EVOLUTION OF BLADDER CANCER INVESTIGATED USING EXOME SEQUENCING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

EZGİ ÖZKURT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOLOGY

JUNE 2015

Approval of the thesis:

EVOLUTION OF BLADDER CANCER INVESTIGATED USING EXOME SEQUENCING

submitted by **EZGİ ÖZKURT** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Orhan Adalı _____
Head of Department, **Biology, METU**

Assist. Prof. Dr. Mehmet Somel _____
Supervisor, **Biology Dept., METU**

Examining Committee Members:

Prof. Dr. Ufuk Gündüz _____
Biology Dept., METU

Assist. Prof. Dr. Mehmet Somel _____
Biology Dept., METU

Assoc. Prof. Dr. Yeşim Aydın Son _____
Dept. of Health Informatics, METU

Assoc.Prof. Dr. Özgür Şahin _____
Molecular Biology and Genetics Dept., Bilkent University

Assoc. Prof. Dr. Ergi Deniz Özsoy _____
Biology Dept., Hacettepe University

Date: 15.06.2015

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Ezgi Özkurt

Signature:

ABSTRACT

EVOLUTION OF BLADDER CANCER INVESTIGATED USING EXOME SEQUENCING

Özkurt, Ezgi
M.S., Department of Biology
Supervisor : Assist. Prof. Dr. Mehmet Somel

June 2015, 86 pages

New genome sequencing technologies today allow the study of cancer evolution within individual tissues. In bladder cancer, it is commonly observed that multiple tumours co-occur in a tissue. However, whether these tumours are related (clonal hypothesis) or develop independently but synchronously (field effect hypothesis), was yet unknown. In this study, exome sequencing data was utilized to reveal the origin of multifocal tumours. The data was generated in an experiment where samples from bladder tumour (3 tumour samples per patient) and neighbouring normal mucosa (1 normal sample per patient) from 3 patients were collected and sequenced. The tumour samples were composed of apex and base sections of the tumours. Thousands of single nucleotide variants (SNV) were called in each patient. The phylogenetic trees constructed by SNV datasets of the 2 patients showed a topology consistent with clonal origin hypothesis, with a long shared tumour branch, indicating that the tumours derive from the same origin. The third patient's samples were suspected to be contaminated with neoplastic material, and thus not included in the rest of the analysis. An analysis of SNV types with respect to sequence context revealed that TpC* mutations were particularly enriched on the shared tumour branch, indicative of activity of APOBEC enzymes (single stranded DNA/RNA editing

proteins) causing accumulation of TpC* mutations. Thus, it is hypothesized that a period of APOBEC activity led to accumulation of TpC* mutations, some of which included driver mutations that led to tumour formation, and subsequent separation of the tumours in 2 patients.

Keywords: Next-generation sequencing, population genetics, cancer evolution, phylogeny, genome editing, tumour multifocality

ÖZ

EKZOM DİZİLEME YOLUYLA MESANE KANSERİ EVRİMİNİN ARAŞTIRILMASI

Özkurt, Ezgi
Yüksek Lisans, Biyoloji Bölümü
Tez Yöneticisi : Assist. Prof. Dr. Mehmet Somel

Haziran 2015, 86 sayfa

Günümüzde yeni nesil genom dizileme teknolojileri kanser evrimini bireysel örneklemlerle çalışmaya olanak tanımaktadır. Mesane kanseri vakalarının %30'unda, birden fazla tümör aynı anda gözlemlenmiştir. Fakat, bu tümörlerin bağlantılı olup olmadığı (Klonal Köken Hipotezi) ya da birbirinden bağımsız olarak aynı anda oluşup oluşmadığı (Alan Etkisi Hipotezi) bilinmemektedir. Bu çalışmada, bu tümörlerin kökenini anlamak için ekzom sekanslama yöntemi kullanıldı. Mesane tümörü örnekleri (hasta başına 3 tümör örneği) ve komşu normal mukoza örneği (hasta başına 1 örnek), 3 hastadan toplandı ve dizilendi. Örnekler, tümör kesitlerinin apeks ve baz kısmından alındı. 2 hastanın, tek nükleotit varyant (TNV) datalarından oluşturulan filogenetik ağaçlar, klonal köken hipoteziyle uyumluydu; yani sonuçlar bu hastalarda tümörlerin aynı kökenden geldiği yönündeydi. Öteyandan, üçüncü hastanın örneklerine neoplastik materyal bulaştığı sonucuna varıldı. Dolayısıyla, bu hasta analizden çıkarıldı. Daha sonra, örneklerde TNV motifleri analiz edildi ve daha önceden de literatürde belirtildiği üzere; TpC* mutasyonlarının, özellikle tümör dalında zenginleştiği farkedildi. Bu bulgular ışığında, 2 hastada enfeksiyonu takiben APOBEC enzimleri (tek sarmallı DNA/RNA editleme proteinleri) aktivitesinin TpC* mutasyonu birikimine sebep olduğundan şüphelenildi. Bu mutasyonların

bir kısmının kanseri tetikleyici bölgelerde yer alarak, tümörleşmeye sebep olduğu ve bunu takriben tümörlerin birbirinden ayrıldığı hipotezi öne sürüldü.

Anahtar kelimeler: Yeni nesil dizileme, popülasyon genetiği, kanser evrimi, filogeni, genom editleme, tümör multifokalitesi

ACKNOWLEDGEMENTS

First and foremost, I am deeply thankful to my supervisor Mehmet Somel. He not only encouraged me during my master studies in METU, but also epitomized the idealist in life.

My special thanks go to Nathan Lack and Can Alkan for their great advices and support for this study.

How can I forget about Lab 234 and 205 members. Thank you for sharing your knowledge and friendship with me.

I am more than grateful to my family for their endless support during my life. A special thank you to Emrah Acaröz for being with me at every step of this study and for patience when I descent into pessimism.

Finally, this research was financially supported by Scientific and Technical Research Council of Turkey as well as Turkish Science Academy.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Cancer and Evolution

1.1.1 Heterogeneity & Clonal Evolution

Tumours are any abnormal growth of a group of cells, either confined to their own location or capable of conquering other tissues, classified accordingly either as "benign tumours" or "malignant tumours", respectively (Geoffrey M. Cooper, Elements of Human Cancer, 1992, p.16). Tumour populations are not much different from populations of bacteria, finches, or any other species; where individual units compete with each other for resources, get exposed to predation, sometimes even cooperate, and develop fascinating adaptations (Crespi & Summers, 2005).

A tumour is a mosaic collection of mutant cells, frequently showing large genetic and epigenetic heterogeneity. Tumours actually can be considered as "microcosms of evolution". There exists large phenotypic variation among neoplastic cells within the same tumour, and their fitness (can be defined as average contribution of each tumour cell's genotype to future generation) also differ according to interactions with other cells and with their microenvironment (Merlo, Pepper, Reid, & Maley, 2006) (Crespi & Summers, 2005). The phenotypic variation is mostly heritable, and therefore neoplasms can evolve. Because cancer cells can evolve very rapidly, many cancers are very challenging to cure. In his paper published in 1976, Nowell was the first to suggest a model of evolution for tumour cells (Nowell, 1976). He proposed *clonal evolution* for tumour cell populations and described a sequential

selection process from the very first neoplastic cell formation to the malignancy stage. Since its introduction in 1976, the description of cancer as an evolutionary process gained wide support and was further expanded through other studies. The "clonal evolution" model proposed by Nowell is summarized in Figure 1:

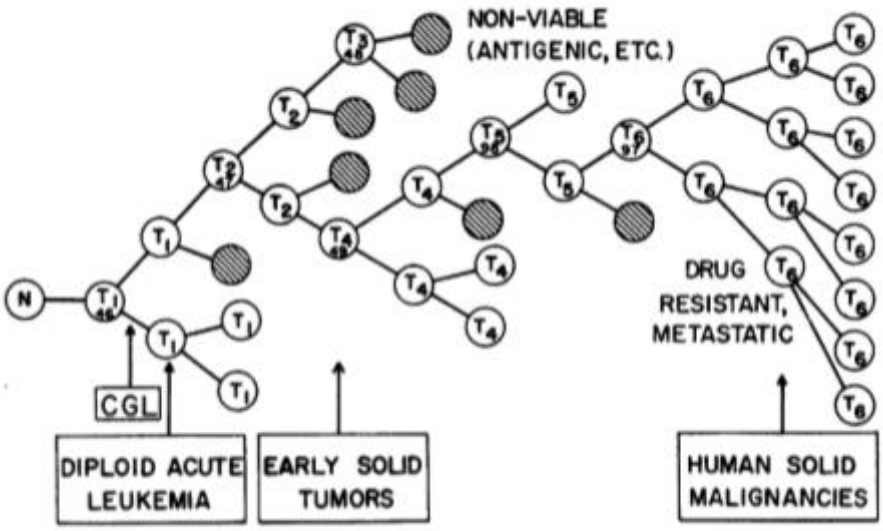

Figure 1: The "clonal evolution" model for tumour cell populations suggested by Nowell.
N represents a progenitor normal cell and $T_{1\text{-}6}$ represent different clones. The numbers inside the circles are chromosome numbers. The hatched circles show variants that died because of metabolic or immunologic disadvantage. Taken from Nowell, 1976.

In the Figure 1, tumour initiation occurs in a single cell (N), which was previously a normal cell. However, it transforms into a neoplastic cell. The neoplastic cell begins to proliferate, but because of genetic instability among the neoplastic cells, different subclones are formed (T1, T2, T4 and T6). For

example, T1 and T6 maintain different chromosome numbers. Some of these clones, T3 for example, is eliminated because of metabolic or immunologic disadvantage. On the other hand, T2 continues to survive until an even fitter clone appears (T4). This selection process results in the fixation of an aneuploid karyotype (n=97) and in malignancy (T6).

1.1.2 Natural Selection in Cancer

Proliferative capacity, orchestrated primarily by growth factors (Witsch, Sela, & Yarden, 2011) is a hallmark of cancer cells (Hanahan & Weinberg, 2011), providing extensive selective advantage to them. While normal cells meticulously control the production and release of growth-promoting signals and thereby maintain homeostasis of cell numbers and tissue architecture; cancer cells acquire the ability to deregulate these signals and develop proliferative capacity in several ways, such as producing their own growth factor ligands or increasing the number of receptor proteins at the surface; thus becoming hypersensitive to limited numbers of growth factor ligands (Paolo et al., 1987). Genetic and epigenetic traits conferring proliferative capacity to tumour cells will thereby be selected and lead to clonal expansion.

In organismal populations, ecological interactions within and among species are driving components of natural selection. Particularly, *predation* and *competition* are such driving interactions. The same ecological interactions also act on somatic evolution of cancer (Crespi & Summers, 2005).

The cellular form of *competition* in nature exists also within the body, among tumour cells. Tumour cells survive in a complex environment; where a clone of cells competes with other clones for resources (Crespi & Summers, 2005). "Glycolytic phenotype" of some tumour cells can be given as an example of the cellular form of adaptation to competition. After a certain stage in their

development, some invasive tumour cells develop this phenotype, which leads to local acidosis that is harmless to the cell itself but toxic to competing cells. Also, the acidified microenvironment facilitates the destruction of the neighbouring normal cell populations and the extracellular matrix (Gatenby & Gillies, 2004). The phenotype can thus be considered as "adaptive", providing a proliferative advantage to the cell.

The cellular analogue of *predation* is immune attack where there is continual immunosurveillance for cells recognized as aberrant (Jakóbisiak, Lasek, & Gołb, 2003). Moreover, cancer cells, in just the same way as natural populations' adaptation to their predators, can evolve adaptations against immunosurveillance. Inadvertently, however, the immune system will select for tumour cells that are less immunogenic (Dunn et al., 2004). For example, some tumours develop in hypoxic (low oxygen) environments before getting vascularized, and this represents an adaptation as such environments are not reachable by the immune system (Gatenby & Gillies, 2004).

Sometimes, tumours can even develop *maladaptations.* "Hypertumours" can be given as example to maladaptation, where a tumour proliferates so aggressively that after some time, it cannot support additional angiogenesis and goes extinct (Nagy, 2004).

1.1.3 Genetic Drift

As noted by Fisher in 1930 for normal populations of the nature (Fisher, 1930), *selection* is not always necessarily the most important evolutionary force in tumour cell populations. The "chance" effect should never be underestimated. Like selection, stochastic events can also be a reason for changes in allele frequencies, called *genetic drift* in evolutionary biology terminology. Effective

population size (N_e, the number of cells contributing to next generation) is the main determinant of the strength of drift.

Genetic drift also applies to healthy tissues and tumours. The random loss or fixation of alleles is especially more effective in "population bottlenecks", which can occur inside the body (Merlo et al., 2006). For instance, the apoptosis of breast epithelium during menstrual cycle forces the cell population through a bottleneck. Mutations occurring in the early stages of development can also go to fixation easily, forming large clones, which are named as "jackpots" (Hutchinson et al., 2003). Cancer therapy might also lead population bottlenecks. This aspect of some type of cancer therapies will be further discussed in the coming paragraphs.

Mutations that do not have a fitness effect and are not under selection are called *neutral mutations*. The majority of the mutations seen in neoplasms are thought to be neutral mutations. Some neutral mutations are linked to adaptive mutations and increase in frequency along with them, and these neutral mutations are called *hitchhiker mutations* (sometimes called as *passenger mutations* in cancer terminology). Distinguishing neutral mutations might allow using them as a *molecular clock*. Molecular clocks are used to help determine how much time has passed since the initiation of neoplasm (Tsao, Yatabe, Salovaara, Ja, & Shibata, 2000). However, molecular clocks in cancer can be inaccurate, as the rate of mutation is highly variable during cancer.

1.1.4 Artificial Selection

One of the major problems about cancer therapeutics is the "recurrence" of the tumours. Tumours are highly adaptive systems; they can develop excellent adaptations, and can overcome chemotherapy (Crespi & Summers, 2005). Although the treatment will cause a dramatic reduction in the tumour size at

the beginning, this also will cause an *artificial selection* for tumour cells having resistance to the therapy (Casás-Selves & Degregori, 2011). Consistent with this idea, a study on acute lymphoblastic leukemia therapy, which analyzed pretherapy and relapse samples, showed that relapse clones were just a selected minority of pretherapy clones (Mullighan et al., 2009).

As already mentioned, cancer is an evolutionary process; thus therapeutic interventions should be designed taking into account evolutionary dynamics. For instance, some interventions ameliorating progression instead of direct killing of the tumour cells could be more effective in controlling tumour size (Pepper, Scott Findlay, Kassen, Spencer, & Maley, 2009). "Tamoxifen", which is a cytostatic rather than cytotoxic drug, and proven to be effective in breast cancer therapy (Robertson, 2004), can exemplify this type of therapeutics. Figure 2 shows a representation of an alternative type of therapeutics that reads well the evolutionary dynamics of cancer:

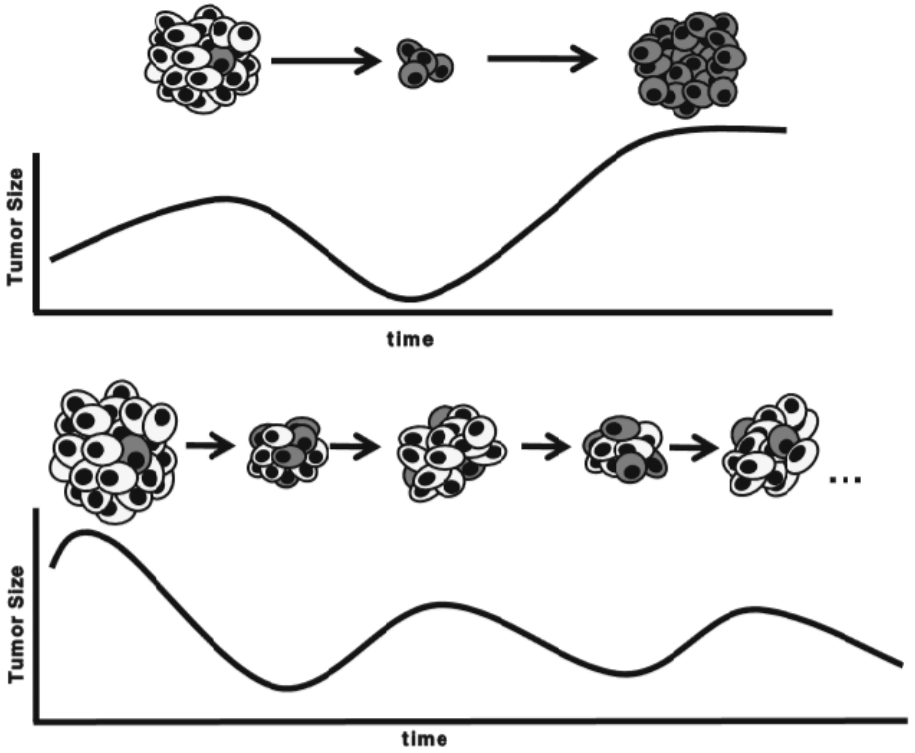

Figure 2: Illustration of effect of classical chemotherapy and a mild chemotherapy on tumour size

Top panel: **Chemotherapy leads to an initial killing of chemo-sensitive cells.** However, this benefits chemo-resistant clones, which are inadvertently selected by the therapy. Thus, chemo-resistant clones repopulate later. ***Bottom panel:*** **A suggested way of chemotherapy, where a mild chemotherapy regime is administrated; a portion of the chemo-sensitive cells are still alive.** These alive cells oppose the growth of the chemo-resistant cells. Hence, the tumour size follows a sinusoid growth. Although the tumour is not fully eradicated, at least the tumour size can be kept under control (Taken from Selves and DeGregori, 2011).

1.1.5 Phylogenetic Cancer Trees

The phylogenetic tree of a tumour is a good view of its evolutionary history. It can provide information about how the tumour is formed, key points of the tumourogenesis and genetic diversity of the clones (Yates & Campbell, 2012).

The phylogenies have a "trunk" branch, representing the complement of the mutations shared by all tumours. The length of the branches represents the 'molecular clock': the number of mutations that occured on that branch. However, this does not correlate with a chronological clock, as the mutation rate per time is not necessarily constant (Yates & Campbell, 2012).

Here, in the Figure 3, a tree representation of *linear evolution* in acute myeloid leukaemia (AML) is shown. It is linearly evolved, because it is formed as a result of post-treatment relapse. Thus, it is like a direct descendant of the major clone.

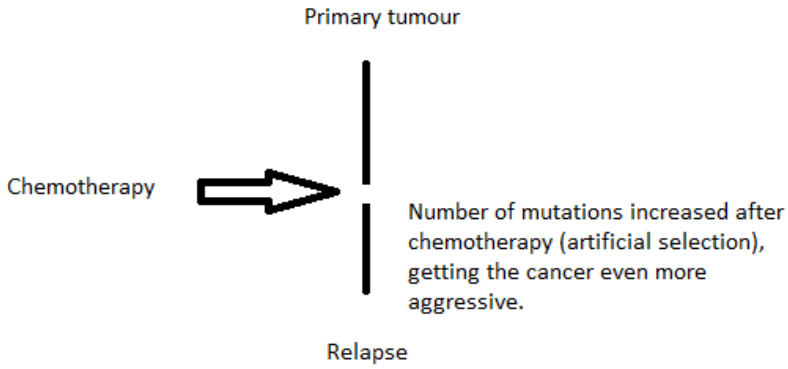

Figure 3: Tree representation of *linear evolution* in AML tumour as a result of post-treatment relapse. [Inspired from Yates & Campbell, 2012, originally adapted from (Ding et al., 2010)].

