. 2011). Corroborating this prediction, Parker *et al* have shown that the application of Her2 in the development of anti-breast cancer treatment reduced the clinical trial risk by 50% (Parker et al. 2012). On a global scale, it has been quantitatively demonstrated that probability of phase to phase transition is 15 to 19 percent higher for anti-cancer treatments if their trials include biomarker programs compared to those without (Hayashi et al. 2013). All these findings indicate that early application of stratified molecular biomarker may significantly improve the success of clinical development. As representative picture of worldwide clinical trials with stratified biomarker, our analysis on ClinicalTrial.gov shows that 21% of the stratified trials are done in the later stages (phase III and IV, Figure 22) with significant efforts in post-marketing research. A more comprehensive picture of worldwide clinical trials can be drawn by inclusion of proprietary clinical trial registries such as Trialtrove and PharmaProjects into our analysis; however the proprietary laws do not allow us to publish these results. ClinicalTrials.gov itself has reported issues with the updates, consistency and completeness of the data particularly in those trial files collected before the database launch (Wadman 2006; Innocenzi et al. 1984). Surely a complete manual curation of all 150.000 trial registry files may enhance the accuracy of our analysis however such an effort is immensely time consuming involving multiple scientific annotators which is beyond the scope of this research. Hence our semi-automatic approach in identifying trials with stratified biomarker from the oldest and largest clinical trial registry and successive analysis is a representative study of worldwide trend in clinical trials.

According to our analysis (Figure 21), oncology represents more than 75% of all the trials with stratified biomarker. So cancer being at the forefront of stratified medicine will dictate

the success rate of future clinical trials with stratified biomarker. The analysis also shows with 95% confidence interval that less than 5% of all interventional trials are using molecular biomarker for patient stratification. Some reasons for slow adoption of biomarker approaches in the clinics are summarized as follows: first of all stratification biomarker discovery requires an in depth understanding of disease mechanism and drug mode of action. Demonstration of clear relationship of biomarker changes and disease progression or treatment success requires substantial effort and resources spent early in pre-clinical research. In this direction, a number of system biology approaches based on quantitative modelling has been suggested to be of use for biomarker prediction reviewed by Kreeger *at al* (Kreeger & Lauffenburger 2009). However, labour-intensive collection of quantitative data as well as limitation of current computational power to model complex biological systems containing over hundred molecules hinders current use of quantitative modelling for biomarker prediction. Qualitative modelling approaches can provide an alternative for prospective biomarker prediction. Quite a few qualitative modelling are based on boolean networks and able to simulate the dynamics of signalling pathways. It has been employed for the discovery of novel oncological biomarkers as well as used to develop robust clinical treatment decisions (Sahoo 2012). An example of another type of modelling is integrative model (Figure 44) i.e. literature-derived knowledge and OMICS data together, reflecting 'cause and effect' relationships into an integrated biomarker discovery platform (Martin et al. 2012).

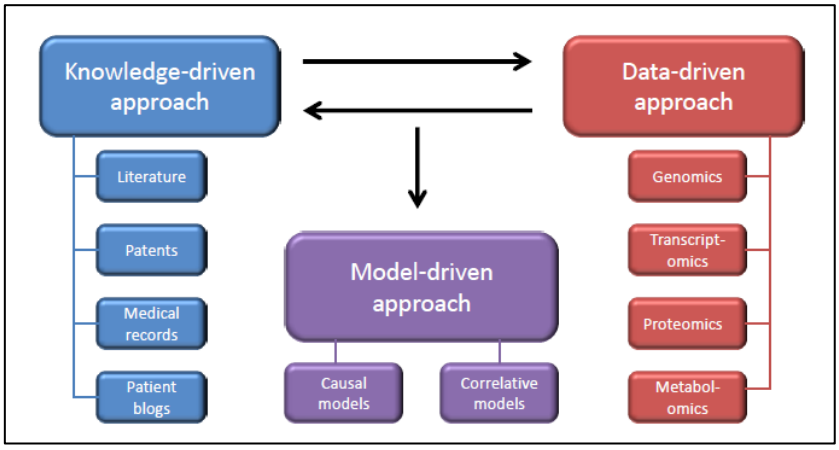


Figure 44: Integrative model driven approach to identifying candidate biomarkers (Younesi et al, 2013).

The figure proposed a model driven approach for the identification of biomarker by integrating omics data and textual knowledge.

A variety of OMICS technologies have been developed in recent years with the aim to contribute detailed understanding of disease pathophysiology and drug mode of action.

However neither OMICS data nor the knowledge accumulated in the text can be automatically translated into clinical advances. Knowledge capturing technologies combined with pathway analytics provide a great framework for OMICS data interpretation. The lack of standardized translational algorithms allowing the use of OMICS data along with the knowledge derived from the scientific literature hampers biomarker prediction. Hence any improvement is concurrent with improvement of knowledge representation standards to present dynamic interconnectivity of the molecular pathways aided by strong signal integration from experimental data.

The clinical developmental success of stratified therapy in cancer can also be augmented by a fresh approach of selecting new class of biomolecules as candidate biomarker. miRNAs can potentially be one such new class of candidate. miRNA plays an important role in tumorgenesis by regulating expression of oncogenes and tumor suppressors thus affecting cell proliferation, differentiation, apoptosis, invasion and angiogenesis. miRNAs are potential biomarkers for diagnosis, prognosis and therapies of different forms of cancer. We have developed a novel translational algorithm i.e. SMARTmiR to identify crucial miRNA for the efficacy of a targeted therapy. As a use case we have also shown the clinical utility of SMARTmiR to identify and rank miRNAs which can predict the efficacy of cetuximab therapy in one of the most prevalent cancer in western world i.e. colorectal cancer.

In colorectal cancer, the effects of cetuximab are mediated through various molecular pathways, including the Ras-Raf-MAPK, PI3K-AKT, protein kinase C, STAT and SRC pathways (Toni M Brand et al. 2011a; Maragkakis et al. 2009). The efficacy of cetuximab in responsive CRC patients is mainly manifested through reduction of cell proliferation and differentiation, inhibition of angiogenesis, prevention of epithelial to mesenchymal transition (metastasis) and induction of apoptosis (Toni M Brand et al. 2011b). However, cetuximab resistance mechanisms by alternative molecular pathways was recently reviewed (Toni M Brand et al. 2011b) which demonstrated that the reactivation of pro-angiogenic factors (pMAPK, VEGF) leading to increased angiogenesis in CRC is one such resistant mechanism (Ciardiello et al. 2004). In agreement with those results, our top-ranked miRNA, i.e., miR-21, is a well-known angiogenesis regulator in both in vitro and in vivo models (Sabatel et al. 2011). Moreover, cetuximab treatment affects the expression of miR-21 in vitro (du Rieu et al. 2010). Another group suggested that the cetuximab resistance mechanism is a phenomenon caused by an increased rate of EGFR degradation and internalisation; switching

towards alternate pathways for growth and survival of CRC resistant tumor cells (Lu et al. 2007). In resistant cells, EGFR is localised in the sub-cellular compartments i.e., endosome, mitochondria and nucleus. The overexpression of nuclear EGFR is linked to SFKs (SRC-family kinase) expression, modulating the up regulation of the PI3K/AKT pathway in cetuximab resistance (Wheeler et al. 2009). Mutations of KRAS are also connected to the increased activation of SFKs, affecting the MAPK, beta-catenin, STAT and PI3K/AKT pathways in CRC resistant tumors (Dunn et al. 2011). However, the mechanisms of cetuximab resistance in CRC remain poorly understood.

To address the issue, we attempted to create a comprehensive CRC specific molecular network snapshot leading to the four cellular processes of apoptosis, proliferation and differentiation, metastasis and angiogenesis, which are crucial for mode of action of Cetuximab therapy in CRC. Pathways from Metacore were integrated. The Metacore pathway knowledge base was selected because of the high experimental validation of molecular interactions in its pathways (Shmelkov et al. 2011b). Of the assembled *Integrated map*, 96% are differentially expressed in published CRC specific RNASeq data (Anon 2012). Based on the well-accepted node removal algorithm (Bossi & Lehner 2009; Waldman et al. 2010; Lopes et al. 2011), that high percentage makes the *Integrated map* CRC specific. miRNA-target interactions have been screened from prediction algorithms (miRanda, Pictar and DianaMicroT) and literature based knowledgebase (Tarbase, Pathway Studio). TargetScanS, Pictar and miRanda used individually, or in combination, provide a good balance between precision and recall (Witkos et al. 2011). However, Maragkakis et al. demonstrated that Pictar predictions overlap more than 75% of the predictions obtained from TargetsScanS (Maragkakis et al. 2009). Therefore, Pictar, miRanda and DianaMicroT were used. Next, a candidate miRNA was ranked based on its relationship to the four cellular processes and to CRC. The other ranking parameters were miRNA target's relation to cetuximab; the node degree of the target; and the number of targets for each candidate miRNA. A high number of targets of miRNA in a pathway suggests the multi-level regulation of that pathway (Uhlmann et al. 2012; Malumbres 2012). miRNAs also preferentially regulate network hubs that participate in complex dynamic processes, and their expression profile is highly dynamic, thereby requiring tighter regulatory control (Cui, Yu, Pan, et al. 2007; Cui, Yu, Purisima, et al. 2007).

Literature search confirmed that three of our top 10 miRNA in Table 13 i.e. miR-21 (1st), miR-34a (2nd), miR-17 (5th) are published biomarker for cetuximab therapy (du Rieu et al.

2010; Schou et al. 2014; Ragusa et al. 2012). Although no relation between miRNA to cetuximab was incorporated into the SMARTmiR scoring function. Our methodology has successfully predicted possible relationships between a miRNA and cetuximab. In the case of three top ranked miRNA i.e. miR-21, miR-34a and miR-17; the relationship has already been experimentally validated. This result again demonstrates the novelty and applicability of our methodology.

