# Comprehensive analysis of structural genomic alterations in cancer

Computational approaches for identifying cancer driver events

Babak Alaei-Mahabadi

Department of Medical Biochemistry and Cell Biology
Institute of Biomedicine
Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2018

Cover: Catalogue of somatic structural genomic alterations in ten cervical tumors. Human chromosomes (plus HPV) are shown around the outer ring. Inner rings represent different tumors where intra-chromosomal structural alterations are shown. Blue:Deletion, Green:Inversion, Red:Tandem duplication. Light green and orange lines linking two chromosomes show inter-chromosomal rearrangements and viral integrations respectively.

Comprehensive analysis of structural genomic alterations in cancer

Babak.alaeimahabadi@gu.se

ISBN 978-91-629-0422-7 (PRINT)
ISBN 978-91-629-0423-4 (PDF)

Printed in Gothenburg, Sweden 2018
Printed by BrandFactory

- To my father

# Comprehensive analysis of structural genomic alterations in cancer

Babak Alaei-Mahabadi

Department of Medical Biochemistry and Cell biology, Institute of Biomedicine
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

## ABSTRACT

The transformation of a normal cell into a cancer cell involves the accumulation of somatic DNA alterations that confer growth and survival advantages. These genomic alterations can be different in terms of pattern and size, comprising single nucleotide variants (SNVs), small insertions or deletions (indels), structural variations (SVs) or foreign DNA insertions such as viral DNA. Cancer genomes typically harbor numerous such changes, of which only small fractions are driver events that are positively selected for during the evolution of the tumor. High throughput sequencing has enabled systematic mapping of somatic DNA alterations across thousands of tumor genomes. Mutations in particular have been thoroughly explored in this type of data, and this has implicated many new genes in tumor development. However, our knowledge remains more limited when it comes to the contribution of SVs to cancer. In the present thesis, we made use of publicly available cancer genomics data to gain further insight into the role of structural genomic alterations in tumor development.

Viruses cause 10-15% of all human cancers through multiple mechanisms, one of which is structural genomic changes due to viral DNA being integrated into the human genome. Thus, in the first study, we performed an unbiased screen for viral genomic integrations into cancer genomes. We developed a computational pipeline using RNA-Seq data from ~4500 tumors across 19 different cancer types to detect viral integrations. We found that recurrent events typically involved known cancer genes, and were associated with altered gene expression.

SVs can lead to copy number amplification of specific cancer driver genes, as well as the formation of fusion oncogenes, but their importance in cancer beyond these types of events is underexplored. We mapped SVs to the human genome using whole genome sequencing data from 600 tumors across 18

different cancer types and investigated the global relationship between SVs and mRNA changes. We found that such events often contribute to altered gene expression in human tumors, but we were not able to detect novel recurrent driver events. To increase the cohort size, we used a larger but lower resolution and more limited dataset, comprising of microarray based DNA copy number profiles from ~10,000 tumors across 32 cancer types, with the aim of identifying recurrent SV driver events in tumors. Specifically, we investigated SVs predicted to result in promoter substitution events, a known mechanism for gene activation in cancer, and found several recurrent activating events with potential cancer driver roles. Notable among our findings in all the studies were human papillomavirus integrations in *RAD51B* and *ERBB2* and gene fusions involving *NFE2L2, TIAM2* and *SCARB1,* all being known cancer genes.

Taken together, massive amounts of genomic and transcriptomic sequencing data allowed us to comprehensively map viral integrations and structural variations in cancer, which led to the identification of several genes with potential roles in tumor development.

**Keywords**: Somatic structural variations, viral integrations, gene fusions

ISBN 978-91-629-0422-7 (PRINT)
ISBN 978-91-629-0423-4 (PDF)

# SAMMANFATTNING PÅ SVENSKA

Transformationen från en vanlig cell till en cancercell involverar somatiska DNA-förändringar som ger tillväxt- och överlevnadsfördelar. Dessa DNA-förändringar kommer i många olika former, och innefattar typer som SNV (från eng. *single nucleotide variants*), insertioner och deletioner (gemensamt benämnda indels), strukturella variationer (SV), samt insertioner av främmande DNA, såsom viralt DNA. Ett cancergenom bär vanligtvis på många sådana förändringar, men bara ett fåtal av dessa är cancerdrivande och har selekteras fram under tumörens utveckling. High throughput sequencing har möjliggjort systematisk kartläggning av somatiska DNA-förändringar i tusentals tumörgenom. Mutationer har undersökts särskilt noga i denna typ av data, med resultatet att många nya gener har knutits till tumörutveckling. Till skillnad från mutationer så är kunskapen om hur SV bidrar till cancer mer begränsad. I denna avhandling har vi använt oss av publikt tillgängliga cancergenomikdata för att fördjupa vår förståelse av strukturella genomförändringars roll i tumörutveckling.

Virus orsakar 10–15 % av alla cancerfall hos människor genom flera mekanismer, varav en är strukturella genomförändringar orsakade av integrering av viralt DNA i det mänskliga genomet. Därför utförde vi i den första studien en sökning efter integrerat viralt DNA i cancergenom. Vi utvecklade en beräkningspipeline som använder sig av RNA-Seq-data från ~4500 tumörer från 19 olika cancertyper för att detektera virala integrationer. Vi fann att återkommande integrationer vanligtvis involverade kända cancergener, samt var associerade med förändrat genuttryck.

SV kan leda till kopietalsökning av specifika drivande cancergener samt bildning av fusions-onkgener, men utöver detta är det mer oklart i vilken utsträckning de bidrar till cancer. Vi kartlade SV i det mänskliga genomet med hjälp av helgenomsekvenseringsdata från 600 tumörer från 18 olika cancertyper, och undersökte sambandet mellan SV och mRNA-förändringar. Vi fann att SV ofta bidrar till förändrat genuttryck i mänskliga tumörer, men hittade inga nya cancerdrivande förändringar. Med målet att finna nya cancerdrivande SV ökade vi kohortstorleken genom att använda ett större men mer lågupplöst och begränsat dataset, bestående av microarray-baserade DNA-kopietalsprofiler från ~10 000 tumörer från 32 cancertyper. Vi undersökte SV som förväntades orsaka växling av promotorer gener emellan - en känd mekanism för genaktivering i cancer - och hittade flera återkommande aktiverande SV med potentiellt cancerdrivande egenskaper.

Av särskilt intresse i våra resultat var integrationer av humant papillomvirus i *RAD51B* och *ERBB2*, samt fusionsgener som involverar *NFE2L2*, *TIAM2* och *SCARB1* – alla kända cancergener.

Sammantaget möjliggjorde enorma mängder genom- och transkriptsekvenseringsdata en omfattande kartläggning av virala integrationer och strukturella variationer i cancer, vilket resulterade i identifiering av ett flertal gener med potentiella roller i tumörutveckling.

# LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. **The landscape of viral expression and host gene fusion and adaptation in human cancer**
Tang KW, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E.
*Nature Commun. 2013;4:2513*.

II. **Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers**
Alaei-Mahabadi B, Bhadury J, Karlsson JW, Nilsson JA, Larsson E.
*Proc Natl Acad Sci U S A (PNAS). 2016;113(48):13768-13773.*

III. **Systematic investigation of promoter substitutions resulting from somatic intrachromosomal structural alterations in diverse human cancers**
Alaei-Mahabadi B, Larsson E.
*Manuscript*

Papers not included in this thesis:

I. **Limited evidence for evolutionarily conserved targeting of long non-coding RNAs by microRNAs**
Alaei-Mahabadi B, Larsson E.
*Silence. 2013;4(1):4.*

II. **Simultaneous DNA and RNA Mapping of Somatic Mitochondrial Mutations across Diverse Human Cancers**
Stewart JB, Alaei-Mahabadi B, Sabarinathan R, Samuelsson T, Gorodkin J, Gustafsson CM, Larsson E.
*PLoS Genet. 2015;11(6):e1005333.*

III. **Temporal separation of replication and transcription during S-phase progression**
Meryet-Figuiere M, Alaei-Mahabadi B, Ali MM, Mitra S, Subhash S, Pandey GK, Larsson E, Kanduri C.
*Cell Cycle. 2014;13(20):3241-8.*

# CONTENT


**ABBREVIATIONS ..... V**

**1 INTRODUCTION ..... 1**

**1.1 BIOLOGY OF CANCER ..... 1**
1.1.1 PROTO-ONCOGENE AND ONCOGENES ..... 1
1.1.2 TUMOR SUPPRESSORS ..... 3
1.1.3 HALLMARKS OF CANCER ..... 4
1.1.4 CELL SIGNALING AND CANCER ..... 8
1.1.5 TUMOR VIRUSES ..... 8
**1.2 THE CANCER GENOME ..... 11**
1.2.1 POINT MUTATIONS AND INDELS ..... 12
1.2.2 STRUCTURAL VARIATIONS ..... 13
1.2.3 RELEVANCE OF STRUCTURAL VARIATIONS IN CANCER ..... 15
1.2.4 UNDERLYING MOLECULAR MECHANISMS OF SVS ..... 18
**1.3 HIGH THROUGHPUT GENOMIC TECHNOLOGIES ..... 22**
1.3.1 ARRAY-BASED TECHNOLOGIES ..... 22
1.3.2 SEQUENCING TECHNOLOGIES ..... 22
**1.4 COMPUTATIONAL CANCER GENOMICS ..... 24**
1.4.1 OVERVIEW ..... 24
1.4.2 BIOINFORMATICS CHALLENGES ..... 25
1.4.3 APPROACHES TO SV DETECTION ..... 26

**2 AIMS ..... 31**

**3 RESULTS AND DISCUSSION ..... 33**

**3.1 TUMOR-VIRUS ASSOCIATIONS (PAPER I) ..... 33**
**3.2 MAPPING OF SOMATIC SVS ACROSS MULTIPLE CANCER TYPES (PAPERS II, III) ..... 36**
3.2.1 MAPPING SVS USING WGS DATA ..... 36
3.2.2 CNVS AS A SUBSET OF SVS ..... 37
**3.3 IMPACT OF SOMATIC SVS ON TUMOR RNA (PAPERS II, III) ..... 38**
3.3.1 PROMOTER SUBSTITUTIONS ..... 38
3.3.2 ENHANCER HIJACKING ..... 41
3.3.3 FUSION GENES ..... 41

**4 CONCLUSIONS AND FUTURE PERSPECTIVES ..... 45**

**ACKNOWLEDGEMENTS** ..... **47**

**REFERENCES** ..... **49**

**APPENDIX** ..... **67**

# ABBREVIATIONS

| | |
|---|---|
| DNA | Deoxyribonucleic acid |
| RNA | Ribonucleic acid |
| SV | Structural variation |
| CNV | Copy number variation |
| DM | Double Minute |
| HSR | Homogeneously staining region |
| RTK | Receptor tyrosine kinase |
| TS | Tumor suppressor |
| HPV | Human papilloma virus |
| HBV | Hepatitis B virus |
| HCV | Hepatitis C virus |
| EBV | Epstein-Barr virus |
| DSB | Double-strand break |
| NAHR | Non-allelic homologous recombination |
| NHEJ | Non homologous end joining |
| MMEJ | Micro-homology mediated end joining |
| RBM | Replication based mechanism |
| HR | Homologous recombination |
| LCR | Low copy repeat |
| BIR | Break induced replication |
| ddNTP | dideoxynucleotides |
| NGS | Next generation sequencing |
| HTS | High throughput sequencing |
| WGS | Whole genome sequencing |
| WES | Whole exome sequencing |
| PR | Read pair |
| SR | Split read |
| RD | Read depth |
| CA | Contig assembly |

# 1 INTRODUCTION

## 1.1 Biology of Cancer

Normal cell division is a tightly regulated and highly coordinated process. When cells break free of these controls, they can begin to divide uncontrollably resulting in accumulation of cells, which if left to grow continuously, forms a tumor. There are two main classifications of tumors: a benign tumor, which does not attack the neighboring cells or tissues, and malignant tumors that are highly invasive and may eventually spread throughout the body (Weinberg 2007). Benign tumors are rarely life threatening, unless they block a vital access path such as a blood vessel, whereas malignant cells can infiltrate other organs and more readily cause fatal damage. Typically, "cancer cells" would refer to malignant rather than benign cells.

The transformation a normal cell to a cancer cell is an evolutionally process. It includes continuous acquisition of alterations in the cellular DNA of somatic cells and selection acting on alterations that confer fitness advantages to cells (Stratton, Campbell and Futreal 2009). Genomic alterations occur randomly all over the genome, however only a small fraction of alterations become beneficial for tumor growth, typically affecting two groups of genes known as oncogenes and tumor suppressor genes (Yarbro 1992).

### 1.1.1 Proto-oncogene and Oncogenes

Cells contain many proteins that promote cell division. As known from the central dogma of molecular biology, the code for creating these proteins is in the sequences in the cellular DNA called genes (**Box 1**). The normal forms of genes coding for such proteins are called proto-oncogenes. Alterations in these genes may further activate them, stimulating excessive division in the cell. Proto-oncogenes with "gain of function" alterations are called oncogenes (Anderson et al. 1992). They are involved in multiple hallmarks of cancer (see **section 1.1.3**). One of the most frequently activated oncogenes in malignant cells is *TERT*, a telomerase subunit, which plays an important role in cellular immortalization (Heidenreich et al. 2014).

Oncogenes can be categorized into several groups: (1) Growth factors, which can increase the cell proliferative capabilities (Witsch, Sela and Yarden 2010); (2) Receptor tyrosine kinases (RTKs)**,** which are part of many key cell signaling pathways regulating cell proliferation. Several RTK subfamilies are

known for their direct contribution to cancer development, one of which are the epidermal growth factor receptors (EGFRs). Overexpression of genes in this family including *HER1* (also known as *EGFR*) and *HER2* (also known as *ERBB2*) have been seen in wide range of cancers, as a result of both activating mutations and amplifications (Voldborg et al. 1997, McKay et al. 2002, Mitri, Constantine and O'Regan 2012); (3) Transcription factors, which are responsible for the regulation of genes involved in several cellular pathways including proliferation. The ETS factor gene family is one of the largest families of transcription factors that are crucial for tumor development. They are involved in several cellular mechanisms such as cell proliferation, apoptosis, and angiogenesis, all of which are key hallmarks of cancer. ETS factors are sometimes activated in tumors by hijacking the strong promoters of highly expressed genes as a consequence of genomic rearrangements (Ida et al. 1995, Peeters et al. 1997, Tomlins et al. 2005). Additionally, another well established transcription factor involved in cancer is *MYC* that plays a critical role in cell cycle progression, apoptosis and cellular transformation (Dang 2012). Several types of genomic alterations, including point mutations, amplifications, and structural alterations (see **section 1.2**) contribute to the activation of *MYC* in cancer (Finver et al. 1988, Escot et al. 1986, Gabay, Li and Felsher 2014, Affer et al. 2014); (4) GTPases, which play a major role in cell signaling transduction. The *Ras* gene family, which is frequently activated in cancer, is responsible for switching on cell growth independent from growth factors (Goodsell 1999). *Ras* genes including *KRAS*, *NRAS*, and *HRAS* are mainly activated in cancer through point mutations (Fernandez-Medarde and Santos 2011).

**Box 1. Central dogma of molecular biology (Crick 1958)**

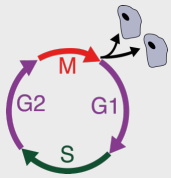
DNA stores all the genetic information in human cells. DNA is made of billions of chemical building blocks called nucleotides. **Replication** is the mechanism of making two near-identical copies of the DNA as cells divide into two daughter cells. Sections of the DNA, known as genes, contain the instructions for making different proteins in the cell. Genes are first synthesized into molecules called RNA through a process called **transcription**. RNA, like DNA, is made of nucleotides. RNAs then are **translated** into proteins that are the fundamental components of the cell. Proteins are made of chains of amino acids, each one determined by a group of three nucleotides in the corresponding gene.

### 1.1.2 Tumor suppressors

Tumor suppressor (TS) genes are defined as genes that inhibit the growth and division of the cell (Friend et al. 1986). They code for proteins whose function is to act as a "brake" in the cell cycle **(Box 2)**. Mutations in these genes may lead to the production of proteins that have lost the "brake" function, allowing the cell to continue to grow. These mutations are known as "loss of function" mutations. Each cell in human body contains two copies of each gene. One functional copy of a tumor suppressor gene is normally enough to regulate cell division. However, once both copies become mutated, the cell cycle brakes no longer work and therefore, the cell can start to proliferate excessively.

**Box 2. Cell cycle**

The cell cycle is a chain of events in the cells resulting in division of one cell into two daughter cells with two identical DNAs copies of its own cellular DNA. The eukaryotic cell cycle has four different phases gap 1, synthesis, gap 2, and mitosis (G1, S, G2, and M respectively). The S phase is when the DNA is duplicated and two newly synthesized cellular DNAs are produced, the M is when the cell physically divides into two daughter cells, and the G1, G2 phases are gap phases involving several checkpoints in which cells ensure that they are ready to enter the S and M phases respectively.




In 1961, Knudson discovered the first TS gene, *RB1*, and proposed the "two-hit" model (Knudson 1971) that was ultimately established in 1986 (Friend et al. 1986). *RB1* is responsible for preventing unnecessary cell growth during cell cycle. Loss of function mutations in *RB1* are associated with tumor growth in many cancer types (Sherr and McCormick 2002). Another predominant tumor suppressor is *TP53*, which is mutated in around 50% of all cancers. *TP53* stops cells with damaged DNA from growing by two key mechanisms, either by halting the cell cycle or by initiating apoptosis (Olivier, Hollstein and Hainaut 2010).

TS genes can be classified into three classes based on the primary function of the proteins they encode: (1) Anti-oncogenes, such as *CDKN2A* and *RB1*, which inhibit the pro-growth activities of oncogenes like *CDK4* and *CCND1* (Serrano, Hannon and Beach 1993); (2) DNA damage checkpoint genes such as *TP53*; (3) Caretaker genes, such as *BRCA1* that help to maintain genomic stability (Yoshida and Miki 2004). Many TS genes have more than one function and could be classified in more than one of the categories mentioned above.

### 1.1.3 Hallmarks of Cancer

Several distinctive biological machineries are accountable for the transformation of a normal cell to a tumor cell. These mechanisms can be summarized into 10 biological hallmarks known as the hallmarks of cancer (Hanahan and Weinberg 2000, Hanahan and Weinberg 2011b) shown in **Fig. 1**. There are six primary hallmarks, two enabling hallmarks and two emerging hallmarks, which will be described in more detail below.