On the other hand, in the Figure 4, you see a tree representation of childhood acute lymphoblastic leukemia tumour (CALL), where a *branching evolution*

pattern, and specifically, "convergent evolution" is observed. The same genetic states are independently achieved in different clades of the tree. The blue boxes contain the recurrently mutated genes. Brown circles stand for cytogenetically distinct populations and the numbers for the number of the copies of the genes. Solid lines represent probable origin of the clones, whereas dashed lines suggest alternative origins.

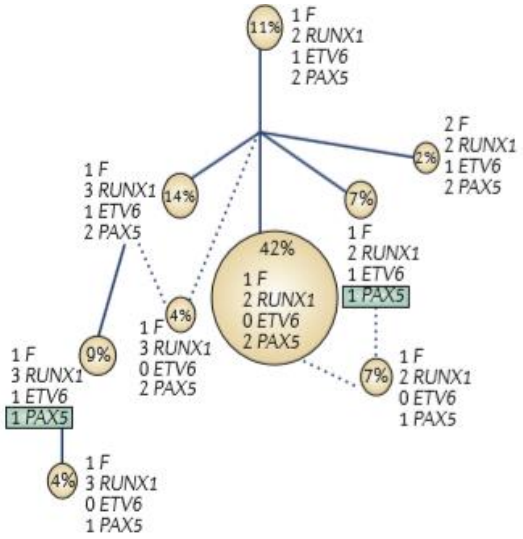

Figure 4: Tree representation of *branching evolution* in CALL tumour. [Taken from Yates & Campbell, 2012, originally from (Anderson et al., 2011)].

1.1.6 Tumour Evolution & Organismal Evolution

Tumour evolution differs in some important ways from organismal evolution, although they are very similar in many aspects. Tumour cells are not sexually reproducing; they are like asexual, single-celled organisms. As a result, there is no meiotic recombination, no Hardy-Weinberg equilibrium and no sexual selection within the tumour cell populations. In contrast to sexual populations, tumour populations are descendant of a single ancestor, a progenitor cell, which was originally a normal somatic cell. Hence, in cancer studies there is

accession to the ancestral genome of the tumour in the healthy somatic tissues of the body (Merlo et al., 2006). This enables comparisons of tumour genotypes with the normal genotype and making inferences about the tumour evolution process. In our study, we profit also from accession to ancestral genome aspect of tumour cells, by comparing the normal genotype with the tumour samples' genotypes, constructing oncogenetic trees, and reaching conclusions about the tumourogenesis process and the order of the events during development of the tumours.

Besides the similar evolutionary forces and dynamics mentioned in previous sections, tumour and organismal evolution also share a question mark about the mechanism and rate of evolutionary change: The longstanding conflict about gradualism versus punctualism in organismal evolution(Gould & Eldredge, 1993) is maintained also in neoplasmic evolution (Greaves & Maley, 2012). It is still under discussion if tumour clones evolve gradually by a sequence of genetic alterations and accumulate lesions or have a few, large-scale mutations caused by an insult (Greaves and Maley, 2012).

1.2 Bladder Cancer:

In our study, tumour development in nonmuscle invasive bladder cancer patients is investigated. Because of this reason, in this section, general information about nonmuscle invasive bladder carcinoma will be presented.

Nonmuscle invasive bladder cancer (NMIC) is the most common bladder cancer case, occurring in about ~60% of all bladder cancer patients (Knowles & Hurst, 2014). It has characteristics of recurrence. 50-70% of NMIC recurs within 5 years. However, only 10-30% progress into muscle invasive bladder cancer. Because of these facts, patients with NMIC need to be regularly

followed in case of recurrence and progression with cystoscopy and urine cytology.

Bladder cancer has many risk factors. "Cigarette smoking" is the primary risk factor known in bladder cancer. The population "attributable risk" (a disease risk proportion in a population, caused by a risk factor, defined by Levin, 1953) for smoking is 46% (American Cancer Society, 2009). Cigarette-smokers are approximately three-fold more prone to bladder cancer than non-smokers (Zeegers, Tan, Dorant, & van Den Brandt, 2000). Aromatic amines in cigarettes are thought to be responsible to trigger bladder cancer. "Age" is the other important parameter accounting for bladder cancer. The median age for men developing bladder cancer is 69, while for women it is 71 (Volanis et al., 2010). "Sex" is another factor in bladder cancer; males being 3 times more prone to bladder cancer than females (Knowles & Hurst, 2014). According to a study conducted in Taiwan in 2004 (Chen, Su, Guo, Houseman, & Christiani, 2005), exposure to certain industrial chemicals used in occupational settings is also a risk factor in bladder cancer, being marginally significantly ($p=0.055$) greater in male exposed workers than non-exposed ones (18% versus 7%). Benzidine, used in dye production and rubber industry, as well as 4-aminobiphenyl used also in rubber industry, can be given as example to these chemicals. The workers of these industries are under high risk. It is also strongly evident that exposure to "arsenic" in drinking water at concentrations exceeding 300 - 500 µg/L is directly linked to bladder cancer risk (Meliker and Nriagu, 2007). A dramatically high level of bladder cancer incidence is seen in northeastern Taiwan, where high levels of arsenic occur in drinking waters (Chiou *et al*, 2001).

Bladder cancer can be detected by several symptoms. "Hematuria" is the most common symptom, occurring up to 85% of the patients. Irritative voiding symptoms and dysuria are also symptoms of bladder cancer (Wakui and Shigai, 2000).

There exist various types of biomarkers of bladder cancer. Epigenetic changes, specifically "DNA methylation" seems to be a good biomarker for detection of bladder cancer. Detection for aberrant DNA methylation level in just a few loci, either from urothelium or biopsy samples of the patient, is sufficient for diagnosis (Chihara et al., 2013). *BCL2*, *CDKN2A*, *APC* can be given as example for these loci. MicroRNA (miRNA) expression levels may also be used as biomarker in bladder cancer (Scher et al., 2012). Specifically, miR-145, miR-143 and miR-125b are known to be down-regulated miRNAs in bladder cancer, while miR-183, miR-96, miR17-5p and miR-20a are upregulated ones (Yoshino et al., 2013).

1.3 RNA Editing and Cancer

1.3.1 RNA Editing

In addition to environmental factors, RNA editing is also emerging as an important driver in cancer initiation and progression. Because of the relation of RNA editing to tumourogenesis, section 1.3 is devoted to RNA editing, RNA editing enzymes and its relation to cancer.

RNA editing is the process where nucleotide sequence of a transcript, originally based on the genomic DNA sequence, is post-transcriptionally modified (Anant & Davidson, 2003). RNA editing is thought to have evolved to expand the genetic repertoire, regulate gene expression, work as a defence system against infection, and plays role in development (Avesson & Barry, 2014).

In mammals, two classes of RNA editing exist: One is via deamination of cytidine (C) to uridine (U), the other is involves conversion of adenosine (A) to inosine (I) within nuclear mRNAs (Savva, Rieder, & Reenan, 2012).

A-to-I editing is the most frequent editing mechanism and is performed by ADARs (Adenosine Deaminases). There are three ADARs: ADAR1, ADAR2 and ADAR3. ADARs require a pre-mRNA template containing intronic regions as a target to edit (Gerber, 2001). Inosine is recognized as guanosine (G), thus the ultimate result of A-to-I editing is the change from A to G base in the open reading frame (Gerber, 2001). One of the examples for A-to-I editing is the inhibition of hepatitis delta virus (HDV) replication in a process that requires RNA editing of endogenous mRNA (Jayan & Casey, 2002). There are two forms of hepatitis delta antigen (HDAg): the small form HDAg-S and the large form HDAg-L (Lai, 1995). The HDAg-L is edited from U-to-C at nucleotide 1012, by changing the stop codon of HDAg-S to tryptophan codon. This change in sequence get HDAg-L form 19 nucleotides larger than the small form. Moreover, this mutation is reported to occur at a 500-fold higher rate than other mutations in other positions of the delta genome (Luo et al., 1990). This single-base change, shown to be mediated by ADARs (Casey, Bergmann, Brown, Purcell, & Gerin, 1992) is neccessary for the formation of HDAg-L that is required to form infectious HDV virions with a hepatitis B virus surface antigen envelope (Glenn & White, 1991).

The best-characterized example of C-to-U editing is mRNA encoding apolipoprotein B (apoB) (Anant & Davidson, 2003). C-to-U editing of apoB requires a single-stranded RNA as target. ApoB100 and its shorter isoform apoB48 differ due to site-specific C-to-U editing which creates a translational terminal codon in apoB48. This C-to-U editing occurs primarily in human small intestine, as apoB48 isoform is necessary for the absorption of dietary lipid. ApoB100 is encoded by hepatic apoB RNA, which is expressed in human liver and is not edited; thus encodes a larger protein. ApoB48 and ApoB100 share the N-terminal sequence, however ApoB48 does not include ApoB100's C-terminal low-density-lipoprotein receptor binding domain (Young, Hubl, Smith, Snyder, & Terdiman, n.d.). C-to-U RNA editing is mediated by

APOBEC1 enzyme (Mehta, Kinter, Sherman, & Driscoll, 2000), a member of APOBEC protein family that will be explained further in the next section.

1.3.2 APOBECs

The APOBEC (apolipoprotein B mRNA editing enzyme catalytic polypeptide-like) protein family are deaminases that are expressed in vertebrates and are able to edit DNA and/or RNAs. The human APOBEC family is composed of 11 members, each with distinct functions (Vieira & Soares, 2013) (Figure 5):

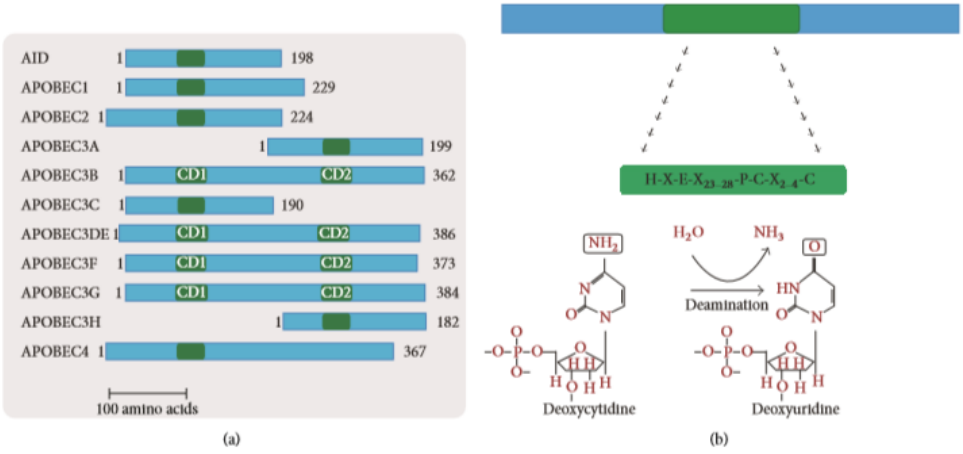

Figure 5: Human APOBEC proteins

a) CD represents catalytic domains and is depicted in green. Proteins containing 2 CD copies, are named as CD1 and CD2, accordingly N- and C-terminal domains respectively. The number next to each bar shows the number of amino acids for each APOBEC. **b)** The conserved amino acid sequence of the APOBEC family is shown. Below that, the hydrolytic deamination reaction at the C4 position of a cytidine or deoxycytidine is shown (Taken from Vieira and Soares, 2013).

APOBEC3 enzymes (specifically APOBEC3Bs), which merit special attention in our study, have 7 members in humans: APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3DE, APOBEC3F, APOBEC3G and APOBEC3H.

The *APOBEC3* gene emerged after the divergence of marsupial and placental lineages. After a duplication event, two ancestral *APOBEC3* genes evolved in the placental mammal lineage. From these two ancestral APOBEC3s, through a complex history of duplications and fusions, the other present APOBECs have evolved (Conticello, Thomas, Petersen-Mahrt, & Neuberger, 2005). In some species like rodents, cattle, and pigs, these two original genes merged and formed a single gene while in others such as primates, horses, bats and felines these two genes have been duplicated to form the APOBEC family. Exclusively in primates, the APOBEC3 locus has rapidly expanded, which is thought to be a consequence of selective pressure from their targets (Sawyer, Emerman, & Malik, 2004) (Zhang & Webb, 2004). A study about evidence for strong positive selection on the *APOBECs* has reported that these genes are selected for genome defence (Sawyer et al., 2004). The existence of recurrent positive selection on protein coding regions can be determined by calculating the ratio of number of changes per non-synonymous (i.e. altering amino acid sequence) sites (K_a) and per synonymous (i.e. not altering amino acid sequence) change ratio (K_s), and testing whether this ratio is greater than 1. In the case of a neutrally evolving protein, a non-synonymous mutation has the same chance of reaching fixation with a synonymous mutation. However, if positive selection is acting on the protein, K_a is expected to be higher as positive selection will favour the diversity at the amino acid level. To investigate the mode of selection on *APOBEC3G*, Sawyer *et al*, sequenced the *APOBEC3G* gene from ten primate species, and found K_a/K_s ratio greater than 1 for the majority of phylogeny. This is signalling for the evidence of positive selection acting on the *APOBEC3G* gene throughout the primate evolution history. Then, they calculated K_a/K_s ratio for other members of *APOBEC*

family and found signals of positive selection on *APOBEC1*, *APOBEC3B*, *APOBEC3C*, *APOBEC3D*, *APOBEC3E*, *APOBEC3DE*, *APOBEC3F*, *APOBEC3G* genes. The selection for *APOBECs* on the primate lineage is speculated to be because of the host defence ability conferred by them.

APOBEC3 enzymes function as a part of the innate immune system, protecting against retroviruses and retrotransposons mainly by DNA editing mechanism (Harris & Liddament, 2004). While APOBEC3B preferentially deaminates cytosine residues in TpC* context (i.e. when cytosine is adjacent to a 5' thymine) and GpA* on the complementary strand), APOBEC3G and APOBEC3F recognize and edit at CpC* and TpC* dinucleotide context (GpG* and GpA* in the complementary strand) (Armitage et al., 2008). Moreover, current studies claim that APOBEC3B discriminates for even a more stringent motif; trinucleotide TpC*Ap and TpC*Tp (TCW motif) for editing (S. A. Roberts et al., 2013). This more stringent motif allows differentiating APOBEC3B-induced mutations, preceded by a thymine (S. A. Roberts et al., 2013), from mutations of the highly mutable CpG* motif, linked to aging-related mutagenesis (Alexandrov & Stratton, 2014).

1.3.3 APOBECs and HPV

It is already known that APOBEC3 enzymes participate in immune response to DNA virus infections in human, such as human papillomavirus (HPV), human immunodeficiency virus (HIV) (Harris et al., 2003) hepatitis virus B (HBV) (Vartanian et al., 2010), and simian immunodeficiency virus (SIV) (Yu et al., 2004).

APOBEC mutation pattern (TCW mutations) and high APOBEC3B mRNA levels occur very frequently in cervical cancers (Burns et al., 2013) and over 99% of cervical cancers occur as a result of HPV (Studies, 1999). It is known

that HPV- HNSCC (HPV negative head and neck squamous cell carcinoma) typically occurs in heavy smokers while HPV+ HNSCC (HPV positive head and neck squamous cell carcinoma usually occurs in non-smokers (Agrawal et al., 2011). Henderson *et al* (Henderson, Chakravarthy, Su, Boshoff, & Fenton, 2014), tested whether there is a relationship between APOBEC-mediated mutagenesis and HPV-driven tumourogenesis. To test this, they compared the exomes of HPV+ and HPV- HNSCCs. They confirmed the existence of such relationship by showing the high association of APOBEC mutation pattern per sample with HPV status, but not with age and smoking status. Also, there was significantly higher APOBEC3B expression in HPV+ patients than HPV- patients, but expression of APOBEC3B was only weakly, but significantly associated TCW mutations.

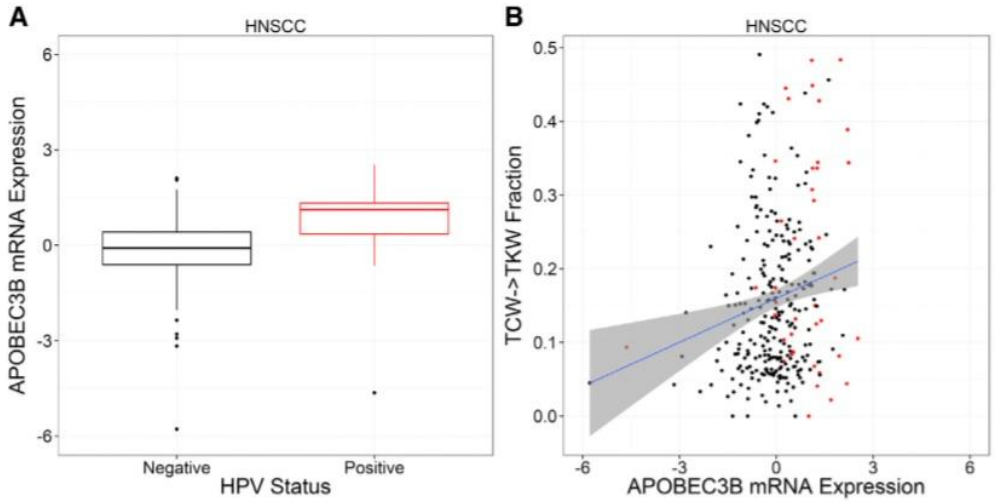

Figure 6: APOBEC3B mRNA levels in HPV- and HPV+ HNSCC and Relationship between APOBEC mRNA levels and TCW mutations
A) APOBEC3B mRNA levels in HPV- and HPV+ HNSCC ($p = 3.66 \times 10^{-11}$ by Wilcoxon). **B)** Relationship between APOBEC mRNA levels and TCW mutations ($p = 0.0127$, by Spearman's rho; $r = 0.144$). *Taken from Henderson et al, 2014.*

1.3.4 APOBECs and Cancer

A number of recent studies suggest that a class of APOBEC enzymes may inadvertently edit host genome, leading to tumourogenesis. Roberts *et al* (2013) analyzed 954,247 mutations in 2,680 exomes from 14 cancer types and revealed an APOBEC-carcinogenesis relationship in bladder, cervical, breast, head and neck and lung cancers. In some samples, APOBEC mutations were even the majority; some reaching 68% of mutations in the exome. Also, the possibility that this result was an artifact of motif-specific functional selection could be ruled out, since the same calculations were done with for silent and noncoding mutations in each sample and similar results were obtained.

In support of the idea of APOBEC-mediated carcinogenic mutagenesis, the existence of a phenomenon termed *kataegis*, referring to mutations clustered on the same strand, was recently detected through next-generation sequencing in multiple myeloma, prostate cancer and HNSCC (S. A Roberts et al., 2012). *Kataegis* is expected in APOBEC-driven mutagenesis, because APOBECs could simultaneously create many mutations within a strand. Such clustered mutations were enriched at trinucleotide motifs recognized by APOBECs: TCW. Moreover, those clusters that included only one mutation pattern were often gathered at rearrangement breakpoints. This suggests that the mutations could occur on single-stranded DNA (ssDNA) regions formed by aberrant DNA double-strand break repair. The observation that ssDNAs are the good substrates for APOBEC enzymes (Smith, 2012) further supports the role of APOBECs in *kataegis* formation.