Genome-wide miRNA and mRNA expression profiles of cetuximab-sensitive and cetuximab-resistant mCRC patients and PLS regression/Pearson's correlation of significant differentially expressed miRNAs and target mRNAs followed by pathway-centric interpretation could be another approach to discover the role of miRNAs in cetuximab resistance mechanisms in mCRC. However, due to a lack of such experimental data, the SMARTmiR algorithm utilised existing resources to predict crucial miRNAs as therapeutic biomarkers for cetuximab treatment in CRC patients.

The accurate prediction of miRNAs that might serve as potential therapeutic biomarkers would be of great importance for patients. Herein, we developed a novel algorithm that facilitates the prediction of potential miRNA biomarkers based on the knowledge accumulated in the public domain. To our knowledge, there is only one published alternative methodology that uses literature-based evidence for drug-associated miRNA predictions. Recently, Rukov *et al*. launched PharmacomiR, a miRNA Pharmacogenomics database that uses the triplet sets consisting of a miRNA, a target gene and a drug associated with the gene to predict miRNAs that could serve as potential therapeutic biomarkers (Rukov et al. 2013). We compared the performance of PharmacomiR to that of our methodology in predicting the pharmacogenomics role of miRNAs in cetuximab treatment. PharmacomiR predicted 1102 unique miRNAs (6975 redundant miRNAs as the initial output). Next, we calculated the overlap between the number of predicted miRNAs with the published differentially upregulated miRNAs by Mosakhani *et al.* (Mosakhani et al. 2012) in cetuximab-sensitive and cetuximab-resistant CRC patients. Clearly, our methodology has higher prediction accuracy, with 27.1% (51 of 188) compared with 4.44% (49 of 1102) for PharmacomiR, in identifying experimentally validated potential miRNAs as therapeutic biomarkers for cetuximab therapy in CRC patients. Unlike PharmacomiR, SMARTmiR could also rank each candidate miRNA based on a novel disease specific score, making our methodology more advanced in prioritising candidate miRNAs for further experimental validation. A methodological comparison SMARTmiR and PharmacomiR has been provided in Table 14.

Methodology	SMARTmiR	Pharmaco-miR
Relation to Disease	Relation to Disease is achieved	No relations to disease
Relation to Drug	Pathway centric relation to Drug	Gene Centric relation to Drug
miRNA Prioritization	Pharmacogenomic role of a miRNA to a drug's action in a specific disease can be evaluated and ranked	Pharmacogenomic role of a miRNA to a drug's action in a specific disease cannot be evaluated and ranked
miRNA prediction Algorithms	Pictar, miRanda, DianaMicroT	Verse, TargetSCan, miRTarBase, miRecords, miRanda,Pita

Table 14: Methodological comparison between SMARTmiR and Pharmaco-miR

However, SMARTmiR algorithm has limitations. The inherent issue of pathway map building is that its completeness is limited to valid and available data. Therefore, the sensitivity and specificity of the translational methodology is a function of completeness of the presented interactome. Additionally, the developed pathway maps could not integrate the dynamic nature of miRNA regulations. Annotations and encoding of the transcripts of the genes in the pathway maps are absent. However, these are outstanding challenges in representing dynamic biological systems, and the scientific community must act to solve these issues.

Biological data and knowledge is growing exponentially. Efficient management of those resources can be invaluable to measure the plausibility of a hypothesis as a prospective biomarker before designing the experiment to validate the hypothesis. In that direction, the thesis has demonstrated the power of efficiently managing existing data and knowledge in oncology to predict the therapeutic biomarker. Finally a prospective plan on future scenario of biomarker research during drug development has been drawn in the next section focusing to reduce the risk of most expensive phase III drug failures.

Building a Disease Mechanism: A concrete understanding of the tissue specific targeted cancer mechanism is of prime importance. A knowledge map representing the disease mechanism has the conceptual benefit to cover the biomolecular space within a tissue where a drug treating the cancer has to work. Omics data sets on targeted cancer without any therapeutic intervention from GEO and data from ICGC/TCGA projects can be analysed, interpreted and further merged with existing knowledge on the targeted cancer from

biomedical literature to deliver a knowledge map representing the targeted cancer mechanisms.

Building Drug Mode of Action (MoA): Next deciphering the drug mode of action holds the key to understand the consistency of the treatment outcome. The data from preclinical, phase I can be analysed, interpreted and represented through a knowledge map. The comparison of drug mode of action map with reference to the disease mechanism map can detect the potential pitfalls of a drug. If the targeted drug is a small molecule then analysis of cancer cell line profiling data (http://www.broadinstitute.org/ctrp/) from Broad Institute can potentially provide vital clues on the drug mode of action and molecules that can potentially detect therapeutic efficacy. Currently targeted cancer drug falls into one of ten categories that have been described in Figure 4 of the thesis. If the drug being investigated falls in one of those ten groups then biomedical literature and ongoing clinical trials on same category of drugs can provide crucial knowledge on trial designing and applied biomarker.

Selection of biomarker: Conventional gene centric biomarker is unlikely to ascertain the therapeutic efficacy of a treatment as complex as cancer. Cancer is a system disorder disrupting the hallmarks of healthy cells (see Figure 3) i.e. severely damaging the normal expression of multiple molecular pathways. Hence transcriptomics profiles of a panel of biomarkers instead of a single one holds the promise for better stratification benefitting from targeted therapy. Especially, a panel of microRNAs which can regulate the expression of several genes and are essential for the drug's mode of action, can be a more potent cancer therapeutic biomarker.

Streamline the biomarker panel: The analysis followed by interpretation of preclinical, phase I data along with the literature mining (described in previous steps) can potentially generate novel hypothesis on a panel of therapeutic biomarkers. The panel of biomarkers from these analyses can be further tested in phase II trials. Supervised modeling can be applied first to train and then to test the stratification power of the panel of biomarkers in phase II trials. A knowledge based score similar to that produced by SMARTmiR algorithm for each of biomarker can be calculated and then applied in the supervised model along with expression profiles of the biomarker. This will ensure that statistically significance of expression along with parameters important for the drug MoA i.e. biological point views have been incorporated in building the model. Ideally this should improve the stratification power of the model compared to a model which is only built on the expression signature of

the biomarker. If the panel of biomarker shows considerable promise to stratify the patient population to predict the therapeutic efficacy of the drug, this panel of biomarker can be applied in designing the phase III trials.

12. Conclusion

Our meta-analysis of stratified molecular biomarker trials registered in ClinicalTrials.gov database clearly shows that the molecular biomarker trend is rapidly being adopted in clinical research with oncology being at the forefront of the personalized medicine. However percentage of the trials including patients' stratification based on molecular differentiation is still very low (less than 5%) reflecting all the challenges of biomarker discovery and development. A variety of OMICS technologies have been developed in recent years with the aim to identify biomarker by detailed understanding of disease pathophysiology and drug mode of action. However neither OMICS data nor the knowledge accumulated in the scientific literature can be automatically translated into clinical advances for biomarker discovery. The lack of standardized translational algorithms allowing the use of OMICS data along with the knowledge derived from the scientific literature hampers endeavours to predict biomarkers with greater confidence. Thus, any improvement of the current situation depends on the improvements in knowledge representation standards enabling to present interconnectivity of the molecular pathways supported by integration of strong signal from experimental data and enriched granular knowledge. The overall trend indicates that there is a drive away from correlative biomarkers towards causative biomarkers. Therefore, the aim of next-generation integrative models is to capture causal relationships between the candidate biomarker and clinical outcome.

A new class of stratified biomarker i.e. miRNA is also rapidly emerging in cancer treatment with the promise of stratifying the patient population with greater confidence and ease of detection. But lack of translational algorithm which can integrate OMICS data and knowledge to predict the causal relationship between candidate miRNA and clinical outcome of a treatment in a disease condition potentially hamper the discovery of miRNA as stratified biomarker. In this direction a novel integrative algorithm i.e. SMARTmiR has been invented by combining granular knowledge and available OMICS data. The algorithm is a pathway-centric methodology that facilitates the prediction of pharmacogenomics role of miRNA. The method is generic and can be applied to model the role of miRNA as a therapeutic biomarker for targeted therapy in any diseases. We are optimistic that the application of an optimised and fully automated version of the algorithm has the potential to be used as clinical decision support tool. Moreover this research will also provide a comprehensive and valuable knowledge map demonstrating functional bimolecular interactions in colorectal cancer to scientific community for future experimental studies in CRC. We have also detected seven

miRNA i.e. hsa-miR-145, has-miR-27a, has-miR-155, hsa-miR-182, hsa-miR-15a, hsa-miR-96 and hsa-miR-106a as top stratified biomarker candidate for cetuximab therapy in CRC which were not reported previously. Hence this research had drawn the attention of scientific community to investigate the pharmacogenomics role of those seven top ranked miRNA when treating CRC patients with cetuximab. In addition the research has identified 188 miRNA in total thus demonstrating an overall miRNA regulatory mechanism in CRC.

References

Abbott, A.L. et al., 2005. The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans. *Developmental Cell*, 9, pp.403–414.

Abdel-Rahman, W.M. & Peltomäki, P., 2008. Lynch syndrome and related familial colorectal cancers. *Critical reviews in oncogenesis*, 14(1), pp.1–22; discussion 23–31.

Addo, S., Yates, R.A. & Laight, A., 2002. *A phase I trial to assess the pharmacology of the new oestrogen receptor antagonist fulvestrant on the endometrium in healthy postmenopausal volunteers.*,

Altar, C.A. et al., 2008. A prototypical process for creating evidentiary standards for biomarkers and diagnostics. *Clinical pharmacology and therapeutics*, 83, pp.368–371.

Amado, R.G. et al., 2008. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 26(10), pp.1626–34.

Ambros, V. et al., 2003. A uniform system for microRNA annotation. *RNA (New York, N.Y.)*, 9, pp.277–279.