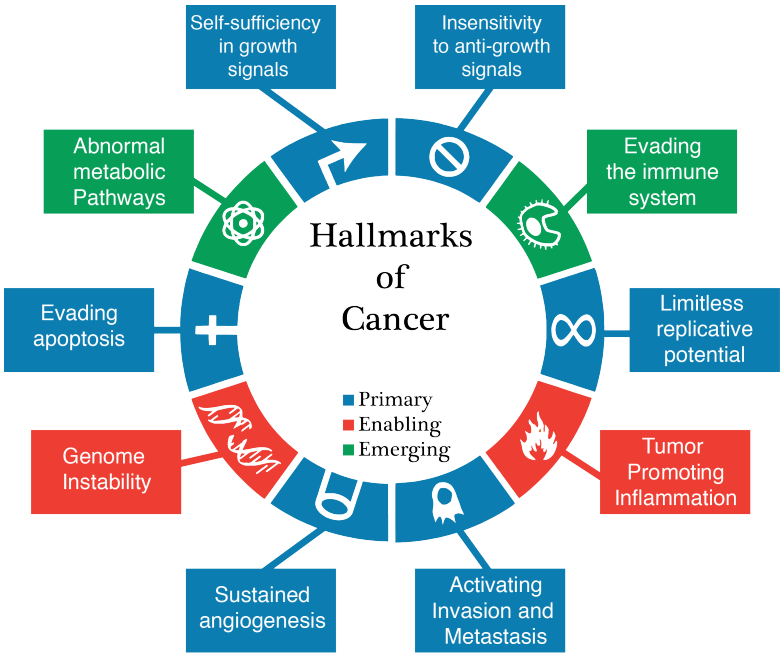



*Figure 1: Hallmarks of cancer. Adapted from (Hanahan and Weinberg 2011b)*

#### Self-sufficiency in Growth Signals

Cell division is controlled by growth factors. Growth factors are proteins in the cell that are responsible for sending signals to other cells to start the

replication process (Paul et al. 1978). They bind to growth factor receptors, which are proteins sitting in the cell membrane. Growth factors activate the receptors by binding to them, leading to a cascade of signals within the cell, signaling that it should divide. The cascade is a series of interactions between numerous proteins in the cell. In a cancer cell, genetic alterations in the genes coding for these receptors can disrupt this highly regulated process. Such alterations can result in the increased activation of a number of genes leading to excessive transcription and increased signaling from the receptors. Alternatively, alterations may result in the formation of new receptors, which activates themselves without the presence of growth factors (Normanno et al. 2006). Growth factor-independent signaling in cancer cells causes uncontrolled cell division and therefore may result in tumor formation.

### Insensitivity to Anti-growth Signals

There are multiple checkpoints at the end of each phase in the cell cycle, where any cell with damaged DNA is detected. Normal cells with defective DNA usually activate the cell death mechanism before they enter mitosis (Cuddihy and O'Connell 2003). TS genes code for the proteins that are responsible for stopping cells with damaged DNA from dividing. In a cancer cell, alterations in TS genes may inactivate these checkpoints, allowing the damaged cells to divide and pass their mutated DNA to their daughter cells.

### Limitless Replicative Potential

Most cells are limited to 40-60 replication cycles (Hayflick 1965). This is regulated through a mechanism called telomere shortening. Telomeres are long repetitive sequences located at the ends of each chromosome which protect the chromosomes from nucleolytic degradation and inter-chromosomal fusions (Witzany 2008). In a normal cell, telomere ends become shorter after each replication cycle, and once it reaches a critical limit, the cell usually undergoes cellular senescence, a mechanism by which cells stop diving (Hayflick and Moorhead 1961). However, by maintaining the length of their telomers, cancer cells can evade the Hayflick limit. This typically happens through the activation of a protein called telomerase (Nosek, Kosa and Tomaska 2006), which adds DNA bases to the telomeres. Telomerase is typically inactive in normal differentiated cells, whereas in cancer cells it may become activated, for example by mutations or SVs (see **section 3.3.2**).

### Evading Apoptosis

The word apoptosis comes from the Greek meaning “falling off”. It is a normal process, in which cells deliberately kill themselves for the good of the

organism (Green 2011). Apoptosis can be triggered by an intrinsic pathway inside the cell like DNA damage, or by extrinsic events such as the lack of nutrients and growth factors outside the cell. Like all the other biological processes, apoptosis involves many proteins with both pro- and anti-apoptotic properties (Silke and Meier 2013).

Cancer cells need to avoid apoptosis to ensure their survival (Fernald and Kurokawa 2013). Genetic alterations in cancer cells not only increase cellular growth, but may also lead to the loss of apoptosis. In some cancer cells, there is a resistance to apoptosis due to activation of anti-apoptotic genes, for example due to mutations in these genes (Yip and Reed 2008). Conversely, deactivating mutations in pro-apoptotic proteins could potentially prevent the cell from entering apoptosis (Lee et al. 2004). The *TP53* gene, known as "the guardian of the genome" plays an important role in detecting DNA damage and signaling to the cell to initiate the repair. Apoptosis is induced in cases where the DNA could not be repaired. *TP53* deactivation through genomic alterations is the most frequent driving event in cancer (Olivier et al. 2010).

### Activating Invasion and Metastasis

Normal cells grow in a well-organized manner where they form tissues and ultimately organs with specific functions. Conversely, malignant cells typically invade the surrounding tissues to find the nutrients they need to survive and sustain their growth. The ability of the cancer cells to break free of their own tissue, enter the blood vessels and reside in another tissue is called metastasis (Gupta and Massague 2006). This is a very complex process, which involves interaction between several proteins. Dysregulation of such proteins by genomic alterations could potentially give cells metastatic capabilities.

### Sustained Angiogenesis

As the tumor grows it needs additional nourishments to maintain its proliferation. Therefore, it requires the recruitment of new blood vessels as they are the main oxygen and nutrients supply. This can be achieved by a mechanism that stimulates the growth of the blood vessels into and around the tissue called angiogenesis (Nishida et al. 2006, Carmeliet and Jain 2000). This machinery is active at specific times in normal conditions such as wound healing. One of the major stimulators of angiogenesis is a protein called vascular endothelial growth factor known as VEGF (Leung et al. 1989). Overexpression of this protein, acquired by genomic alterations, is one of the mechanisms activating angiogenesis.

## Genome Instability

The six characteristics mentioned so far, known as the “primary hallmarks” of cancer, allow cancer cells to survive, proliferate and transfer irregularly within the body. The mechanisms that allow cancer cells to acquire these primary hallmarks are known as “enabling hallmarks". One of these enabling characteristics is genomic instability, which results in a large number of genomic alterations. These alterations could become beneficial for tumors by orchestrating the primary hallmarks of cancer (Negrini, Gorgoulis and Halazonetis 2010).

Mutations in genes involved in the DNA maintenance machinery, recognized as caretakers, have often been observed in context of cancer. These caretaker genes are involved in several mechanisms, one of which is to detect DNA damage and activate the repair mechanism. Inactivating mutations in these genes are associated with increased genomic instability and therefore play an important role in cancer progression (Barnes and Lindahl 2004, Korkola and Gray 2010).

## Tumor Promoting Inflammation

Another enabling hallmark of cancer is tumor-promoting inflammation. Inflammation is a complex biological response triggered in the presence of harmful stimuli. There are two types of inflammation: acute and chronic. While acute inflammation is typically protective, chronic inflammation caused by the continuous persistence of an infectious agent is associated with cancer development. Chronic inflammation can contribute to cancer progression by affecting multiple hallmarks of cancer. These include providing growth factor to sustain proliferative capabilities, pro-angiogenesis enzymes (Grivennikov, Greten and Karin 2010), and inducing cellular stress that can damage the DNA (Visconti and Grieco 2009).

## Evading the Immune System

Progress in cancer research in the last decades has added two emerging hallmarks to the list of general characteristics of cancer cells, which are beneficial to the growth, and survival of the cancer. The immune system has many mechanisms that prevent it from attacking self cells. However, the immune response to cancer cells involves self-attacking cells with abnormal metabolism and growth, a mechanism called cancer immunoediting (Dunn et al. 2002). Immunoediting encompasses three phases: elimination, equilibrium and escape (Dunn, Old and Schreiber 2004). Elimination refers to the first phase also known as immune surveillance, during which the tumor cells, by releasing tumor associated antigen, can be recognized by immune system and

are therefore eradicated. In the equilibrium phase, cells with continuous DNA alterations, eventually acquire a non-immunogenic phenotype and are positively selected for during the evolution of the tumor. Finally, during the escape phase, those cells that survived the elimination and equilibrium phases grow uncontrollably leading to the formation of noticeable tumors.

### Abnormal Metabolic Pathways

Energy and nutrients are necessary for cells to grow. Cancer cells grow uncontrollably and to sustain their proliferation capacity, they need an increased uptake of nutrients such as glucose, which can be achieved by adjusting their metabolism (Lunt and Vander Heiden 2011). Cancer cells, unlike most normal cells, tend to metabolize glucose and produce energy through biochemical pathways that do not involve oxygen even when it is available. This phenomenon is known as Warburg effect (Warburg, Wind and Negelein 1927). While this is an inefficient metabolic pathway, malignant cells typically produce ATP that is the primary energy carriers, up to 100 times faster than healthy cells.

## 1.1.4 Cell signaling and cancer

Cell signaling is a part of a complex communication process that manages basic cellular activities. Three stages are involved in cell signaling: reception, transduction and response. Reception is when the cell recognizes the signaling molecule through proteins called receptors. Transduction is when the receptor protein transmits the signal further through a series of molecular events, thereby initiating a cellular response, and response is when different cellular activities are triggered such as cell growth, expression, cell death and so on. Errors in signaling pathways may result in diseases such as cancer.

All the hallmarks of cancer discussed above arise as modifications in several signaling pathways that are responsible for the regulation of diverse cellular activities in normal cells (Martin 2003). One of the key protein families involved in such cellular processes including cell growth is receptor tyrosine kinases (RTKs). Abnormal signaling by RTKs, such as *EGFR* and *HER2*, (see **section 1.1.2**) have been shown to be critically involved in cancer progression (Zwick, Bange and Ullrich 2001).

## 1.1.5 Tumor viruses

Rous sarcoma virus (RSV) was the first virus discovered to be associated with cancer development. In 1911, Peyton Rous injected a cell free extract of a chicken sarcoma tumor into healthy chickens and observed that they

developed tumors (Rous 1911). He concluded that the carcinogenic agent passed on to the healthy chickens might have been a virus, which was later established and named RSV. Since the discovery of RSV, seven types of viruses have been found to be responsible for 10-15% of all human cancers (**Table 1**). Viral oncogenicity involves multiple mechanisms. Direct mechanisms include expression of viral oncogenes (EVO) and integration of viral DNA (IVD) into the host DNA by which they either facilitate the expression of their own oncogenes or promote the expression of already existing proto-oncogenes in the host DNA. Additionally, viruses can induce chronic inflammation (ICI) sometimes even after decades of acute infection indirectly enabling tumor growth.

*Table 1: Human tumor associated viruses. DS: double strand. SS: single strand. C: circular. L: linear*

| Virus | Cancer type | Mechanism | Virus Type |
|---|---|---|---|
| **HBV** | Hepatocellular | ICI, IVD, EVO | DS C DNA |
| **HCV** | Hepatocellular | ICI | SS L RNA |
| **EBV (HHV4)** | Subset of lymphomas | EVO | DS L DNA |
| **HPV** | Cervical, Oral cavity | IVD, EVO | DS C DNA |
| **HTLV-1** | T-cell leukemia | ICI, IVD | SS L RNA |
| **KSHV (HHV8)** | Sarcoma, lymphoma | EVO | DS C DNA |
| **MCV** | Merkel cell | IVD | DS C DNA |

DNA viruses store their genetic materials in DNA. 5/7 reported human tumor viruses are DNA viruses. They typically utilize the cellular replication mechanism to ensure their own replication occurs (Munger et al. 2004). Human papilloma virus (HPV), a DNA virus typically infecting the genital tissues, promotes carcinogenesis by integrating its DNA into the human genome leading to the expression of the E6 and E7 viral oncogenes that inactivate the *TP53* and *RB* tumor suppressor genes respectively (zur Hausen 2002). Additionally, *MYC* activation is shown to be associated with HPV viral integration into the cellular DNA particularly in the *MYC* region (Peter et al. 2006). Another DNA tumor virus is Epstein-Barr virus (EBV), which normally infects B-cells of the immune system. It encodes three viral oncogenes, *LMP1, EBNA-2* and *EBNA-3C* that are essential for B-cell growth, transformation and the disruption of cellular signaling pathways (Arvanitakis, Yaseen and Sharma 1995). Similarly, Human herpes virus 8 (KSHV or HHV8) encodes many oncogenic viral homologues to host proteins, which can potentially drive cell proliferation, immune evasion and

angiogenesis (Bais et al. 1998, Yang et al. 2000). The two latter viruses, unlike HPV, do not integrate into the human genome, but instead they are maintained as circular episomes that replicate independently from the host cellular chromosomes.

Hepatitis B and C viruses (HBV, HCV) usually cause hepatocellular liver cancer by inducing chronic inflammation in liver cells leading to cirrhosis (Ganem and Prince 2004, Colombo et al. 1989). Cirrhosis is a condition in which the liver cells are damaged and scared and can no longer function properly. Additionally, HBV expresses X antigen (HBx), which promotes cell proliferation, and integrates its DNA into the host genome, inducing proto-oncogene activation and chromosomal instability (Sung et al. 2012). HCV, unlike HBV, is a single stranded RNA virus that uses RNA instead of DNA to store its genetic material.

Human T-lymphotropic virus (HTLV-1) is a retrovirus responsible for a subset of adult T-cell leukemias and lymphomas (Boxus and Willems 2009). Retroviruses are a subset of RNA viruses that use their own reverse transcriptase enzyme to synthesis double strand DNA from RNA and then integrate into the host cell using another enzyme called integrase.

## 1.2 The Cancer Genome

Over a century ago, Theodor Boveri hypothesized that chromosomal aberrations may be the underlying factor driving cancer (Boveri 1914). Following the discovery of DNA as the genetic material (Avery, Macleod and McCarty 1944) and its structure (Watson and Crick 1953), it was shown that alterations in DNA could potentially be the driving force behind cancer development. The first evidence of a cancer driver alteration was found nearly 50 years after Boveri's hypothesis with the identification of "The Philadelphia Chromosome", as a translocation between chromosome 9 and 22 in leukemia tumors (Rowley 1973, Nowell 1962). A new protein with oncogenic properties was produced fusing two genes, *BCR* and *ABL*, as a result of this chromosomal rearrangement (**Fig. 2**). Subsequently it was shown that the activation of the *h-ras* oncogene was associated with a point mutation (Reddy et al. 1982). These discoveries led to further investigation of cancer-associated genomic alterations, which were functionally important for the development of the tumor.

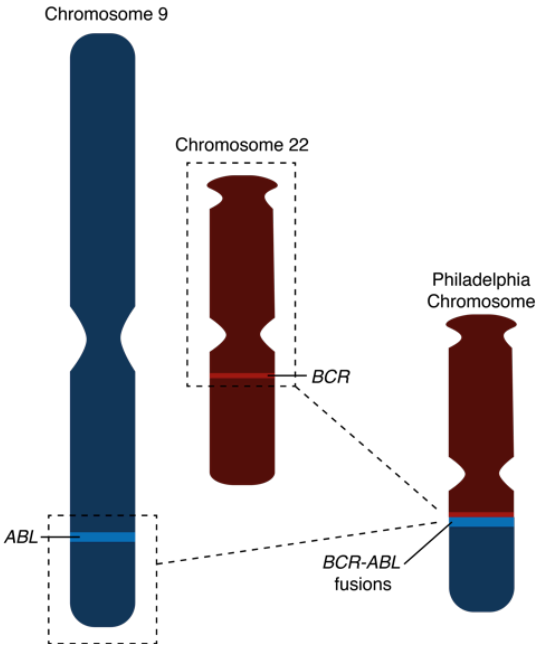



*Figure 2: The Philadelphia chromosome*

DNA alterations occur frequently in the human body, where most are repaired through a mechanism called DNA repair. However, a small fraction of these alterations avoid being repaired, and some of them will give the cell

certain characteristics outlined previously (see **section 1.1.3**) as the hallmarks of cancer (Hanahan and Weinberg 2011a). These alterations that are beneficial for the transformation of the normal cell to the tumor cell are called "driver events" whereas all the other random alterations in the cellular DNA are likely to be "passengers" for the cancer development.

Somatic alterations in cancer genomes can be divided into several distinct classes in terms of size and type. These include point mutations, insertions or deletions of small DNA segments (indels), structural variations (SVs), and insertions of non-endogenous sequences such as viral DNA **(Fig. 3)**.

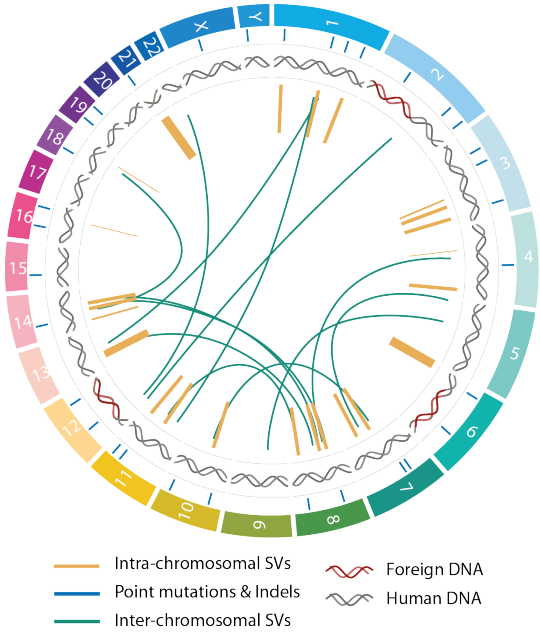



*Figure 3: Catalogue of somatic alterations in the cancer genome*

## 1.2.1 Point mutations and Indels

The human genome is made up of billions of pairs of nucleotides. Point mutations are defined as the substitution of one base pair for another. Additionally, indels are defined as small insertions and deletions in the cellular DNA. Although point mutations in coding genes result in altered DNA sequences, they don't necessarily change the resulting amino acid sequences of the proteins, as multiple nucleotide sequences code for the same amino acid. To date, several driver somatic mutations are known to be associated with multiple different cancer types **(Table 2),** sometimes

affecting as much as 80% of tumors in a given cancer type (Rubio-Perez et al. 2015, Gonzalez-Perez et al. 2013).