Roberts *et al.* (2013) have additionally studied available RNA-sequencing (RNA-seq) data from 2,048 tumours of 14 cancer types, and claimed that *APOBEC3B* mRNA level was strongly correlated with TCW mutation motif per exome, when compared across samples. In addition, for bladder cancer and lung squamous cell cancers, median *APOBEC3B* expression was increased

nearly 3-fold compared to median *APOBEC3B* expression of all samples across all 14 cancer types. In another study, Sasaki *et al.* (2014) investigated *APOBEC3B* mRNA expression level in 88 non-small-cell lung cancer patients. *APOBEC3B/β-actin* level was significantly higher in lung tumours compared to neighbouring normal tissue ($p < 0.0001$). Besides, the ratio of expression levels of *APOBEC3B/β-actin* between tumour and normal cells did not differ with gender, age, smoking status and pathological stages. Meanwhile, Burns *et al.* (Burns et al., 2013) quantified mRNA levels for each member of the APOBEC family in breast cancer cell lines. It was only *APOBEC3B* that tended to be upregulated. Overall, these studies support a relationship between APOBEC-mediated mutagenesis and development of several cancer types.

1.4 Multifocality & Origin of Multifocality Problem

The simultaneous or metachronous occurrence of more than one tumour within the tissue of a patient, multifocality, is seen in some cancers like prostate, breast and micropapillary thyroid cancer (Ruijter *et al*, 1996), (Pedersen, Gunnarsdottir, Rasmussen, Moeller, & Lanng, 2004), (Ross et al., 2009). Multifocality is also a prominent characteristic feature of urothelial carcinoma, observed on urothelial tract of a patient (Kakizoe, 1991). To explain this situation, two hypotheses were proposed (Harris and Neal, 1992). One is the *field hypothesis*, which suggests that multiple tumour cells arise independently by a carcinogen insult on the tissue. Thus, these types of tumours should acquire independent genetic alterations. The other is the *monoclonality hypothesis*, which suggests that multifocal tumours originate from a single transformed cell, which then migrates to form other urothelial tumours either by intraluminal seeding or intraepithelial spread (Garcia, 1999). Tumours originated in this way are accordingly expected to share some mutations,

including drivers and passengers. Although these two hypotheses are not mutually exclusive, it is still under debate which model is more common among multifocal bladder tumour cases.

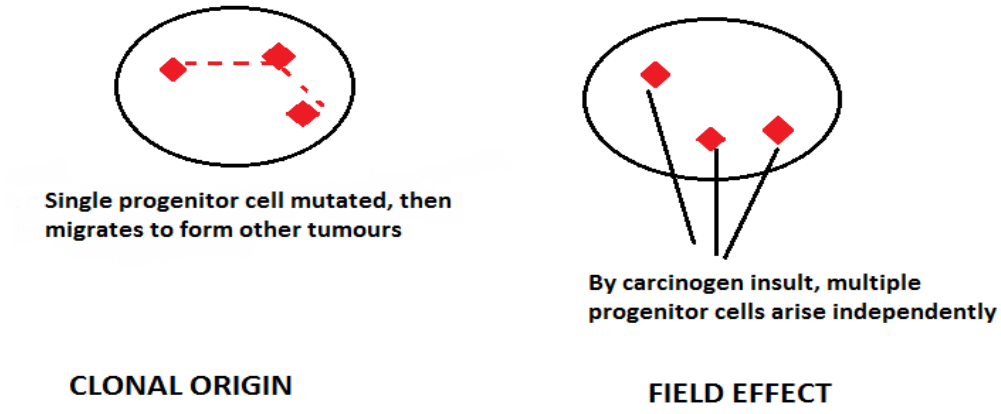

Figure 7: Representation of the two hypotheses: *Clonal Origin* and *Field Effect*
On the *left*, the clonal origin hypothesis is described: a single transformed cell migrates to form other tumour cells. On the *right*, the field effect hypothesis is described: multiple tumour cells occur independently. Each red square represents a single tumour cell, while dashed lines represent migration events.

Several molecular genetics techniques have been used to resolve this long-debated problem on the origin of multifocal urothelial tumours. In the tables below, an overview of molecular studies on this question is provided:

Table 1: Overview of molecular studies supporting *monoclonal origin* of multifocal urothelial carcinomas.

A part of the table is adopted from Hafner *et al*, 2002 (Christian Hafner, Knuechel, Stoehr, & Hartmann, 2002). X-inact: X-chromosome inactivation, LOH: loss of heterozygosity, FISH: fluorescence *in situ* hybridization, p53: *p53* gene mutation analysis, CA: cytogenetic analysis, Rb: *Rb* gene mutation analysis, CGH: comparative genomic hybridization.

Authors	**Patient**	**Tumour**	**Methods**
(Sidransky, 1992)	3	10	X-inact, LOH
(Habuchi, 2005)	4	11	p53
(Miyao et al., 1993)	6	13	p53
(Xu, 1996)	5	16	p53
(Chern et al., 1996)	5	10	p53, Rb
(Takahashi et al., 1998)	20	67	LOH
(M. Li & Cannizzaro, 1999)	10	35	X-inact
(Fadl-Elmula et al., 1999)	6	21	CA
(Hartmann et al., 2000)	9	47	LOH, FISH
Louhelainen, 2000	5	32	LOH
(Simon et al., 2001)	6	32	CGH
(Takahashi et al., 2001)	23	73	LOH
(C Hafner et al., 2001)	10	55	LOH, p53
(Dalbagni, Ren, Herr, Cordon-cardo, & Reuter, 2001)	7	23	p53
Vriesema, 2001	6	20	p53
(Kawanishi et al., 2007)	5	24	CGH
(Wang, Lang, Pin, & Izawa, 2013)	4	32	LOH

Table 2: Overview of molecular studies supporting *field effect origin* of multifocal urothelial carcinomas.
A part of the table is adopted from Hafner *et al*, 2002. MSI: detection of high grade microsatellite instability

Authors	Patient	Tumour	Methods
(Miyao, 1993)	1	4	p53
(Petersen, 1993)	1	2	p53
Spruck, 1994	3	7	LOH, p53
(Yoshimura, Kudoh, Saito, Tazaki, & Shimizu, 1995)	1	2	p53
Goto, 1997	13	36	p53
(Hartmann et al., 2000)	1	5	LOH, FISH
(Takahashi et al., 2001)	9	19	LOH
(C Hafner et al., 2001)	5	23	LOH, MSI
(Hartmann et al., 2000)	5	20	LOH, FISH, p53
(Jones et al., 2005)	21	58	LOH, X-inact, p53

One of the most frequently used methods in this area is *X-inactivation pattern* detection (Christian Hafner et al., 2002). If the tumours are coming from the same origin, which means from a *monoclonal origin*, all should have the same X chromosome inactivated in the tumours. However, if these tumours are coming from a different precursor cell, they have 50% probability of having the same X chromosome to be inactivated (Sidransky *et al*, 1992). In the first

such study, Sidransky *et al* (1992) used this molecular genetics tool to resolve the multifocality problem. In their study with four females with bladder carcinoma, after digestion with methylation-sensitive endonucleases, normal bladder mucosa showed a polyclonal pattern of X-inactivation. In contrast, tumours of the same patient showed a monoclonal pattern after digestion.

Loss of heterozygosity (LOH) analysis at multiple marker loci is another molecular genetics method to address the origin of multifocal tumours problem. Especially, deletions on chromosome 9 are of special interest as such deletions are usually an early event in bladder carcinoma (Miyao, 1993). In Sidransky's study (1992), the authors examined the multifocal bladder tumours also for loss of chromosome 9q, 17p and 18q sequences. DNA from the tumours of the 3 patients showed the loss of the same 9q allele for each patient. However, for 17p and 18q allelic losses, the pattern was not consistent among the tumours of a given patient. This may be expected, because 17p and 18q losses are usually considered to occur in high-grade tumours. In conclusion, Sidransky's study provided strong evidence that the studied multifocal tumours are genetically related, and thus arose from a single precursor cell.

Microsatellite instability could also be utilized for clonal analysis. Jones *et al* (2005), for example, examined microsatellite alterations in 21 urothelial carcinoma patients, and supplemented their results by X-chromosome inactivation data. This study found evidence for the field effect in the coexisting bladder tumours for the majority of the patients (18/21) they studied.

Mutations in *oncogenes*, such as *p53* are also frequently investigated to answer the clonal origin question. One such early study, conducted in 4 patients, found evidence for bladder tumours coming from a single progenitor cell (Habuchi, 1993), whereas later studies using the same method found evidence for the field effect (Petersen, 1993) (Yoshimura et al., 1995) (Hartmann et al., 2000).

The *Comparative Genomic Hybridization* (CGH) technique, which compares allelic gains and losses at each chromosome, has also been used to address this question. Simon *et al* (2001) utilized CGH to detect unbalanced chromosomal aberrations, and then constructed "cytogenic pedigrees" to infer shared aberrations. This study, conducted in 6 patients, reached the conclusion of monoclonal origin in their patients. Similar to CGH, *cytogenetic analysis* can be also utilized, by comparing karyotype of different tumours (Fadl-Elmula, 1999). However, reaching a conclusion about the origin of multifocal tumours by using cytogenetic analysis technique can be inaccurate, as during cancer development, lots of karyotypic change take place in tumour clones.

1.5 Exome sequencing

Exome sequencing is a next-generation "targeted sequencing" approach, which focuses on the protein-coding regions of the genome. Although the exomic region covers ~1% of the genome, it is worth to note that it harbours 85% of mutations with large phenotypic effect (Bamshad et al., 2011).

Exome sequencing is carried out firstly by randomly shearing DNA into very small fragments. Several micrograms of the sample is used to construct a DNA library. Then, the fragments are attached to special adaptors. After that step, exomic regions are "captured" and enriched in biotinylated DNA or RNA baits by oligonucleotides in these baits complementary to the adaptors. The capture step is followed by amplification and repeatedly parallel sequencing of the fragments (Bamshad et al., 2011). The term "read" is used for the result of a sequenced fragment. Finally, the millions of reads are aligned to the reference genome and variant calling is achieved. In high coverage sequencing, each nucleotide is sequenced repeatedly many times; therefore included in many reads (Bick & Dimmock, 2011).

In this study, exome sequencing was utilized to study genetic relatedness of multifocal bladder tumours. Below, an illustration of the two hypothetical phylogenies expected to be found under the monoclonal origin and field effect hypotheses is presented:

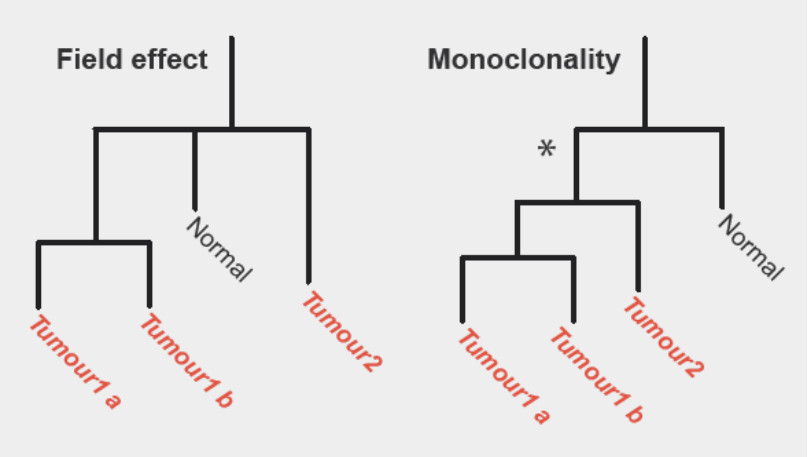

Figure 8: Two hypothetical phylogenetic trees, expected to be observed in the case of field effect hypothesis and monoclonal origin hypothesis, respectively
Branch with "*" shows the tumourogenesis branch (Acar & Ozkurt, 2015), under review in BMC cancer.

In the field effect phylogeny, tumour samples are not grouped together as they have evolved independently from each other. On the other hand, in the monoclonal phylogeny, tumour samples are grouped together as they are originating from a one single progenitor cell. Hence, in the latter hypothesis, a longer tumourogenesis branch (i.e. the branch where mutations shared by all tumours are included) than recent branches is expected.

If there is low number of mutations in tumourogenesis branch (the branch where mutations shared by all tumours are included) but tumours are coming from the same origin, it could be difficult to manage to observe the mutations

on this branch by using low number of variants. Tumours may separated just after tumourogenesis, so the tumourogenesis branch could be short. Another possibility is that after separation of tumours, the mutation rate could increased greatly, thus tumourogenesis branch again appeared as short. Hence, there is need to analyze high number of variants to be able to observe the shared mutations on the tumourogenesis branch. Otherwise, the results could lead to misinterpretation of the cases as coming from polyclonal origin. However, exome-sequencing allows to high number of variants; thus increasing the possibility to observe the prospective shared mutations on a branch.

1.6 Aim of the Study:

As mentioned already, various studies using diverse techniques have been conducted to address the origin of multifocal bladder tumours. However, there is no consensus yet about which hypothesis is more valid about the evolutionary relationship among multifocal tumours. In our study, we try to investigate the same question with a different approach: exome-sequencing. Exome sequencing, which is an economically efficient technique, yields a high number of variants, increasing statistical power compared to previous studies. Thus, it can be very effective in solving the still open question on the origins of multifocal bladder tumours.

In the study, we aimed to study the type (in dinucleotide and trinucleotide context), function (degree of the effect on amino acid sequence) of the mutations and mutated bladder cancer driver genes by analyzing exome-sequencing data. The ultimate goal was to be able to construct a hypothetical timeline about the evolutionary processes of the multifocal bladder tumours and to show the utility of exome-sequencing and population genetic analysis

on creating this timeline. All these results could contribute greatly to therapeutic interventions for bladder cancer.

CHAPTER 2

MATERIALS AND METHODS

2.1 Clinical Sample Collection and Sequencing

The study was designed following ethical guidelines and the protocol was approved by the Koç University Institutional Review Board. During transurethral resection, all visible tumours were resected. Four samples, three from the tumours and one from normal mucosa, were collected from each of three patients. The location of the normal mucosa sample was at equal distance from the tumoural foci as the two tumours were from each other (for further information see Acar, Özkurt *et al*, 2015, under review in BMC Cancer).

2.1.1 Clinical Information About Patients

Patient 1: A 62-year-old male patient who has presented because of showing painless macroscopic hematuria symptom. There was not any remarkable point in his past medical history. A tumoural lesion (approximately 2.5 cm) was detected on the right bladder wall. The lesion was lying just superior and lateral to the ureteric orifice. Additionally, a smaller (approximately 0.5 cm) lesion is also detected, settled 1-2 cm laterally to the index tumoural focus. Histopathological examinations allowed classifying the case as high grade, pTa, papillary urothelial carcinoma.

Patient 2: A 64-year-old male patient who has presented because of lower urinary tract symptoms and macroscopic hematuria. Multiple tumoural lesions were detected inside the bladder by ultrasonography. As remarkable points in his medical past, he had diabetes mellitus and he was an active smoker (20

packs/year). 2 physically distinct tumoural foci (index lesion ≈ 3 cm, smaller lesion ≈ 1 cm) that are settled on the left lateral wall, are observed during cystoscopic detection. Histopathological examinations documented the case as low grade, pTa, papillary urothelial carcinoma.

Patient 3: A 72-year-old male patient who has presented because of macroscopic hematuria symptom. A solid lesion on the right lateral wall of the bladder is observed by ultrasonography. As remarkable points in his medical past, he has been an active smoker (25 packs/year), had hypertension and suffered from pericarditis in the past. By cystoscopic examinations, 2 physically separated tumoural foci (index lesion ≈ 2.5 cm, smaller lesion ≈ 1.5 cm) that are harboured on the right lateral wall, are detected. Histopathological examinations allowed to classify the case as high grade, pTa, papillary urothelial carcinoma.

2.1.2 Molecular biology & Sequence Alignment

Genomic DNA was isolated from the samples by Nathan Lack's laboratory (Koç University, School of Medicine) by utilizing Qiagen QIAmp DNA kit. Agilent SureSelect v5 capture kit is used to target coding regions. There are actually 2 other whole-exome capture platforms, NimbleGen's Sequence Capture Array and SeqCap EZ, but all demonstrated to show roughly the same exome SNP calling efficiency (Asan et al., 2011). Exomic regions are sequenced on an Illumina HiSeq 2000 high-throughput sequencer with 100bp paired-end read mode at 100x coverage by Centrillion BioScience, which is private biotechnology company in California, providing genomic and bioinformatics solution.

The exome data generated was aligned to the reference human genome (hg19 / NCBI GRCh37) by BWA aligner (version 0.7.10) (Table 4 and Table 5). By default, "mem" algorithm is used on paired-end mode (H. Li, 2013). After this step, the GATK tool is used to realign indel (insertion and deletion)-containing reads (McKenna et al., 2010), thus to correct for mapping biases. GATK UnifiedGenotyper in multi-sample mode was utilized to generate Single Nucleotide Variant (SNV) and insetion and deletion (indel) callsets by merging read data from the four samples of a patient (Table 6).

SNV (Single Nucleotide Variant) term is used instead of *SNP* (Single Nucleotide Variant), because they refer to different things. SNPs can be defined basically as inheritable, well-validated common variations (>1%) within a population However, SNVs are not well-validated variations; they are private to the individual. Because of this fact, variations of cancer tissues are named as SNV.

Several filtering processes are implemented. Firstly, The Variant Quality Score Recalibration (VQSR) filter, downloaded from GATK resource bundle version 2.5, enabled to minimize false positives; thus to generate accurate call sets. The VQSR filters call sets, generating for each variant a well-calibrated probability whether a position is a true genetic variant or just a data-processing artefact. This estimation is inferred from SNP call annotations like RMSMappingQuality (Root Mean Square of the mapping quality of reads across all samples) or HaplotypeScore, taking place in VCF (Variant Call Format) files. Secondly, the variants that are represented in dbSNP (Single Nucleotide Polymorphism Database) version 138 are filtered from the dataset; because dbSNP archives common SNPs within the species and we are actually interested in cancer-specific variants. Also, segmental duplications are excluded from the analysis as reading two copies of the same polymorphic region could lead to misinterpretation of them, as being heterozygous alleles. Finally, the SNV dataset only included SNVs and indels that passed the GATK

quality filter and that was read at least 4 times in all four samples. The information about the number of raw and filtered SNVs and indels during filtering processes are given in the results part (Table 6). The same procedure was followed independently for each patient.

2.2 Downstream Bioinformatics Analysis

2.2.1 Analysis of SNVs and Indels

R and Python programming languages are utilized for population genetics analysis of the tumour samples.