Andersen, J.N. et al., 2010. Pathway-based identification of biomarkers for targeted therapeutics: personalized oncology with PI3K pathway inhibitors. *Science translational medicine*, 2, p.43ra55.

Aoki, K. & Taketo, M.M., 2007. Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. *Journal of cell science*, 120(Pt 19), pp.3327–35.

Aravin, A.A. et al., 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Current Biology*, 11, pp.1017–1027.

Aravin, A.A. et al., 2003. The small RNA profile during Drosophila melanogaster development. *Developmental Cell*, 5, pp.337–350.

Arrowsmith, J., 2011a. Trial watch: Phase II failures: 2008-2010. *Nature reviews. Drug discovery*, 10(5), pp.328–9.

Arrowsmith, J., 2011b. Trial watch: phase III and submission failures: 2007-2010. *Nature reviews. Drug discovery*, 10(2), p.87.

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1), pp.25–29.

ASCO, Colorectal Cancer: Treatment Options. Available at: http://www.cancer.net/cancer-types/colorectal-cancer/treatment-options [Accessed January 2, 2015b].

Avisek Deyati, Rama Devi Sanam, Sreenivasa Rao Guggilla, V.R.P.& N.N., 2014. Molecular biomarkers in clinical development: what could we learn from the clinical trial registry? *Personalized Medicine*, 11(4), pp.381–393.

Bader, G.D., Betel, D. & Hogue, C.W. V, 2003. BIND: The Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31(1), pp.248–250.

Baker, S.J., Preisinger, A.C., et al., 1990. p53 gene mutations occur in combination with 17p allelic deletions as late events in colorectal tumorigenesis. *Cancer research*, 50(23), pp.7717–22.

Baker, S.J., Markowitz, S., et al., 1990. Suppression of human colorectal carcinoma cell growth by wild-type p53. *Science (New York, N.Y.)*, 249(4971), pp.912–915.

Bakhtiar, R., Biomarkers in drug discovery and development. *Journal of pharmacological and toxicological methods*, 57(2), pp.85–91.

Bandyopadhyaya, R. et al., 2002. Stabilization of Individual Carbon Nanotubes in Aqueous Solutions. *Nano Letters*, 2(1), pp.25–28.

Bartel, D.P., 2004. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116, pp.281–297.

Bartel, D.P., 2009. MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136, pp.215–233.

Bauer-Mehren, A., Furlong, L.I. & Sanz, F., 2009. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular systems biology*, 5, p.290.

Bazensky, I., Shoobridge-Moran, C. & Yoder, L.H., 2007. Colorectal cancer: an overview of the epidemiology, risk factors, symptoms, and screening guidelines. *Medsurg nursing : official journal of the Academy of Medical-Surgical Nurses*, 16(1), pp.46–51; quiz 52.

Beard, D.A. et al., 2009. CellML metadata standards, associated tools and repositories. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367(1895), pp.1845–1867.

Bekaii-Saab, T. et al., 2009. *A multi-institutional phase II study of the efficacy and tolerability of lapatinib in patients with advanced hepatocellular carcinomas.*,

Bell, S.A. & Tudur Smith, C., 2014. A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of ClinicalTrials.gov. *Orphanet journal of rare diseases*, 9, p.170.

Benjamini Yoav, H.Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1), pp.289–300.

Berezikov, E., 2011. Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics*, 12, pp.846–860.

Bhattacharya, S. & Mariani, T.J., 2009. Array of hope: expression profiling identifies disease biomarkers and mechanism. *Biochemical Society transactions*, 37(Pt 4), pp.855–62.

Biomarkers Definitions Working Group, 2001. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology and therapeutics*, 69(3), pp.89–95.

Bogaert, J. & Prenen, H., 2014a. Molecular genetics of colorectal cancer. *Annals of Gastroenterology*, 27, pp.9–14.

Bogaert, J. & Prenen, H., 2014b. Molecular genetics of colorectal cancer. *Annals of gastroenterology : quarterly publication of the Hellenic Society of Gastroenterology*, 27(1), pp.9–14.

Boland, C.R. & Goel, A., 2010. Microsatellite Instability in Colorectal Cancer. *Gastroenterology*, 138(6), pp.2073–2087.e3.

Bollag, G. et al., 2012. Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nature Reviews Drug Discovery*.

Bosetti, C. et al., 2011. Recent trends in colorectal cancer mortality in Europe. *International journal of cancer. Journal international du cancer*, 129(1), pp.180–91.

Bossi, A. & Lehner, B., 2009. Tissue specificity and the human protein interaction network. *Molecular systems biology*, 5, p.260.

Bradley, C.J. et al., 2011. Productivity savings from colorectal cancer prevention and control strategies. *American journal of preventive medicine*, 41(2), pp.e5–e14.

Brand, T.M., Iida, M. & Wheeler, D.L., 2011. Molecular mechanisms of resistance to the EGFR monoclonal antibody cetuximab. *Cancer Biology and Therapy*, 11, pp.777–792.

Brand, T.M., Iida, M. & Wheeler, D.L., 2011a. Molecular mechanisms of resistance to the EGFR monoclonal antibody cetuximab. *Cancer biology & therapy*, 11(9), pp.777–92.

Brand, T.M., Iida, M. & Wheeler, D.L., 2011b. Molecular mechanisms of resistance to the EGFR monoclonal antibody cetuximab. *Cancer biology & therapy*, 11(9), pp.777–92.

Brazma, A. et al., 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics*, 29(4), pp.365–371.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45(1), pp.5–32.

Brenner, H., Kloor, M. & Pox, C.P., 2014. Colorectal cancer. *Lancet*, 383(9927), pp.1490–502.

Brentnall, T.A. et al., 2009. Proteins that underlie neoplastic progression of ulcerative colitis. *Proteomics - Clinical Applications*, 3, pp.1326–1337.

Bui, T. V. & Mendell, J.T., 2010. Myc: Maestro of MicroRNAs. *Genes & Cancer*, 1, pp.568–575.

Bundschus, M. et al., 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics*, 9, p.207.

Bushati, N. & Cohen, S.M., 2007. microRNA functions. *Annual review of cell and developmental biology*, 23, pp.175–205.

Butcher, E.C., Berg, E.L. & Kunkel, E.J., 2004. Systems biology in drug discovery. *Nature biotechnology*, 22, pp.1253–1259.

Calin, G.A. et al., 2002. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, 99, pp.15524–15529.

Calvert, P.M. & Frucht, H., 2002. The genetics of colorectal cancer. *Annals of internal medicine*, 137(7), pp.603–12.

Cao, X., Maloney, K.B. & Brusic, V., 2008. Data mining of cancer vaccine trials: A bird's-eye view. *Immunome Research*, 4(1).

Cancer Genome Atlas Network, 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), pp.330–7.

Carmeliet, P. & Jain, R.K., 2011. Molecular mechanisms and clinical applications of angiogenesis. *Nature*, 473, pp.298–307.

Center, M.M. et al., 2009. Worldwide variations in colorectal cancer. *CA: a cancer journal for clinicians*, 59(6), pp.366–78.

Chapman, P.B. et al., 2011. *Improved survival with vemurafenib in melanoma with BRAF V600E mutation.*,

Chen, R. et al., 2007. Quantitative proteomics analysis reveals that proteins differentially expressed in chronic pancreatitis are also frequently involved in pancreatic cancer. *Molecular & cellular proteomics : MCP*, 6, pp.1331–1342.

Chen, X. et al., 2008. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell research*, 18, pp.997–1006.

Chen, X., Slack, F.J. & Zhao, H., 2013. Joint analysis of expression profiles from multiple cancers improves the identification of microRNA-gene interactions. *Bioinformatics*, 29(17), pp.2137–2145.

Chiang, H.R. et al., 2010. Mammalian microRNAs: Experimental evaluation of novel and previously annotated genes. *Genes and Development*, 24, pp.992–1009.

Chung, K.Y., Shia, J., Kemeny, N.E., et al., 2005. Cetuximab shows activity in colorectal cancer patients with tumors that do not express the epidermal growth factor receptor by immunohistochemistry. *Journal of Clinical Oncology*, 23, pp.1803–1810.

Chung, K.Y., Shia, J., Kemeny, N.E., et al., 2005. Cetuximab shows activity in colorectal cancer patients with tumors that do not express the epidermal growth factor receptor by immunohistochemistry. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23(9), pp.1803–10.

Ciardiello, F. et al., 2004. Antitumor activity of ZD6474, a vascular endothelial growth factor receptor tyrosine kinase inhibitor, in human cancer cells with acquired resistance to antiepidermal growth factor receptor therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 10(2), pp.784–93.

Colburn, W.A., 2003. Biomarkers in drug discovery and development: from target identification through drug marketing. *Journal of clinical pharmacology*, 43(4), pp.329–41.

Croft, D. et al., 2014. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1).

Cui, Q., Yu, Z., Purisima, E.O., et al., 2007. MicroRNA regulation and interspecific variation of gene expression. *Trends in genetics : TIG*, 23(8), pp.372–5.

Cui, Q., Yu, Z., Pan, Y., et al., 2007. MicroRNAs preferentially target the genes with high transcriptional regulation complexity. *Biochemical and biophysical research communications*, 352(3), pp.733–8.

Cunningham, D. et al., 2004a. *Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer.*,

Cunningham, D. et al., 2004b. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *The New England journal of medicine*, 351(4), pp.337–45.

Dabney, A.R., 2005. Classification of microarrays to nearest centroids. *Bioinformatics*, 21(22), pp.4148–4154.

Dang, N.H. et al., 2007. Phase II trial of the combination of denileukin diftitox and rituximab for relapsed/refractory B-cell non-Hodgkin lymphoma. *British Journal of Haematology*, 138, pp.502–505.

Davies, H. et al., 2002. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892), pp.949–954.

Davis, J.C. et al., 2009. The microeconomics of personalized medicine: today's challenge and tomorrow's promise. *Nature reviews. Drug discovery*, 8, pp.279–286.