*Table 2: The most recurrently mutated **cancer driver genes.** Genes with a gain of function mutation (oncogenes) are shown in red whereas the loss of function mutated genes (tumor suppressors) are in blue*

| Symbol | %Mutated (Cancers) | %Mutated in all cancers |
|---|---|---|
| TP53 | > 80 (Ovarian, Lung) | > 30 |
| PIK3CA | > 50 (Uterine) | > 10 |
| KRAS | > 45 (pancreas, Colorectal) | > 5 |
| BRAF | > 50 (Thyroid, Melanoma) | > 5 |
| PTEN | > 60 (Uterine) | > 5 |
| MLL3 | > 20 (Bladder) | > 5 |
| APC | > 75 (Colorectal) | > 4 |
| MLL2 | > 20 (Bladder, Lung) | > 4 |
| ARID1A | > 25 (Uterine, Bladder) | > 4 |
| NF1 | > 10 (Lung, Melanoma) | > 4 |

### 1.2.2 Structural Variations

Structural variations (SVs) are defined as alterations in chromosomal DNA typically larger than 1 Kb. Structural variations consist of copy number imbalance events such as deletions and duplications, inversions, interchromosomal translocations **(Fig. 4)**, transposon insertions, or foreign DNA insertions such as viral DNA (Feuk, Carson and Scherer 2006). Traditionally, the two later are not classified as SVs even though by definition they are variations in the chromosomal structure.

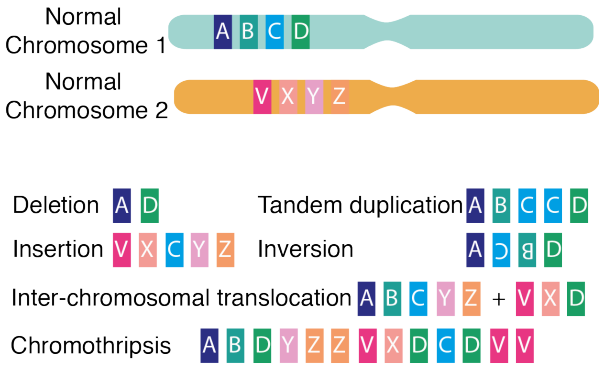



*Figure 4: Different Types of SVs*

## Deletions and duplications

Deletions and duplications are two classes of structural variations that are copy number unbalanced. Deletions result in the loss of a genomic region whereas an extra copy of a DNA segment is added to the genome through duplication (Feuk et al. 2006). Duplications typically happen in two forms: (1) Through DNA insertions, in which one fragment of DNA is duplicated and inserted into another genome region, as a result of both inter or intra chromosomal translocation and (2) through tandem duplications, by which a DNA fragment is placed adjacent to itself (McBride et al. 2012). While deletions and duplications contribute to cancer development mainly by altering copy number of oncogenes and tumor suppressors leading to their deregulation, they may also cause gene fusions with novel properties that are potentially important in cancer. The most common case of such events is a deletion in chromosome 17 causing the activation of the *ERG* oncogene through fusion with the *TMPRSS2* gene (Linn et al. 2016).

## Inversions

Not all SVs lead to DNA copy number alterations (Feuk et al. 2006). Inversions are copy number neutral rearrangements in which a segment of DNA is reversed end to end within the same chromosome. Inversions will usually not influence the genes within the boundaries of the inverted region. However, the genes that span the DNA break junctions might be deregulated through, for example, the creation of gene fusions. Recurrent inversion events involving the *RET* oncogene, a RTK, in thyroid cancer has previously been reported as a mechanism to activate this gene (Cinti et al. 2000). Due to the complex nature of these events, not being detectable by CNV detection approaches, many potentially important events in cancer are still yet to be found.

## Inter-chromosomal translocations

As discussed above, the first genomic alteration known to be functional in cancer was an inter-chromosomal translocation leading to a *BCR-ABL* oncogenic gene fusion (Nowell 1962). Inter-chromosomal translocation is a type of rearrangement where two chromosomes break and are fused with each other. They can be reciprocal, in which they exchange the broken segments and therefore no genomic region is lost or gained, or non-reciprocal where only one of the broken segments of DNA in the two chromosomes are fused. The result is a DNA copy number loss of the other broken ends of the two chromosomes (Rabbitts 1994). Soon after the discovery of the Philadelphia chromosome, several other inter-translocations were found to be associated with the development of various tumor types (Zech et al. 1976,

Oshimura, Freeman and Sandberg 1977, Rowley, Golomb and Dougherty 1977, Fukuhara et al. 1979).

### Viral Integrations

As discussed in **section 1.1.5**, one of the ways that viruses cause cancer is by integrating their own DNA into human DNA. HPV and HBV, two big classes of oncoviruses, frequently integrate their genome into the human cellular DNA. These integrations may lead to the activation of proto-oncogenes such as *MYC* and *TERT* (Ferber et al. 2003), as well as the expression of viral oncogenes including E6 and E7 in HPV (Finzer, Aguilar-Lemarroy and Rosl 2002).

### Chromothripsis and Chromoplexy

All SVs mentioned so far were considered to be simple SVs, corresponding to one rearrangement in one single event. Chromothripsis on the other hand, is a phenomenon whereby a cluster of SVs occurs in a single catastrophic event, resulting in highly rearranged chromosomal region (Stephens et al. 2011). The initial observation was made in myeloid leukemia (Stephens et al. 2011), but additional chromothripsis cases have been reported in almost all cancer types (Rode et al. 2016). Additionally, chromothripsis has been linked to poor prognosis, indicating that it may play an important role in tumorigenisis (Rode et al. 2016). A relevant phenomenon is chromoplexy where random broken chromosome fragments rejoin and result in a balanced chain of rearrangements (Shen 2013).

## 1.2.3 Relevance of structural variations in cancer

Somatic SVs may result in amplifications, deletions or rearrangements of genomic features such as genes and regulatory elements, all of which could alter gene expression and therefore contribute to cancer progression. Chromosomal translocations can promote tumor growth through multiple mechanisms (1) Creation of novel fusion genes with oncogenic properties; (2) Rearrangements of regulatory elements such as gene promoters and enhancers leading to abnormal expression of normal cellular genes such as proto-oncogenes; (3) Silencing tumor suppressor genes by inducing a premature stop codon **(Fig. 5)**.

### Copy number alterations

SVs often result in copy number variations (CNVs). In cancer genomes, tumor suppressor genes are often lost while oncogenes are often copy number amplified, sometimes even as much as 1000-fold. High DNA copy number amplifications in cancer normally occur in the form of double minute (DM)

chromosomes or intra-chromosomal homogeneously staining regions (HSR) (Storlazzi et al. 2010). DMs are small fragments of chromosomal DNA forming a small circular extrachromosome with no centromere or telomere. DMs are not distributed evenly into the daughter cell after mitosis; whereas HSR are chromosomal segments that are duplicated many times in a normal chromosome and are replicated like the rest of the chromosomal DNA. Both typically contain oncogenes that give a selective advantage to the development of the tumor. Three frequently amplified oncogenes in cancer, *MYC*, *EGFR* and *ERBB2*, are often amplified through the creation of DMs and HSRs in various cancer types (Savelyeva and Schwab 2001, Vogt et al. 2004, Vicario et al. 2015). While tandem duplications and insertions also lead to an altered copy number, they are only limited to a one copy increase of the amplified DNA.

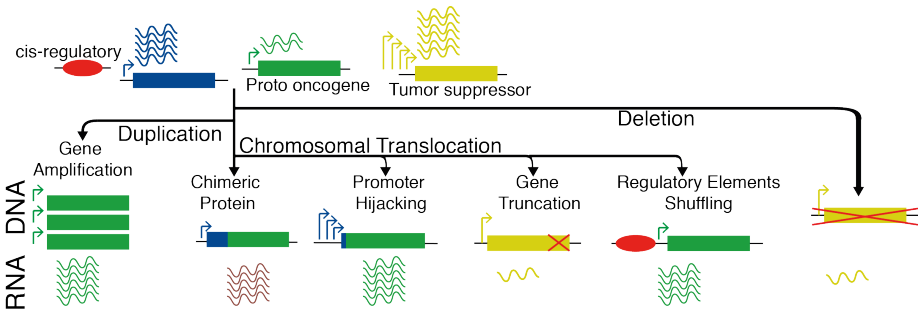



*Figure 5: Mechanisms by which SVs contribute to cancer development. Wavy lines represent the amount of mRNA.*

## Transcriptional deregulation

The relocation of regulatory elements, particularly promoters, to the vicinity of proto-oncogenes is known to be a driving mechanism in cancer. One example is "promoter insertion", observed for the first time in 1981 (Hayward, Neel and Astrin 1981, Neel et al. 1981), where the activation of cellular proto-oncogenes was caused by the insertion of a strong promoter **(Box 3)** from viral DNA into their proximity. A related effect is when a genomic translocation leads to a strong promoter in the genome being juxtaposed to a weaker promoter of a proto-oncogene (Leder et al. 1983, Grimaldi and Meeker 1989, Erikson et al. 1986). These events are typically a result of the creation of gene fusions in which the 3' partner is dysregulated by hijacking the 5' fusion partner promoter (Mertens et al. 2015). The most frequent example of upregulation of the 3' partner by such "promoter substitution" is the fusion of several ETS factor proto-oncogenes with

*TMPRSS2* which occurs in more than 50% of prostate tumors leading to strong transcriptional activation of the ETS genes (Tomlins et al. 2005).

Recent studies have shown that enhancers **(Box 3)**, as another class of regulatory elements, could in fact have the same consequence in cancer genomes (Northcott et al. 2014). This was recently observed in medulloblastoma tumors where the activation of several members of the *GFI1* oncogene family was associated with the juxtaposition of enhancers to these genes (Northcott et al. 2014). Both inter-chromosomal and intra-chromosomal translocations have been shown to contribute to this mechanism (Groschel et al. 2014, Weischenfeldt et al. 2017).

**Box 3. Regulatory elements (Maston, Evans and Green 2006)**

Regulatory elements (REs) are non-coding regions of DNA, which play an important role in regulating the transcription process. REs are typically upstream of transcription start sites. They include promoters, activators and enhancer sequences, all of which promote the expression of genes, as well as silencer sequences that inhibit expression.

Promoters are short sequences located near the transcription start site indicating where the transcription of the genes should start. RNA polymerase binds to this region and initiates the transcription. Enhancers are distal regulatory elements that, by binding to proteins called transcription factors, can boost the transcription of a specific gene. Enhancers can be located thousands of base pairs away from the promoters but, since the DNA is folded and coiled, end up adjacent to the promoter in the folded state.

## Chimeric genes

As mentioned in the previous section, gene fusions may result in the upregulation of proto-oncogenes; in fact, it was initially believed that the functional outcome of the Philadelphia chromosome was the activation of the *ABL1* gene acting as an oncogene through swapping its promoter with *BCR* gene. However, it was later shown that the result of this translocation was a new chimeric gene, which coded a hybrid protein with abnormal oncogenic activity (Shtivelman et al. 1985, Stam et al. 1985). Creation of an oncogenic fusion protein through the joining of two genes that originally coded for different proteins is now a well-known mechanism for the development of cancer (Sorensen and Triche 1996, Mertens et al. 2015).

### Gene truncation

Point mutations and copy number losses have been discussed as two ways that a TS gene can be deactivated. Another mechanism by which a gene can become silenced is to manipulate the structure of the gene by introducing SV breakpoints leading to the creation of a dysfunctional truncated protein. Deactivation of several tumor suppressor genes such as *CDKN2A* and *NF1* were shown to be through this mechanism (Duro et al. 1996, Storlazzi et al. 2005). In some cases the structural breakpoint results in the creation of a fusion gene, but typically the resulting protein has either a frame shift in the reading frame, known as an out-of-frame fusion, or a premature stop codon in the novel fusion transcript, both resulting in a dysfunctional protein (Cancer Genome Atlas Research et al. 2013).

## 1.2.4 Underlying molecular mechanisms of SVs

Cellular DNA gets damaged at least 10,000 times per day in a given cell (De Bont and van Larebeke 2004). These errors include nucleotide damage, nucleotide mismatches, and single and double strand breaks. While most of these damages gets fixed through multiple mechanisms called DNA repair, a small fraction of them, due to imperfect repair, cause mutations and genomic rearrangements in the genome. Double strand breaks (DSBs) in particular are harmful for the cells since they can lead to creation of SVs. Four major mechanisms involving DSB repair may cause SVs in the genome: non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), microhomology mediated end joining (MMEJ), and replication based mechanisms (RBMs).

### Non-allelic Homologous Recombination

Homologous recombination (HR) is a mechanism by which two highly similar chromosomes or DNA fragments are exchanged (Capecchi 1989). HR based DSB repair involves two identical alleles in which the homologous region in one allele is used to repair the same region of the broken DNA of the other allele. *BRCA1* and *BRCA2*, two well-known tumor suppressors, are required for the HR mediated DSB repair (Venkitaraman 2002) and therefore the loss of function mutations in these genes leads to genomic instability and eventually cancer.

Most recurrent SVs here defined as SVs sharing the same exact genomic interval and content, are caused by non-allelic homologous recombination (NAHR) (Gu, Zhang and Lupski 2008, Liu et al. 2012). NAHR is a type of homologous recombination that connects two highly similar fragments of DNA in one allele known as low copy repeats (LCRs) (Shaw and Lupski 2004) resulting in chromosomal rearrangements. Recurrent SVs are flanked by LCRs, which is typically indicative of high homology at the breakpoints. Depending on the location and orientation of the LCRs, different types of SVs can be introduced in the genome. Recombination between the directly oriented LCRs on the same chromosome leads to deletions or duplications, whereas inversions happen when two LCRs are on the same chromosome but in the opposite direction (Lupski 1998). Additionally, LCRs on different chromosomes lead to chromosomal translocations **(Fig. 6).**

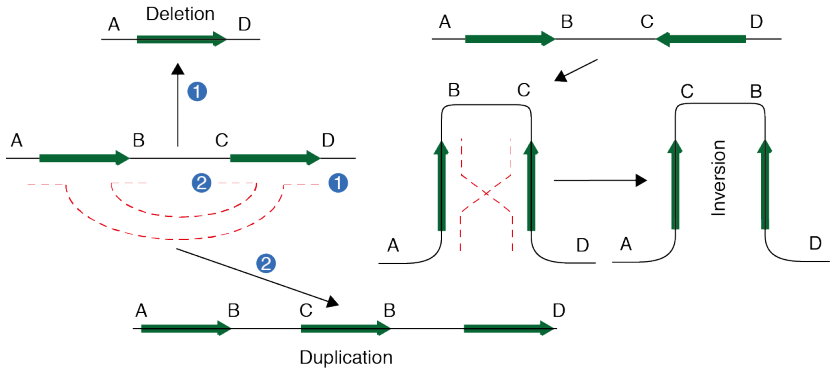



*Figure 6: Intrachromosomal SVs resulted from NAHR. LCRs are shown with green arrow where the orientation is shown by the arrowhead.*

## Non-homologous end joining

Non-recurrent SVs often have microhomology or small insertions or deletions at the breakpoint junctions, which is in contrast with the main characteristic of recurrent SVs having extensive homologous sequences (up to 10kb) at their breakpoint (Ottaviani, LeCain and Sheer 2014, Carvalho and Lupski 2016).

Non-homologous end joining (NHEJ) is one of the mechanisms used for DSB repair, which may result in the creation of non-recurrent SVs (Gu et al. 2008). In contrast to HR repair, the two broken ends of the DNA are joined without relying on a homologous sequence as a template (Moore and Haber 1996). DSBs typically result in a single stranded DNA overhang on one side of the double strand DNA. Incompatible overhang sequences are modified at the broken DNA ends, normally causing small deletions or insertions (1-4 bp)

at the joint region **(Fig. 7)** (Lieber 2008). Finally the two broken DNA strands are joined together using a ligase enzyme.

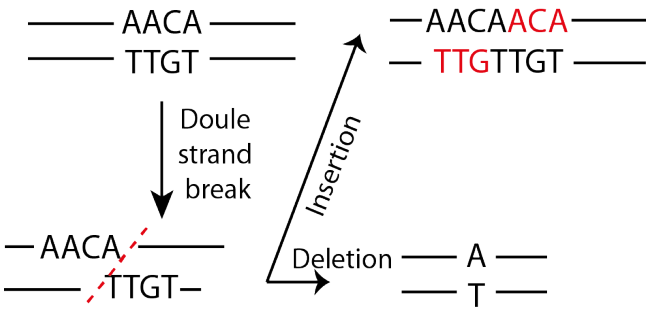



*Figure 7: NHEJ may result in small insertion and deletions*

## Micro-homology mediated end joining

In the absence of NHEJ mechanism in the cell, a more error-prone pathway, known as micro-homology mediated end joining (MMEJ) is used to repair the induced DSB (McVey and Lee 2008). When a DSB occurs in the cell, MMEJ uses 5-25 bp homologous sequences to align two broken DNA fragments, therefore a deletion of the same size is introduced at the original break site.

## Replication-based mechanisms

More complex SVs, defined as series of rearrangements which occur in a single catastrophic event, cannot be explained by neither NAHR nor NHEJ, but replication-based mechanisms are able to explain such events. Break induced replication (BIR), is one of these mechanisms that significantly contribute to the formation of SVs (Carvalho and Lupski 2016). It is a homologous recombination pathway used to repair DSB with only a single end, as it happens during the DNA synthesis. Template switching is the main mechanism in BIR, where the broken chromosome end invades another homologous template and resumes the replication until the next replication fork or the end of the chromosome. Defects in this machinery will essentially give rise to the creation of different SVs in the genome. Additionally, complex SVs can be caused by BIR, given the fact that multiple strand invasions can occur during replication (Lee, Carvalho and Lupski 2007, Smith, Llorente and Symington 2007, Tsaponina and Haber 2014).

## Breakage-fusion-bridge

Telomeres are located at the two ends of the chromosomes and are responsible for protecting them from degradation or fusion (McEachern, Krauskopf and Blackburn 2000). In normal cells, once the telomeres become abnormally too short or broken, apoptosis is triggered by p53. However in

absence of p53 signaling, the two ends of the unprotected chromosomes may fuse together causing genomic instability (Bailey and Murnane 2006). The resulting fused chromosomes would have two centromeres and therefore, as they are pulled in opposite directions during cell division, the chromosome breaks at a random position. This leads to translocations and new uncapped chromosome ends. As this process repeats, amplifications and rearrangements are formed that accelerate tumor growth (DePinho and Polyak 2004).

## 1.3 High Throughput genomic Technologies

### 1.3.1 Array-based Technologies

High throughput microarray-based technologies have revolutionized the genetics field (Heller 2002). DNA hybridization, which is the property of two complementary DNA strands from different sources binding together, is the main principle behind these methods. An array chip is a solid surface in which hundreds of single strand DNA probes are spotted. Each spot corresponds to a specific DNA fragment and contains millions of copies. A fluorescently labeled DNA sample is added to the surface. Different DNA fragments in the sample bind to the relevant complementary DNA probes, leading to the formation of hybridized double strand DNA molecules. Special scanners are then used to quantify the amount of DNA as a measure of light emitted from the fluorescently labeled DNA molecules. Array-based methods, depending on the probe types, can have distinct applications of which the most common application is gene expression profiling (Schena et al. 1995).