By conducting R programming, The SNV data (1628-1733 SNVs per patient, Table 6) are converted into binary form: heterozygous (represented as 0/1) or homozygous (1/1) non-reference alleles called “1”, and homozygous reference alleles (0/0) called “0”. Heterozygous and homozygous non-reference alleles are treated in the same way. Normally a mutation affects only one allele. However, if both alleles are mutated, this may be either because of parent alleles having one alternative allele or just simply because of technical error in sequencing. Thus, as we are only interested in mutations occurred because of cancer, it is more convenient to call homozygous non-reference positions as "1". To construct a reference genome, "0" is assigned to all variants. Monoallelic positions are removed from the data. Then, these SNV datasets, transformed into numeric form, were used to construct Euclidean distance matrices among samples using the R “dist” function. R "ape" package's (Paradis, Claude, & Strimmer, 2004) “bionj” algorithm (a neighbour-joining phylogenetic tree construction algorithm) allowed to construct rooted phylogenetic trees (e.g. rooted by the reference genome) among samples of the patient. "Bionj" algorithm is claimed to perform better than other algorithms when branch lengths are variable. As will be shown later, the phylogenetic

trees' branch lengths also showed high variance, because of surplus mutation accumulation on tumour branches (Gascuel, 1997). 10,000 bootstraps is performed using the "boots.phylo" function of the R "ape" package.

2.2.2 Non-parametric Bootstrapping

Bootstrapping statistics (Efron, 1979) is analogous to jackknife statistics where the data, whose distribution is not known, is resampled to infer the variability of the estimate. Felsenstein (1985) proposed to use bootstrapping to place confidence limits (bootstrap proportions) to internal branches of phylogenetic trees.

In bootstrapping method, the data is resampled many times (10,000 times in our case) with replacement to form a simulated dataset. Thus, the simulated data contains the same set of species but some characters are duplicated while some others are dropped and a collection of number of resampling times of estimates of the parameter is obtained. This simulated data approximates the actual distribution (Felsenstein, 1985).

Actually, Felsenstein (Felsenstein, 1983) first proposed bootstrapping proportions to measure the "repeatability" of the phylogenetic tree. However, bootstrapping can be used both to support "repeatability" and "accuracy" of the inferred tree (Hillis and Bull, 1993). In statistical terms, we can denote "repeatability" as P_n ($c \varepsilon T^* | c \varepsilon T^*_0$), while "accuracy" as P ($c \varepsilon T | c \varepsilon T^*_n$) (Holmes, 2003). In this notation, "T^*" denotes metric estimate, "T" denotes true estimate and "c" for clade. The "repeability" refers to the probability that another sample shares the clade with the true sample, thus the terms repeability and accuracy are linked. In the study, bootstrap proportions are used as a measure of accuracy, by 10,000 times running of pseudosampling of the SNV data.

The data set included also 2130-2555 indels in the three patients by GATK analysis. (Table 6). The same Euclidean distance, bioinj algorithm phylogenetic tree construction and bootstrapping analysis was applied to the indel data as well. However, the tumour branches of the tree are not resolved clearly, as indicated by low bootstrapping values.

2.2.3 Functional analysis

The "SnpEff" open source software (Cingolani et al., 2012) is used to annotate variants' effect according to their impact on protein structure based on genomic locations. Mutations are classified as having "high", "moderate", "low" and "modifier" effects on protein structure by the software. The "high" or "moderate" effect mutations (e.g. loss of splice sites, non-synonymous substitutions, stop-codon insertions) are considered as being "functional" and others as being "nonfunctional". Then, ratios of functional vs. non-functional mutations between SNVs shared by all tumours and SNVs shared by all samples including normal mucosa (e.g. representing individual's unique genotype) are compared by Fisher's Exact Test.

2.2.4 Mutation type analysis

SNVs are classified based on dinucleotide and trinucleotide sequence context, inspiring from Lawrence *et al*. (Lawrence et al., 2013) work. To obtain dinucleotide sequence context, the nucleotide preceding a given SNV is noted, based on the human reference genome (hg19). To obtain trinucleotide sequence context, the nucleotide preceding the given SNV, the SNV itself and the following nucleotide again are noted based on hg19. Following Nordentoft et

al. (2014), only A and C positions on each strand are took account to simplify the analysis. Thus, all the positions on both strand are considered, as T is complementary to A and C is to G. The same analysis is repeated in dinucleotide frequency context as well as trinucleotide frequency context.

2.2.5 Candidate driver gene analysis

92 candidate driver genes that are frequently mutated in bladder cancer are compiled from the COSMIC (Forbes et al., 2010) and ATLAS ("Comprehensive molecular characterization of urothelial bladder carcinoma," 2014) databases. The potential driver gene list is displayed in the table below (Table 3):

Table 3: Compiled list of potential driver genes for bladder cancer

Associated Gene Name	Ensembl Gene ID	Associated Gene Name	Ensembl Gene ID
TARBP2	ENSG00000139546	CREBBP	ENSG00000005339
RB1	ENSG00000139687	TACC3	ENSG00000013810
NCOR1	ENSG00000141027	SERPINB1	ENSG00000021355
TP53	ENSG00000141510	CDH1	ENSG00000039068
ERBB2	ENSG00000141736	TPR	ENSG00000047410
AKT1	ENSG00000142208	KMT2C	ENSG00000055609
EPHA2	ENSG00000142627	LZTS1	ENSG00000061337
HMCN1	ENSG00000143341	RASSF1	ENSG00000068028
FLG	ENSG00000143631	FGFR3	ENSG00000068078
AFF3	ENSG00000144218	SMC1A	ENSG00000072501
ANK2	ENSG00000145362	XRCC1	ENSG00000073050
KDM6A	ENSG00000147050	FGFR1	ENSG00000077782
CDKN2A	ENSG00000147889	BRINP1	ENSG00000078725
ATM	ENSG00000149311	LRP2	ENSG00000081479
FRG1B	ENSG00000149531	GSTP1	ENSG00000084207
DAB2	ENSG00000153071	FXYD3	ENSG00000089356
CD109	ENSG00000156535	MAP3K1	ENSG00000095015
KALRN	ENSG00000160145	MYH9	ENSG00000100345
MAPKAPK2	ENSG00000162889	EP300	ENSG00000100393

Table 3 (cont'd): Compiled list of potential driver genes for bladder cancer

Associated Gene Name	Ensembl Gene ID	Associated Gene Name	Ensembl Gene ID
ELF3	ENSG00000163435	STAG2	ENSG00000101972
TGFBR2	ENSG00000163513	MMP15	ENSG00000102996
NIPBL	ENSG00000164190	TSC2	ENSG00000103197
HCN1	ENSG00000164588	CTSH	ENSG00000103811
CSMD3	ENSG00000164796	ERCC2	ENSG00000104884
TSC1	ENSG00000165699	PRX	ENSG00000105227
KMT2D	ENSG00000167548	PTPRS	ENSG00000105426
KLK5	ENSG00000167754	FBXW7	ENSG00000109670
CTNNB1	ENSG00000168036	CCND1	ENSG00000110092
LRP1B	ENSG00000168702	E2F3	ENSG00000112242
MUC17	ENSG00000169876	LAMA4	ENSG00000112769
PTEN	ENSG00000171862	SF3B1	ENSG00000115524
ID4	ENSG00000172201	NFE2L2	ENSG00000116044
SYNPO2	ENSG00000172403	ARID1A	ENSG00000117713
CKS1B	ENSG00000173207	KMT2A	ENSG00000118058
HRAS	ENSG00000174775	PIK3CA	ENSG00000121879
PDE4DIP	ENSG00000178104	CHD6	ENSG00000124177
MUC16	ENSG00000181143	SNAI1	ENSG00000124216
TRAK1	ENSG00000182606	SYNE1	ENSG00000131018
PTCH1	ENSG00000185920	PDZD2	ENSG00000133401
DCC	ENSG00000187323	KRAS	ENSG00000133703
FAT4	ENSG00000196159	PDGFRA	ENSG00000134853
TRPV1	ENSG00000196689	APC	ENSG00000134982
NF1	ENSG00000196712	ESPL1	ENSG00000135476
DAPK1	ENSG00000196730	MDM2	ENSG00000135679
RYR2	ENSG00000198626	MYC	ENSG00000136997
NRAS	ENSG00000213281	ANG	ENSG00000214274

The driver genes overlapping with genes containing functional SNVs and indels are determined. Then, it is checked whether the overlapping genes are including mutations in TpC* or TpG* context and if they are not mutated in normal mucosa sample, but at least in one tumour sample.

2.2.6 Statistical Tests

The *Fisher's Exact Test* is used to calculate (by using "fisher.test" function in R) whether there is a significant effect for the categorized data by a contingency table: functional & nonfunctional mutations among "all samples

shared" (e.g. shared by all samples of a patient, including normal sample) & "all tumours shared" (e.g. shared by all tumours of a patient, excluding normal sample) mutations, as well as among all samples & tumour-private mutations (e.g. mutations that are specific to one tumour sample). The Fisher's Test, but not the Chi-square Test is used in the analysis, because Chi-square Test is used only for categorical data with large sample sizes, assuming each cell of the table is greater than 5. However, Fisher's Exact Test do not have such an assumption, could be used for any value for each cell. The test gives results about the significance of the difference among groups (e.g. p-value). Fisher's Exact Test is also used to check the amount and significance of the difference for TpC* and non-TpC* mutations among "all samples" & "all tumours shared" mutations (See Results, Section 3.4). Dinucleotide and trinucleotide context mutations' frequencies are also compared by Fisher's Exact Test for both "all tumours shared" and "tumour-private mutations" & "all samples shared" mutations (Table 9, Table 10, Table 11, Table 12 and Table 13)

Permutation tests were applied to check whether the overlap between the potential driver genes list and specific sets of variants identified in the experiments was statistically meaningful or not. Actually, permutation is done to simulate the null model by randomly mixing N variants (N being the number of "all tumours shared" mutations) 10,000 times, and each time checking the number of genes both including these randomly chosen N variants and overlapping with the potential driver gene list. Then, the result gives whether the observed data significantly differs from the simulated data in terms of overlapping with candidate driver genes list. p-value is.

2.2.7 Kataegis detection

Kataegis term is already mentioned in the introduction part, as being clusters of the mutations on the same strand. It is checked whether kataegis is seen in the samples within 10,000 nucleotide distance.

CHAPTER 3

RESULTS

3.1 Sequencing Results: Number of Reads, SNVs and Indels

To address the origin of the tumours question, exome sequencing of multifocal nonmuscle invasive urothelial bladder tumours was conducted at high depth, using tumour and normal mucosa samples collected from three patients during transurethral resection. The quality results of the sequencing are summarized in Table 4 and Table 5.

Table 4: Target size, total number of reads, read length and total sequenced bases for each sample

Sample	Target size	Total Reads	Read length	Total bases
1st patient-Tumour1 Base	50,000,000	54,022,290	100	5,402,229,000
1st patient-Tumour2 Apex	50,000,000	51,415,608	100	5,141,560,800
1st patient-Tumour2 Base	50,000,000	52,056,638	100	5,205,663,800
1st patient-Normal	50,000,000	56,497,854	100	5,649,785,400
2nd patient-Tumour1 Apex	50,000,000	52,336,578	100	5,233,657,800
2nd patient-Tumour1 Base	50,000,000	60,280,740	100	6,028,074,000
2nd patient-Tumour2 Apex	50,000,000	56,414,382	100	5,641,438,200
2nd patient-Normal	50,000,000	66,765,540	100	6,676,554,000
3rd patient-Tumor1 Apex	50,000,000	58,541,416	100	5,854,141,600
3rd patient-Tumor1 Base	50,000,000	61,787,922	100	6,178,792,200
3rd patient-Tumor2 Base	50,000,000	57,071,432	100	5,707,143,200
3rd patient-Normal	50,000,000	51,860,180	100	5,186,018,000

In Table 4, "Total bases" is calculated as the multiplication of the number of reads with read length.

In Table 5, "Expected coverage" describes the average number of times that a nucleotide is expected to be sequenced given the target size and the total number of sequenced bases (Lander *et al*, 2001). "Effective coverage"

represents the exact number of times that a nucleotide in the reference genome is covered by aligned reads of the sequencing experiment (Sims, Sudbery, Ilott, Heger, & Ponting, 2014). Expected coverage is calculated as the ratio of total bases to target size. Meanwhile, effective coverage is calculated as the ratio of mapped bases (mapped reads X read length) to target size.

Normally, 35X coverage would be needed to reliably call SNVs and small indels across 95% of the genome or 95% of the exome (Ajay, Parker, Abaan, Fajardo, & Margulies, 2010). Here, in Table 5, high effective coverage is found for each sample.

Table 5: Expected Coverage (shown as "Expected Cov."), Mapped Reads and Effective Coverage (shown as "Effective Cov.") for each sample

Sample	Expected Cov (X)	Mapped reads+	Effective Cov (X)
1st patient-Tumour1 Base	108.04	49,738,705	99.48
1st patient-Tumour2 Apex	102.83	45,235,390	90.47
1st patient-Tumour2 Base	104.11	46,177,418	92.35
1st patient-Normal	113.00	50,135,166	100.27
2nd patient-Tumour1 Apex	104.67	49,420,093	98.84
2nd patient-Tumour1 Base	120.56	57,033,018	114.07
2nd patient-Tumour2 Apex	112.83	53,200,303	106.40
2nd patient-Normal	133.53	62,527,762	125.06
3rd patient-Tumor1 Apex	117.08	55,805,883	111.61
3rd patient-Tumor1 Base	123.58	58,330,200	116.66
3rd patient-Tumor2 Base	114.14	53,567,052	107.13
3rd patient-Normal	103.72	49,072,940	98.15

The power of discovering variants is lowered by low base quality and by non-uniformity of the coverage (Sims et al., 2014). To handle this, some variants are eliminated according to what we call here "strict filtering criteria". The "strict filtering criteria" involve the GATK quality score (“PASS”), filtering out segmental duplications, and also including SNVs only consistently read in all 4 samples and identified at minimum depth 4. This latter rule means the SNV is covered by at least in 4 reads per sample. Note that we also exclude any SNVs in dbSNPv.138 (see section 2.1.2). Table 6 illustrates the number of variants remained after strict filtering:

Table 6: Number of SNVs and indels before (raw SNVs and indels) and after filtering (strict Filter SNVs and indels) for each sample

Sample	Raw SNVs	Raw INDELs	*Strict Filter SNVs	*Strict Filter INDELs
1st patient-Tumour1 Base	225,116	18,022	1,628	2,130
1st patient-Tumour2 Apex	216,393	17,838	1,628	2,130
1st patient-Tumour2 Base	213,508	17,747	1,628	2,130
1st patient-Normal	232,006	18,255	1,628	2,130
2nd patient-Tumour1 Apex	545,565	28,369	1,733	2,278
2nd patient-Tumour1 Base	585,967	29,126	1,733	2,278
2nd patient-Tumour2 Apex	571,255	28,595	1,733	2,778
2nd patient-Normal	618,910	28,894	1,733	2,778
3rd patient-Tumor1 Apex	690,221	30,147	1,628	2,555
3rd patient-Tumor1 Base	653,140	30,148	1,628	2,555
3rd patient-Tumor2 Base	664,479	29,874	1,628	2,555
3rd patient-Normal	600,240	29,479	1,628	2,555

As the result of filtering, the bulk of raw variants were eliminated and 1628-1733 SNVs and 2130-2555 indels remained for each sample.

3.2 Phylogenetic Trees

After transforming the SNV data into binary form (0 or 1), the filtered SNV datasets were used to calculate Euclidean distance matrices among samples of each patient. Then, these Euclidean distance matrices were used to construct neighbour-joining phylogenetic trees, reflecting genetic relationship of the tumour and normal samples relative to reference genome.

Figure 9 shows the reconstructed SNV phylogenies of the tumour and normal mucosa samples of Patient 1 and Patient 2:

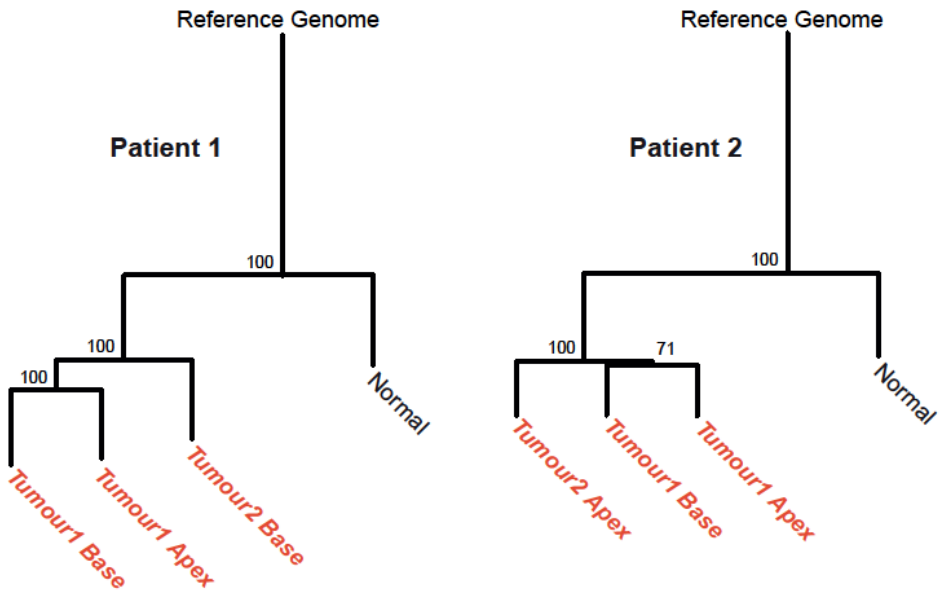

Figure 9: Phylogenetic trees constructed from SNV data for each patient. The bootstrap values of each node are shown next to each node. A bootstrap value of 100 means that, among 10,000 bootstraps, the same result was obtained 10,000 times. The Reference Genome sequence ("0" assigned to all positions) was used as outgroup (Acar, Özkurt *et al*, 2015; under review in BMC Cancer).

The phylogenetic trees of the 2 patients were reconstructed. Both trees were consistent with the topology of the hypothetical monoclonal phylogeny in Figure 8, with all tumours being grouped together. The normal tissue samples are expected to appear as the closest branches to the reference genome, as the tumour samples accumulate high number of mutations. As expected, in both trees, the normal tissue branch was the closest branch to the reference genome. High bootstrap values show that the tree topologies are robust to random sequencing and sampling errors; that is, the same topology is seen independent of which set of SNVs are chosen randomly.

The bootstrap value for the node resolving Tumour1 Apex and Tumour1 Base samples of the Patient 2 is 71, meaning that among 10,000 bootstraps, Tumour1 Apex and Tumour1 Base cluster together to the exclusion of Tumour2 Base only 71% of the time. This suggests that these two tumour branches are not resolved very robustly. However, this weak bootstrap support does not affect the conclusion that the tumours are clustered to the exclusion of the "normal sample" in the trees.

In contrast to Patient 1 and 2, in the tree of Patient 3, the normal sample is grouped together with 2 other tumour samples with high confidence bootstrap value (Figure 10). This is unexpected. One possible explanation is mix up of samples during sample preparation. Arguing against this, however, it is found that a large number of driver mutations for bladder carcinoma are shared by all samples of this patient, as will be discussed later (Table 14).