Dean, E. & Lorigan, P., 2012. Advances in the management of melanoma: targeted therapy, immunotherapy and future directions. *Expert review of anticancer therapy*, 12(11), pp.1437–48.

DeCosse, J.J. et al., 1993. Gender and colorectal cancer. *European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP)*, 2(2), pp.105–15.

Demir, E. et al., 2010. The BioPAX community standard for pathway data sharing. *Nature biotechnology*, 28(9), pp.935–942.

DeRisi, J.L., Iyer, V.R. & Brown, P.O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science (New York, N.Y.)*, 278(5338), pp.680–686.

Derynck, R., Akhurst, R.J. & Balmain, A., 2001. TGF-beta signaling in tumor suppression and cancer progression. *Nature genetics*, 29(2), pp.117–129.

Deutsch, E.W. et al., 2015. Development of data representation standards by the human proteome organization proteomics standards initiative. *Journal of the American Medical Informatics Association : JAMIA*, 22(3), pp.495–506.

DeVita VT, Lawrence TS, Rosenberg SA, 2008. *Cancer: Principles and Practice of Oncology* 1st ed., Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins.

Deyati, A. et al., 2013. Challenges and opportunities for oncology biomarker discovery. *Drug discovery today*, 18(13-14), pp.614–24.

Deyati, A. et al., micro-RNAs as therapeutic biomarkers of. , pp.1–9.

Diez, D. et al., 2010. The use of network analyses for elucidating mechanisms in cardiovascular disease. *Molecular bioSystems*, 6(2), pp.289–304.

Di Nicolantonio, F. et al., 2008. Wild-type BRAF is required for response to panitumumab or cetuximab in metastatic colorectal cancer. *Journal of Clinical Oncology*, 26, pp.5705–5712.

Doniger, S.W. et al., 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome biology*, 4(1), p.R7.

Dove-Edwin, I. et al., 2006. Prospective results of surveillance colonoscopy in dominant familial colorectal cancer with and without Lynch syndrome. *Gastroenterology*, 130(7), pp.1995–2000.

Draghici, S. et al., 2003. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Research*, 31(13), pp.3775–3781.

Druker, B.J. et al., 2001. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *The New England journal of medicine*, 344(14), pp.1031–7.

Ducray, F. et al., 2008. Anaplastic oligodendrogliomas with 1p19q codeletion have a proneural gene expression profile. *Molecular cancer*, 7, p.41.

Dunn, E.F. et al., 2011. Dasatinib sensitizes KRAS mutant colorectal tumors to cetuximab. *Oncogene*, 30(5), pp.561–74.

Dupart, J., Zhang, W. & Trent, J.C., 2011. Gastrointestinal stromal tumor and its targeted therapeutics. *Chinese Journal of Cancer*, 30, pp.303–314.

Eberhard, D.A. et al., 2005. Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23(25), pp.5900–9.

Edgar, R., Domrachev, M. & Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1), pp.207–210.

Edge, S.B. & Compton, C.C., 2010. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of surgical oncology*, 17(6), pp.1471–1474.

Eisen, M.B. et al., 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25), pp.14863–14868.

Ekins, S. et al., 2007. Pathway mapping tools for analysis of high content data. *Methods in molecular biology (Clifton, N.J.)*, 356, pp.319–50.

Ekstrøm, C.T. et al., 2004. Spot shape modelling and data transformations for microarrays. *Bioinformatics*, 20(14), pp.2270–2278.

Elias, T. et al, 2006. Why products fail in phase III. *In Vivo 24*.

Enot, D.P. et al., 2006. Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. *Proceedings of the National Academy of Sciences of the United States of America*, 103(40), pp.14865–14870.

Evan, G.I. & Vousden, K.H., 2001. Proliferation, cell cycle and apoptosis in cancer. *Nature*, 411, pp.342–348.

Fazi, F. et al., 2007. Epigenetic Silencing of the Myelopoiesis Regulator microRNA-223 by the AML1/ETO Oncoprotein. *Cancer Cell*, 12, pp.457–466.

Fearon, E.R., 2011. Molecular genetics of colorectal cancer. *Annual review of pathology*, 6, pp.479–507.

Fearon, E.R. & Vogelstein, B., 1990. A genetic model for colorectal tumorigenesis. *Cell*, 61(5), pp.759–67.

Felli, N. et al., 2009. MicroRNA 223-dependent expression of LMO2 regulates normal erythropoiesis. *Haematologica*, 94, pp.479–486.

Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F., 2012. *GLOBOCAN 2012—cancer incidence and mortality worldwide*,

Fichtner, I. et al., 2008. Establishment of patient-derived non-small cell lung cancer xenografts as models for the identification of predictive biomarkers. *Clinical Cancer Research*, 14, pp.6456–6468.

Flynt, A.S. et al., 2007. Zebrafish miR-214 modulates Hedgehog signaling to specify muscle cell fate. *Nature genetics*, 39, pp.259–263.

Frank, R. & Hargreaves, R., 2003a. Clinical biomarkers in drug discovery and development. *Nature reviews. Drug discovery*, 2(7), pp.566–80.

Frank, R. & Hargreaves, R., 2003b. Clinical biomarkers in drug discovery and development. *Nature reviews. Drug discovery*, 2, pp.566–580.

Frattini, M. et al., 2007. PTEN loss of expression predicts cetuximab efficacy in metastatic colorectal cancer patients. *British journal of cancer*, 97(8), pp.1139–45.

Freeman, D.J. et al., 2009a. Activity of panitumumab alone or with chemotherapy in non-small cell lung carcinoma cell lines expressing mutant epidermal growth factor receptor. *Molecular cancer therapeutics*, 8(6), pp.1536–46.

Freeman, D.J. et al., 2009b. Activity of panitumumab alone or with chemotherapy in non-small cell lung carcinoma cell lines expressing mutant epidermal growth factor receptor. *Molecular cancer therapeutics*, 8, pp.1536–1546.

Friedman, Lawrence M., Furberg, Curt D., DeMets, D., 2010. *Fundamentals of Clinical Trials* 4th ed.,

Frueh, F.W. et al., 2008. Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. *Pharmacotherapy*, 28(8), pp.992–8.

Fu, Z. et al., 2006. Metastasis suppressor gene Raf kinase inhibitor protein (RKIP) is a novel prognostic marker in prostate cancer. *The Prostate*, 66, pp.248–256.

Galiatsatos, P. & Foulkes, W.D., 2006. Familial adenomatous polyposis. *The American journal of gastroenterology*, 101(2), pp.385–398.

Gill, G.N. et al., 1984. Monoclonal anti-epidermal growth factor receptor antibodies which are inhibitors of epidermal growth factor binding and antagonists of epidermal growth factor binding and antagonists of epidermal growth factor-stimulated tyrosine protein kinase activity. *The Journal of biological chemistry*, 259, pp.7755–7760.

Gilmartin, A.G. et al., 2011. GSK1120212 (JTP-74057) is an inhibitor of MEK activity and activation with favorable pharmacokinetic properties for sustained in vivo pathway inhibition. *Clinical Cancer Research*, 17, pp.989–1000.

Gold, D.L., Wang, J. & Coombes, K.R., 2005. Inter-gene correlation on oligonucleotide arrays: How much does normalization matter? *American Journal of PharmacoGenomics*, 5(4), pp.271–279.

Grady, W.M. et al., 1998. Mutation of the type II transforming growth factor-beta receptor is coincident with the transformation of human colon adenomas to malignant carcinomas. *Cancer research*, 58(14), pp.3101–4.

Grady, W.M. et al., 1999. Mutational inactivation of transforming growth factor beta receptor type II in microsatellite stable colon cancers. *Cancer research*, 59(2), pp.320–4.

Greene FL, Page DL, Fleming ID, 2002. *AJCC Cancer Staging Manual* 6th ed., New York: Springer Verlag.

Guo, C.J. et al., 2009. Effects of upregulated expression of microRNA-16 on biological properties of culture-activated hepatic stellate cells. *Apoptosis*, 14, pp.1331–1340.

Guo, P. et al., 2012. Trends in cancer mortality in China: an update. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, 23(10), pp.2755–62.

H. Chang, C. E. Horak, P. Mukhopadhyay, C. Lowery, J.B. and J.A.S., 2011. No Title. *Journal of Clinical Oncology*, 29, p.1064.

Ha, M. & Kim, V.N., 2014. Regulation of microRNA biogenesis. *Nature reviews. Molecular cell biology*, 15, pp.509–524.

Haggar, F.A. & Boushey, R.P., 2009. Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 22(4), pp.191–7.

Hall, M.A., 2000. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In *Seventeenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, pp. 359–366.

Hanahan, D. & Weinberg, R.A., 2011. Hallmarks of cancer: The next generation. *Cell*, 144, pp.646–674.

Hanahan, D. & Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5), pp.646–74.

Harley, C.B., 2008. Telomerase and cancer therapeutics. *Nature reviews. Cancer*, 8, pp.167–179.

Harsha, H.C. et al., 2009. A compendium of potential biomarkers of pancreatic cancer. *PLoS Medicine*, 6.

Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and J.D., 2000. *Molecular Cell Biology* 4th ed.,

Hatziapostolou, M., Polytarchou, C. & Iliopoulos, D., 2013. MiRNAs link metabolic reprogramming to oncogenesis. *Trends in Endocrinology and Metabolism*, 24, pp.361–373.

Hayashi, K., Masuda, S. & Kimura, H., 2013. Impact of biomarker usage on oncology drug development. *Journal of Clinical Pharmacy and Therapeutics*, 38, pp.62–67.

Hayes, J., Peruzzi, P.P. & Lawler, S., 2014. MicroRNAs in cancer: Biomarkers, functions and therapy. *Trends in Molecular Medicine*, 20, pp.460–469.

He, Z. et al., 2014. A method for analyzing commonalities in clinical trial target populations. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2014, pp.1777–86.

Heinrich, M.C. et al., 2003. Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. *Journal of Clinical Oncology*, 21, pp.4342–4349.