Arrays can also be applied to detect CNVs through a technique called array comparative genomic hybridization (aCGH) (Solinas-Toldo et al. 1997). CGH is a method to quantify CNVs as a measure of DNA content differences in a test sample versus a control sample. The intensity signal from the differentially labeled test and control samples can be used to identify unbalanced chromosomal regions (Kallioniemi et al. 1992). However, CGH methods are only capable of identifying big CNVs (> 5 Mb) in the genome. Array-based CGH uses the same principle as traditional CGH methods, but uses diversely located cloned DNA fragments across the genome (Shaw-Smith et al. 2004), which lead to the identification of CNVs at higher resolutions (> 10 Kb). Multiple aCGH platforms have been developed one of which is the Affymetrix “SNP6 array”, with 1 million probes each representing a unique position in the genome.

### 1.3.2 Sequencing Technologies

Sequencing is the process of digitally reading the exact order of nucleotides in DNA or RNA molecules. The first widely used sequencing approach, Sanger sequencing, was developed in 1977 (Sanger, Nicklen and Coulson 1977). It is based on a method called “chain termination” in which chain-terminating dideoxynucleotides (ddNTPs) are used to color code each newly synthesized DNA with variable length generated from the initial DNA molecule. The resulting DNA molecules are then separated by their size and

read based on their color representing different nucleotides. This technology enabled determination of DNA sequences from any organism, and therefore was widely adapted by scientists around the world. However, it is limited in regards to speed and scalability, which forced the development of larger scale sequencing technologies, later known as next generation sequencing (NGS) (Brenner et al. 2000). While NGS allows the sequencing of thousands of DNA or RNA molecules simultaneously, Sanger methods are still being used as the "golden standard" particularly for the validation of NGS data.

NGS, also known as high throughput sequencing (HTS), utilizes massively parallel approaches to sequence up to millions of bases at the same time. In contrast with Sanger sequencing in which only one read is generated for a typically large DNA fragment, several reads are generated simultaneously by HTS from millions of overlapping small DNA pieces. Nowadays, HTS is widely used to sequence the whole genome (WGS), whole exome (WES), transcriptome (RNA-Seq), DNA protein interaction (Chip-Seq) and targeted genomes (de Magalhaes, Finch and Janssens 2010). The most commonly used HTS principle is Illumina/Solexa sequencing, which involves multiple steps: First, the DNA sample is sheared into small pieces. Then, special adapters are added to the two sides of the DNA fragments allowing them to attach to a solid surface. Once the fragments are attached to the surface they become amplified, resulting in clonal amplification of all the fragments. Fluorescently tagged nucleotides are then used to identify the nucleotide sequence producing one read per each fragment. The method is called sequencing by synthesis (Goodwin, McPherson and McCombie 2016).

HTS produces two main types of sequencing data: Single end (SE) and paired end (PE). In single end sequencing platforms, DNA fragments are sequenced from one end, whereas paired end reads provide reads from both ends of the fragments (Buermans and den Dunnen 2014). The relative directionality and the probability distribution of the distance between the two reads is predefined, which could aid in the prediction of SVs in the genome. Furthermore, PE reads are more reliable in the quantification of gene expression derived from RNA-Seq since the read pairs can provide more precise information about how the genes are spliced.

# 1.4 Computational cancer genomics

## 1.4.1 Overview

The assembly of the first human reference genome in 2001 (Venter et al. 2001) started a new period in biomedical research known as the genomic era. The last 15 years have witnessed a drastic increase in the amount of "omics" (in particular, genomics and transcriptomics) data being produced, while the cost was significantly reduced (**Fig. 8**). It is estimated that within the next decade, this amount would aggregate to 40 exabytes annually ($10^{18}$ bytes or 1 million terabytes), much of which is cancer related. Prior to the genomic era, all cancer studies used low throughput sequencing in which only a small number of genes and mutations were studied together. However, HTS provided the opportunity to explore cancer genomes on a much larger scale, with many thousands of genes being surveyed together. This has provided many new insights into how cancers develop.

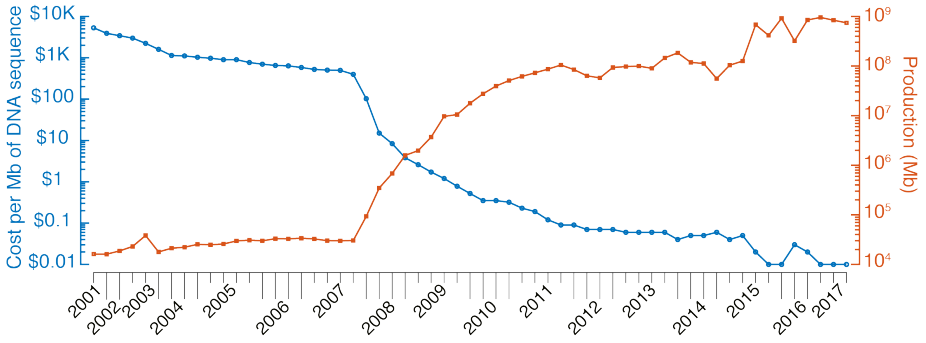



*Figure 8: DNA sequencing cost and production. Data is taken from the National Human Genome Research Institute (https://www.genome.gov/sequencingcostsdata).*

So far, numerous cancer genome projects have been initiated, of which the most widely used in cancer research is The Cancer Genome Atlas (TCGA). TCGA comprises sequencing data for more than 10,000 patient tumors across 33 cancer types. To date, several thousand studies have been published using these large repositories of sequencing data, leading to the identification of several novel driver genomic alterations in cancer (Samuels et al. 2004, Levine et al. 2005). For example, pathways not previously associated with oncogenesis were also observed to be altered in tumors, one of which was the activation of *KEAP1*/*NRF2*-signaling pathway, a pathway regulating the oxidative stress response in all cells, through the identification of frequent mutations in this pathway in lung cancer (Shibata et al. 2008). Additionally,

large-scale cancer genome investigations revealed important insights into general patterns of somatic alterations in cancer (Alexandrov et al. 2013, Greenman et al. 2007). Similarities and differences between multiple tumor types were studied using a combined analysis of large pan-cancer data. For example, tumors were classified into copy number or mutation driven subtypes based on their molecular profiles (Ciriello et al. 2013).

Most importantly, these large-scale molecular profiling studies have shown promising results in regards to cancer prognosis and diagnosis. Traditionally, prognosis of cancer outcome relied on clinical variables such as age and tumor stage. Recently, extensive efforts have been made to improve cancer prognosis by leveraging molecular information such as tumor genetic profiles. This has led to the identification of several biomarkers in different cancer types with clinical implications.

The ultimate goal of cancer genomic studies is to improve treatment and diagnostics of cancer. In fact, the impact of different cancer therapies on tumor genomes has been investigated extensively during the last decade (Hunter et al. 2006, Cahill et al. 2007, Noorani et al. 2017). This has led to identification of mechanisms responsible for drug resistance during and after different cancer therapies. For example, *NRAS* and *MEK1* activating mutations have been shown to be associated with relapsed melanoma tumors initially treated with *RAF* inhibitors (Emery et al. 2009, Nazarian et al. 2010).

### 1.4.2 Bioinformatics challenges

The drastic increase of HTS data has driven the rapid development of computational and mathematical approaches by adapting to the increased complexity that comes with it. Several challenges have been identified in the analysis of large scale sequencing data, one of which is the need for the development of specialized tools to detect different classes of genomic alterations. During the last decade, various computational methods were developed specifically for this purpose. Well-established methods now exist for the detection of point mutations and indels (Koboldt et al. 2012, Cibulskis et al. 2013), copy number changes (Zare et al. 2017, Li and Olivier 2013), genomic rearrangements (Chen et al. 2009, Rausch et al. 2012), and gene fusions (Kim and Salzberg 2011, Benelli et al. 2012).

Another general problem in sequencing data analysis is to distinguish the true biologically relevant signal from the technical noise introduced by experimental or computational pipelines, such as sequencing artifacts or mapping issues. The very first step in analyzing the HTS data is usually to

map the sequencing reads to the relevant reference genome. Several methods have been developed for such purpose, differing based on the complexity of the reference genomes and the quality and type of the sequencing data (Li and Durbin 2010, Langmead et al. 2009, Trapnell, Pachter and Salzberg 2009).

Additionally, certain challenges are specifically related to cancer genomes. Cancer is typically driven by somatic genomic alterations and therefore simultaneous analysis of tumor and patient-matched normal pairs are needed to identify such events. However, not all somatic alterations are involved in cancer development, and most of them are so called "passengers" with no impact on the tumor. Identification of driver events that contribute to cancer development is yet another challenge in large-scale cancer genomic studies. Many mathematical and probabilistic models have been developed to detect somatic events in tumor genomes (Meyerson, Gabriel and Getz 2010), all based on the presence of the event in the tumor cells and absence in the paired matching normal.

Generally, experimental validation is needed to ensure the functional relevance of genomic events in cancer. However, computational approaches have been employed to identify potential driver somatic events. These methods mainly rely on the recurrence of genomic events across tumor types as an indication of positive selection, and the functional impact of individual mutations or clusters of mutations within the same cancer pathway (Dees et al. 2012, Gonzalez-Perez and Lopez-Bigas 2012, Tamborero, Gonzalez-Perez and Lopez-Bigas 2013, Mermel et al. 2011).

### 1.4.3 Approaches to SV detection

Initially, genomic SVs were detected using cytogenetic approaches such as fluorescence *in situ* hybridization (FISH) (Feuk et al. 2006). However, the low resolution and throughput of this method only allowed the identification of specifically large and simple SVs in the complex tumor genomes. While later microarray-based approaches such as aCGH made significant advances in the detection of CNVs (Olshen et al. 2004, Yau et al. 2010), they were still not capable of detecting balanced SVs such as inversions and translocations. Furthermore, they had low resolution in regards to the breakpoint locations.

The development of sequencing technologies provided a unique opportunity to map SVs with the highest resolution possible, base pair resolution. During the last decade, sequencing data, including WES and more recently WGS, has been the primary tool for SV detection. While WES is restricted to the detection of rearrangements within or near exons, WGS is the most revealing but costly approach to detect such events. Furthermore, RNA-Seq can be used for the same purpose, but is limited to the detection of alterations involving transcribed regions, and would typically fail to detect cases involving noncoding or not expressed parts of the genome. Four main approaches have been developed to identify SVs using HTS data (Medvedev, Stanciu and Brudno 2009): Read pairs (RP), Split reads (SR), Read depth (RD), and Contig assembly (CA) **(Fig. 9)**.

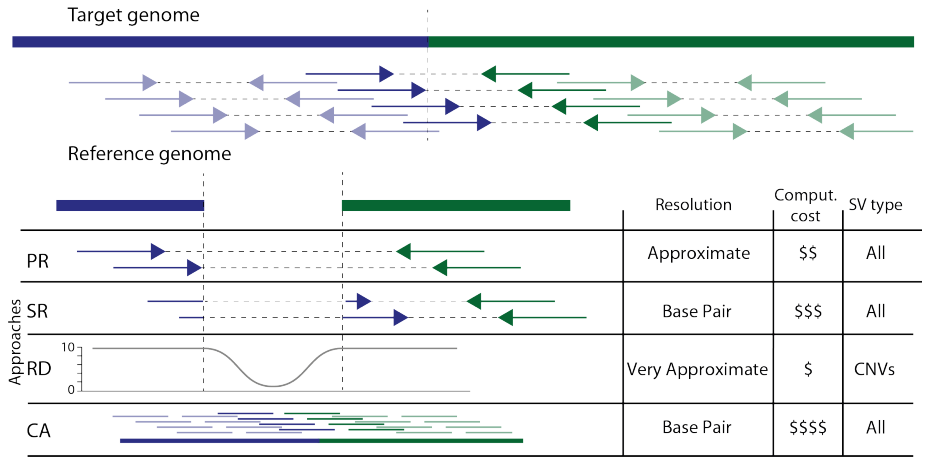



*Figure 9: Approaches to detect SVs. Dashed lines connects two mates in a pair.*

## Paired-end reads approach

Once the short reads are mapped to the reference genome, the chromosome positions and strands of the short reads are determined. Depending on the sequencing platform, read pairs should have a fixed directionality and insertion size. For example, conventional paired-end Illumina sequencing reads are typically aligned in forward reverse (FR) order. The short insertion size means that the forward read is aligned at a lower coordinate than the reverse read, and the mate reads in a pair are usually < 1 kb apart from each other, which should be the case for a concordantly mapped pair. However, anomalously mapped read pairs in the genome often have an unusual signature, such as incorrect mates orientation (e.g. RF or RR for Illumina) or abnormal insertion size (e.g. mates mapped to different chromosomes) indicative of possible SVs in the genome.

Discordant read pairs create different unique signatures corresponding to different classes of SVs. The most commonly detected signature is the "simple deletion" signature, where the mates are mapped in the right orientation but with a larger mapping distance than the expected insertion size (Medvedev et al. 2009). Conversely, smaller mapping distance corresponds to a genomic insertion **(Fig 9)**. Additionally, if the mates in a pair are mapped to different chromosomes, it is considered as an inter-chromosomal translocation signature. All the signatures mentioned so far rely solely on abnormal mapping distance. However, tandem duplications and inversions would have more complex signatures, where the orientation of the mates is also taken into account. Assuming the reads are derived from FR sequencing platform, a RF mapped pair can be indicative of tandem duplication, whereas FF or RR pairs could correspond to inversions in the genome **(Fig. 10) (Medvedev et al. 2009)**. The combination of these simple signatures can be used to detect more complex SVs (Yang et al. 2013a).

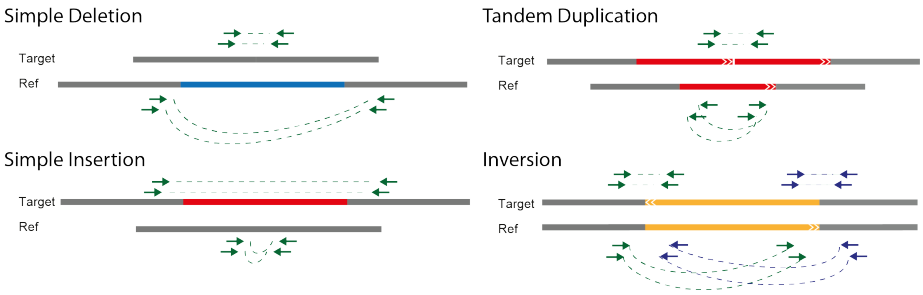



*Figure 10: RP signatures corresponding to different intra-chromosomal SVs adapted from Yang et al (Yang et al. 2013a)*

Several tools have been developed based on the RP approach (Sindi et al. 2009, Sindi et al. 2012, Chen et al. 2009). While SV detection using this approach is relatively reliable and fast, it is unable to detect the exact breakpoints of the SVs. Therefore, combining it with other approaches capable of precise breakpoint identification could be ideal for this purpose.

## Split reads approach

SV in the genome are not only supported by read pairs bridging the two breakpoints of the event, but also by single reads spanning each breakpoint with a "split mapping" signature. Split mapping reads are defined as reads that are partially mapped to one location and partially to the other. They can be used to detect SVs with base pair precision. However, split read alignment would typically require more computational time and resources compared to normal short read alignments, therefore it is less feasible to do on a large

scale. Due to this limitation, only a few tools have been developed using merely SR approach (Suzuki et al. 2011, Wang et al. 2011). However, hybrid approaches using the two, PR and SR, have been widely used and have shown promising results. Typically, PR approaches are used to identify candidate SVs that are later validated and fine tuned using the SR approach (Yang, Chockalingam and Aluru 2013b, Rausch et al. 2012, Yang et al. 2013a, Newman et al. 2014).

### Read depth approach

With the assumption that the genome is sequenced uniformly, the number of reads mapped to different regions should be proportional to the actual copy number of that region. For example, a deleted region would typically have less reads compared to a neutral region, whereas duplicated regions would be associated with a higher number of reads. Thus, the read-depth feature can be used to detect SVs (Bailey et al. 2002, Campbell et al. 2008), but it is extremely limited compared to the two approaches mentioned above. First, it can only detect unbalanced SVs such as deletions and duplications. Second, the exact structural basis is not always evident. For example, the duplication signature does not specify where the duplication happens but rather what the duplicated sequence is. Additionally, the RD approach is incapable of high-resolution identification of SV breakpoints. Even though the RD approach is poorer compared to the other methods for SV detection, it is sometimes used in combination with the other tools to better annotate the predicted SVs (Sindi et al. 2012).

### Contig assembly approach

All the methods discussed so far rely on the direct mapping of short reads to the reference genome. However, due to the complexity of the human genome and specifically cancer genomes, it is not always possible to accurately map short reads to the reference genome. Another approach that has been proposed for SV detection is to assemble the short reads into contigs or longer sequences independent of the reference genome, and then map them to the genome. Several tools have been implemented using this approach, which are often computationally inefficient (Mohiyuddin et al. 2015, Chen et al. 2014). However, similar to SR approach, base pair resolution of SV breakpoints is achievable by CA and therefore can be combined with more efficient methods for better characterization of SVs in the genome (Schroder et al. 2014).

# 2 AIMS

The main objective of this thesis was to comprehensively investigate SVs such as viral integrations and inter- and intra-chromosomal rearrangements, and their impact on the tumor transcriptome. Computational approaches were applied to cancer genomics data to answer biologically relevant questions. RNA-Seq and SNP6 based copy number data from ~10,000 tumors, and WGS data of 600 tumor normal pairs (each > 75 Gb) were used in this thesis, all of which aggregated to >200 Tb of data.

More specifically, the objectives were:

- To provide a complete map of different classes of SVs in cancer genomes with the help of WGS data **(Papers I, II, III)**
- To investigate the association between SVs and gene expression levels **(Papers I, II, II)**
- To highlight specific cases with potential functional implications in cancer **(Papers I, II, II)**
- To identify viral integration sites in cancer genomes **(Paper I)**
- To explore the relationship between SVs and CNAs **(Papers II, III)**
- To use a dual DNA/RNA approach to provide a high confidence set of gene fusions in cancer **(Papers II, III)**
- To identify intra-chromosomal SVs in a larger cohort with the help of array-based copy number data **(Paper III)**
- To find potentially functional intra-chromosomal SVs that lead to oncogene activation through promoter substitution **(Paper III)**

# 3 RESULTS AND DISCUSSION

## 3.1 Tumor-virus associations (Paper I)

Tumor-virus associations have been extensively studied in a few cancer types mostly using low throughput approaches. The availability of tumor sequencing data provides an opportunity to survey the relevance of viral associations in diverse cancer types. Here, a systematic screen for viral expression was performed using 4,433 tumors across 19 tumor types, and viral integration sites were identified in virus positive tumors.