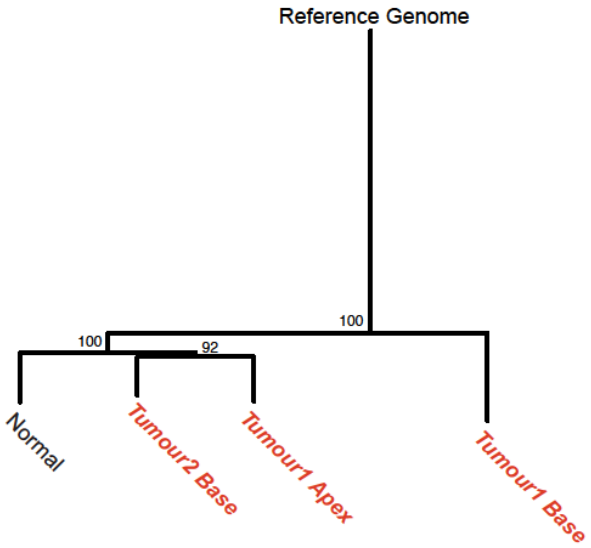

Figure 10: Phylogenetic tree constructed from SNV data for the 3rd patient.
The bootstrap values of each node are shown next to each node. The Reference Genome sequence ("0" is assigned to all positions) is used as outgroup (Acar, Özkurt *et al*, 2015; under review in BMC Cancer).

Therefore, it is suspected that the normal sample of the 3rd patient was contaminated with neoplastic material. In addition, in contrast to Patient 1 and Patient 2 trees, neither tumour branch appeared longer than the "normal sample" branch. Furthermore, the Tumor1 Apex and Tumor2 Base samples are clustered together, meaning that they are coming from the same monophyletic clade. This situation suggests that the Patient 3's phylogeny could also be explained by monoclonal origin hypothesis, rather than field effect hypothesis. As a result, these 3 SNV phylogenetic trees strongly support the monoclonal origin hypothesis.

Indels have been suggested to have higher resolution power than SNVs for reconstructing phylogenies than SNVs, because indels are less prone to homoplasia (Rokas & Holland, 2000). Therefore, indel data was also used to construct the phylogenies of the samples from 3 patients (Figure 11).

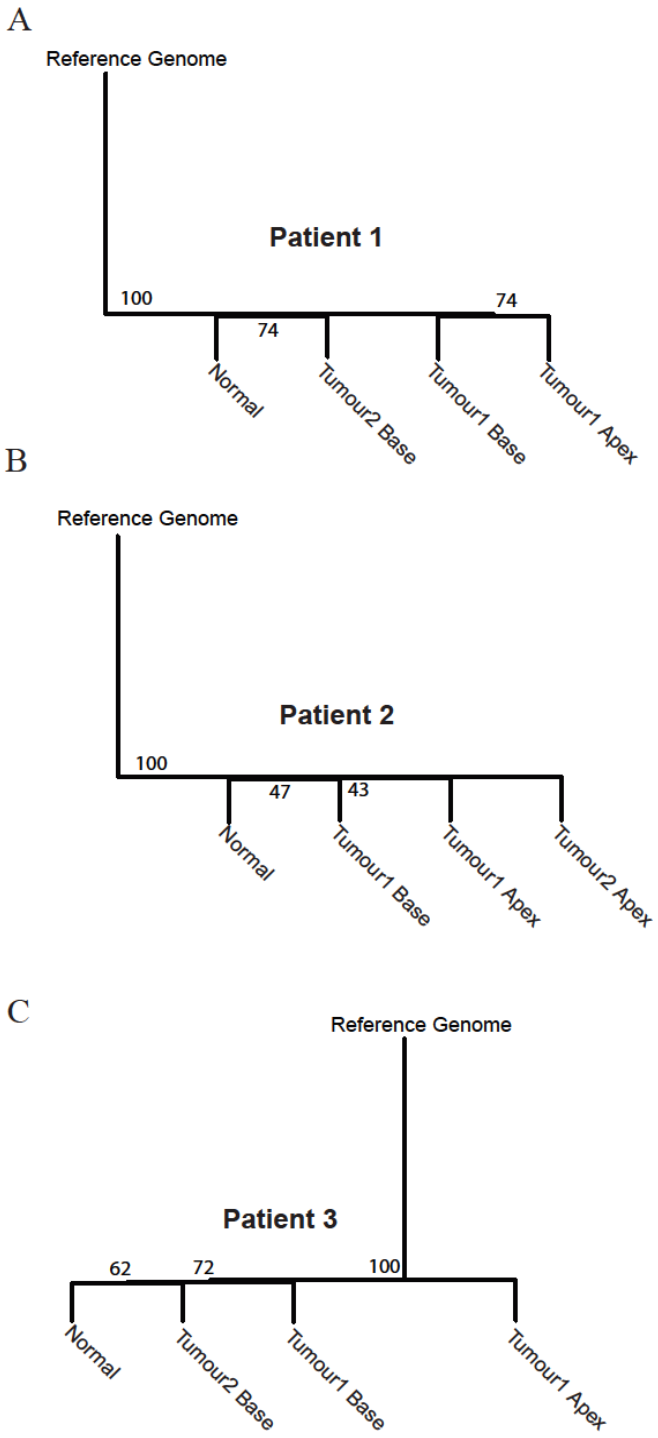

Figure 11: Indel tree of Patient 1 (A), Patient 2 (B) and Patient 3 (C)
The bootstrap values of each node are shown next to each node. 100 bootstrap value means, for 10,000 bootstraps, 10,000 times the same result is obtained.

The Reference Genome sequence is used as outgroup (Acar, Özkurt *et al*, 2015; under review in BMC Cancer).

As can be observed in Figure 11, indel phylogenies show low bootstrap values, implying that the tree topologies are not reliable. Only the branches separating the reference genome shows 100 bootstrap values for each tree, while the branches resolving the tumours and normal mucosa show low bootstrap values. Based on this result, indel datasets were excluded from the rest of the analysis.

3.3 Distribution of SNVs among samples and their functionality

Under the *monoclonal origin* hypothesis the tumours should share the same origin, and therefore higher frequency of "functional" mutation accumulation on the tumourogenesis branch could be expected. In contrast, under the *field effect* hypothesis, tumours should have evolved independently, and there will be a higher number of "tumour private" mutations than "all tumours shared" mutations. The topology of the Patient 1 and Patient 2 shows already that there is high mutation accumulation on the tumourogenesis branch, as expected in monoclonal origin hypothesis. However, in this section, we checked the exact number of the SNVs among different classes and detected whether mutations on tumourogenesis branch are more "functional" than "all samples shared" mutations.

SNVs of each patient were categorized as "Non-tumour associated" & "Tumour-associated" with respect to occurrence among samples.

Table 7: The SNV distribution of Patient 1 and Patient 2.

PATIENT 1			
	Non-tumour associated		**Tumour-associated**
All samples shared	842	**All tumours shared**	428
Normal private	3	**Tumour1 Base private**	80
Other	49	**Tumour1 Apex private**	49
-	-	**Tumour2 Base private**	57
-	-	**Other**	120
PATIENT 2			
	Non-tumour associated		**Tumour-associated**
All samples shared	1059	**All tumours shared**	473
Normal private	12	**Tumour1 Apex private**	14
Other	98	**Tumour1 Base private**	11
-	-	**Tumour2 Apex private**	18
-	-	**Other**	48

Table 8: The SNV distribution of the Patient 3.

PATIENT 3			
	Non-tumour associated		**Tumour-associated**
All samples shared	1304	**All tumours shared**	30
Normal private	20	**Tumour1 Base private**	178
Other	55	**Tumour1 Apex private**	20
-	-	**Tumour2 Base private**	8
-	-	**Other**	13

"Tumour1 Base Private", "Tumour1 Apex Private", "Tumour2 Base/Apex Private", "Normal private" terms represent SNVs *only* in that sample. The SNVs falling into the private category, except "Normal private", are grouped as "Tumour-associated" SNVs. "Normal private" SNVs are considered "Non tumour-associated" SNVs. "All samples shared" describes SNVs shared by all 4 samples, including normal mucosa. Thus, these SNVs represent the individual genotype, falling into the "Non tumour-associated" category. "All tumours shared" describes SNVs shared by all 3 tumour samples, but not by normal mucosa. Therefore, these SNVs represent the mutations related with tumourogenesis. "Other non tumour-associated" represents SNVs found in the normal sample, but not falling into the previous categories. Other "tumour-associated" refers to SNVs not found in the normal sample and not falling into the previous categories.

In Table 7, the distribution of SNVs to the samples of the patients (Patient 1 and Patient 2) are summarized. Table 7 shows that number of "all tumours shared" SNVs (n=842 to 1059) are higher than "all samples shared" SNVs (n=428 to 473). Moreover, the majority of the tumour-associated SNVs were found among all 3 tumours in each patient (58 to 84%), only 0.6 to 5% of them

(n=11-80 SNVs) were private to each tumour. This result supports again the *monoclonal origin* hypothesis in these patients. Table 8, however, shows that in Patient 3, 80% of SNVs were of the category "all samples shared". In contrast, only 15% of the SNVs were "tumour-associated", only 2% of these being "all tumours shared" SNVs. All of these results are already reflected on the phylogenies of the patients: Patient 1 and Patient 2's tree showing long tumourogenesis branch, while Patient 3's tree showing very short one (Figure 9 and Figure 10).

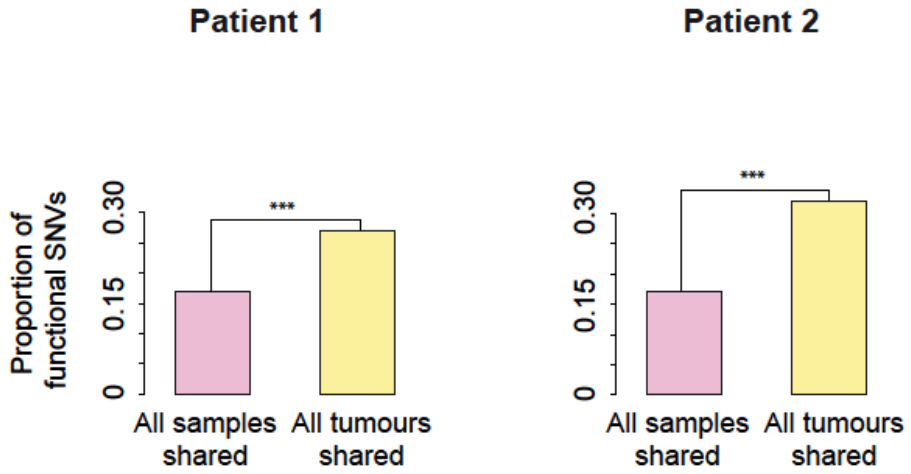

Figure 12: Proportion of functional SNVs in Patients 1 and 2.

SNVs were grouped as "all samples shared" (n=143 to 182) and "all tumours shared" (n=116 to 152) for Patient 1 and Patient 2. Functionality is defined as a SNVs putative effect on protein sequence (see text for details). *** indicates Fisher's exact test p<0.001 (Acar, Özkurt *et al*, 2015; under review in BMC Cancer).

After having determined the number of "all samples shared" and "all tumours shared" type of SNVs (Table 7), we investigated differences in the functional properties of SNVs in each of these classes for Patient 1 and Patient 2. To do this, from 1628 to 1733 SNVs were classified by the SnpEff software,

according to their impact on protein sequence. We found that from 350 to 354 (20 to 22%) of these SNVs were "potentially functional" (See Materials and Methods, section 2.2.3). Furthermore, these functional mutations were about 1.8 to 2.3 times more common among “all tumours shared" SNVs than among “all samples shared" SNVs (two- sided Fisher's exact test $p<10^{-5}$). Hence, besides the majority of "tumour-associated" mutations being "all tumours shared", a higher proportion of the latter mutations were functional than "all samples shared" mutations.

3.4 Distribution of SNV frequencies, Candidate Driver Gene Analysis, Permutation and Kataegis Results

After classifying SNVs with respect to occurrence among samples and their effect on protein sequence, we searched for signs of APOBEC enzyme activity (*i.e.* TpC* dinucleotide mutation pattern) among the patients’ tumour sample sequences (See Materials and Methods, section 2.2.4).

As can be observed in Figure 13, the category of “all tumours shared” SNVs had 6-6.3 fold higher frequency of TpC* mutations compared to "all samples shared" (e.g. SNVs representing the individual's genotype) in Patient 1 and Patient 2 (Fisher's exact test $p<10^{-41}$). To be more specific, TpC*->TpT* and TpC*-> TpG* mutation patterns, especially the former, is at higher frequency than TpC*->TpA*, which is consistent with APOBEC mutation activity. Figure 14 shows the frequency of mutations for Patient 3. The TpC* frequency of the Patient 3 does not show any particular elevation.

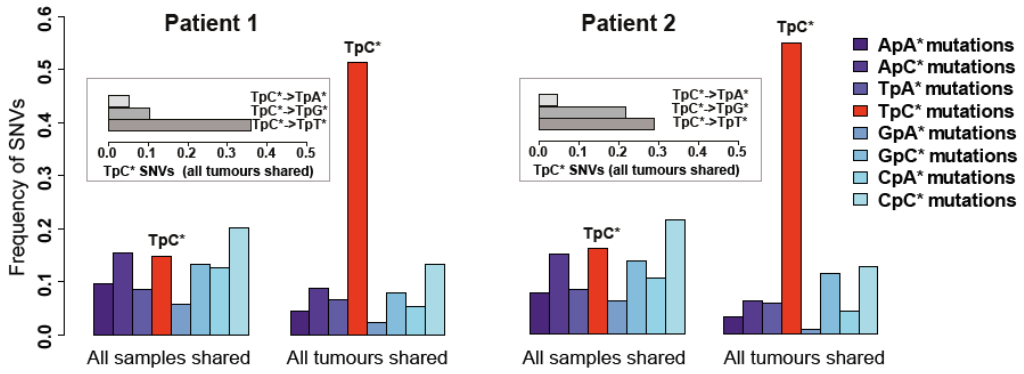

Figure 13: SNVs in dinucleotide context for Patient 1 and Patient 2.
SNV frequencies between "all samples shared" and "all tumours shared" categories are compared. Only A or C mutations in either strand are considered, in the manner of Nordentoft *et al*, 2014. "*" shows the base that is mutated; *e.g.* TpC* stands for the mutations that is from TpC to TpA or to TpG or to TpT: TpC->TpA, TpC->TpG, or TpC->TpT. TpC* mutation bars are red colored. Insets show the distribution of TpC* mutations frequency in "all tumours shared" SNVs (Acar, Özkurt *et al*, 2015; under review in BMC Cancer).

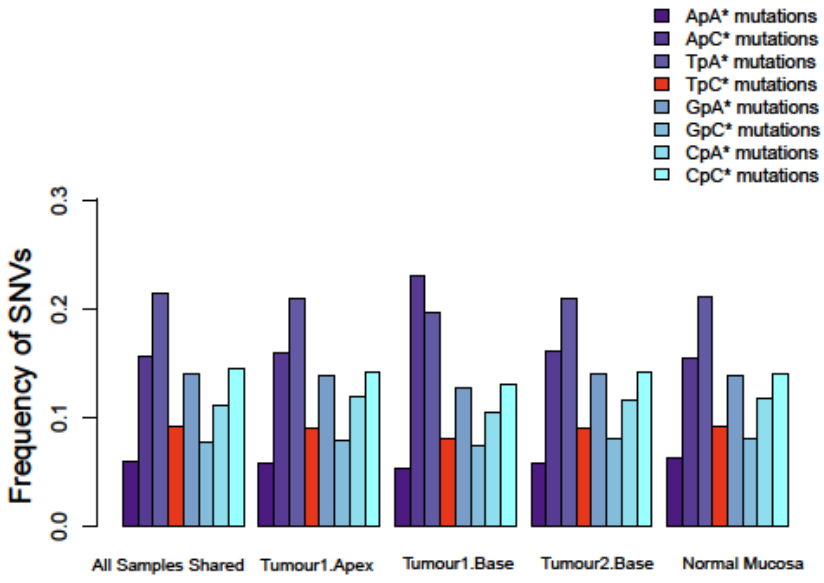

Figure 14: Frequencies of SNVs in dinucleotide context for Patient 3.

TpC* mutations bar is marked as red coloured. (Acar, Özkurt *et al*, 2015; under review in BMC Cancer)

We then compared the frequencies in dinucleotide and trinucleotide context between "all tumours shared" and "all samples shared" SNVs (Tables 9, 10, 12 and 13), as well as between "all tumours shared" and "all tumour private" SNVs (Table 11). In the tables, "E" refers to exponential notation. To give an example, 7.60E-01 is the same as the frequency of TpA* mutations in "all tumours shared" category being 0.76 of that of "all samples shared" category. As can be seen from Table 9, TpC* mutation is the most common dinucleotide mutation type (specifically TpC* to TpT*) when “all tumours shared” SNVs are compared with “all samples shared” SNVs in Patient 1 and Patient 2 (Fisher's exact test $p<10^{-41}$). No other mutation type in dinucleotide pattern is significantly common with odds ratio larger than 1 in these patients.

Table 9: Frequency differences between "all tumours shared" and "all samples shared" SNVs in dinucleotide context.

The odds ratio and two-sided Fisher's exact test p-values are shown for Patients 1 and 2. "E" is the exponential notation (power of 10). Bold lines mark significant results with odds ratio > 1 (Acar, Özkurt *et al*, 2015; under review in BMC Cancer).

Patient 1		
Mutated reference sequence	Odds Ratio	p-value
TpC*	**6.06E+00**	**4.41E-42**
TpA*	7.60E-01	2.69E-01
GpC*	5.63E-01	5.08E-03
GpA*	3.87E-01	4.56E-03
CpC*	6.12E-01	3.12E-03
CpA*	3.90E-01	3.27E-05
ApC*	5.23E-01	7.81E-04
ApA*	4.43E-01	1.24E-03
Patient 2		
Mutated reference sequence	Odds Ratio	p-value
TpC*	**6.29E+00**	**3.53E-52**
TpA*	6.86E-01	9.63E-02
GpC*	8.00E-01	2.19E-01
GpA*	1.28E-01	2.17E-07
CpC*	5.24E-01	2.16E-05
CpA*	3.93E-01	4.82E-05
ApC*	3.81E-01	5.85E-07
ApA*	4.12E-01	7.01E-04

Table 10 shows that TpC* is again the most common mutation type in dinucleotide context, all tumours shared SNVs being 2.5 to 4.6 times higher than the tumour private SNVs (p-value<10^{-4}). Indeed, TpCpA* and TpCpT* mutation types show highest frequency among trinucleotide mutation types

(Table 12 and Table 13), which is consistent with APOBEC3B mutation context (S. A. Roberts et al., 2013).

Table 10: The odds ratio and Fisher's exact test p-values for Patient 1 and Patient 2's frequency difference between "all tumours shared" and "all samples shared" SNVs in dinucleotide pattern and with resulting mutations.

(Acar, Özkurt *et al*, 2015; under review in BMC Cancer)

"E" is the exponential notation (power 10). Bold lines mark significant results with odds ratio > 1. NAs indicate no observation in that category.