Hochhaus, A. et al., 2008. Dasatinib induces durable cytogenetic responses in patients with chronic myelogenous leukemia in chronic phase with resistance or intolerance to imatinib. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K*, 22, pp.1200–1206.

Hsu, S.-D. et al., 2011. miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic acids research*, 39(Database issue), pp.D163–9.

Huang, S.M., Bock, J.M. & Harari, P.M., 1999. Epidermal growth factor receptor blockade with C225 modulates proliferation, apoptosis, and radiosensitivity in squamous cell carcinomas of the head and neck. *Cancer research*, 59, pp.1935–1940.

Hucka, M. et al., 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, 19(4), pp.524–31.

Innocenzi, A. et al., 1984. Proposal of a new cutaneous incision in radical mastectomy. *Minerva chirurgica*, 39, pp.785–790.

Irizarry, R.A. et al., 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4), p.e15.

Jackson-Thompson, J. et al., 2006. Descriptive epidemiology of colorectal cancer in the United States, 1998-2001. *Cancer*, 107(5 Suppl), pp.1103–11.

Jacobsen, A. et al., 2013. Analysis of microRNA-target interactions across diverse cancer types. *Nature structural & molecular biology*, 20(11), pp.1325–32.

Jagarlapudi, S.A.R.P. & Kishan, K.V.R., 2009. Database systems for knowledge-based discovery. *Methods in molecular biology (Clifton, N.J.)*, 575, pp.159–172.

Janoueix-Lerosey, I. et al., 2008. Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature*, 455, pp.967–970.

Janout, V. & Kollárová, H., 2001. Epidemiology of colorectal cancer. *Biomedical papers of the Medical Faculty of the University Palacký, Olomouc, Czechoslovakia*, 145(1), pp.5–10.

Jansson, M.D. & Lund, A.H., 2012. MicroRNA and cancer. *Molecular Oncology*, 6, pp.590–610.

Jasperson, K.W. et al., 2010. Hereditary and familial colon cancer. *Gastroenterology*, 138(6), pp.2044–58.

Jass, J.R., 2004. Limitations of the adenoma-carcinoma sequence in colorectum. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 10(17), pp.5969–70; author reply 5970.

Jemal, A. et al., 2011. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2), pp.69–90.

Jessen, W. et al., 2012. Mining PubMed for Biomarker-Disease Associations to Guide Discovery. *Nature Precedings*.

Jiang, Q. et al., 2009. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic acids research*, 37(Database issue), pp.D98–104.

Jørgensen, J.T., 2013. Companion diagnostics in oncology - Current status and future aspects. *Oncology (Switzerland)*, 85, pp.59–68.

Jovanovic, M. & Hengartner, M.O., 2006. miRNAs and apoptosis: RNAs to die for. *Oncogene*, 25, pp.6176–6187.

Jung, S.-B. et al., 2012. Clinico-pathologic Parameters for Prediction of Microsatellite Instability in Colorectal Cancer. *Cancer research and treatment : official journal of Korean Cancer Association*, 44(3), pp.179–86.

Kaminski, M.S. et al., 2005. *131I-tositumomab therapy as initial treatment for follicular lymphoma.*,

Kanehisa, M. & Goto, S., 2000. Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28, pp.27–30.

Karube, Y. et al., 2005. Reduced expression of Dicer associated with poor prognosis in lung cancer patients. *Cancer Science*, 96, pp.111–115.

Kawamoto, T. et al., 1983. Growth stimulation of A431 cells by epidermal growth factor: identification of high-affinity receptors for epidermal growth factor by an anti-receptor monoclonal antibody. *Proceedings of the National Academy of Sciences of the United States of America*, 80, pp.1337–1341.

Keshava Prasad, T.S. et al., 2009. Human Protein Reference Database - 2009 update. *Nucleic Acids Research*, 37(SUPPL. 1), pp.D767–D772.

Khambata-Ford, S., Garrett, C.R., et al., 2007. Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *Journal of Clinical Oncology*, 25, pp.3230–3237.

Khambata-Ford, S., Garrett, C.R., et al., 2007. Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 25(22), pp.3230–7.

Khanna, I., 2012. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug discovery today*, 17(19-20), pp.1088–102.

Khatri, P., Sirota, M. & Butte, A.J., 2012. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 8.

Kimura, H. et al., 2007. Antibody-dependent cellular cytotoxicity of cetuximab against tumor cells with wild-type or mutant epidermal growth factor receptor. *Cancer Science*, 98, pp.1275–1280.

Kingsford, C. & Salzberg, S.L., 2008. What are decision trees? *Nature Biotechnology*, 26(9), pp.1011–1013.

Kinzler, K.W. & Vogelstein, B., 1998. Landscaping the cancer terrain. *Science (New York, N.Y.)*, 280(5366), pp.1036–7.

Kinzler, K.W. & Vogelstein, B., 1996. Lessons from hereditary colorectal cancer. *Cell*, 87(2), pp.159–70.

Kitano, H. et al., 2005. Using process diagrams for the graphical representation of biological networks. *Nature biotechnology*, 23(8), pp.961–966.

Knowlton, M.N. et al., 2008. A PATO-compliant zebrafish screening database (MODB): management of morpholino knockdown screen information. *BMC bioinformatics*, 9, p.7.

Korinek, V. et al., 1997. Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma. *Science (New York, N.Y.)*, 275(5307), pp.1784–7.

Korkontzelos, I., Mu, T. & Ananiadou, S., 2012. ASCOT: a text mining-based web-service for efficient search and assisted creation of clinical trials. *BMC Medical Informatics and Decision Making*, 12(Suppl 1), p.S3.

Korol, A.B., 2003, Microarray cluster analysis and applications. http://www.science.co.il/enuka/Essays/Microarray-Review.pdf [Accessed June 25, 2015]

Kreeger, P.K. & Lauffenburger, D.A., 2009. Cancer systems biology: A network modeling perspective. *Carcinogenesis*, 31, pp.2–8.

Kreitman, R.J., 2006. Immunotoxins for targeted cancer therapy. *The AAPS journal*, 8(3), pp.E532–51.

Kressner, U. et al., 1999. Prognostic value of p53 genetic changes in colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 17(2), pp.593–9.

Krutovskikh, V.A. & Herceg, Z., 2010. Oncogenic microRNAs (OncomiRs) as a new class of cancer biomarkers. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 32(10), pp.894–904.

Kumar, A. et al., 2008. Structure and clinical relevance of the epidermal growth factor receptor in human cancer. *Journal of Clinical Oncology*, 26, pp.1742–1751.

Kwak, E.L. et al., 2010. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *The New England journal of medicine*, 363(18), pp.1693–703.

Lagos-Quintana, M. et al., 2001. Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294, pp.853–858.

Larsson, S.C. & Wolk, A., 2006. Meat consumption and risk of colorectal cancer: a meta-analysis of prospective studies. *International journal of cancer. Journal international du cancer*, 119(11), pp.2657–64.

Lau, N.C. et al., 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294, pp.858–862.

Le, T.D. et al., 2014. From miRNA regulation to miRNA-TF co-regulation: computational approaches and challenges. *Briefings in Bioinformatics*, p.bbu023–.

Le, T.D. et al., 2013. Inferring microRNA-mRNA causal regulatory relationships from expression data. *Bioinformatics*, 29(6), pp.765–771.

Lee, K.M., Choi, E.J. & Kim, I.A., 2011. microRNA-7 increases radiosensitivity of human cancer cells with activated EGFR-associated signaling. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*, 101(1), pp.171–6.

Lee, Y.H. et al., 2012. The CTLA-4 +49 A/G and -318 C/T polymorphisms and susceptibility to asthma: A meta-analysis. *Molecular Biology Reports*, 39, pp.8525–8532.

Van Leeuwen, I.M.M. et al., 2006. Crypt dynamics and colorectal cancer: Advances in mathematical modelling. *Cell Proliferation*, 39(3), pp.157–181.

Lehmann, E.L, Romano, J.., 2006. Testing statistical hypotheses. *Metrika*, 64(2), pp.255–56.

Lengauer, C., Kinzler, K.W. & Vogelstein, B., 1997. Genetic instability in colorectal cancers. *Nature*, 386, pp.623–627.

Lewis Phillips, G.D. et al., 2008. Targeting HER2-positive breast cancer with trastuzumab-DM1, an antibody-cytotoxic drug conjugate. *Cancer research*, 68(22), pp.9280–90.

Li, C. & Wong, W.H., 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, 98(1), pp.31–36.

Li Gong, X.J.L., EGFR Inhibitor Pathway, Pharmacodynamics. Available at: https://www.pharmgkb.org/pathway/PA162356267 [Accessed January 5, 2015].

Li, J. & Lu, Z., 2012. Systematic identification of pharmacogenomics information from clinical trials. *Journal of Biomedical Informatics*, 45(5), pp.870–878.

Li, J.-M. et al., 2012. Down-regulation of fecal miR-143 and miR-145 as potential markers for colorectal cancer. *Saudi medical journal*, 33(1), pp.24–9.

Li, S. et al., 2005. Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. *Cancer Cell*, 7, pp.301–311.

Liang, Z. et al., 2011. MirAct: A web tool for evaluating microRNA activity based on gene expression data. *Nucleic Acids Research*, 39(SUPPL. 2).

Licata, L. et al., 2012. MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research*, 40(D1).

Lièvre, A. et al., 2006. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer research*, 66(8), pp.3992–5.

Lim, L.P. et al., 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027), pp.769–73.

Lim, L.P. et al., 2003. Vertebrate microRNA genes. *Science (New York, N.Y.)*, 299, p.1540.

Lin, S.-L. et al., 2008. Mir-302 reprograms human skin cancer cells into a pluripotent ES-cell-like state. *RNA (New York, N.Y.)*, 14, pp.2115–2124.