The discordant read pair approach has been widely used to detect SVs in the human genome. In principle, the same approach can be used to detect foreign DNA insertions, such as viral DNA, into the human genome. In this context, discordant read pairs refer to pairs that have one mate mapping to the human genome and the other to the viral DNA. Integrated viral DNA with a possible functional role is often expressed, and therefore discordant reads should also reveal themselves in RNA-Seq data **(Fig. 11)**.

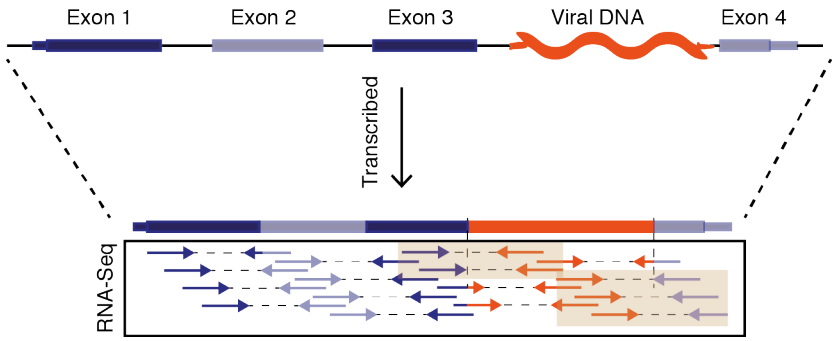



*Figure 11: Schematic of viral-human fusion; Discordant read pairs are shown in the brown boxes*

A pipeline was first developed to identify virus positive tumors by measuring the viral expression in tumors. This was performed by counting the number of reads that mapped to the viral genome and not to the human genome, normalized for the library size and indicated as parts per million (p.p.m.). Viral expression was observed in 178 tumors (> 2 p.p.m), of which 150 cases (84.2%) were HPV and HBV positive. As expected, HPV was dominantly observed in cervical (Clifford et al. 2006) followed by head and neck tumors (Mork et al. 2001) (96.6% and 14.1% respectively), whereas HBV infected tumors were only seen in hepatocellular cancer (11 tumors, 32.4%) at the

selected level of stringency. The remaining infected viruses such as herpes viruses were mostly non-driver events with no active role in tumor development. However, one bladder tumor showed high expression of BK virus (BKV), specifically its known oncogene (*Tag*), which further supported the previously proposed aetiological role of this virus in tumor formation (Abend, Jiang and Imperiale 2009).

Next, a viral integration pipeline was developed based on the discordant read pair principle, and applied to the virus positive cases. Only integrations supported by multiple discordant read pairs where the human mates were clustered together in a genomic region were considered. Confirming previous studies (Schmitz et al. 2012, Sung et al. 2012), viral integrations were observed in most HPV and HBV positive tumors (104 tumors; 70%). Additionally one BKV positive bladder tumor had evidence of viral integration on chromosome 2.

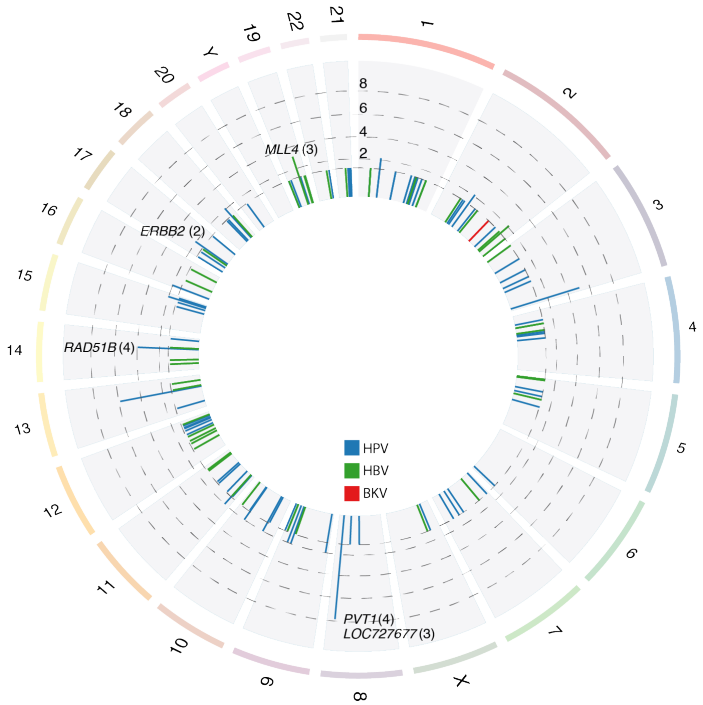



*Figure 12: Viral integration sites randomly distributed across the genome. Number of integrations for each chromosomal cytoband is shown here. Selected genes are shown for cytobands with recurrent integrations.*

While viral integration sites were widely spread across the genome, regions with recurrent viral integrations mostly contained known cancer genes or previously described fusions sites **(Fig. 12)**. The most recurrent integration was HPV insertion in the *MYC* locus, previously described as a known HPV fusion site in cervical cancer (Peter et al. 2006). HPV was recurrently fused with *PVT1* and *LOC727677*, two long non-coding RNAs downstream and upstream of *MYC* (four and three cases respectively), all of which were associated with elevated expression levels of these genes. Another confirmatory observation was the recurrent HBV integration with *MLL4* in hepatocellular cancer (n = 3), which was associated with overexpression of this gene (Sung et al. 2012).

Notable among the novel findings, two cervical tumors had an HPV insertion in *ERBB2*, one of which showed strong *ERBB2* induction. Moreover, several cervical tumors had *RAD51B-HPV* fusion (n = 4), resulting in weak reduction of *RAD51B* expression, a gene involved in DNA DSB repair. Loss of RAD51B protein function has been shown to contribute to tumor growth by inducing genomic instability (Suwaki, Klare and Tarsounas 2011). Our novel findings regarding HPV integration sites in cervical and head and neck tumors using HTS data, show that HPV integration into cellular DNA may play a role in oncogenesis, not only by expressing HPV viral oncogenes including E6 and E7, but also by altering the activity of cellular oncogenes and tumors suppressors such as *ERBB2* and *RAD51B*.

## 3.2 Mapping of Somatic SVs across multiple cancer types (Papers II, III)

In this part of the thesis, intra-cellular SVs were identified using WGS data **(Paper II)** and SNP6 copy number profiles **(Paper III)**.

### 3.2.1 Mapping SVs using WGS data

The availability of a large amount of WGS data in TCGA made it possible to carefully map somatic SVs in several hundreds of tumor genomes. First and foremost, a computationally and biologically robust WGS-based pipeline capable of high-resolution identification of these events was needed. Unlike inter-cellular SVs such as viral insertions, several tools had previously developed for this purpose. Thus, we decided to employ the already available tools instead of implementing our own method. However, large discrepancies between different tools were observed after applying them to WGS data, and therefore a careful assessment of different SV callers was required. Benchmarking SVs has been challenging for several reasons, one of which is the lack of golden standard data to objectively evaluate different SV callers. While simulated human genomes have been widely used for this purpose (Qin et al. 2015, Bartenhagen and Dugas 2013, Hu et al. 2012, Korbel et al. 2009), the true complexity of cancer genomes is not fully reflected using simulated data. As copy number changes are ideally a subset of SVs, a perfect SV detection tool should be able to detect them. Thus, array-based CNV data could be used as true positive set to assess WGS based SVs obtained from different SV callers.

Four SV caller tools – **SVDetect** (Zeitouni et al. 2010), **BreakDancer** (Chen et al. 2009), **Delly** (Rausch et al. 2012), and **Meerkat** (Yang et al. 2013a) - were primarily selected for further evaluation. All four made use of the PR approach to identify regions with potential SVs. While SVDetect purely relies on this method, Meerkat and Delly combine it the SR approach for a more precise identification of SVs with base pair resolution. Additionally, BreakDancer uses the RD approach specifically for accurate characterization of copy number imbalance SVs.

Next, the sensitivity and specificity of the four SV callers were measured using true and randomized CNV breakpoints as true and false positive sets respectively. **Meerkat** had the best performance considering both sensitivity and specificity scores, and therefore was selected as the primary SV detection tool. Finally, somatic SVs were mapped in 600 tumors across 18 cancer types

using high coverage WGS data. This provided the most comprehensive map of somatic SVs in cancer to date.

### 3.2.2 CNVs as a subset of SVs

Having access to base-pair resolution SVs from WGS as well as CNV data derived from the Affymetrix SNP6 platform for 600 tumor genomes provided a unique opportunity to systematically investigate the relationship between SVs and CNVs, which have typically been studied in isolation (**Paper II**). At the breakpoint level, the two correlated considerably in terms of absolute number of breakpoints within different tumors (Pearson's r = 0.81). Additionally, even though a small fraction of CNV events had a correspondence in SV data (~10%), the overlapping set was mostly classified correctly where 97% and 90% of copy number losses and gains were categorized as deletions and tandem duplications respectively. As array-based CNVs data is still considerably more abundant than WGS data, it is temping to use it as a substitute to identify genomic rearrangement caused by SVs in the genome. However, it should be noted that this only represents a small fraction of SVs that are copy number imbalanced, and therefore WGS data, when available, is highly favorable over array-based data.

In TCGA, high coverage WGS data is available for 600 tumors, while SNP6 copy number profiles are available for ~10,000 tumors. Thus, a comprehensive analysis of a subset of SVs in an even a larger set of tumors was possible. In **Paper III**, deletions and tandem duplications were categorized as CNVs with a clear interpretation in terms of their structural basis. For example, only genomic regions with one extra copy and no gained adjacent regions were considered as tandem duplication. Using this strict filtering on CNV data, we were able to provide a catalogue of simple deletions and tandem duplications in a large cancer cohort.

## 3.3 Impact of somatic SVs on tumor RNA (Papers II, III)

As it follows from the central dogma of molecular biology, a genomic alteration that has a functional role during tumor development should also have an impact on the RNA produced by the cell. SVs contribute to cancer development by multiple different mechanisms (see **section 1.2.3**), all of which have a direct impact on tumor RNA by either altering the mRNA levels, for example by promoter substitution; or altering the mRNA structure by forming new chimeric transcripts. In this part of the thesis, the global impact of somatic SVs on tumor RNA, using both expression levels and structure, was explored. Candidate tumor driver events with an impact on tumor transcriptional output were highlighted **(Papers II, III)**.

### 3.3.1 Promoter substitutions

A well-established mechanism for gene activation by SVs is to substitute a weak promoter of a gene for a stronger promoter of another, which usually occurs as a result of the two genes being fused together. As mentioned in **section 1.2.3**, activation of individual oncogenes through promoter substitution (PS) has been previously described for several cases in cancer (Tomlins et al. 2005, Oliveira et al. 2005). However, it is still unclear how often such events occur in tumors and to what extent they have an impact on tumor transcriptional output. To systematically investigate these cases in cancer, we used SV calls, from both WGS **(Paper II)** and SNP6 **(Paper III)** data, to identify cases that may result in the creation of PS events. Only SVs resulting in chimeric gene regions where the promoter of the 3' partner was substituted for the 5' partner promoter were considered. In both studies, we observed that mRNA of the 3' partner was more in induced when the 5' partner had the stronger promoter rather than the weaker one.

In **Paper II**, 62 PS events were found where the 3' partner was significantly induced (>3 fold), of which only two cases were recurrent within one cancer type, both being previously described. These cases included activation of *ERG* and *RET* in prostate and thyroid cancers, through the hijacking of the strong promoters of *TMPRSS2* and *CCDC6,* respectively. Notable among the non-recurrent cases was strong induction of *PRKCB*, a protein kinase C gene, observed in one colorectal tumor through a PS event with *USP7*, a gene typically highly expressed in colorectal carcinoma. This resulted from an inversion on chromosome 16, creating a fusion transcript where the two promoters are swapped (**Fig. 13**). Frequent fusions involving genes in protein

kinase C family such as *PRKCA* and *PRKCB*, have been previously described in different tumors, and are mostly associated with gene activation (Stransky et al. 2014, Bridge et al. 2013, Plaszczyca et al. 2014). Taken together, although it remains unclear to what extent such events are under positive selection and functional in cancer, results from this study suggest that PS often contribute to gene activations in cancer.

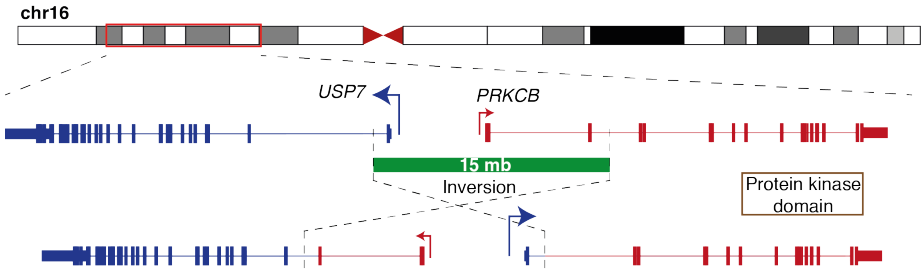



*Figure 13 Strong promoter of USP7 (indicated by big blue arrow) is swapped with the weak promoter of PRKCB (indicated by small red arrow) through an inversion in chromosome 16.*

Rare driver events in cancer occur at low rate and therefore are only observable when a large enough set of tumors is being analyzed. This motivated us to make use of SVs calls derived from SNP6 copy number profiles (**Paper III**), available for a larger set of tumors (~10,000), to identify more recurrent PS affected cases. 126 repeated (n >=2) cases showed evidence of PS where the 3' partner was induced within the same cancer type (2-fold).

Notable among the significantly induced cases ($n$ = 8; FDR 10%), was strong induction of *TIAM2* in five ovarian and one uterine tumors through PS with *SCAF8*, a nearby gene that shows strong promoter activity in ovary and uterus tissue types. This resulted from a genomic deletion on chromosome 6, juxtaposing the *SCAF8* promoter to the *TIAM2* promoter region, upstream of the transcription start site **(Fig 14)**. T-cell lymphoma invasion and metastasis genes (*TIAM1*, *TIAM2*) act as regulators in the Rac GTPase pathway, an important signaling pathway in cancer (Parri and Chiarugi 2010). While the significance of TIAM genes in cancer is well established (Liu et al. 2007, Wang and Wang 2012, Zhao et al. 2013), the underlying molecular mechanism of their activation is poorly understood. Here, we propose a novel mechanism for *TIAM2* activation; however, further investigation is needed to establish that the fusion transcript is translatable into *TIAM2* protein and to determine the functional relevance of increased *TIAM2* protein levels in these tumors.

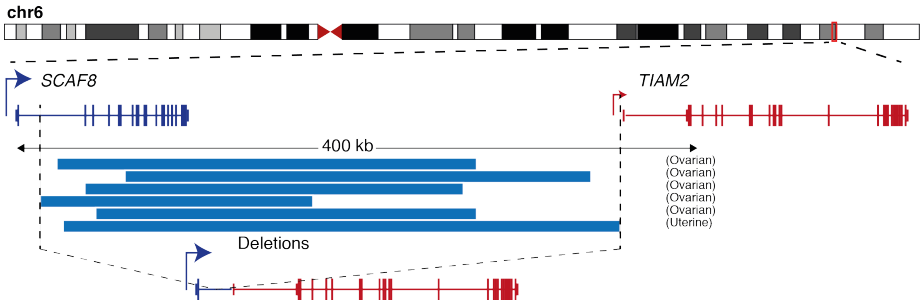



*Figure 15: Strong promoter of SCAF8 is brought to the proximity of TIAM2 by genomic deletions. Blue boxes correspond to deletions in different tumors.*

Additionally, we found that *SCARB1* mRNA was induced through hijacking of the promoter of *NCOR2,* an adjacent gene with high expression in these tumors, in stomach, esophageal, and lung adenocarcinoma. On the DNA level, this results from a tandem duplication by which an additional fusion transcript is formed. While the functional domain of *SCARB1* (CD36) is maintained in the new chimeric gene, the 5' end including the promoter region is replaced with the 5' end of *NCOR2* **(Fig 15)**. Overexpression of scavenger receptor class B (*SCARB1)* is known to be associated cancer development. Additionally, a recent study has shown that CD36 is required for the acquisition of metastatic phenotypes in the cell, therefore could be used as a possible target for anti-cancer drugs (Pascual et al. 2017).

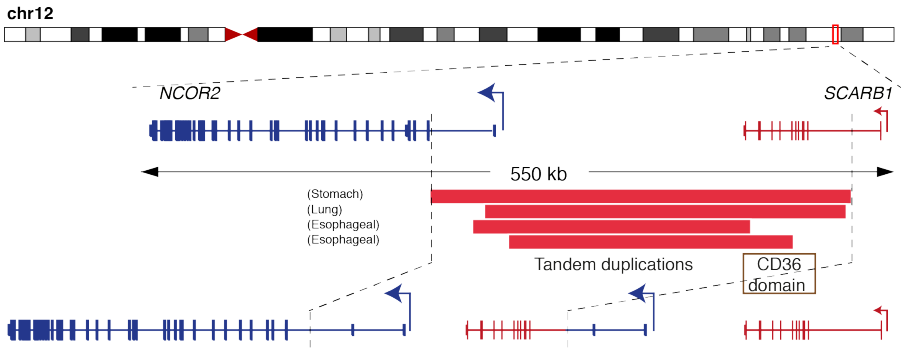



*Figure 14: SCARB1 overexpression by hijacking the strong promoter of NCOR2. The red boxes represent the tandem duplications in different tumors.*

Taken together, we observed that all classes of SVs, including deletions, tandem duplications and inversions, are involved in transcriptional activation of cancer-relevant genes through hijacking of strong promoters from other cellular genes.

### 3.3.2 Enhancer Hijacking

PS is not the only way that SVs cause transcriptional deregulation in tumors. Recent studies have demonstrated that mRNA induction can occur as a consequence of rearrangements of noncoding cis-regulatory elements such as enhancers, without altering the mRNA structure (Northcott et al. 2014, Affer et al. 2014). During the last decade, with the availability of more WGS data, noncoding genomic alterations were studied more comprehensively as a way of activating cancer driver genes. The most frequent case to date is the *TERT* promoter mutation described in several cancer types, which is typically associated with *TERT* elevated expression (Huang et al. 2013, Fredriksson et al. 2014). Additionally, activating SVs upstream of the *TERT* transcription start site resulting in juxtaposition of enhancers to the *TERT* promoter region has been described in kidney chromophobe tumors (Davis et al. 2014). This motivated us to systematically investigate the impact of upstream SVs on the tumor transcriptome. In **Paper II**, elevated expression levels associated with upstream SVs were observed in 39 genes. Among those were *TERT* promoter proximal SVs identified in kidney renal, colorectal and melanoma, most of which led to *TERT* overexpression. Our results further established the role of SVs involving regulatory elements in tumor transcriptional induction.