Patient 1			**Patient 2**		
Base Change	Odds Ratio	p-value	Base Change	Odds Ratio	p-value
TpC*->TpT*	**5.20E+00**	**2.98E-28**	**TpC*->TpT***	**3.59E+00**	**8.55E-19**
TpC*->TpG*	**4.08E+00**	**4.32E-08**	**TpC*->TpG***	**6.65E+00**	**2.72E-25**
TpC*->TpA*	**2.23E+00**	**1.22E-02**	**TpC*->TpA***	**2.20E+00**	**1.30E-02**
TpA*->TpT*	1.76E+00	3.01E-01	TpA*->TpT*	7.44E-01	8.08E-01
TpA*->TpG*	6.54E-01	1.55E-01	TpA*->TpG*	6.98E-01	1.81E-01
TpA*->TpC*	4.35E-01	3.52E-01	TpA*->TpC*	6.08E-01	5.69E-01
GpC*->GpT*	6.17E-01	5.38E-02	GpC*->GpT*	6.64E-01	5.81E-02
GpC*->GpG*	9.64E-02	3.86E-03	GpC*->GpG*	9.21E-01	1.00E+00
GpC*->GpA*	9.35E-01	1.00E+00	GpC*->GpA*	1.20E+00	5.34E-01
GpA*->GpT*	4.35E-01	3.52E-01	GpA*->GpT*	3.71E-01	2.49E-01
GpA*->GpG*	4.50E-01	5.40E-02	GpA*->GpG*	1.03E-01	2.05E-05
GpA*->GpC*	1.95E-01	1.11E-01	GpA*->GpC*	NA	NA
CpC*->CpT*	5.81E-01	6.35E-03	CpC*->CpT*	5.07E-01	1.38E-04
CpC*->CpG*	6.48E-01	3.08E-01	CpC*->CpG*	2.95E-01	6.10E-03
CpC*->CpA*	9.35E-01	1.00E+00	CpC*->CpA*	1.21E+00	6.04E-01
CpA*->CpT*	7.34E-01	6.51E-01	CpA*->CpT*	1.80E+00	2.09E-01
CpA*->CpG*	3.23E-01	1.34E-04	CpA*->CpG*	2.44E-01	6.43E-06
CpA*->CpC*	4.41E-01	1.02E-01	CpA*->CpC*	3.68E-01	6.29E-02
ApC*->ApT*	4.79E-01	1.38E-03	ApC*->ApT*	3.59E-01	9.59E-06
ApC*->ApG*	7.54E-01	8.02E-01	ApC*->ApG*	2.88E-01	3.19E-02
ApC*->ApA*	6.78E-01	4.43E-01	ApC*->ApA*	7.43E-01	6.80E-01
ApA*->ApT*	3.25E-01	1.59E-01	ApA*->ApT*	9.32E-01	1.00E+00
ApA*->ApG*	2.96E-01	2.48E-04	ApA*->ApG*	2.30E-01	1.09E-04
ApA*->ApC*	1.44E+00	4.67E-01	ApA*->ApC*	NA	NA

We also compared "all tumours shared" SNVs with "all tumour private" SNVs. The results presented in Table 11 show that higher amount of mutations accumulated on the tumourogenesis branch, compared to the more recent, "tumour-private" branches.

Table 11: The odds ratio and Fisher's exact test p-values for the frequency difference between "all tumours shared" and "all tumour private" SNVs in dinucleotide context, for Patient 1 and Patient 2.
(Acar, Özkurt *et al*, 2015; under review in BMC Cancer)
"E" is the exponential notation (power 10). Bold lines mark significant results with odds ratio > 1.

Patient 1		
Mutated reference sequence	Odds Ratio	p-value
TpC*	**2.52E+00**	**6.56E-07**
TpA*	1.11E+00	8.58E-01
GpC*	4.16E-01	1.03E-03
GpA*	5.33E-01	1.98E-01
CpC*	7.99E-01	3.79E-01
CpA*	6.48E-01	2.06E-01
ApC*	7.06E-01	2.34E-01
ApA*	6.19E-01	2.35E-01
Patient 2		
Mutated reference sequence	Odds Ratio	p-value
TpC*	**4.60E+00**	**2.00E-05**
TpA*	8.39E-01	7.36E-01
GpC*	1.26E+00	8.06E-01
GpA*	8.41E-02	2.30E-03
CpC*	2.97E-01	8.81E-03
CpA*	1.47E+00	1.00E+00
ApC*	4.24E-01	3.36E-02
ApA*	6.20E-01	4.40E-01

Table 12: The odds ratio and Fisher's exact test p-values of frequency difference between "all tumours shared" and "all samples shared" SNVs in trinucleotide pattern for Patient 1.

(Acar, Özkurt *et al*, 2015; under review in BMC Cancer)

"E" is the exponential notation (power 10). Bold lines mark significant results with odds ratio > 1.

Mutated reference sequence	Odds Ratio	p-value
TC*T	**3.97E+00**	**1.29E-10**
TC*G	1.54E+00	1.38E-01
TC*C	**1.95E+00**	**6.14E-03**
TC*A	**1.19E+01**	**3.22E-30**
TA*T	1.12E+00	7.54E-01
TA*G	6.99E-01	6.28E-01
TA*C	7.36E-01	7.59E-01
TA*A	3.06E-01	6.57E-02
GC*T	5.22E-01	1.39E-01
GC*G	4.93E-01	4.58E-02
GC*C	8.71E-01	8.64E-01
GC*A	4.86E-01	1.99E-01
GA*T	8.93E-01	1.00E+00
GA*G	0.00E+00	2.15E-03
GA*C	1.62E-01	7.19E-02
GA*A	7.85E-01	7.83E-01
CC*T	5.43E-01	9.23E-02
CC*G	4.72E-01	6.99E-03
CC*C	5.03E-01	5.25E-02
CC*A	1.58E+00	1.43E-01
CA*T	3.43E-01	6.41E-03
CA*G	4.21E-01	3.83E-02
CA*C	4.32E-01	1.71E-01
CA*A	5.41E-01	2.70E-01
AC*T	5.27E-01	9.78E-02
AC*G	2.95E-01	1.39E-03
AC*C	7.81E-01	5.12E-01
AC*A	8.40E-01	8.45E-01
AA*T	3.20E-01	1.70E-02
AA*G	4.21E-01	1.04E-01
AA*C	3.43E-01	9.46E-02
AA*A	1.18E+00	7.92E-01

Table 13: The odds ratio and Fisher's exact test p-values of frequency difference between "all tumours shared" and "all samples shared" SNVs in trinucleotide pattern for Patient 2.

(Acar, Özkurt *et al*, 2015; under review in BMC Cancer)

"E" is the exponential notation (power 10). Bold lines mark significant results with odds ratio > 1.

Mutated reference sequence	Odds Ratio	p-value
TC*T	**4.77E+00**	**1.61E-17**
TC*G	**1.95E+00**	**7.83E-03**
TC*C	1.50E+00	1.09E-01
TC*A	**7.15E+00**	**8.62E-29**
TA*T	1.06E+00	8.77E-01
TA*G	3.91E-01	1.48E-01
TA*C	1.03E+00	1.00E+00
TA*A	2.64E-01	2.14E-02
GC*T	7.98E-01	5.47E-01
GC*G	6.92E-01	2.26E-01
GC*C	1.25E+00	5.00E-01
GC*A	6.65E-01	3.80E-01
GA*T	8.42E-02	1.08E-03
GA*G	1.48E-01	3.02E-02
GA*C	1.38E-01	3.09E-02
GA*A	2.47E-01	1.89E-01
CC*T	6.61E-01	1.73E-01
CC*G	4.74E-01	2.97E-03
CC*C	4.91E-01	1.53E-02
CC*A	8.36E-01	7.41E-01
CAA	1.16E-01	1.23E-02
CA*T	4.27E-01	4.32E-02
CA*G	4.53E-01	5.91E-02
CA*C	5.79E-01	3.10E-01
AC*T	1.76E-01	7.22E-04
AC*G	3.08E-01	7.76E-04
AC*C	5.36E-01	1.06E-01
AC*A	7.41E-01	4.90E-01
AA*T	2.83E-01	9.39E-03
AA*G	7.98E-01	8.05E-01
AA*C	0.00E+00	4.55E-03
AA*A	7.08E-01	5.44E-01

We find that in Patient 1 and Patient 2, TpC* mutations are much more accumulated on the tumourogenesis branch (represented by the "all tumours shared" category), than the individual's unique genotype (represented by the "all samples shared" category) (Table 9). When SNVs are considered with the original and derived substitution (Table 10), TpC*->Tp*T mutations appear as the most common mutation pattern, consistent with APOBEC activity signature on genome. Morever, the most common type of mutation in trinucleotide context (Table 12, Table 13) appears as TpCpT* in both Patient 1 and Patient 2. TpCpT* is already claimed to be the most common motif of APOBEC3B activity (S. A. Roberts et al., 2013).

To check whether TpC* mutations could be the inducer of tumour related mutations, we checked TpC* mutations overlapping with potential driver genes (Table 14). Half of the "all tumours shared" mutations affecting bladder cancer driver genes appear to be in TpC* context.

Table 14: The overlap between candidate driver genes and functional SNVs for each patient.

(Acar, Özkurt *et al*, 2015; under review in BMC Cancer)

Mutation Effect and Mutation Type are annotated with SnpEff software. Mutated Samples shows the samples that show the mutation. Base Conversion shows the insertions, deletions, SNVs and their resulting mutations in dinucleotide context. TpC* mutations are marked as bold.

PATIENT 1				
GENE NAME	MUTATION EFFECT	MUTATION TYPE	MUTATED SAMPLES	BASE CONVERSION
KMT2A	MODERATE	MISSENSE	all tumours shared	**TpC*>TpT***
KMT2D	MODERATE	MISSENSE	all tumours shared	**TpC*>TpT***
FAT4	MODERATE	MISSENSE	all tumours shared	**TpC*>TpT***
ERCC2	MODERATE	MISSENSE	all tumours shared	GpA*>GpC*
ANK2	MODERATE	MISSENSE	all tumours shared	TpA*>TpG*
KDM6A	MODERATE	MISSENSE	all tumours shared	TpA*>TpG*
PIK3CA	MODERATE	MISSENSE	Tumour1.Base private	GpC*>GpG*,TpT*>TpA*
FGFR3	MODERATE	MISSENSE	Tumour1.Base private	**TpC*>TpT***
TSC1	HIGH	FRAMESHIFT	Tumour1.Apex private	C deletion
HMCN1	MODERATE	MISSENSE	all samples shared	T insertion
ARID1A	MODERATE	MISSENSE	all samples shared	ApA*>ApG*
HCN1	MODERATE	MISSENSE	all samples shared	ApA*>ApG*
PDE4DIP	HIGH	FRAMESHIFT	all samples shared	A insertion
MAP3K1	MODERATE	CODON DELETION	all samples shared	CAA deletion
PATIENT 2				
GENE NAME	MUTATION EFFECT	MUTATION TYPE	MUTATED SAMPLES	BASE CONVERSION
RB1	HIGH	NON SENSE	all tumours shared	ApC*>ApG*
ATM	MODERATE	MISSENSE	all tumours shared	TpC*>TpG*, CpC*>CpT*
LRP1B	MODERATE	MISSENSE	all tumours shared	**TpC*>TpG***
MUC16	MODERATE	MISSENSE	all tumours shared Tumour1.Apex,Tumour1.Base	**TpC*>TpT***
KMT2C	MODERATE	MISSENSE	all samples shared	GpC*>GpG*
HMCN1	MODERATE	MISSENSE	all samples shared	GpC*>GpT*
PATIENT 3				
GENE NAME	MUTATION EFFECT	MUTATION TYPE	MUTATED SAMPLES	BASE CONVERSION
KDM6A	HIGH	NON SENSE	all samples	**TpC*>TpT***
RYR2	MODERATE	MISSENSE	all samples	**TpC*>TpT***
KMT2D	MODERATE	CODON DELETION	all samples	TGCTGCTGT deletion
NF1	MODERATE	MISSENSE	all samples	**TpC*>TpG***
LRP1B	MODERATE	MISSENSE	all samples	ApA*>ApG*
SYNE1	MODERATE	MISSENSE	private to one sample	**TpC*>TpG***
STAG2	HIGH	SPLICE SITE DONOR	shared by 3 samples	ApC*>ApA*

Moreover, to check whether observing that number of both "functional" and "all tumours shared" mutations with bladder cancer driver genes was meaningful or not, we performed a simulation test using the R programming language, where we randomly sampled N variants (N being the number of "functional" and "all tumours shared" mutations) among all SNVs 10,000 times. In other words, we permuted the "all tumours shared" label among SNVs. The permutation test p-values was significant for Patient 1 and was not

for Patient 2 (p-value = 0.087 and 0.172 for Patient 1 and 2, respectively), implying that the number of both "functional" and "all tumours shared" mutations overlapping with potential driver genes is not significantly higher than the number of mutations in the randomized data.

As noted earlier, Table 14 shows that Patient 3 has a large number of missense mutations in known bladder cancer genes in the normal mucosa sample as well. This situation enhances the possibility that the normal mucosa is contaminated with neoplastic material.

Network analysis of potential driver genes showing TpC* mutation pattern is also performed in Patient 1 and Patient 2 (Figure 15) (Warde-Farley et al., 2010). As expected, the network implies that the potential driver genes of 2 independent cases share some pathways. Some genes in the network are co-expressed or co-localized, while some genes interact with each other.

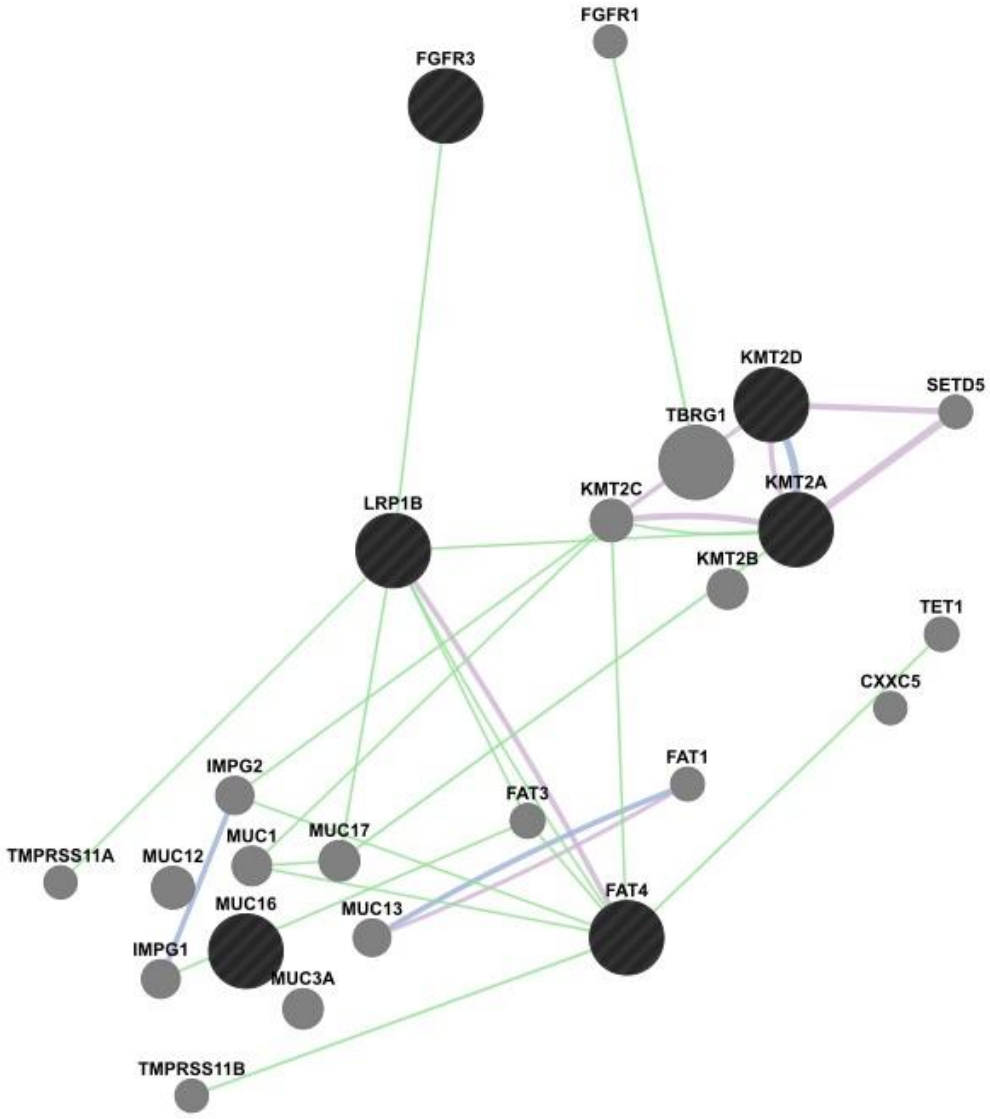

Figure 15: Network analysis of driver genes showing TpC* mutations in Patient 1 and Patient 2.

Black circles shows the potential driver genes showing TpC* mutation in Patient 1 and Patient 2. Purple , blue and green lines represent co-expression, co-localization and genetic interactions, respectively.

It can be inferred when these TpC* mutations occurred, by looking at the distribution of TpC* mutations within the phylogenetic trees. As can be seen from Table 9, Table 10 and Figure 13, TpC* mutations are 6 times more

common on the tumourogenesis branch than the ancestral branch. In addition, the tumourogenesis branch includes several TpC* mutations occurring in potential driver genes in bladder cancer (Table 14). Also, TpC* mutations are almost 2 to 5 times higher on the tumourogenesis branch than on more recent branches (Table 11). This suggests that TpC* mutations occurred early in development, during tumourogenesis. Later, TpC* accumulation must have subsided. This description was summarized in Figure 16:

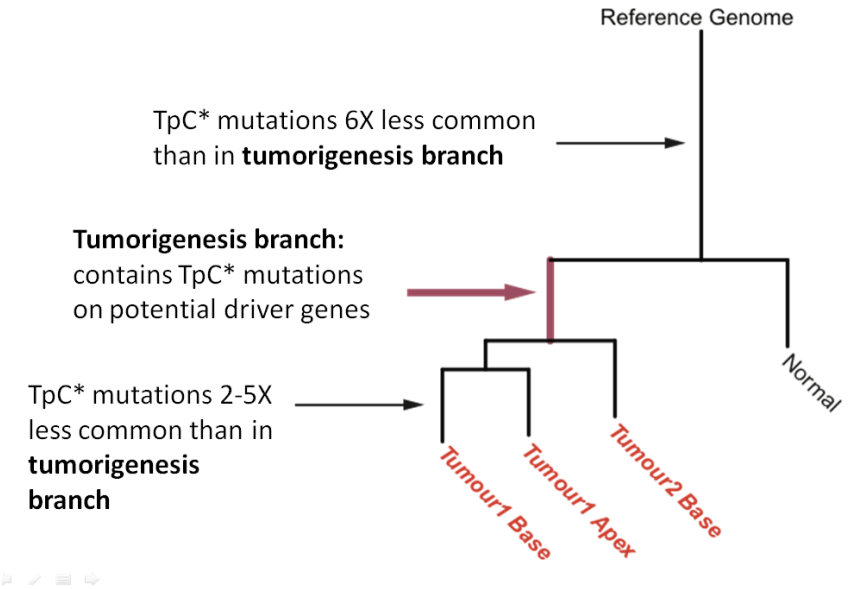

Figure 16: TpC* mutation frequencies reflected on SNV phylogenetic tree. The red colored branch shows the tumourogenesis branch**.**

Finally, the existence of *kataegis* (see section 1.3.4), that is, an aggregation of APOBEC-induced mutations previously detected in the context of bladder and other cancers, was investigated in data from Patients 1 and 2. To do this, we measured the clustering of TpC* mutations (the numbers of TpC*s within 10,000 nucleotide distance of each other in each sample) on the same strand. Then a permutation test was performed by 10,000 times mixing up the positions of mutations randomly and checking the number of TpC* mutations

in proximity in this randomized data. Comparing the results of the kataegis test with the permutated kataegis test allowed testing whether the number of TpC* mutations in proximity in the original test is significant or not. The result was insignificant for both patients.