Lito, P., Rosen, N. & Solit, D.B., 2013. Tumor adaptation and resistance to RAF inhibitors. *Nature medicine*, 19, pp.1401–9.

Liu, B. et al., 2009. Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy. *BMC bioinformatics*, 10, p.408.

Liu, B. et al., 2000. Induction of apoptosis and activation of the caspase cascade by anti-EGF receptor monoclonal antibodies in DiFi human colon cancer cells do not involve the c-jun N-terminal kinase activity. *British journal of cancer*, 82, pp.1991–1999.

Liu, C.-G. et al., 2008. MicroRNA expression profiling using microarrays. *Nature protocols*, 3(4), pp.563–78.

Liu, K. et al., 2011. Increased expression of microRNA-21and its association with chemotherapeutic response in human colorectal cancer. *The Journal of international medical research*, 39(6), pp.2288–95.

Liu, N. & Zhao, H., 2006. A non-parametric approach to population structure inference using multilocus genotypes. *Human genomics*, 2(6), pp.353–364.

Lockhart, D.J. et al., 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13), pp.1675–1680.

Lopes, T.J.S. et al., 2011. Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics (Oxford, England)*, 27(17), pp.2414–21.

Loupakis, F. et al., 2008. EGF-receptor targeting with monoclonal antibodies in colorectal carcinomas: rationale for a pharmacogenomic approach. *Pharmacogenomics*, 9, pp.55–69.

Lu, Y. et al., 2007. Epidermal growth factor receptor (EGFR) ubiquitination as a mechanism of acquired resistance escaping treatment by the anti-EGFR monoclonal antibody cetuximab. *Cancer research*, 67(17), pp.8240–7.

Ludwig, W.-D., 2012. [Possibilities and limitations of stratified medicine based on biomarkers and targeted therapies in oncology]. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 106, pp.11–22.

Lynch, H.T. & de la Chapelle, A., 2003. Hereditary colorectal cancer. *The New England journal of medicine*, 348(10), pp.919–32.

Lynch, T.J. et al., 2004. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England journal of medicine*, 350(21), pp.2129–39.

Ma, S.-F. et al., 2005. Bioinformatic identification of novel early stress response genes in rodent models of lung injury. *American journal of physiology. Lung cellular and molecular physiology*, 289(3), pp.L468–L477.

Majewski, I.J. & Bernards, R., 2011. Taming the dragon: genomic biomarkers to individualize the treatment of cancer. *Nature medicine*, 17(3), pp.304–12.

Mal, M. et al., 2012. Metabotyping of human colorectal cancer using two-dimensional gas chromatography mass spectrometry. *Analytical and Bioanalytical Chemistry*, 403, pp.483–493.

Malumbres, M., 2012. miRNAs versus oncogenes: the power of social networking. *Molecular systems biology*, 8, p.569.

Maragkakis, M. et al., 2009. Accurate microRNA target prediction correlates with protein repression levels. *BMC bioinformatics*, 10, p.295.

Markowitz, S.D. et al., 2002. Focus on colon cancer. *Cancer cell*, 1(3), pp.233–6.

Markowitz, S.D. & Bertagnolli, M.M., 2009. Molecular origins of cancer: Molecular basis of colorectal cancer. *The New England journal of medicine*, 361(25), pp.2449–60.

Martin, F. et al., 2012. Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC Systems Biology*, 6, p.54.

Masui, H. et al., 1984. Growth inhibition of human tumor cells in athymic mice by anti-epidermal growth factor receptor monoclonal antibodies. *Cancer Research*, 44, pp.1002–1007.

McDermed, J.E. et al., 2012. Nucleic acid detection immunoassay for prostate-specific antigen based on immuno-PCR methodology. *Clinical chemistry*, 58(4), pp.732–40.

Melo, S.A. et al., 2009. A TARBP2 mutation in human cancer impairs microRNA processing and DICER1 function. *Nature genetics*, 41, pp.365–370.

Mertens-Talcott, S.U. et al., 2007. The oncogenic microRNA-27a targets genes that regulate specificity protein transcription factors and the G2-M checkpoint in MDA-MB-231 breast cancer cells. *Cancer Research*, 67, pp.11001–11011.

Minami, K. et al., 2014. miRNA expression atlas in male rat. *Scientific Data*, 1, pp.1–9.

Moroni, M. et al., 2005. Gene copy number for epidermal growth factor receptor (EGFR) and clinical response to antiEGFR treatment in colorectal cancer: a cohort study. *The lancet oncology*, 6(5), pp.279–86.

Moroni, M. et al., 2005. Somatic mutation of EGFR catalytic domain and treatment with gefitinib in colorectal cancer [4]. *Annals of Oncology*, 16, pp.1848–1849.

Mosakhani, N. et al., 2012. MicroRNA profiling predicts survival in anti-EGFR treated chemorefractory metastatic colorectal cancer patients with wild-type KRAS and BRAF. *Cancer genetics*, 205(11), pp.545–51.

Muzny, D.M. et al., 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), pp.330–337.

National Institutes of Health, 2006. *What You Need To Know About Cancer of the Colon and Rectum*, Bethesda, MD.

Nautiyal, J. et al., 2012. EGFR regulation of colon cancer stem-like cells during aging and in response to the colonic carcinogen dimethylhydrazine. *American journal of physiology. Gastrointestinal and liver physiology*, 302(7), pp.G655–63.

Niault, T.S. & Baccarini, M., 2010. Targets of Raf in tumorigenesis. *Carcinogenesis*, 31, pp.1165–1174.

Niemeyer, C.M., Adler, M. & Wacker, R., 2005. Immuno-PCR: high sensitivity detection of proteins by nucleic acid amplification. *Trends in biotechnology*, 23(4), pp.208–16.

NIH, Colon and Rectal Cancer: Treatment. Available at: http://www.cancer.gov/cancertopics/treatment/colon-and-rectal [Accessed May 5, 2014a].

Nikitin, A. et al., 2003. Pathway studio--the analysis and navigation of molecular networks. *Bioinformatics (Oxford, England)*, 19(16), pp.2155–7.

Niu, C. et al., 1999. Studies on treatment of acute promyelocytic leukemia with arsenic trioxide: remission induction, follow-up, and molecular monitoring in 11 newly diagnosed and 47 relapsed acute promyelocytic leukemia patients. *Blood*, 94(10), pp.3315–24.

Nosho, K. et al., 2008. Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample. *PloS one*, 3(11), p.e3698.

Le Novère, N. et al., 2009. The Systems Biology Graphical Notation. *Nature biotechnology*, 27(8), pp.735–741.

Nyga, A. et al., 2013. A novel tissue engineered three-dimensional in vitro colorectal cancer model. *Acta Biomaterialia*, 9(8), pp.7917–7926.

Obermeier, M., Symons, J. & Wensing, A.M.J., 2012. HIV population genotypic tropism testing and its clinical significance. *Current Opinion in HIV and AIDS*, 7, pp.470–477.

Obulkasim, A. & van de Wiel, M.A., 2015. HCsnip: An R Package for Semi-supervised Snipping of the Hierarchical Clustering Tree. *Cancer informatics*, 14, pp.1–19.

Ongenaert, M. et al., 2008. PubMeth: A cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Research*, 36.

Orchard, S. et al., 2007. The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature biotechnology*, 25(8), pp.894–898.

Orchard, S. et al., 2014. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1).

Ou, S.-H.I. et al., 2012. Crizotinib for the Treatment of ALK-Rearranged Non-Small Cell Lung Cancer: A Success Story to Usher in the Second Decade of Molecular Targeted Therapy in Oncology. *The Oncologist*.

Ou, S.-H.I. et al., 2012. Crizotinib for the treatment of ALK-rearranged non-small cell lung cancer: a success story to usher in the second decade of molecular targeted therapy in oncology. *The oncologist*, 17(11), pp.1351–75.

Ouellet D, Grossmann KF, Limentani G, Nebot N, Lan K, Knowles L, Gordon MS, Sharma S, Infante JR, Lorusso PM, Pande G, Krachey EC, Blackman SC, C.S., 2013. Effects of particle size, food, and capsule shell composition on the oral bioavailability of dabrafenib, a BRAF inhibitor, in patients with BRAF mutation-positive tumors. *J Pharm Sci*, 102(9), pp.3100–9.

P. W. Faber, S. A. Vaziri, L. Wood, C. Nemec, P. Elson, J. A. Garcia, B. I. Rini, R. M. Bukowski, M.K.G. and R.G., 2008. Potential non-synonymous single nucleotide polymorphisms (nsSNPs) associated with toxicity in metastatic clear cell renal cell carcinoma (MCCRCC) patients (pts) treated with sunitinib. *Journal of Clinical Oncology*, 26.

Paez, J.G. et al., 2004. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science (New York, N.Y.)*, 304(5676), pp.1497–500.

Paik, S. et al., 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine*, 351(27), pp.2817–26.

Pammolli, F., Magazzini, L. & Riccaboni, M., 2011. The productivity crisis in pharmaceutical R&D. *Nature reviews. Drug discovery*, 10, pp.428–438.

Pampalakis, G. et al., 2010. Down-regulation of dicer expression in ovarian cancer tissues. *Clinical Biochemistry*, 43, pp.324–327.

Parker, J.L. et al., 2012. Impact of biomarkers on clinical trial risk in breast cancer. *Breast Cancer Research and Treatment*, 136, pp.179–185.

Penaloza, A., Roy, P.-M. & Kline, J., 2012. Risk stratification and treatment strategy of pulmonary embolism. *Current Opinion in Critical Care*, 18, pp.318–325.

Pennings, J.L.A. et al., 2009. Discovery of novel serum biomarkers for prenatal down syndrome screening by integrative data mining. *PLoS ONE*, 4.

Perrone, F. et al., 2009. PI3KCA/PTEN deregulation contributes to impaired responses to cetuximab in metastatic colorectal cancer patients. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, 20(1), pp.84–90.