### 3.3.3 Fusion genes

SVs may lead to the juxtaposition of two genes and hence the creation of a fusion gene. Gene fusions are now recognized as one of the most common driver alterations in cancer. Another way by which these events could contribute to tumorigenesis, in addition to promoter substitution, is to create a chimeric transcript with a novel oncogenic function that is different from the individual fusion partner genes (see **section 1.2.3**).

Large-scale gene fusion investigations are mostly based on RNA-Seq data, due to the low cost of RNA-Seq in comparison to WGS data. Functional gene fusions in cancer are usually expressed in the tumor and theoretically should be detectable by RNA-Seq data. However, that is not always the case in reality and many fusion detection tools fail to identify expressed fusion transcripts. One study demonstrated that large variability between different RNA-Seq based algorithms exist (Carrara et al. 2013), indicative of the large number of false positives and negatives in the data. Additionally, even for

correctly characterized fusion genes, the genomic structural basis of such events cannot be determined using RNA-Seq data. Having access to SV calls for many tumors provided a unique opportunity to investigate gene fusions using a more informative dataset. However, similar to RNA-Seq data, the sensitivity and specificity of such data is low (see **section 3.2.1**). Nevertheless, as the two data types are generated independently, presumably they do not share the same artifacts. Thus, the integration of both DNA-based and RNA-based fusion calls could in principle provide a more sensitive and specific set of gene fusions **(Paper II)**.

First, fusions were identified in tumors with both WGS and RNA-seq data (431 samples) independently. SVs where the two breakpoints spanned two distinct genes with the correct orientation based on the SV type were considered as valid gene fusions (5,209 events). In parallel, fusion transcripts were identified using an RNA-Seq fusion detection tool (9,657 events). Surprisingly, only 299 gene pairs were found to be common in DNA and RNA calls. We next assessed if the sensitivity was reduced in the combined set in regard to identifying potential driver events in cancer. Strong enrichment of known cancer-associated genes was observed in the combined approach compared to the two approaches separately. These results indicated that the dual approach could be beneficial in identifying cancer driver events. Taken together, we show that a more reliable set of fusions in cancer can be achieved by using both DNA and RNA data together.

Notable among the high confidence set of fusions was a fusion involving *NRF2* gene in one thyroid tumor. *NRF2* is a transcription factor, which is a key regulator in the cellular oxidative stress response pathway. In normal cells, NRF2 protein remains inactive due to binding to a repressor protein called KEAP1. Conversely, in cancer cells, the NRF2-KEAP1 interaction is frequently disrupted leading to the activation of NRF2 (Kansanen et al. 2013). Several mechanisms are known to be involved in the deregulation of the NRF2 signaling pathway in cancer including mutations in *KEAP1* (Ohta et al. 2008), mutations in the *KEAP1* binding domain of *NRF2* (Shibata et al. 2008), or *KEAP1* hypermethylation (Hanada et al. 2012). Here we observed transcriptional induction of *NRF2* through fusion with *PAX8,* as a consequence of a tandem duplication in chromosome 2 **(Fig. 16)**. Additional observations, including the loss of the *KEAP1* interaction domain of NRF2 in the new fusion transcript, and strong induction of several known *NRF2* targets, further supported the functional role of this fusion in the activation of NRF2 pathway.

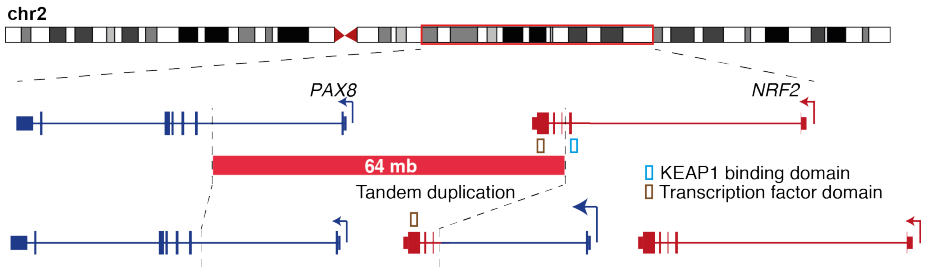



*Figure 16: Activation of NRF2 by PAX8-NRF2 fusion as a result of a tandem duplication*

In summary, through integrative analysis of genomic and transcriptomic data from a large cancer cohort, multiple insights into the relevance of SVs in cancer including the relationship between genomic SVs, CNVs and gene expression were provided.

# 4 CONCLUSIONS AND FUTURE PERSPECTIVES

This thesis provides a comprehensive catalogue of simple structural genomic alterations in cancer. Bioinformatical approaches were developed to systematically investigate the role of such alterations in cancer, and several potentially functional events were highlighted.

Despite unprecedented progress in the identification and analysis of SVs in cancer, providing a complete map remains challenging. One of the main challenges is to identify more complex types of SVs such as choromothripsis and chromplexy, where large number of structural changes arises in a single event (see **section 1.2.2**). Both chromoplexy and chromothripsis have been observed in individual cancer types, but it is not yet fully understood how these events contribute to tumor progression. The main limitation here is the imperfect mapping of these complex events, which can be improved by obtaining longer reads. This can be partially solved by assembly of short read sequences into longer contigs, as this data is already available for thousands of tumors. Furthermore, while recent studies have tried to infer the underlying mechanisms of somatic SVs in the genome, the molecular mechanism for many of these events is still yet to be determined. Given the distinct genomic signatures of such mechanisms of SV formation, investigating general patterns of SVs in several tumor genomes could aid to understand the mechanisms by which these genomics alterations occur.

Numerous clinically relevant driver events have been found during recent decades, using both low and high-throughput approaches. HTS has empowered the detection of less frequent genomic events that could not be detected with the small number of samples being analyzed. By applying methods such as those described in this thesis to even larger cohorts, it is likely that additionally functional events, that may not be frequent enough to identify as recurrent in currently available datasets, can be pinpointed.

# ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, **Erik**, for letting me join his group and become his first "own" PhD student, even though I wasn't a "Star Wars" fan! I doubt that many PhD students are as lucky as me to have a caring supervisor who is fully involved in their project. Thanks for giving me freedom when I needed it and letting me grow as an independent researcher. For inspiring me when nothing seemed to work. For teaching me how to write a manuscript, I know it was sometimes painful to comment "italized gene names" on "alaei_mahabadi_et_al_v30"! Basically, for being an incredible supervisor! Also, thanks to my co-supervisor, **Claes Gustafsson.** It was always nice talking to you.

I would like to express my sincere gratitude to my colleagues in the Larsson Lab. My time here would not have been as fun without you all. **Arghavan**, thanks for always taking the time to answer my basic biology questions. It's been great having the Iranian corner in the lab; we had so many constructive "scientific" discussions! ;) **Jimmy** for all the nice late evening talks in the lab - both scientific and non-scientific of course! Most importantly, for being the only other person in the lab interested in football! And FYI, I am still not convinced that Belgium is the best football team in the world, I might change my mind this summer though. ;) **Swaraj,** it has been extremely fun working with you. Your jokes and stories (including your very famous buffalo joke:D) made my time in the lab a lot more enjoyable. Thanks for always helping me even when you were busy with your own projects. **Kerryn**, you are one of the most positive people I know. Thanks for taking care of the "real science" in a "bioinformatics lab" ;) Also, for proof reading my thesis. Now you know about the law of Babak's thesis: "THE" is neither added nor removed; rather, it is only moved from one place to another. **Markus,** I enjoyed your company a lot. I nominate you as the next pinbot and rudy caretaker, as I got the impression that you liked it a lot! **Martin** thanks for proof reading my **"**Sammanfattning på svenska**"** I hope that there is no hidden joke in it now. **Joakim**, we had great scientific collaborations over the past years, especially on the SV project, thanks a lot. Thanks to our former postdoc in the group, **Johan**. It was fun sharing an office with you for a while.

Thanks to our collaborators, **Jonas** and **Joydeep**. Your thoroughly conducted experiments added so much value to my favorite paper during my PhD!

Everyone else in our corridor (2nd floor), especially **Clausen(ian)s** and **Kanduri(ian)s**, had so much fun over so many lunches and fikas with all of you guys.

**Zhiyuan**, as one of your lunch buddies, I am also grateful for all the fun discussions we had over lunch! ;)

**Ka-Wei**, it was nice working with you on the virus project; it turned out to be a great paper.

Special thanks to my family. **Baba** without your enthusiasm for science, I would not have been interested in it in the first place. **Maman**, thanks for your endless love and support. For always believing in me and encouraging me towards new challenges in life. **Mehraveh**, for being the kindest and most supportive sister. Love you all!

Thanks to all my friends outside work: **Saghi**, I guess I beat you this time ;) without our deep Swedish discussions everyday, I would not have been this fluent in Swedish, tackaar! The last episode of "ghors of a PhD student" is yet to come for you! **Amin**, Don't remember if it was an odd or an even day when I decided to start a PhD, but I guess we are both happy with the decisions we made after all those long overnight talks about the future, you know what I am talking about! **Maral**, for all the lunches we had together at the medicinareberget. **Kambiz**, **Nazanin** for all the super fun weekends (and weekdays) together. Lets keep it up! Look forward to seeing "Taha baba" soon after my defense. All my other friends whom I enjoyed your company a lot during the last 5 years!

**Sheedeh**, not everyone has the luxury of having a sister in the same city when living abroad! Thanks for all the fun we had together especially the trip to the North Pole! ;) **Payam** for helping me out with the venue arrangements for my defense party.

Last but in no way the least, my sincere gratitude to my very own family, **Anna**, for always being so supportive. Thanks for your infinite patience in putting up with my "PhD student lifestyle". I know you always wished that you could sleep as much as me in the mornings. ;) Getting married to you during my PhD studies, made this period even more unforgettable for me! Looking forward to all the new adventures ahead of us. Love you always and forever.

# REFERENCES


Abend, J. R., M. Jiang & M. J. Imperiale (2009) BK virus and human cancer: innocent until proven guilty. *Semin Cancer Biol,* 19**,** 252-60.

Affer, M., M. Chesi, W. G. Chen, J. J. Keats, Y. N. Demchenko, A. V. Roschke, S. Van Wier, R. Fonseca, P. L. Bergsagel & W. M. Kuehl (2014) Promiscuous MYC locus rearrangements hijack enhancers but mostly super-enhancers to dysregulate MYC expression in multiple myeloma. *Leukemia,* 28**,** 1725-1735.

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, P. J. Campbell & M. R. Stratton (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep,* 3**,** 246-59.

Anderson, M. W., S. H. Reynolds, M. You & R. M. Maronpot (1992) Role of proto-oncogene activation in carcinogenesis. *Environ Health Perspect,* 98**,** 13-24.

Arvanitakis, L., N. Yaseen & S. Sharma (1995) Latent membrane protein-1 induces cyclin D2 expression, pRb hyperphosphorylation, and loss of TGF-beta 1-mediated growth inhibition in EBV-positive B cells. *J Immunol,* 155**,** 1047-56.

Avery, O. T., C. M. Macleod & M. McCarty (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med,* 79**,** 137-58.

Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li & E. E. Eichler (2002) Recent segmental duplications in the human genome. *Science,* 297**,** 1003-7.

Bailey, S. M. & J. P. Murnane (2006) Telomeres, chromosome instability and cancer. *Nucleic Acids Res,* 34**,** 2408-17.

Bais, C., B. Santomasso, O. Coso, L. Arvanitakis, E. G. Raaka, J. S. Gutkind, A. S. Asch, E. Cesarman, M. C. Gershengorn & E. A. Mesri (1998) G-protein-coupled receptor of Kaposi's sarcoma-associated herpesvirus is a viral oncogene and angiogenesis activator. *Nature,* 391**,** 86-9.

Barnes, D. E. & T. Lindahl (2004) Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu Rev Genet,* 38**,** 445-76.

Bartenhagen, C. & M. Dugas (2013) RSVSim: an R/Bioconductor package for the simulation of structural variations. *Bioinformatics,* 29**,** 1679-81.

Benelli, M., C. Pescucci, G. Marseglia, M. Severgnini, F. Torricelli & A. Magi (2012) Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics,* 28**,** 3232-9.

Boveri, T. 1914. *Zur frage der entstehung maligner tumoren*. Jena,: G. Fischer.

Boxus, M. & L. Willems (2009) Mechanisms of HTLV-1 persistence and transformation. *Br J Cancer,* 101**,** 1497-501.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao & K. Corcoran (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol,* 18**,** 630-4.

Bridge, J. A., X. Q. Liu, J. Sumegi, M. Nelson, C. Reyes, L. A. Bruch, M. Rosenblum, M. J. Puccioni, B. S. Bowdino & R. D. McComb (2013) Identification of a novel, recurrent SLC44A1-PRKCA fusion in papillary glioneuronal tumor. *Brain Pathol,* 23**,** 121-8.

Buermans, H. P. & J. T. den Dunnen (2014) Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta,* 1842**,** 1932-1941.

Cahill, D. P., K. K. Levine, R. A. Betensky, P. J. Codd, C. A. Romany, L. B. Reavie, T. T. Batchelor, P. A. Futreal, M. R. Stratton, W. T. Curry, A. J. Iafrate & D. N. Louis (2007) Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clin Cancer Res,* 13**,** 2038-45.

Campbell, P. J., P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. Edwards, G. R. Bignell, M. R. Stratton & P. A. Futreal (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet,* 40**,** 722-9.

Cancer Genome Atlas Research, N., T. J. Ley, C. Miller, L. Ding, B. J. Raphael, A. J. Mungall, A. Robertson, K. Hoadley, T. J. Triche, Jr., P. W. Laird, J. D. Baty, L. L. Fulton, R. Fulton, S. E. Heath, J. Kalicki-Veizer, C. Kandoth, J. M. Klco, D. C. Koboldt, K. L. Kanchi, S. Kulkarni, T. L. Lamprecht, D. E. Larson, L. Lin, C. Lu, M. D. McLellan, J. F. McMichael, J. Payton, H. Schmidt, D. H. Spencer, M. H. Tomasson, J. W. Wallis, L. D. Wartman, M. A. Watson, J. Welch, M. C. Wendl, A. Ally, M. Balasundaram, I. Birol, Y. Butterfield, R. Chiu, A. Chu, E. Chuah, H. J. Chun, R. Corbett, N. Dhalla, R. Guin, A. He, C. Hirst, M. Hirst, R. A. Holt, S. Jones, A. Karsan, D. Lee, H. I. Li, M. A. Marra, M. Mayo, R. A. Moore, K. Mungall, J. Parker, E. Pleasance, P. Plettner, J. Schein, D. Stoll, L. Swanson, A. Tam, N. Thiessen, R. Varhol, N. Wye, Y. Zhao, S. Gabriel, G. Getz, C. Sougnez, L. Zou, M. D. Leiserson, F. Vandin, H.

T. Wu, F. Applebaum, S. B. Baylin, R. Akbani, B. M. Broom, K. Chen, T. C. Motter, K. Nguyen, J. N. Weinstein, N. Zhang, M. L. Ferguson, C. Adams, A. Black, J. Bowen, J. Gastier-Foster, T. Grossman, T. Lichtenberg, L. Wise, T. Davidsen, J. A. Demchok, K. R. Shaw, M. Sheth, H. J. Sofia, L. Yang, J. R. Downing, et al. (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med,* 368**,** 2059-74.

Capecchi, M. R. (1989) Altering the genome by homologous recombination. *Science,* 244**,** 1288-92.

Carmeliet, P. & R. K. Jain (2000) Angiogenesis in cancer and other diseases. *Nature,* 407**,** 249-57.

Carrara, M., M. Beccuti, F. Lazzarato, F. Cavallo, F. Cordero, S. Donatelli & R. A. Calogero (2013) State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed Res Int,* 2013**,** 340620.

Carvalho, C. M. & J. R. Lupski (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet,* 17**,** 224-38.

Chen, K., L. Chen, X. Fan, J. Wallis, L. Ding & G. Weinstock (2014) TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res,* 24**,** 310-7.

Chen, K., J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, X. Shi, R. S. Fulton, T. J. Ley, R. K. Wilson, L. Ding & E. R. Mardis (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods,* 6**,** 677-81.

Cibulskis, K., M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander & G. Getz (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol,* 31**,** 213-9.

Cinti, R., L. Yin, K. Ilc, N. Berger, F. Basolo, S. Cuccato, R. Giannini, G. Torre, P. Miccoli, P. Amati, G. Romeo & R. Corvi (2000) RET rearrangements in papillary thyroid carcinomas and adenomas detected by interphase FISH. *Cytogenet Cell Genet,* 88**,** 56-61.

Ciriello, G., M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz & C. Sander (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat Genet,* 45**,** 1127-33.

Clifford, G., S. Franceschi, M. Diaz, N. Munoz & L. L. Villa (2006) Chapter 3: HPV type-distribution in women with and without cervical neoplastic diseases. *Vaccine,* 24 Suppl 3**,** S3/26-34.

Colombo, M., G. Kuo, Q. L. Choo, M. F. Donato, E. Del Ninno, M. A. Tommasini, N. Dioguardi & M. Houghton (1989) Prevalence of antibodies to hepatitis C virus in Italian patients with hepatocellular carcinoma. *Lancet,* 2**,** 1006-8.

Crick, F. H. (1958) On protein synthesis. *Symp Soc Exp Biol,* 12**,** 138-63.

Cuddihy, A. R. & M. J. O'Connell (2003) Cell-cycle responses to DNA damage in G2. *Int Rev Cytol,* 222**,** 99-140.

Dang, C. V. (2012) MYC on the path to cancer. *Cell,* 149**,** 22-35.

Davis, C. F., C. J. Ricketts, M. Wang, L. Yang, A. D. Cherniack, H. Shen, C. Buhay, H. Kang, S. C. Kim, C. C. Fahey, K. E. Hacker, G. Bhanot, D. A. Gordenin, A. Chu, P. H. Gunaratne, M. Biehl, S. Seth, B. A. Kaipparettu, C. A. Bristow, L. A. Donehower, E. M. Wallen, A. B. Smith, S. K. Tickoo, P. Tamboli, V. Reuter, L. S. Schmidt, J. J. Hsieh, T. K. Choueiri, A. A. Hakimi, N. The Cancer Genome Atlas Research, L. Chin, M. Meyerson, R. Kucherlapati, W. Y. Park, A. G. Robertson, P. W. Laird, E. P. Henske, D. J. Kwiatkowski, P. J. Park, M. Morgan, B. Shuch, D. Muzny, D. A. Wheeler, W. M. Linehan, R. A. Gibbs, W. K. Rathmell & C. J. Creighton (2014) The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell,* 26**,** 319-330.