CHAPTER 4

DISCUSSION

4.1 Comments on Previous Studies

Studying cancer as a Darwinian process can help understand evolutionary dynamics of cancer and figure out its progression, which could have clinical and precautionary implications. Here, a still-debated question is investigated: the origin of multifocal bladder tumours. This problem has been studied many times in the past decades (since 1992, by Sidransky *et al*) and conflicting results have been reached (Table 1 and Table 2). The reason for these conflicting results may be that the previous studies included a limited number of genotypic variants, and thus could have had limited statistical power (variant n=1 to 800. Note however that number of tumours collected was usually high: n=2 to 73; Table 1 and Table 2). Thus, it could have been power issues that was leading to inconsistent results. Also, some of these studies investigated late stage invasive carcinomas, where high mutation accumulation can further obscure genetic relationships among tumours.

Moreover, the methods used to detect clonality could have limitations. In some of the studies (Sidransky *et al*, 1992; Li & Cannizzaro, 1999; Jones *et al* 2005) X chromosome inactivation pattern was utilized to evaluate clonality, as always the same X chromosome is expected to be methylated in the case of monoclonal origin of the tumours. However, X inactivation pattern is limited to female patients, whereas the majority of bladder patients are males (Knowles & Hurst, 2014). This may lead to limitations in sample availability (Christian Hafner et al., 2002).

Using X inactivation is also complicated because of unstable methylation in tumours (Jones and Buckley, 1990) and preferential amplification of the allele having lower molecular weight (Mutter & Boynton, 1995). The human androgen receptor (*HUMARA*) gene is usually used to detect X-inactivation pattern of the patients. For example, Jones *et al* (2005) used *HUMARA* gene alleles to infer X inactivation pattern of the bladder tumours, and their results supported the field effect hypothesis. Low quantity and quality of template can cause imbalances in PCR (Polymerase Chain Reaction) products, preferentially amplifying the *HUMARA* allele having lower molecular weight. This bias can distort the results in favour of monoclonal origin hypothesis, leading to inaccurate X-inactivation signals. However, even if a study supports monoclonal origin hypothesis by detecting X-inactivation pattern, the result could be unreliable.

Another method used to detect the origin of multifocality is LOH analysis. This method could also have caveats. The analysis is usually done for several loci. However, not observing a common deletion among the tumours compared could be because of neglecting the right marker loci whose deletion is actually shared among tumours (Christian Hafner et al., 2002). If there is limited number of mutations shared by all tumours, the tumourogenesis branch will be short (the common ancestry did not extend long enough), and the evolutionary relatedness of the tumours could remain unnoticed (see Figure 8). Hence, it is really crucial to detect a high number of variants to address the origin of multifocal tumours. Exome sequencing provides a high number of variants, and could resolve the evolutionary relatedness even if the tumourogenesis branch is fairly short.

4.2 Interpretation of SNV data

Here, this long-debated problem was investigated with a different approach: exome sequencing. Exome sequencing yields high number of variants (n=1628 to 1733 in our case) and it is a cost-effective next-generation sequencing technique. For each patient, the statistical power of the technique used to reveal the origin of multifocal tumours problem is very high, owing to the substantially higher number of variants compared to previous techniques that utilized a small number of loci. Even if there occur problems in detection of some of the variants, exome sequencing, yielding thousands of variants, provides enough data that the analysis is minimally affected. The high quality of the SNV data also allows avoiding caveats of the previous techniques, such as obtaining the wrong signal from X-inactivation detection.

Tumour DNA sequencing requires higher sequence depth than normal DNA does in order to identify accurately variants specific to these tumours (i.e. SNVs). Compared to whole genome sequencing, exome-sequencing can be more advantageous in cancer studies as it generates high-depth data. However, exome-sequencing does not provide information about mutations in non-coding regions and genetic alterations at chromosomal level. These can be limitations of exome-sequencing to study tumourogenesis mechanism. However, to reveal evolutionary relationships of tumours, exome-sequencing could be optimal method for now.

Using this approach, we generated SNV-based phylogenies for three male patients. Phylogenies of two of these, Patients 1 and 2 (Figure 9) showed similar topology to the hypothetical monoclonal origin tree (Figure 8), and their branches separated with high confidence. In addition, in both trees, the tumour branches appeared evidently longer than the normal sample branch.

This indicates the accumulation of excess number of mutations in tumour lineages.

For the Patients 1 and 2, 52 to 61% of the mutations were shared among all samples, and these were considered representing individual's unique genotype. A clear majority of the tumour-associated mutations were "all tumours shared" mutations (58 to 84%). Only 0.6 to 5% (n=11 to 80 SNVs) of these "tumour-associated" mutations were private to each tumour (Table 7). This implies much higher mutation accumulation on the putative tumourogenesis branch than private branches.

SNVs of the Patient 1 and 2 were classified according to their "functionality", i.e. changing protein sequence or not. "All tumours shared" SNVs were found to be 1.8 to 2.3 times more "functional" than "all samples shared" SNVs (Fisher's exact test $p<10^{-5}$) (Figure 12). This indicated that the tumour lineage branch, tumourogenesis branch, includes high number of "functional" mutations. This situation increases the possibility of tumourogenesis occurred on this branch, hence the tumours coming from the same origin. As a result, the result supports again the monoclonal origin hypothesis.

Exosomes, being small (50-90 nm) plasma membrane vesicles, are released from cell to extracellular matrix when endosomes fuse with plasma membrane (van Niel, Porto-Carreiro, Simoes, & Raposo, 2006). Exosomes are also observed in tumour cells . Valadi et al (2007), showed that exosomes can carry both mRNA and microRNA, deliver these RNAs to a recipient cell; furthermore delivered RNAs can be functional in recipient cells (Valadi et al., 2007). Exosome sharing could be possible also in multifocal bladder tumours, leading to some shared genomic regions among tumours, thus misinterpretation of the tumours as coming from monoclonal origin. However, exome-sequencing data provides high number variants (~2000 variants), hence high statistical power. Thus, it is robust to such later sharing events. Also, high

bootstrap values in the phylogenies of Patient 1 and Patient 2 shows that at each resampling, the same tree topology is reconstructed, being independent of the exact set of SNVs included.

4.3 Patient 3

Meanwhile, the phylogeny of the Patient 3 appeared bizarre (Figure 10); where the normal sample is grouped with tumour samples. In addition, neither tumour branch appeared longer than the normal sample branch. Moreover, Tumour2 Base was grouped with Tumour1 Apex. The simplest explanation for this pattern is that the Patient 3 normal sample is contaminated with neoplastic material.

Also, Patient 3's SNV distribution profile looked different from Patients 1 and 2. The great majority of SNVs, 80% of them, were shared among all samples, whereas only 15% of them were "tumour-associated". Only 2% of these "tumour-associated" mutations were "all tumours shared" mutations (Table 8). Furthermore, Patient 3's normal sample contains several functional mutations on driver genes known for bladder carcinoma (Table 14), much more than observed for the other patients. Thus, it can be concluded that the Patient 3's normal mucosa sample looks very similar to the tumour sample profile; it is most probably contaminated with neoplastic material.

On the other hand, the Patient 3 phylogenetic tree topology might still be explained best by the clonal hypothesis, rather than the field effect hypothesis. This is because Tumour1 and Tumour2 samples do not form separate monophyletic clades, which would be expected under the field effect hypothesis. In all, it can be claimed that the third patient's SNV dataset also supports the clonal hypothesis, where tumours are evolutionary related rather than having evolved independently.

4.4 Indel Trees & Possible Reasons for Low Bootstrap Values

The same phylogenetic relationships among tumours are reconstructed using indel trees. However, indel trees had low bootstrap support values and their branches remained unresolved. This result can arise because of several reasons. First of all, indel calling is a process more prone to error than SNV calling (O'Rawe et al., 2013). Low concordance of different variant calling pipelines for indels has been previously reported. For example, indels called by GATK Unified Genotyper (v1.5), SOAPindel (v1.0) and SAMtools (v0.1.18), showed only 26.8 % of overlap, while SNV overlapping was 57.4% (O'Rawe et al, 2013). Moreover, Fang *et al* (Fang et al., 2014) reported that whole genome sequencing-specific indels are validated at much higher rates than exome sequencing indels (84% & 57%), with exome sequencing missing large indels.
Secondly, we did not apply most filters considered for SNVs on the indel data; and only the GATK quality score ("PASS") was used in indel filtering. Only SNVs consistently read in all 4 samples and identified at least in 4 samples are included in SNV datasets. Also, segmental duplications are filtered out from SNV datasets. However, none of these filtering were applied to the indel data. Thus, our indel datasets may not be as reliable as our SNV datasets. It remains possible that if we were to use algorithms that reduce the false positive and/or negative discovery rate in indel calling and filter the data more strictly, we could obtain phylogenies with higher bootstrapping values.

In short, a conclusion cannot be reach about the origin of these tumours from indel data for neither of the patients, because indel trees show low bootstrapping values most probably as the consequence of high rate of false positive and/or negative rates affecting indel calling. Another possible explanation for the unresolved branches in indels tree could be homoplasia (convergent changes). However, it is very unlikely that the same indel occurs twice exactly at the same position. In fact, homoplasia are so rare among indels

that they are claimed to be more reliable as a tool for phylogenomics than nucleotide substitutions (Rokas & Holland, 2000). Also, it can be speculated that indels could arise at later stages in carcinogenesis; thus they would be underrepresented in the tumourigenesis branch and appear private to tumours.

On the other hand, SNV datasets should be more reliable, first because they were subjected to stricter filtering criteria, and also because the trees show high bootstrap values, that is they give the same tree topology at each resampling, independent of the SNVs included. Therefore, they allow us to reach a conclusion about the origin of these tumours.

4.5 Detection of APOBEC activity:

Finally, the existence of TpC* substitution accumulation in the tumours of the first two patients was studied, in order to investigate recent claims for a relationship between bladder carcinoma and APOBEC-mediated RNA editing as immune response (S. A. Roberts et al., 2013). Indeed, this was the case in Patients 1 and 2: Compared to "all samples shared SNVs" (ancestral branch), "all tumours shared" SNVs (tumourogenesis branch) included 6 to 6.3 fold higher TpC* substitution frequencies, particularly TpC*->TpT*/TpG*, relative to "all samples shared" category mutations (Fisher's exact test $p<10^{-41}$) (Figure 13) (Table 9, Table 10). Furthermore, when the same data was analyzed in trinucleotide context, TpC*pA and TpC*pT mutations were higher in frequency, accordant with the APOBEC3B mutational signature (Burns et al., 2013) (Table 12, Table 13). On the other hand, in Patient 3, TpC* substitution frequency did not show any increase in any of the samples, each sample showing more or less the same frequency of SNVs in dinucleotide context (Figure 14).

As explained in section 1.3.3 of the Introduction, there is strong evidence for the APOBEC3B enzyme participating in the response to HPV infection. Thus, our results can be explained by Patients 1 and 2 having been infected with HPV before tumourogenesis started.

TpC* mutation frequencies of "all samples shared" SNVs (ancestral branch) were also compared to SNVs private to each tumour (more recent branches) for Patient 1 and 2. Compared to ancestral branch, more recent branches included 2.5 to 4.6 fold higher TpC* substitution frequencies (p-value$<10^{-4}$) (Table 11). It can be concluded that APOBEC enzyme activity increased early in the development of carcinoma, and subsequently diminished in the tumours of these two patients. However, the same pattern is not observed in Patient 3.

The overlap between the bladder cancer driver genes list for bladder cancer (Table 3) and functional mutations is compared for Patient 1 and 2. The majority of overlapping mutations were "tumour-associated" for both patients (n1=9/14, n2=4/6). Moreover, approximately half of these were TpC* substitutions (Table 12). However, the results of permutation test, simulating the significance of the overlap between the number of both "functional" and "all tumours shared" mutations and the driver gene list, was significant for Patient 1 while not for Patient 2. This indicates that the number of "all tumours shared" mutations occurring on bladder cancer driver genes is beyond the chance for Patient 1 but not for Patient 2.

To detect signs of APOBEC enzyme activity on the genome, we further investigated the so-called *kataegis* signature in SNV data from Patient 1 and Patient 2. However, we did not detect clustering of TpC mutations on the same strand. This result may not be surprising as we were analyzing exome data, and therefore missing any SNV clustering that overlaps introns and intergenic regions. If the same test would have been applied to whole genome sequences of these patients, we might have found a positive signal for *kataegis*.

In addition, we detected also if Ras oncogenes are also mutated in the patients or not. Ras isoforms (H-Ras, K-Ras, N-Ras) are proto-oncogenes belonging to GTPase protein family, can potentially trigger tumourogenesis when activated at codons 12, 13 or 61 by a single mutation (Quinlan & Settleman, 2009). K-Ras isoform is by far more frequently detected than other isoforms in different cancer types. However, H-Ras is the most common mutated member of Ras family in urinary tract carcinoma contrary to other cancer types, appearing 11% of urinary tract carcinoma cases (Prior, Lewis, & Mattos, 2012). Incidence of H-Ras mutation in urinary tract is relatively high compared to its incidence in other cancer types. Actually, the highest incidence of H-Ras mutation is observed in salivary gland cancer, being observed 15% of all patients. H-Ras usually favours GGC to GTC mutation context in bladder carcinoma (Prior et al., 2012). However, potential driver gene mutation analysis of the 3 patients showed none of Ras isoforms are mutated in the tumour of these patients.

4.6 Conclusion & Therapeutic Interventions

In conclusion, this study demonstrates the utility of exome sequencing and population genetics analysis for answering the origin of multifocal tumours question. Also, the study could have ramifications for bladder cancer therapy. Determining the origin of multifocal tumours and the timing of genetic alterations can provide new therapeutic interventions and change in therapy strategy. For example, use of gene therapy could be very effective against bladder carcinoma if tumours are coming from a clonal origin (Duggan et al., 2004). If tumours are sharing the same oncogenic mutation in a gene, at the early stages of cancer, a therapy fixing the mutation on this gene could be more effective, compared to the situation where tumours include different oncogenic mutations.

Additionally, the finding that APOBEC3B activity is increased in the early of the tumour development, but calms down during later cancer development suggests that targeting APOBEC3B activity may not be proper way for bladder cancer therapy. Of course, this pattern is yet demonstrated only in 2 patients, and needs to be replicated in larger samples. Also notably, our observation on the rate of APOBEC3B activity is different from recent studies. Zhang *et al* (2014) and Bruin *et al* (2015) , studying lung cancer, reports not calming down but accelerated accumulation of APOBEC-mediated mutations. This can be interpreted as different cancers behaving differently in terms of APOBEC activity. TpC* mutations, being potentially damaging, can impair tumour cell survival. In this case, decrease in of TpC* mutation accumulation rate after tumourogenesis could also be explained by negative selection acting on these deleterious sites.

CHAPTER 5

CONCLUSION

In the study, we investigated the evolutionary history of multifocal bladder tumours by using a statistically powerful method: exome-sequencing. 3/3 patients' multifocal bladder tumour sequence analysis suggested monoclonal origin hypothesis. Then, we returned to the very beginning of the story, and asked but how the very first progenitor cell is formed. According to TpC* substitutions analysis, we proposed APOBEC-mediated RNA editing to be tumourogenesis trigerring mechanism. All the evolutionary history of the multifocal bladder tumours is summarized in the Figure 17:

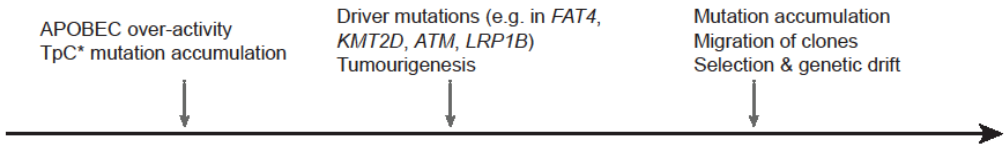

Figure 17: Hypothetical timeline summarizing the evolutionary history of the multifocal bladder tumours
(Acar, Özkurt *et al*, 2015; under review in BMC Cancer)

According to the results, as illustrated in Figure 17, while APOBEC3B enzymes editing virus genome (most probably HPV genome) as an immune response to infection, inadvertently edited host genome, specifically in TpC*pA trinucleotide context. However, editing acted on genes that are driver for bladder carcinoma, thus led to tumourogenesis. Finally, high mutation accumulated on the neoplasm and clones migrated and formed other evolutionarily related tumours. These newly arised tumours continued to be exposed to evolutionary forces, especially selection and genetic drift, as the

result of high competition among tumour populations. Finally, the tumours diverged from each other in the wake of evolutionary forces but still sharing some mutations.

Here, we developed the hypothetical timeline by considering the result of SNV analysis of spatially collected multifocal tumour samples. It could also be possible to develop the timeline by analyzing chronologically sampled tumours. Early and late sample collection from the same multifocal tumours could lead to higher resolution of the timeline, allowing to draw the sequence of driver gene mutations and identify the mutations that are eliminated by selection or genetic drift. In this case, it would be possible to figure out the reason of decrease in of TpC* mutation accumulation, either by calming down of APOBEC enzyme activity or by negative selection acting on tumours.

Also, it should be kept on mind that, this timeline is constructed by analyzing just 12 bladder samples exome data from 3 patients, thus it should not be generalized. To further support these results, more multifocal bladder tumours should be sequenced and expression analysis of the APOBECs (transcriptome and immunohistochemistry analysis) should be done.

REFERENCES

Acar & Ozkurt. (2015). Determining the origin of synchronous multifocal bladder cancer by exome sequencing. *BMC Cancer*.