Personeni, N. et al., 2008. Clinical usefulness of EGFR gene copy number as a predictive marker in colorectal cancer patients treated with cetuximab: a fluorescent in situ hybridization study. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14, pp.5869–5876.

Polakis, P., 2007. The many ways of Wnt in cancer. *Current opinion in genetics & development*, 17(1), pp.45–51.

Popovic, R. et al., 2009. Regulation of mir-196b by MLL and its overexpression by MLL fusions contributes to immortalization. *Blood*, 113, pp.3314–3322.

Prenen, H. et al., 2009. PIK3CA mutations are not a major determinant of resistance to the epidermal growth factor receptor inhibitor cetuximab in metastatic colorectal cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 15(9), pp.3184–8.

Ptolemy, A.S. & Rifai, N., 2010. What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. *Scandinavian journal of clinical and laboratory investigation. Supplementum*, 242, pp.6–14.

Quackenbush, J., 2001. Computational analysis of microarray data. *Nature reviews. Genetics*, 2(6), pp.418–427.

Ragusa, M. et al., 2012. Specific alterations of the microRNA transcriptome and global network structure in colorectal cancer after treatment with MAPK/ERK inhibitors. *Journal of Molecular Medicine*, 90(12), pp.1421–1438.

Rhodes, D.R. et al., 2003. Multiplex biomarker approach for determining risk of prostate-specific antigen-defined recurrence of prostate cancer. *Journal of the National Cancer Institute*, 95, pp.661–668.

Ries L AG, Melbert D, Krapcho M, 2008. *SEER cancer statistics review, 1975–2005*, Bethesda, MD.

Du Rieu, M.C. et al., 2010. MicroRNA-21 is induced early in pancreatic ductal adenocarcinoma precursor lesions. *Clinical chemistry*, 56(4), pp.603–12.

Ringnér, M., 2008. What is principal component analysis? *Nature biotechnology*, 26(3), pp.303–304.

Ritchie, M.E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*.

Rodriguez, A. et al., 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Research*, 14, pp.1902–1910.

Rojas, J. et al., 2009. Development of predictive models of proliferative vitreoretinopathy based on genetic variables: The retina 4 project. *Investigative Ophthalmology and Visual Science*, 50(5), pp.2384–2390.

Roznovăţ, I.A. & Ruskin, H.J., 2013. A computational model for genetic and epigenetic signals in colon cancer. *Interdisciplinary Sciences: Computational Life Sciences*, 5(3), pp.175–186.

Ruan, L. et al., 2011. Analysis of EGFR signaling pathway in nasopharyngeal carcinoma cells by quantitative phosphoproteomics. *Proteome science*, 9, p.35.

Rukov, J.L. et al., 2013. Pharmaco-miR: linking microRNAs and drug effects. *Briefings in bioinformatics*.

Rustgi, A.K., 2007. The genetics of hereditary colon cancer. *Genes & development*, 21(20), pp.2525–38.

Ruzzo, A. et al., 2010a. Molecular predictors of efficacy to anti-EGFR agents in colorectal cancer patients. *Current cancer drug targets*, 10(1), pp.68–79.

Ruzzo, A. et al., 2010b. Molecular predictors of efficacy to anti-EGFR agents in colorectal cancer patients. *Current cancer drug targets*, 10, pp.68–79.

Sabatel, C. et al., 2011. MicroRNA-21 exhibits antiangiogenic function by targeting RhoB expression in endothelial cells. *PloS one*, 6(2), p.e16979.

Sahoo, D., 2012. The power of Boolean implication networks. *Frontiers in Physiology*, 3 JUL.

Sahu, A. et al., 2013. Crizotinib: A comprehensive review. *South Asian journal of cancer*, 2, pp.91–7.

Sakurai, T. & Kudo, M., 2011. Signaling pathways governing tumor angiogenesis. *Oncology*, 81 Suppl 1, pp.24–9.

Saltz, L.B. et al., 2004. Phase II trial of cetuximab in patients with refractory colorectal cancer that expresses the epidermal growth factor receptor. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 22(7), pp.1201–8.

Santarelli, R.L., Pierre, F. & Corpet, D.E., 2008. Processed meat and colorectal cancer: a review of epidemiologic and experimental evidence. *Nutrition and cancer*, 60(2), pp.131–44.

Sartore-Bianchi, A., Di Nicolantonio, F., et al., 2009. Multi-determinants analysis of molecular alterations for predicting clinical benefit to EGFR-targeted monoclonal antibodies in colorectal cancer. *PloS one*, 4(10), p.e7287.

Sartore-Bianchi, A., Martini, M., et al., 2009. PIK3CA mutations in colorectal cancer are associated with clinical resistance to EGFR-targeted monoclonal antibodies. *Cancer research*, 69(5), pp.1851–7.

Sato, J.D. et al., 1983. Biological effects in vitro of monoclonal antibodies to human epidermal growth factor receptors. *Molecular biology & medicine*, 1, pp.511–529.

Sawyers, C., 2004. Targeted cancer therapy. *Nature*, 432(7015), pp.294–7.

Scardoni, G., Petterlini, M. & Laudanna, C., 2009. Analyzing biological network parameters with CentiScaPe. *Bioinformatics (Oxford, England)*, 25(21), pp.2857–9.

Schaefer, C.F. et al., 2009. PID: The pathway interaction database. *Nucleic Acids Research*, 37(SUPPL. 1).

Schou, J. V. et al., 2014. miR-345 in metastatic colorectal cancer: A non-invasive biomarker for clinical outcome in Non-KRAS mutant patients treated with 3rd line cetuximab and irinotecan. *PLoS ONE*, 9(6).

Sebolt-Leopold, J.S. & Herrera, R., 2004. Targeting the mitogen-activated protein kinase cascade to treat cancer. *Nature reviews. Cancer*, 4, pp.937–947.

Segditsas, S. & Tomlinson, I., 2006. Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene*, 25(57), pp.7531–7.

Selventa, 2011. *Reverse Causal Reasoning Methods Whitepaper*,

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., et al., 2003. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), pp.2498–2504.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11), pp.2498–504.

Shmelkov, E. et al., 2011a. Assessing quality and completeness of human transcriptional regulatory pathways on a genome-wide scale. *Biology direct*, 6, p.15.

Shmelkov, E. et al., 2011b. Assessing quality and completeness of human transcriptional regulatory pathways on a genome-wide scale. *Biology direct*, 6, p.15.

Sjöblom, T. et al., 2006. The consensus coding sequences of human breast and colorectal cancers. *Science (New York, N.Y.)*, 314(5797), pp.268–74.

Skibber J, Minsky B, H.P., 2001. Cancer of the colon and rectum. In R. S. DeVita VT Jr, Hellmann S, ed. *Cancer: principles & practice of oncology*. Philadelphia: Lippincott Williams & Wilkins, pp. 1216–1271.

Slamon, D.J. et al., 2001. *Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.*,

Smith, L.P. et al., 2014. SBML and CellML translation in Antimony and JSim. *Bioinformatics*, 30(7), pp.903–907.

Sobin LH, Wittekind C, 2002. *TNM: Classification of Malignant Tumours* 6th ed., New York: Wiley-Liss.

Society, A.C., Treatment of colon cancer by stage. Available at: http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-treating-by-stage-colon [Accessed January 2, 2015].

Sottoriva, A. et al., 2015. A Big Bang model of human colorectal tumor growth. *Nature genetics*, 47(3), pp.209–16.

Stahlhut, C. & Slack, F.J., 2013. MicroRNAs and the cancer phenotype: profiling, signatures and clinical implications. *Genome medicine*, 5, p.111.

Stark, C. et al., 2006. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue), pp.D535–D539.

Stock, C. et al., 2012. Subsite-specific colorectal cancer risk in the colorectal endoscopy era. *Gastrointestinal endoscopy*, 75(3), pp.621–30.

Strömbäck, L. & Lambrix, P., 2005. Representations of molecular pathways: An evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, 21(24), pp.4401–4407.

Sunada, H. et al., 1986. Monoclonal antibody against epidermal growth factor receptor is internalized without stimulating receptor phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, 83, pp.3825–3829.

Szalma, S. et al., 2010. Effective knowledge management in translational medicine. *Journal of translational medicine*, 8, p.68.

Szklarczyk, D. et al., 2014. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1), pp.D447–D452.

Takei, H. et al., 2008. Multicenter phase II trial of neoadjuvant exemestane for postmenopausal patients with hormone receptor-positive, operable breast cancer: Saitama Breast Cancer Clinical Study Group (SBCCSG-03). *Breast Cancer Research and Treatment*, 107, pp.87–94.

Tamayo, P. et al., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6), pp.2907–2912.

Tang, F. et al., 2006. MicroRNA expression profiling of single whole embryonic stem cells. *Nucleic acids research*, 34(2), p.e9.

Tannapfel, A. et al., 2003. Mutations of the BRAF gene in cholangiocarcinoma but not in hepatocellular carcinoma. *Gut*, 52(5), pp.706–12.

Tasneem, A. et al., 2012. The database for aggregate analysis of Clinicaltrials.gov (AACT) and subsequent regrouping by clinical specialty. *PLoS ONE*, 7(3).

Taylor, D.R., 2011. Using biomarkers in the assessment of airways disease. *Journal of Allergy and Clinical Immunology*, 128, pp.927–934.

Tessari, A., Palmieri, D. & Di Cosimo, S., 2013. Overview of diagnostic/targeted treatment combinations in personalized medicine for breast cancer patients. *Pharmacogenomics and Personalized Medicine*, 7.

Tibshirani, R. et al., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), pp.6567–6572.

Tortora, G. et al., 1999. Cooperative inhibitory effect of novel mixed backbone oligonucleotide targeting protein kinase A in combination with docetaxel and anti-epidermal growth factor-receptor antibody on human breast cancer cell growth. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 5, pp.875–881.

Tracey A. Martin, Lin Ye, Andrew J. Sanders, Jane Lane, and W.G.J., 2013. *Cancer Invasion and Metastasis: Molecular and Cellular Perspective* R. Jandial, ed., Landes Bioscience.