De Bont, R. & N. van Larebeke (2004) Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis,* 19**,** 169-85.

de Magalhaes, J. P., C. E. Finch & G. Janssens (2010) Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing Res Rev,* 9**,** 315-23.

Dees, N. D., Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson & L. Ding (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res,* 22**,** 1589-98.

DePinho, R. A. & K. Polyak (2004) Cancer chromosomes in crisis. *Nat Genet,* 36**,** 932-4.

Dunn, G. P., A. T. Bruce, H. Ikeda, L. J. Old & R. D. Schreiber (2002) Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol,* 3**,** 991-8.

Dunn, G. P., L. J. Old & R. D. Schreiber (2004) The three Es of cancer immunoediting. *Annu Rev Immunol,* 22**,** 329-60.

Duro, D., O. Bernard, V. Della Valle, T. Leblanc, R. Berger & C. J. Larsen (1996) Inactivation of the P16INK4/MTS1 gene by a chromosome translocation t(9;14)(p21-22;q11) in an acute lymphoblastic leukemia of B-cell type. *Cancer Res,* 56**,** 848-54.

Emery, C. M., K. G. Vijayendran, M. C. Zipser, A. M. Sawyer, L. Niu, J. J. Kim, C. Hatton, R. Chopra, P. A. Oberholzer, M. B. Karpova, L. E. MacConaill, J. Zhang, N. S. Gray, W. R. Sellers, R. Dummer & L. A. Garraway (2009) MEK1 mutations confer resistance to MEK and B-RAF inhibition. *Proc Natl Acad Sci U S A,* 106**,** 20411-6.

Erikson, J., L. Finger, L. Sun, A. ar-Rushdi, K. Nishikura, J. Minowada, J. Finan, B. S. Emanuel, P. C. Nowell & C. M. Croce (1986) Deregulation of c-myc by translocation of the alpha-locus of the T-cell receptor in T-cell leukemias. *Science,* 232**,** 884-6.

Escot, C., C. Theillet, R. Lidereau, F. Spyratos, M. H. Champeme, J. Gest & R. Callahan (1986) Genetic alteration of the c-myc protooncogene

(MYC) in human primary breast carcinomas. *Proc Natl Acad Sci U S A,* 83**,** 4834-8.

Ferber, M. J., D. P. Montoya, C. Yu, I. Aderca, A. McGee, E. C. Thorland, D. M. Nagorney, B. S. Gostout, L. J. Burgart, L. Boix, J. Bruix, B. J. McMahon, T. H. Cheung, T. K. Chung, Y. F. Wong, D. I. Smith & L. R. Roberts (2003) Integrations of the hepatitis B virus (HBV) and human papillomavirus (HPV) into the human telomerase reverse transcriptase (hTERT) gene in liver and cervical cancers. *Oncogene,* 22**,** 3813-20.

Fernald, K. & M. Kurokawa (2013) Evading apoptosis in cancer. *Trends Cell Biol,* 23**,** 620-33.

Fernandez-Medarde, A. & E. Santos (2011) Ras in cancer and developmental diseases. *Genes Cancer,* 2**,** 344-58.

Feuk, L., A. R. Carson & S. W. Scherer (2006) Structural variation in the human genome. *Nat Rev Genet,* 7**,** 85-97.

Finver, S. N., K. Nishikura, L. R. Finger, F. G. Haluska, J. Finan, P. C. Nowell & C. M. Croce (1988) Sequence analysis of the MYC oncogene involved in the t(8;14)(q24;q11) chromosome translocation in a human leukemia T-cell line indicates that putative regulatory regions are not altered. *Proc Natl Acad Sci U S A,* 85**,** 3052-6.

Finzer, P., A. Aguilar-Lemarroy & F. Rosl (2002) The role of human papillomavirus oncoproteins E6 and E7 in apoptosis. *Cancer Lett,* 188**,** 15-24.

Fredriksson, N. J., L. Ny, J. A. Nilsson & E. Larsson (2014) Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet,* 46**,** 1258-63.

Friend, S. H., R. Bernards, S. Rogelj, R. A. Weinberg, J. M. Rapaport, D. M. Albert & T. P. Dryja (1986) A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature,* 323**,** 643-6.

Fukuhara, S., J. D. Rowley, D. Variakojis & H. M. Golomb (1979) Chromosome abnormalities in poorly differentiated lymphocytic lymphoma. *Cancer Res,* 39**,** 3119-28.

Gabay, M., Y. Li & D. W. Felsher (2014) MYC activation is a hallmark of cancer initiation and maintenance. *Cold Spring Harb Perspect Med,* 4.

Ganem, D. & A. M. Prince (2004) Hepatitis B virus infection--natural history and clinical consequences. *N Engl J Med,* 350**,** 1118-29.

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

Gonzalez-Perez, A., C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos & N. Lopez-Bigas (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods,* 10**,** 1081-2.

Goodsell, D. S. (1999) The molecular perspective: the ras oncogene. *Stem Cells,* 17**,** 235-6.

Goodwin, S., J. D. McPherson & W. R. McCombie (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-51.

Green, D. R. 2011. *Means to an end : apoptosis and other cell death mechanisms*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Greenman, C., P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y. E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M. H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal & M. R. Stratton (2007) Patterns of somatic mutation in human cancer genomes. *Nature,* 446**,** 153-8.

Grimaldi, J. C. & T. C. Meeker (1989) The t(5;14) chromosomal translocation in a case of acute lymphocytic leukemia joins the interleukin-3 gene to the immunoglobulin heavy chain gene. *Blood,* 73**,** 2081-5.

Grivennikov, S. I., F. R. Greten & M. Karin (2010) Immunity, inflammation, and cancer. *Cell,* 140**,** 883-99.

Groschel, S., M. A. Sanders, R. Hoogenboezem, E. de Wit, B. A. M. Bouwman, C. Erpelinck, V. H. J. van der Velden, M. Havermans, R. Avellino, K. van Lom, E. J. Rombouts, M. van Duin, K. Dohner, H. B. Beverloo, J. E. Bradner, H. Dohner, B. Lowenberg, P. J. M. Valk, E. M. J. Bindels, W. de Laat & R. Delwel (2014) A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell,* 157**,** 369-381.

Gu, W., F. Zhang & J. R. Lupski (2008) Mechanisms for human genomic rearrangements. *Pathogenetics,* 1**,** 4.

Gupta, G. P. & J. Massague (2006) Cancer metastasis: building a framework. *Cell,* 127**,** 679-95.

Hanada, N., T. Takahata, Q. Zhou, X. Ye, R. Sun, J. Itoh, A. Ishiguro, H. Kijima, J. Mimura, K. Itoh, S. Fukuda & Y. Saijo (2012) Methylation of the KEAP1 gene promoter region in human colorectal cancer. *BMC Cancer,* 12**,** 66.

Hanahan, D. & R. A. Weinberg (2000) The hallmarks of cancer. *Cell,* 100**,** 57-70.

--- (2011a) Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.

--- (2011b) Hallmarks of Cancer: The Next Generation. *Cell,* 144**,** 646-674.
Hayflick, L. (1965) The Limited in Vitro Lifetime of Human Diploid Cell Strains. *Exp Cell Res,* 37**,** 614-36.
Hayflick, L. & P. S. Moorhead (1961) The serial cultivation of human diploid cell strains. *Exp Cell Res,* 25**,** 585-621.
Hayward, W. S., B. G. Neel & S. M. Astrin (1981) Activation of a cellular onc gene by promoter insertion in ALV-induced lymphoid leukosis. *Nature,* 290**,** 475-80.
Heidenreich, B., P. S. Rachakonda, K. Hemminki & R. Kumar (2014) TERT promoter mutations in cancer development. *Curr Opin Genet Dev,* 24**,** 30-7.
Heller, M. J. (2002) DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng,* 4**,** 129-53.
Hu, X., J. Yuan, Y. Shi, J. Lu, B. Liu, Z. Li, Y. Chen, D. Mu, H. Zhang, N. Li, Z. Yue, F. Bai, H. Li & W. Fan (2012) pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics,* 28**,** 1533-5.
Huang, F. W., E. Hodis, M. J. Xu, G. V. Kryukov, L. Chin & L. A. Garraway (2013) Highly recurrent TERT promoter mutations in human melanoma. *Science,* 339**,** 957-9.
Hunter, C., R. Smith, D. P. Cahill, P. Stephens, C. Stevens, J. Teague, C. Greenman, S. Edkins, G. Bignell, H. Davies, S. O'Meara, A. Parker, T. Avis, S. Barthorpe, L. Brackenbury, G. Buck, A. Butler, J. Clements, J. Cole, E. Dicks, S. Forbes, M. Gorton, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, V. Kosmidou, R. Laman, R. Lugg, A. Menzies, J. Perry, R. Petty, K. Raine, D. Richardson, R. Shepherd, A. Small, H. Solomon, C. Tofts, J. Varian, S. West, S. Widaa, A. Yates, D. F. Easton, G. Riggins, J. E. Roy, K. K. Levine, W. Mueller, T. T. Batchelor, D. N. Louis, M. R. Stratton, P. A. Futreal & R. Wooster (2006) A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res,* 66**,** 3987-91.
Ida, K., S. Kobayashi, T. Taki, R. Hanada, F. Bessho, S. Yamamori, T. Sugimoto, M. Ohki & Y. Hayashi (1995) EWS-FLI-1 and EWS-ERG chimeric mRNAs in Ewing's sarcoma and primitive neuroectodermal tumor. *Int J Cancer,* 63**,** 500-4.
Kallioniemi, A., O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman & D. Pinkel (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science,* 258**,** 818-21.
Kansanen, E., S. M. Kuosmanen, H. Leinonen & A. L. Levonen (2013) The Keap1-Nrf2 pathway: Mechanisms of activation and dysregulation in cancer. *Redox Biol,* 1**,** 45-9.
Kim, D. & S. L. Salzberg (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol,* 12**,** R72.

Knudson, A. G., Jr. (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A,* 68**,** 820-3.

Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding & R. K. Wilson (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res,* 22**,** 568-76.

Korbel, J. O., A. Abyzov, X. J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder & M. B. Gerstein (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol,* 10**,** R23.

Korkola, J. & J. W. Gray (2010) Breast cancer genomes--form and function. *Curr Opin Genet Dev,* 20**,** 4-14.

Langmead, B., C. Trapnell, M. Pop & S. L. Salzberg (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.

Leder, P., J. Battey, G. Lenoir, C. Moulding, W. Murphy, H. Potter, T. Stewart & R. Taub (1983) Translocations among antibody genes in human cancer. *Science,* 222**,** 765-71.

Lee, J. A., C. M. Carvalho & J. R. Lupski (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell,* 131**,** 1235-47.

Lee, J. H., Y. H. Soung, J. W. Lee, W. S. Park, S. Y. Kim, Y. G. Cho, C. J. Kim, S. H. Seo, H. S. Kim, S. W. Nam, N. J. Yoo, S. H. Lee & J. Y. Lee (2004) Inactivating mutation of the pro-apoptotic gene BID in gastric cancer. *J Pathol,* 202**,** 439-45.

Leung, D. W., G. Cachianes, W. J. Kuang, D. V. Goeddel & N. Ferrara (1989) Vascular endothelial growth factor is a secreted angiogenic mitogen. *Science,* 246**,** 1306-9.

Levine, R. L., M. Wadleigh, J. Cools, B. L. Ebert, G. Wernig, B. J. Huntly, T. J. Boggon, I. Wlodarska, J. J. Clark, S. Moore, J. Adelsperger, S. Koo, J. C. Lee, S. Gabriel, T. Mercher, A. D'Andrea, S. Frohling, K. Dohner, P. Marynen, P. Vandenberghe, R. A. Mesa, A. Tefferi, J. D. Griffin, M. J. Eck, W. R. Sellers, M. Meyerson, T. R. Golub, S. J. Lee & D. G. Gilliland (2005) Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell,* 7**,** 387-97.

Li, H. & R. Durbin (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* 26**,** 589-95.

Li, W. & M. Olivier (2013) Current analysis platforms and methods for detecting copy number variation. *Physiol Genomics,* 45**,** 1-16.

Lieber, M. R. (2008) The mechanism of human nonhomologous DNA end joining. *J Biol Chem,* 283**,** 1-5.

Linn, D. E., K. L. Penney, R. T. Bronson, L. A. Mucci & Z. Li (2016) Deletion of Interstitial Genes between TMPRSS2 and ERG Promotes Prostate Cancer Progression. *Cancer Res,* 76**,** 1869-81.

Liu, L., A. G. Xu, Q. L. Zhang, Y. F. Zhang & Y. Q. Ding (2007) [Effect of Tiam1 overexpression on proliferation and metastatic potential of human colorectal cancer]. *Zhonghua Bing Li Xue Za Zhi,* 36**,** 390-3.

Liu, P., C. M. Carvalho, P. J. Hastings & J. R. Lupski (2012) Mechanisms for recurrent and complex human genomic rearrangements. *Curr Opin Genet Dev,* 22**,** 211-20.

Lunt, S. Y. & M. G. Vander Heiden (2011) Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. *Annu Rev Cell Dev Biol,* 27**,** 441-64.

Lupski, J. R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet,* 14**,** 417-22.

Martin, G. S. (2003) Cell signaling and cancer. *Cancer Cell,* 4**,** 167-74.

Maston, G. A., S. K. Evans & M. R. Green (2006) Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet,* 7**,** 29-59.

McBride, D. J., D. Etemadmoghadam, S. L. Cooke, K. Alsop, J. George, A. Butler, J. Cho, D. Galappaththige, C. Greenman, K. D. Howarth, K. W. Lau, C. K. Ng, K. Raine, J. Teague, D. C. Wedge, A. O. Cancer Study Group, X. Caubit, M. R. Stratton, J. D. Brenton, P. J. Campbell, P. A. Futreal & D. D. Bowtell (2012) Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes. *J Pathol,* 227**,** 446-55.

McEachern, M. J., A. Krauskopf & E. H. Blackburn (2000) Telomeres and their control. *Annu Rev Genet,* 34**,** 331-358.

McKay, J. A., L. J. Murray, S. Curran, V. G. Ross, C. Clark, G. I. Murray, J. Cassidy & H. L. McLeod (2002) Evaluation of the epidermal growth factor receptor (EGFR) in colorectal tumours and lymph node metastases. *Eur J Cancer,* 38**,** 2258-64.

McVey, M. & S. E. Lee (2008) MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet,* 24**,** 529-38.

Medvedev, P., M. Stanciu & M. Brudno (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods,* 6**,** S13-20.

Mermel, C. H., S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim & G. Getz (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol,* 12**,** R41.

Mertens, F., B. Johansson, T. Fioretos & F. Mitelman (2015) The emerging complexity of gene fusions in cancer. *Nat Rev Cancer,* 15**,** 371-81.

Meyerson, M., S. Gabriel & G. Getz (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet,* 11**,** 685-96.

Mitri, Z., T. Constantine & R. O'Regan (2012) The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. *Chemother Res Pract,* 2012**,** 743193.

Mohiyuddin, M., J. C. Mu, J. Li, N. Bani Asadi, M. B. Gerstein, A. Abyzov, W. H. Wong & H. Y. Lam (2015) MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics,* 31**,** 2741-4.

Moore, J. K. & J. E. Haber (1996) Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol Cell Biol,* 16**,** 2164-73.

Mork, J., A. K. Lie, E. Glattre, G. Hallmans, E. Jellum, P. Koskela, B. Moller, E. Pukkala, J. T. Schiller, L. Youngman, M. Lehtinen & J. Dillner (2001) Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck. *N Engl J Med,* 344**,** 1125-31.

Munger, K., A. Baldwin, K. M. Edwards, H. Hayakawa, C. L. Nguyen, M. Owens, M. Grace & K. Huh (2004) Mechanisms of human papillomavirus-induced oncogenesis. *J Virol,* 78**,** 11451-60.

Nazarian, R., H. Shi, Q. Wang, X. Kong, R. C. Koya, H. Lee, Z. Chen, M. K. Lee, N. Attar, H. Sazegar, T. Chodon, S. F. Nelson, G. McArthur, J. A. Sosman, A. Ribas & R. S. Lo (2010) Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature,* 468**,** 973-7.

Neel, B. G., W. S. Hayward, H. L. Robinson, J. Fang & S. M. Astrin (1981) Avian leukosis virus-induced tumors have common proviral integration sites and synthesize discrete new RNAs: oncogenesis by promoter insertion. *Cell,* 23**,** 323-34.

Negrini, S., V. G. Gorgoulis & T. D. Halazonetis (2010) Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol,* 11**,** 220-8.

Newman, A. M., S. V. Bratman, H. Stehr, L. J. Lee, C. L. Liu, M. Diehn & A. A. Alizadeh (2014) FACTERA: a practical method for the discovery of genomic rearrangements at breakpoint resolution. *Bioinformatics,* 30**,** 3390-3.

Nishida, N., H. Yano, T. Nishida, T. Kamura & M. Kojiro (2006) Angiogenesis in cancer. *Vasc Health Risk Manag,* 2**,** 213-9.

Noorani, A., J. Bornschein, A. G. Lynch, M. Secrier, A. Achilleos, M. Eldridge, L. Bower, J. M. J. Weaver, J. Crawte, C. A. Ong, N. Shannon, S. MacRae, N. Grehan, B. Nutzinger, M. O'Donovan, R. Hardwick, S. Tavare, R. C. Fitzgerald, C. Oesophageal Cancer & C. Molecular Stratification (2017) A comparative analysis of whole genome sequencing of esophageal adenocarcinoma pre- and post-chemotherapy. *Genome Res,* 27**,** 902-912.

Normanno, N., A. De Luca, C. Bianco, L. Strizzi, M. Mancino, M. R. Maiello, A. Carotenuto, G. De Feo, F. Caponigro & D. S. Salomon (2006) Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene,* 366**,** 2-16.