Trahey, M. & McCormick, F., 1987. A cytoplasmic protein stimulates normal N-ras p21 GTPase, but does not affect oncogenic mutants. *Science (New York, N.Y.)*, 238, pp.542–545.

Trusheim, M.R. et al., 2011. Quantifying factors for the success of stratified medicine. *Nature Reviews Drug Discovery*, 10, pp.817–833.

Trusheim, M.R., Berndt, E.R. & Douglas, F.L., 2007a. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nature reviews. Drug discovery*, 6(4), pp.287–93.

Trusheim, M.R., Berndt, E.R. & Douglas, F.L., 2007b. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nature reviews. Drug discovery*, 6, pp.287–293.

Tsai, K.-W. et al., 2010. Epigenetic regulation of miR-196b expression in gastric cancer. *Genes, chromosomes & cancer*, 49, pp.969–980.

Tsao, M.-S. et al., 2005. *Erlotinib in lung cancer - molecular and clinical predictors of outcome.*,

Tsong, W.H. et al., 2007. Cigarettes and alcohol in relation to colorectal cancer: the Singapore Chinese Health Study. *British journal of cancer*, 96(5), pp.821–7.

Tsui, I.F.L. et al., 2007. Public databases and software for the pathway analysis of cancer genomes. *Cancer Informatics*, 3, pp.389–407.

Tusher, V.G., Tibshirani, R. & Chu, G., 2001a. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9), pp.5116–21.

Tusher, V.G., Tibshirani, R. & Chu, G., 2001b. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9), pp.5116–21.

Uhlmann, S. et al., 2012. Global microRNA level regulation of EGFR-driven cell-cycle protein network in breast cancer. *Molecular systems biology*, 8, p.570.

US Food and Drug Administration, Table of Pharmacogenomic Biomarkers in Drug Labeling. Available at: http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm [Accessed March 1, 2014].

Vacchi-Suzzi, C. et al., 2013. Heart Structure-Specific Transcriptomic Atlas Reveals Conserved microRNA-mRNA Interactions. *PLoS ONE*, 8.

Vanhoefer, U. et al., 2004. Phase I study of the humanized antiepidermal growth factor receptor monoclonal antibody EMD72000 in patients with advanced solid tumors that express the epidermal growth factor receptor. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 22(1), pp.175–84.

Van ’t Veer, L.J. et al., 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871), pp.530–6.

Vazquez, A. et al., 2008. The genetics of the p53 pathway, apoptosis and cancer therapy. *Nature reviews. Drug discovery*, 7(12), pp.979–87.

Venugopal, S.K. et al., 2010. Liver fibrosis causes downregulation of miRNA-150 and miRNA-194 in hepatic stellate cells, and their overexpression causes decreased stellate cell activation. *American journal of physiology. Gastrointestinal and liver physiology*, 298, pp.G101–G106.

Verducci, J.S. et al., 2006. Microarray analysis of gene expression: considerations in data mining and statistical treatment. *Physiological genomics*, 25(3), pp.355–363.

Vergoulis, T. et al., 2012. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic acids research*, 40(Database issue), pp.D222–9.

Verma, S. et al., 2012. Trastuzumab Emtansine for HER2-Positive Advanced Breast Cancer. *New England Journal of Medicine*, p.121001050017006.

Vogelstein B, K.K., 2002. *The genetic basis of human cancer* 2nd ed., New York: McGraw Hill.

Wadman, M., 2006. Verdict on clinical trials registries? Good, but must do better. *Nature reviews. Drug discovery*, 5, pp.175–176.

Waldman, Y.Y. et al., 2010. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic acids research*, 38(9), pp.2964–74.

Walia, V. et al., 2012. Loss of breast epithelial marker hCLCA2 promotes epithelial-to-mesenchymal transition and indicates higher risk of metastasis. *Oncogene*, 31, pp.2237–2246.

Waltemath, D. et al., 2011. Minimum information about a simulation experiment (MIASE). *PLoS Computational Biology*, 7(4).

Wang, Z. et al., 2004. Identification and utilization of inter-species conserved (ISC) probesets on Affymetrix human GeneChip platforms for the optimization of the assessment of expression patterns in non human primate (NHP) samples. *BMC bioinformatics*, 5, p.165.

Ward, P.S. & Thompson, C.B., 2012. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer cell*, 21(3), pp.297–308.

Webster, R.J. et al., 2009. Regulation of epidermal growth factor receptor signaling in human cancer cells by microRNA-7. *The Journal of biological chemistry*, 284(9), pp.5731–41.

Weisenberger, D.J. et al., 2006. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nature genetics*, 38(7), pp.787–93.

Wheeler, D.L. et al., 2009. Epidermal growth factor receptor cooperates with Src family kinases in acquired resistance to cetuximab. *Cancer biology & therapy*, 8(8), pp.696–703.

Willett, W.C., 2005. Diet and cancer: an evolving picture. *JAMA*, 293(2), pp.233–4.

Witkos, T.M., Koscianska, E. & Krzyzosiak, W.J., 2011. Practical Aspects of microRNA Target Prediction. *Current molecular medicine*, 11(2), pp.93–109.

Wittekind, C. & Neid, M., 2005. Cancer invasion and metastasis. *Oncology*, 69 Suppl 1, pp.14–6.

Wong, Y.F. et al., 2007. Identification of molecular markers and signaling pathway in endometrial cancer in Hong Kong Chinese women by genome-wide gene expression profiling. *Oncogene*, 26, pp.1971–1982.

Wood, L.D. et al., 2007. The genomic landscapes of human breast and colorectal cancers. *Science (New York, N.Y.)*, 318(5853), pp.1108–13.

World Cancer Research Fund, A.I. for C.R., 2007. *Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective*, Washington DC.

Wu, F.-X., Zhang, W.J. & Kusalik, A.J., 2005. Dynamic model-based clustering for time-course gene expression data. *Journal of bioinformatics and computational biology*, 3(4), pp.821–836.

Wu, X. et al., 1995. Apoptosis induced by an anti-epidermal growth factor receptor monoclonal antibody in a human colorectal carcinoma cell line and its delay by insulin. *Journal of Clinical Investigation*, 95, pp.1897–1905.

Wu, X. et al., 1996. Involvement of p27KIP1 in G1 arrest mediated by an anti-epidermal growth factor receptor monoclonal antibody. *Oncogene*, 12, pp.1397–1403.

Xenarios, I. et al., 2000. DIP: the database of interacting proteins. *Nucleic Acids Res*, 28(1), pp.289–291.

Xiao, F. et al., 2009. miRecords: an integrated resource for microRNA-target interactions. *Nucleic acids research*, 37(Database issue), pp.D105–10.

Xu, X.T. et al., 2012. MicroRNA expression profiling identifies miR-328 regulates cancer stem cell-like SP cells in colorectal cancer. *British journal of cancer*, 106(7), pp.1320–30.

Younesi, E. et al., 2012. Mining biomarker information in biomedical literature. *BMC medical informatics and decision making*, 12, p.148.

Yue, D., Liu, H. & Huang, Y., 2009. Survey of Computational Algorithms for MicroRNA Target Prediction. *Current genomics*, 10, pp.478–492.

Zeeberg, B.R. et al., 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome biology*, 4(4), p.R28.

Zeichner, S.B. et al., 2012. A De Novo Germline APC Mutation (3927del5) in a Patient with Familial Adenomatous Polyposis: Case Report and Literature Review. *Clinical Medicine Insights. Oncology*, 6, pp.315–23.

Zheng, H. et al., 2013. Advances in the techniques for the prediction of microRNA targets. *International Journal of Molecular Sciences*, 14, pp.8179–8187.

Zhu, D.-X. et al., 2012. Downregulated Dicer expression predicts poor prognosis in chronic lymphocytic leukemia. *Cancer science*, 103, pp.875–81.

Zhu, H. et al., 2011. EGFR signals downregulate tumor suppressors miR-143 and miR-145 in Western diet-promoted murine colon cancer: role of G1 regulators. *Molecular cancer research : MCR*, 9(7), pp.960–75.

Zisman, A.L. et al., 2006. Associations between the age at diagnosis and location of colorectal cancer and the use of alcohol and tobacco: implications for screening. *Archives of internal medicine*, 166(6), pp.629–34.

AVISEK DEYATI

EDUCATION

- Doctoral Student, Computational Life Sciences, University of Bonn, Germany, Since 2011
- MSc in Bioinformatics, University of Sussex, UK, 2006 - 2008
- B.Tech in Biotechnology, West Bengal University of Technology, India, 2001 - 2005

WORK EXPERIENCES

- R&D Scientist, GlaxoSmithKline Vaccines, Rixensart, Belgium (17.08.2015 to Present)
- Bioinformatics Consultant, GlaxoSmithKline Vaccines and Business & Decision, Rixensart, Belgium (04.2014 – 14.08.2015)
- Doctoral Student, Merck Serono, Darmstadt, Germany (02.2011 to 03.2014)
- Team Leader, Data Base Development, Syntekabio, Seoul, South Korea (04/2010 to 02/2011)

PUBLICATIONS

- Challenges and opportunities for oncology biomarker discovery.
  Avisek Deyati, Erfan Younesi, Martin Hofmann-Apitius, Natalia Novac. Drug Discovery Today, 01.2013
- Biomarker in clinical development: what could we learn from the Past?
  Avisek Deyati, Rama Devi Sanam, Sreenivasa Rao, Guggilla Vijaya Rao Pidugu, Natalia Novac. Personalized Medicine, 08.2014 .
- Systems approach for the selection of micro-RNAs as therapeutic biomarkers of anti-EGFR monoclonal antibody treatment in colorectal cancer. Avisek Deyati, Shweta Bagewadi, Philipp Senger, Martin Hofmann-Apitius, Natalia Novac. Nature Scientific Reports, 27.01.2015.