Northcott, P. A., C. Lee, T. Zichner, A. M. Stutz, S. Erkek, D. Kawauchi, D. J. Shih, V. Hovestadt, M. Zapatka, D. Sturm, D. T. Jones, M. Kool, M. Remke, F. M. Cavalli, S. Zuyderduyn, G. D. Bader, S. VandenBerg, L. A. Esparza, M. Ryzhova, W. Wang, A. Wittmann, S. Stark, L. Sieber, H. Seker-Cin, L. Linke, F. Kratochwil, N. Jager, I. Buchhalter, C. D. Imbusch, G. Zipprich, B. Raeder, S. Schmidt, N. Diessl, S. Wolf, S. Wiemann, B. Brors, C. Lawerenz, J. Eils, H. J. Warnatz, T. Risch, M. L. Yaspo, U. D. Weber, C. C. Bartholomae, C. von Kalle, E. Turanyi, P. Hauser, E. Sanden, A. Darabi, P. Siesjo, J. Sterba, K. Zitterbart, D. Sumerauer, P. van Sluis, R. Versteeg, R. Volckmann, J. Koster, M. U. Schuhmann, M. Ebinger, H. L. Grimes, G. W. Robinson, A. Gajjar, M. Mynarek, K. von Hoff, S. Rutkowski, T. Pietsch, W. Scheurlen, J. Felsberg, G. Reifenberger, A. E. Kulozik, A. von Deimling, O. Witt, R. Eils, R. J. Gilbertson, A. Korshunov, M. D. Taylor, P. Lichter, J. O. Korbel, R. J. Wechsler-Reya & S. M. Pfister (2014) Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature,* 511**,** 428-34.

Nosek, J., P. Kosa & L. Tomaska (2006) On the origin of telomeres: a glimpse at the pre-telomerase world. *Bioessays,* 28**,** 182-90.

Nowell, P. C. (1962) The minute chromosome (Phl) in chronic granulocytic leukemia. *Blut,* 8**,** 65-6.

Ohta, T., K. Iijima, M. Miyamoto, I. Nakahara, H. Tanaka, M. Ohtsuji, T. Suzuki, A. Kobayashi, J. Yokota, T. Sakiyama, T. Shibata, M. Yamamoto & S. Hirohashi (2008) Loss of Keap1 function activates Nrf2 and provides advantages for lung cancer cell growth. *Cancer Res,* 68**,** 1303-9.

Oliveira, A. M., A. R. Perez-Atayde, P. Dal Cin, M. C. Gebhardt, C. J. Chen, J. R. Neff, G. D. Demetri, A. E. Rosenberg, J. A. Bridge & J. A. Fletcher (2005) Aneurysmal bone cyst variant translocations upregulate USP6 transcription by promoter swapping with the ZNF9, COL1A1, TRAP150, and OMD genes. *Oncogene,* 24**,** 3419-26.

Olivier, M., M. Hollstein & P. Hainaut (2010) TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol,* 2**,** a001008.

Olshen, A. B., E. S. Venkatraman, R. Lucito & M. Wigler (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics,* 5**,** 557-72.

Oshimura, M., A. I. Freeman & A. A. Sandberg (1977) Chromosomes and causation of human cancer and leukemia. XXVI. Binding studies in acute lymphoblastic leukemia (ALL). *Cancer,* 40**,** 1161-72.

Ottaviani, D., M. LeCain & D. Sheer (2014) The role of microhomology in genomic structural variation. *Trends Genet,* 30**,** 85-94.

Parri, M. & P. Chiarugi (2010) Rac and Rho GTPases in cancer cell motility control. *Cell Commun Signal,* 8**,** 23.

Pascual, G., A. Avgustinova, S. Mejetta, M. Martin, A. Castellanos, C. S. Attolini, A. Berenguer, N. Prats, A. Toll, J. A. Hueto, C. Bescos, L. Di Croce & S. A. Benitah (2017) Targeting metastasis-initiating cells through the fatty acid receptor CD36. *Nature,* 541**,** 41-45.

Paul, D., K. D. Brown, H. T. Rupniak & H. J. Ristow (1978) Cell-Cycle Regulation by Growth-Factors and Nutrients in Normal and Transformed-Cells. *In Vitro-Journal of the Tissue Culture Association,* 14**,** 76-84.

Peeters, P., S. D. Raynaud, J. Cools, I. Wlodarska, J. Grosgeorge, P. Philip, F. Monpoux, L. Van Rompaey, M. Baens, H. Van den Berghe & P. Marynen (1997) Fusion of TEL, the ETS-variant gene 6 (ETV6), to the receptor-associated kinase JAK2 as a result of t(9;12) in a lymphoid and t(9;15;12) in a myeloid leukemia. *Blood,* 90**,** 2535-40.

Peter, M., C. Rosty, J. Couturier, F. Radvanyi, H. Teshima & X. Sastre-Garau (2006) MYC activation associated with the integration of HPV DNA at the MYC locus in genital tumors. *Oncogene,* 25**,** 5985-93.

Plaszczyca, A., J. Nilsson, L. Magnusson, O. Brosjo, O. Larsson, F. Vult von Steyern, H. A. Domanski, H. Lilljebjorn, T. Fioretos, J. Tayebwa, N. Mandahl, K. H. Nord & F. Mertens (2014) Fusions involving protein kinase C and membrane-associated proteins in benign fibrous histiocytoma. *Int J Biochem Cell Biol,* 53**,** 475-81.

Qin, M., B. Liu, J. M. Conroy, C. D. Morrison, Q. Hu, Y. Cheng, M. Murakami, A. O. Odunsi, C. S. Johnson, L. Wei, S. Liu & J. Wang (2015) SCNVSim: somatic copy number variation and structure variation simulator. *BMC Bioinformatics,* 16**,** 66.

Rabbitts, T. H. (1994) Chromosomal translocations in human cancer. *Nature,* 372**,** 143-9.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stutz, V. Benes & J. O. Korbel (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics,* 28**,** i333-i339.

Reddy, E. P., R. K. Reynolds, E. Santos & M. Barbacid (1982) A Point Mutation Is Responsible for the Acquisition of Transforming Properties by the T24 Human Bladder-Carcinoma Oncogene. *Nature,* 300**,** 149-152.

Rode, A., K. K. Maass, K. V. Willmund, P. Lichter & A. Ernst (2016) Chromothripsis in cancer cells: An update. *Int J Cancer,* 138**,** 2322-33.

Rous, P. (1911) A Sarcoma of the Fowl Transmissible by an Agent Separable from the Tumor Cells. *J Exp Med,* 13**,** 397-411.

Rowley, J. D. (1973) Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature,* 243**,** 290-3.

Rowley, J. D., H. M. Golomb & C. Dougherty (1977) 15/17 translocation, a consistent chromosomal change in acute promyelocytic leukaemia. *Lancet,* 1**,** 549-50.

Rubio-Perez, C., D. Tamborero, M. P. Schroeder, A. A. Antolin, J. Deu-Pons, C. Perez-Llamas, J. Mestres, A. Gonzalez-Perez & N. Lopez-Bigas (2015) In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell,* 27**,** 382-96.

Samuels, Y., Z. Wang, A. Bardelli, N. Silliman, J. Ptak, S. Szabo, H. Yan, A. Gazdar, S. M. Powell, G. J. Riggins, J. K. Willson, S. Markowitz, K. W. Kinzler, B. Vogelstein & V. E. Velculescu (2004) High frequency of mutations of the PIK3CA gene in human cancers. *Science,* 304**,** 554.

Sanger, F., S. Nicklen & A. R. Coulson (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74**,** 5463-7.

Savelyeva, L. & M. Schwab (2001) Amplification of oncogenes revisited: from expression profiling to clinical application. *Cancer Lett,* 167**,** 115-23.

Schena, M., D. Shalon, R. W. Davis & P. O. Brown (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* 270**,** 467-70.

Schmitz, M., C. Driesch, L. Jansen, I. B. Runnebaum & M. Durst (2012) Non-random integration of the HPV genome in cervical cancer. *PLoS One,* 7**,** e39632.

Schroder, J., A. Hsu, S. E. Boyle, G. Macintyre, M. Cmero, R. W. Tothill, R. W. Johnstone, M. Shackleton & A. T. Papenfuss (2014) Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics,* 30**,** 1064-1072.

Serrano, M., G. J. Hannon & D. Beach (1993) A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. *Nature,* 366**,** 704-7.

Shaw, C. J. & J. R. Lupski (2004) Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol Genet,* 13 Spec No 1**,** R57-64.

Shaw-Smith, C., R. Redon, L. Rickman, M. Rio, L. Willatt, H. Fiegler, H. Firth, D. Sanlaville, R. Winter, L. Colleaux, M. Bobrow & N. P. Carter (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J Med Genet,* 41**,** 241-8.

Shen, M. M. (2013) Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer Cell,* 23**,** 567-9.

Sherr, C. J. & F. McCormick (2002) The RB and p53 pathways in cancer. *Cancer Cell,* 2**,** 103-12.

Shibata, T., T. Ohta, K. I. Tong, A. Kokubu, R. Odogawa, K. Tsuta, H. Asamura, M. Yamamoto & S. Hirohashi (2008) Cancer related mutations in NRF2 impair its recognition by Keap1-Cul3 E3 ligase and promote malignancy. *Proc Natl Acad Sci U S A,* 105**,** 13568-73.

Shtivelman, E., B. Lifshitz, R. P. Gale & E. Canaani (1985) Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature,* 315**,** 550-4.

Silke, J. & P. Meier (2013) Inhibitor of apoptosis (IAP) proteins-modulators of cell death and inflammation. *Cold Spring Harb Perspect Biol,* 5.

Sindi, S., E. Helman, A. Bashir & B. J. Raphael (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics,* 25**,** i222-30.

Sindi, S. S., S. Onal, L. C. Peng, H. T. Wu & B. J. Raphael (2012) An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol,* 13**,** R22.

Smith, C. E., B. Llorente & L. S. Symington (2007) Template switching during break-induced replication. *Nature,* 447**,** 102-5.

Solinas-Toldo, S., S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Dohner, T. Cremer & P. Lichter (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer,* 20**,** 399-407.

Sorensen, P. H. & T. J. Triche (1996) Gene fusions encoding chimaeric transcription factors in solid tumours. *Semin Cancer Biol,* 7**,** 3-14.

Stam, K., N. Heisterkamp, G. Grosveld, A. de Klein, R. S. Verma, M. Coleman, H. Dosik & J. Groffen (1985) Evidence of a new chimeric bcr/c-abl mRNA in patients with chronic myelocytic leukemia and the Philadelphia chromosome. *N Engl J Med,* 313**,** 1429-33.

Stephens, P. J., C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M. L. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Morsberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal & P. J. Campbell (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell,* 144**,** 27-40.

Storlazzi, C. T., A. Lonoce, M. C. Guastadisegni, D. Trombetta, P. D'Addabbo, G. Daniele, A. L'Abbate, G. Macchia, C. Surace, K. Kok, R. Ullmann, S. Purgato, O. Palumbo, M. Carella, P. F. Ambros & M. Rocchi (2010) Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res,* 20**,** 1198-206.

Storlazzi, C. T., F. V. Von Steyern, H. A. Domanski, N. Mandahl & F. Mertens (2005) Biallelic somatic inactivation of the NF1 gene through chromosomal translocations in a sporadic neurofibroma. *Int J Cancer,* 117**,** 1055-7.

Stransky, N., E. Cerami, S. Schalm, J. L. Kim & C. Lengauer (2014) The landscape of kinase fusions in cancer. *Nat Commun,* 5**,** 4846.

Stratton, M. R., P. J. Campbell & P. A. Futreal (2009) The cancer genome. *Nature,* 458**,** 719-24.

Sung, W. K., H. Zheng, S. Li, R. Chen, X. Liu, Y. Li, N. P. Lee, W. H. Lee, P. N. Ariyaratne, C. Tennakoon, F. H. Mulawadi, K. F. Wong, A. M. Liu, R. T. Poon, S. T. Fan, K. L. Chan, Z. Gong, Y. Hu, Z. Lin, G. Wang, Q. Zhang, T. D. Barber, W. C. Chou, A. Aggarwal, K. Hao, W. Zhou, C. Zhang, J. Hardwick, C. Buser, J. Xu, Z. Kan, H. Dai, M. Mao, C. Reinhard, J. Wang & J. M. Luk (2012) Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet,* 44**,** 765-9.

Suwaki, N., K. Klare & M. Tarsounas (2011) RAD51 paralogs: roles in DNA damage signalling, recombinational repair and tumorigenesis. *Semin Cell Dev Biol,* 22**,** 898-905.

Suzuki, S., T. Yasuda, Y. Shiraishi, S. Miyano & M. Nagasaki (2011) ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics,* 12 Suppl 14**,** S7.

Tamborero, D., A. Gonzalez-Perez & N. Lopez-Bigas (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics,* 29**,** 2238-44.

Tomlins, S. A., D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin & A. M. Chinnaiyan (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science,* 310**,** 644-8.

Trapnell, C., L. Pachter & S. L. Salzberg (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics,* 25**,** 1105-11.

Tsaponina, O. & J. E. Haber (2014) Frequent Interchromosomal Template Switches during Gene Conversion in S. cerevisiae. *Mol Cell,* 55**,** 615-25.

Venkitaraman, A. R. (2002) Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell,* 108**,** 171-82.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M.

Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, et al. (2001) The sequence of the human genome. *Science,* 291**,** 1304-51.

Vicario, R., V. Peg, B. Morancho, M. Zacarias-Fluck, J. Zhang, A. Martinez-Barriocanal, A. Navarro Jimenez, C. Aura, O. Burgues, A. Lluch, J. Cortes, P. Nuciforo, I. T. Rubio, E. Marangoni, J. Deeds, M. Boehm, R. Schlegel, J. Tabernero, R. Mosher & J. Arribas (2015) Patterns of HER2 Gene Amplification and Response to Anti-HER2 Therapies. *PLoS One,* 10**,** e0129876.

Visconti, R. & D. Grieco (2009) New insights on oxidative stress in cancer. *Curr Opin Drug Discov Devel,* 12**,** 240-5.

Vogt, N., S. H. Lefevre, F. Apiou, A. M. Dutrillaux, A. Cor, P. Leuraud, M. F. Poupon, B. Dutrillaux, M. Debatisse & B. Malfoy (2004) Molecular structure of double-minute chromosomes bearing amplified copies of the epidermal growth factor receptor gene in gliomas. *Proc Natl Acad Sci U S A,* 101**,** 11368-73.

Voldborg, B. R., L. Damstrup, M. Spang-Thomsen & H. S. Poulsen (1997) Epidermal growth factor receptor (EGFR) and EGFR mutations, function and possible role in clinical trials. *Ann Oncol,* 8**,** 1197-206.

Wang, H. M. & J. Wang (2012) Expression of Tiam1 in lung cancer and its clinical significance. *Asian Pac J Cancer Prev,* 13**,** 613-5.

Wang, J., C. G. Mullighan, J. Easton, S. Roberts, S. L. Heatley, J. Ma, M. C. Rusch, K. Chen, C. C. Harris, L. Ding, L. Holmfeldt, D. Payne-Turner, X. Fan, L. Wei, D. Zhao, J. C. Obenauer, C. Naeve, E. R. Mardis, R. K. Wilson, J. R. Downing & J. Zhang (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods,* 8**,** 652-4.

Warburg, O., F. Wind & E. Negelein (1927) The Metabolism of Tumors in the Body. *J Gen Physiol,* 8**,** 519-30.

Watson, J. D. & F. H. C. Crick (1953) Molecular Structure of Nucleic Acids - a Structure for Deoxyribose Nucleic Acid. *Nature,* 171**,** 737-738.

Weinberg, R. A. 2007. *The biology of cancer*. New York: Garland Science.

Weischenfeldt, J., T. Dubash, A. P. Drainas, B. R. Mardin, Y. Chen, A. M. Stutz, S. M. Waszak, G. Bosco, A. R. Halvorsen, B. Raeder, T. Efthymiopoulos, S. Erkek, C. Siegl, H. Brenner, O. T. Brustugun, S.

M. Dieter, P. A. Northcott, I. Petersen, S. M. Pfister, M. Schneider, S. K. Solberg, E. Thunissen, W. Weichert, T. Zichner, R. Thomas, M. Peifer, A. Helland, C. R. Ball, M. Jechlinger, R. Sotillo, H. Glimm & J. O. Korbel (2017) Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat Genet,* 49**,** 65-74.

Witsch, E., M. Sela & Y. Yarden (2010) Roles for growth factors in cancer progression. *Physiology (Bethesda),* 25**,** 85-101.

Witzany, G. (2008) The Viral Origins of Telomeres and Telomerases and their Important Role in Eukaryogenesis and Genome Maintenance. *Biosemiotics,* 1**,** 191-206.

Yang, L., L. J. Luquette, N. Gehlenborg, R. Xi, P. S. Haseley, C. H. Hsieh, C. Zhang, X. Ren, A. Protopopov, L. Chin, R. Kucherlapati, C. Lee & P. J. Park (2013a) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell,* 153**,** 919-29.

Yang, T. Y., S. C. Chen, M. W. Leach, D. Manfra, B. Homey, M. Wiekowski, L. Sullivan, C. H. Jenh, S. K. Narula, S. W. Chensue & S. A. Lira (2000) Transgenic expression of the chemokine receptor encoded by human herpesvirus 8 induces an angioproliferative disease resembling Kaposi's sarcoma. *J Exp Med,* 191**,** 445-54.

Yang, X., S. P. Chockalingam & S. Aluru (2013b) A survey of error-correction methods for next-generation sequencing. *Brief Bioinform,* 14**,** 56-66.

Yarbro, J. W. (1992) Oncogenes and cancer suppressor genes. *Semin Oncol Nurs,* 8**,** 30-9.

Yau, C., D. Mouradov, R. N. Jorissen, S. Colella, G. Mirza, G. Steers, A. Harris, J. Ragoussis, O. Sieber & C. C. Holmes (2010) A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol,* 11**,** R92.

Yip, K. W. & J. C. Reed (2008) Bcl-2 family proteins and cancer. *Oncogene,* 27**,** 6398-406.

Yoshida, K. & Y. Miki (2004) Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Sci,* 95**,** 866-71.

Zare, F., M. Dow, N. Monteleone, A. Hosny & S. Nabavi (2017) An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics,* 18**,** 286.

Zech, L., U. Haglund, K. Nilsson & G. Klein (1976) Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with Burkitt and non-Burkitt lymphomas. *Int J Cancer,* 17**,** 47-56.

Zeitouni, B., V. Boeva, I. Janoueix-Lerosey, S. Loeillet, P. Legoix-ne, A. Nicolas, O. Delattre & E. Barillot (2010) SVDetect: a tool to identify

genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics,* 26**,** 1895-6.

Zhao, Z. Y., C. G. Han, J. T. Liu, C. L. Wang, Y. Wang & L. Y. Cheng (2013) TIAM2 enhances non-small cell lung cancer cell invasion and motility. *Asian Pac J Cancer Prev,* 14**,** 6305-9.

zur Hausen, H. (2002) Papillomaviruses and cancer: from basic studies to clinical application. *Nat Rev Cancer,* 2**,** 342-50.

Zwick, E., J. Bange & A. Ullrich (2001) Receptor tyrosine kinase signalling as a target for cancer intervention strategies. *Endocr Relat Cancer,* 8**,** 161-73.

# APPENDIX