Agrawal, N., Frederick, M. J., Pickering, C. R., Chang, K., Li, R. J., Fakhry, C., … Street, N. W. (2011). NIH Public Access, *333*(6046), 1154–1157. doi:10.1126/science.1206923.Exome

Ajay, S. S., Parker, S. C. J., Abaan, H. O., Fajardo, K. V. F., & Margulies, E. H. (2010). Accurate and comprehensive sequencing of personal genomes, 1498–1505. doi:10.1101/gr.123638.111.Freely

Alexandrov, L. B., & Stratton, M. R. (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics & Development*, *24*, 52–60. doi:10.1016/j.gde.2013.11.014

Anant, S., & Davidson, N. O. (2003). Hydrolytic nucleoside and nucleotide deamination, and genetic instability: a possible link between RNA-editing enzymes and cancer? *Trends in Molecular Medicine*, *9*(4), 147–152. doi:10.1016/S1471-4914(03)00032-7

Anderson, K., Lutz, C., van Delft, F. W., Bateman, C. M., Guo, Y., Colman, S. M., … Greaves, M. (2011). Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*, *469*(7330), 356–61. doi:10.1038/nature09650

Armitage, A. E., Katzourakis, A., de Oliveira, T., Welch, J. J., Belshaw, R., Bishop, K. N., … Iversen, A. K. N. (2008). Conserved footprints of APOBEC3G on Hypermutated human immunodeficiency virus type 1 and human endogenous retrovirus HERV-K(HML2) sequences. *Journal of Virology*, *82*(17), 8743–61. doi:10.1128/JVI.00584-08

Asan, Xu, Y., Jiang, H., Tyler-Smith, C., Xue, Y., Jiang, T., … Zhang, X. (2011). Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biology*, *12*(9), R95. doi:10.1186/gb-2011-12-9-r95

Avesson, L., & Barry, G. (2014). The emerging role of RNA and DNA editing in cancer. *Biochimica et Biophysica Acta*, *1845*(2), 308–16. doi:10.1016/j.bbcan.2014.03.001

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. a, & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews. Genetics*, *12*(11), 745–55. doi:10.1038/nrg3031

Bick, D., & Dimmock, D. (2011). Whole exome and whole genome sequencing. *Current Opinion in Pediatrics*, *23*(6), 594–600. doi:10.1097/MOP.0b013e32834b20ec

Burns, M. B., Lackey, L., Carpenter, M. A., Land, A. M., Leonard, B., Refsland, E. W., … Harris, R. S. (2013). APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*, *494*(7437), 366–370. doi:10.1038/nature11881.APOBEC3B

Casás-Selves, M., & Degregori, J. (2011). How cancer shapes evolution, and how evolution shapes cancer. *Evolution*, *4*(4), 624–634. doi:10.1007/s12052-011-0373-y

Casey, J. L., Bergmann, K. F., Brown, T. L., Purcell, R. H., & Gerin, J. L. (1992). Structural requirements for RNA editing in hepatitis 6, *89*(August), 7149–7153.

Chen, Y.-C., Su, H.-J. J., Guo, Y.-L. L., Houseman, E. A., & Christiani, D. C. (2005). Interaction between environmental tobacco smoke and arsenic methylation ability on the risk of bladder cancer. *Cancer Causes & Control : CCC*, *16*(2), 75–81. doi:10.1007/s10552-004-2235-1

Chern, H., Becich, M. J., Persad, R. A. J. A., Romkes, M., Smith, P., Collins, C., … Branch, R. A. (1996). CLONAL ANALYSIS OF HUMAN RECURRENT SUPERFICIAL BLADDER CANCER BY IMMUNOHISTOCHEMISTRY OF P53 AND i : tzn, 1846–1849.

Chihara, Y., Kanai, Y., Fujimoto, H., Sugano, K., Kawashima, K., Liang, G., … Hirao, Y. (2013). Diagnostic markers of urothelial cancer based on DNA methylation analysis. *BMC Cancer*, *13*(1), 275. doi:10.1186/1471-2407-13-275

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; *Fly*, *6*(2), 80–92.

Comprehensive molecular characterization of urothelial bladder carcinoma. (2014). *Nature*, *507*(7492), 315–322. doi:10.1038/nature12965

Conticello, S. G., Thomas, C. J. F., Petersen-Mahrt, S. K., & Neuberger, M. S. (2005). Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Molecular Biology and Evolution*, *22*(2), 367–77. doi:10.1093/molbev/msi026

Crespi, B., & Summers, K. (2005). Evolutionary biology of cancer. *Trends in Ecology & Evolution*, *20*(10), 545–52. doi:10.1016/j.tree.2005.07.007

Dalbagni, G., Ren, Z., Herr, H., Cordon-cardo, C., & Reuter, V. (2001). Genetic Alterations in TP53 in Recurrent Urothelial Cancer : A Longitudinal Study, *7*(September), 2797–2801.

Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., … Mardis, E. R. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, *464*(7291), 999–1005. doi:10.1038/nature08989

Duggan, B. J., Gray, S. B., McKnight, J. J., Watson, C. J., Johnston, S. R., & Williamson, K. E. (2004). Oligoclonality in bladder cancer: the implication for molecular therapies. *The Journal of Urology*, *171*(1), 419–25. doi:10.1097/01.ju.0000100105.27708.6c

Dunn, G. P., Old, L. J., Schreiber, R. D., Louis, S., Burnet, F. M., & Thomas, L. (2004). of Cancer Immunosurveillance and Immunoediting, *21*, 137–148.

Fadl-Elmula, I., Gorunova, L., Mandahl, N., Elfving, P., Lundgren, R., Rademark, C., & Heim, S. (1999). Cytogenetic Analysis of Upper Urinary Tract Transitional Cell Carcinomas. *Cancer Genetics and Cytogenetics*, *115*(2), 123–127. doi:10.1016/S0165-4608(99)00075-8

Fang, H., Wu, Y., Narzisi, G., O¿Rawe, J. a, Jimenez Barrón, L. T., Rosenbaum, J., … Lyon, G. J. (2014). Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Medicine*, *6*(10), 89. doi:10.1186/PREACCEPT-1179619571327140

Felsenstein. (1983). Statistical Inference of Phylogenies, 246–272.

Forbes, S. a, Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., … Futreal, P. A. (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Research*, *38*(Database issue), D652–7. doi:10.1093/nar/gkp995

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–95.

Gatenby, R. a, & Gillies, R. J. (2004). Why do cancers have high aerobic glycolysis? *Nature Reviews. Cancer*, *4*(11), 891–9. doi:10.1038/nrc1478

Glenn, J. S., & White, J. M. (1991). trans-Dominant Inhibition of Human Hepatitis Delta Virus Genome Replication, *65*(5), 2357–2361.

Gould, S. J., & Eldredge, N. (1993). Punctuated equilibrium comes of age. *Nature*, *366*(6452), 223–7. doi:10.1038/366223a0

Greaves, M., & Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, *481*(7381), 306–13. doi:10.1038/nature10762

Habuchi, T. (2005). Origin of multifocal carcinomas of the bladder and upper urinary tract: molecular analysis and clinical implications. *International Journal of Urology : Official Journal of the Japanese Urological Association*, *12*(8), 709–16. doi:10.1111/j.1442-2042.2005.01155.x

Hafner, C., Knuechel, R., Stoehr, R., & Hartmann, A. (2002). Clonality of multifocal urothelial carcinomas: 10 years of molecular genetic studies. *International Journal of Cancer. Journal International Du Cancer*, *101*(1), 1–6. doi:10.1002/ijc.10544

Hafner, C., Knuechel, R., Zanardo, L., Dietmaier, W., Blaszyk, H., Cheville, J., … Hartmann, a. (2001). Evidence for oligoclonality and tumor spread by intraluminal seeding in multifocal urothelial carcinomas of the upper and lower urinary tract. *Oncogene*, *20*(35), 4910–5. doi:10.1038/sj.onc.1204671

Hanahan, D., & Weinberg, R. a. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–74. doi:10.1016/j.cell.2011.02.013

Harris, R. S., Bishop, K. N., Sheehy, A. M., Craig, H. M., Petersen-mahrt, S. K., Watt, I. N., … Malim, M. H. (2003). to Retroviral Infection, *113*, 803–809.

Harris, R. S., & Liddament, M. T. (2004). Retroviral restriction by APOBEC proteins. *Nature Reviews. Immunology*, *4*(11), 868–77. doi:10.1038/nri1489

Hartmann, A., Ro, U., Schlake, G., Dietmaier, W., Zaak, D., Hofstaedter, F., & Knuechel, R. (2000). Clonality and Genetic Divergence in Multifocal Low-Grade Superficial Urothelial Carcinoma as Determined by Chromosome 9 and p53 Deletion Analysis, *80*(5), 709–718.

Henderson, S., Chakravarthy, A., Su, X., Boshoff, C., & Fenton, T. R. (2014). APOBEC-Mediated Cytosine Deamination Links PIK3CA Helical Domain Mutations to Human Papillomavirus-Driven Tumor Development. *Cell Reports*, *7*(6), 1833–1841. doi:10.1016/j.celrep.2014.05.012

Hutchinson, J. R., Famini, D., Lair, R., Kram, R., Frank, S. A., & Nowak, M. A. (2003). Developmental predisposition to cancer, *422*(April), 10570.

Jakóbisiak, M., Lasek, W., & Gołb, J. (2003). Natural mechanisms protecting against cancer. *Immunology Letters*, *90*(2-3), 103–122. doi:10.1016/j.imlet.2003.08.005

Jayan, G. C., & Casey, J. L. (2002). Inhibition of Hepatitis Delta Virus RNA Editing by Short Inhibitory RNA-Mediated Knockdown of ADAR1 but Not ADAR2 Expression. *Journal of Virology*, *76*(23), 12399–12404. doi:10.1128/JVI.76.23.12399-12404.2002

Jones, T. D., Wang, M., Eble, J. N., MacLennan, G. T., Lopez-Beltran, A., Zhang, S., … Cheng, L. (2005). Molecular evidence supporting field effect in urothelial carcinogenesis. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *11*(18), 6512–9. doi:10.1158/1078-0432.CCR-05-0891

Kawanishi, H., Takahashi, T., Ito, M., Matsui, Y., Watanabe, J., Ito, N., … Ogawa, O. (2007). Genetic analysis of multifocal superficial urothelial cancers by array-based comparative genomic hybridisation. *British Journal of Cancer*, *97*(2), 260–6. doi:10.1038/sj.bjc.6603850

Knowles, M. a., & Hurst, C. D. (2014). Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nature Reviews Cancer*, *15*(1), 25–41. doi:10.1038/nrc3817

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218. doi:10.1038/nature12213

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *00*(00), 3. Genomics. Retrieved from http://arxiv.org/abs/1303.3997

Li, M., & Cannizzaro, L. a. (1999). Identical clonal origin of synchronous and metachronous low-grade, noninvasive papillary transitional cell carcinomas of the urinary tract. *Human Pathology*, *30*(10), 1197–1200. doi:10.1016/S0046-8177(99)90037-0

Luo, G., Chao, M. E. I., Hsieh, S., Sureau, C., Nishikura, K., & Taylor, J. (1990). A Specific Base Transition Occurs on Replicating Hepatitis Delta Virus RNA GACATCAGGGGAACCIGGGAIMICCATIGGATATACTMCCCAGC CGATCCACCCIIIiCTCCCCAGAGTTGTCGACCCCAGTGAT, *64*(3), 1021–1027.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. a. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–303. doi:10.1101/gr.107524.110

Mehta, a., Kinter, M. T., Sherman, N. E., & Driscoll, D. M. (2000). Molecular Cloning of Apobec-1 Complementation Factor, a Novel RNA-Binding Protein Involved in the Editing of Apolipoprotein B mRNA. *Molecular and Cellular Biology*, *20*(5), 1846–1854. doi:10.1128/MCB.20.5.1846-1854.2000

Merlo, L. M. F., Pepper, J. W., Reid, B. J., & Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nature Reviews. Cancer*, *6*(12), 924–35. doi:10.1038/nrc2013

Miyao, N., Tsai, Y. C., Lerner, S. P., Olumi, A. F., Iii, C. H. S., Gonzalez-zulueta, M., … Jones, P. A. (1993). Role of Chromosome 9 in Human Bladder Cancer1, *407*(9), 4066–4070.

Mullighan, C. G., Phillips, L. A., Su, X., Ma, J., Miller, C. B., Shurtleff, S. A., & Downing, J. R. (2009). GENOMIC ANALYSIS OF THE CLONAL ORIGINS OF RELAPSED, *322*(5906), 1377–1380. doi:10.1126/science.1164266.GENOMIC

Mutter, G. L., & Boynton, K. A. (1995). PCR bias in amplification of androgen receptor alleles , a trinucleotide repeat marker used in clonality studies, *23*(8), 1411–1418.

Nagy, J. D. (2004). Competition and natural selection in a mathematical model of cancer. *Bulletin of Mathematical Biology*, *66*(4), 663–87. doi:10.1016/j.bulm.2003.10.001

Nowell, P. C. (1950). The Clonal Evolution of Tumor Cell Populations, 23–28.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., … Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine*, *5*(3), 28. doi:10.1186/gm432

Paolo, P., Fiore, D., Pierce, J. H., Kraus, M. H., Segatro, O., King, C. R., & Aaronson, S. A. (1987). erbB-2 Is a Potent Oncogene When Overexpressed in NIHI/3T3 Cells, *2*(January).

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, *20*(2), 289–290. doi:10.1093/bioinformatics/btg412

Pedersen, L., Gunnarsdottir, K. a, Rasmussen, B. B., Moeller, S., & Lanng, C. (2004). The prognostic influence of multifocality in breast cancer patients. *Breast (Edinburgh, Scotland)*, *13*(3), 188–93. doi:10.1016/j.breast.2003.11.004

Pepper, J. W., Scott Findlay, C., Kassen, R., Spencer, S. L., & Maley, C. C. (2009). Cancer research meets evolutionary biology. *Evolutionary Applications*, *2*(1), 62–70. doi:10.1111/j.1752-4571.2008.00063.x

Prior, I. a, Lewis, P. D., & Mattos, C. (2012). A comprehensive survey of Ras mutations in cancer. *Cancer Research*, *72*(10), 2457–67. doi:10.1158/0008-5472.CAN-11-2612

Quinlan, M. P., & Settleman, J. (2009). Isoform-specific ras functions in development and cancer. *Future Oncology (London, England)*, *5*(1), 105–16. doi:10.2217/14796694.5.1.105

Roberts, S. a, Sterling, J., Thompson, C., Harris, S., Mav, D., Shah, R., … Gordenin, D. a. (2012). Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Molecular Cell*, *46*(4), 424–35. doi:10.1016/j.molcel.2012.03.030

Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Stojanov, P., Kiezun, A., … Getz, G. (2013). NIH Public Access, *45*(9), 970–976. doi:10.1038/ng.2702.An

Rokas, A., & Holland, P. W. H. (2000). Rare genomic changes as a tool for phylogenetics. *Trends in Ecology & Evolution*, *15*(11), 454–459. doi:10.1016/S0169-5347(00)01967-4

Ross, D. S., Litofsky, D., Ain, K. B., Bigos, T., Brierley, J. D., Cooper, D. S., … Sherman, S. I. (2009). Recurrence after treatment of micropapillary thyroid cancer. *Thyroid : Official Journal of the American Thyroid Association*, *19*(10), 1043–8. doi:10.1089/thy.2008.0407

Savva, Y. a, Rieder, L. E., & Reenan, R. a. (2012). The ADAR protein family. *Genome Biology*, *13*(12), 252. doi:10.1186/gb-2012-13-12-252

Sawyer, S. L., Emerman, M., & Malik, H. S. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biology*, *2*(9), E275. doi:10.1371/journal.pbio.0020275

Scher, M. B., Elbaum, M. B., Mogilevkin, Y., Hilbert, D. W., Mydlo, J. H., Sidi, a A., … Trama, J. P. (2012). Detecting DNA methylation of the BCL2, CDKN2A and NID2 genes in urine using a nested methylation specific polymerase chain reaction assay to predict bladder cancer. *The Journal of Urology*, *188*(6), 2101–7. doi:10.1016/j.juro.2012.08.015

Simon, R., Eltze, E., Scha, K., Bu, H., Semjonow, A., Hertle, L., … Bo, W. (2001). Cytogenetic Analysis of Multifocal Bladder Cancer Supports a Monoclonal Origin and Intraepithelial Spread of Tumor Cells 1, 355–362.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews. Genetics*, *15*(2), 121–32. doi:10.1038/nrg3642

Studies, I. (1999). HUMAN PAPILLOMAVIRUS IS A NECESSARY CAUSE, *19*(May), 12–19.

Takahashi, T., Habuchi, T., Kakehi, Y., Mitsumori, K., Akao, T., & Terachi, T. (1998). Clouai and Chronological Genetic Analysis of Multifocal Cancers of the Bladder and Upper Urinary Tract1, 5835–5841.

Takahashi, T., Kakehi, Y., Mitsumori, K., Akao, T., Terachi, T., Kato, T., … Habuchi, T. (2001). DISTINCT MICROSATELLITE ALTERATIONS IN UPPER URINARY TRACT TUMORS AND SUBSEQUENT BLADDER TUMORS, *165*(February), 672–677.

Tsao, J., Yatabe, Y., Salovaara, R., Ja, H. J., & Shibata, D. (2000). Genetic reconstruction of individual colorectal tumor histories.

Valadi, H., Ekström, K., Bossios, A., Sjöstrand, M., Lee, J. J., & Lötvall, J. O. (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nature Cell Biology*, *9*(6), 654–9. doi:10.1038/ncb1596

Van Niel, G., Porto-Carreiro, I., Simoes, S., & Raposo, G. (2006). Exosomes: a common pathway for a specialized function. *Journal of Biochemistry*, *140*(1), 13–21. doi:10.1093/jb/mvj128

Vartanian, J.-P., Henry, M., Marchio, A., Suspène, R., Aynaud, M.-M., Guétard, D., … Wain-Hobson, S. (2010). Massive APOBEC3 editing of hepatitis B viral DNA in cirrhosis. *PLoS Pathogens*, *6*(5), e1000928. doi:10.1371/journal.ppat.1000928

Vieira, V. C., & Soares, M. a. (2013). The role of cytidine deaminases on innate immune responses against human viral infections. *BioMed Research International*, *2013*, 683095. doi:10.1155/2013/683095

Volanis, D., Kadiyska, T., Galanis, A., Delakas, D., Logotheti, S., & Zoumpourlis, V. (2010). Environmental factors and genetic susceptibility promote urinary bladder cancer. *Toxicology Letters*, *193*(2), 131–7. doi:10.1016/j.toxlet.2009.12.018

Wang, Y., Lang, M. R., Pin, C. L., & Izawa, J. I. (2013). Comparison of the clonality of urothelial carcinoma developing in the upper urinary tract and those developing in the bladder. *SpringerPlus*, *2*(1), 412. doi:10.1186/2193-1801-2-412

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., … Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, *38*(Web Server issue), W214–20. doi:10.1093/nar/gkq537

Witsch, E., Sela, M., & Yarden, Y. (2011). Roles for Growth Factors in Cancer Progression, *25*(2), 85–101. doi:10.1152/physiol.00045.2009.Roles

Yates, L. R., & Campbell, P. J. (2012). Evolution of the cancer genome. *Nature Reviews. Genetics*, *13*(11), 795–806. doi:10.1038/nrg3317

Yoshimura, I., Kudoh, J., Saito, S., Tazaki, H., & Shimizu, N. (1995). p53 Gene Mutation in Recurrent Superficial Bladder Cancer. *The Journal of Urology*, *153*(5), 1711–1715. doi:10.1016/S0022-5347(01)67510-4

Yoshino, H., Seki, N., Itesako, T., Chiyomaru, T., Nakagawa, M., & Enokida, H. (2013). Aberrant expression of microRNAs in bladder cancer. *Nature Reviews. Urology*, *10*(7), 396–404. doi:10.1038/nrurol.2013.113

Young, S. G., Hubl, S. T., Smith, R. S., Snyder, S. M., & Terdiman, J. F. (n.d.). Familial Hypobetalipoproteinemia Caused by a Mutation in the Apolipoprotein B Gene That Results in a Truncated Species of Apolipoprotein B (B-31) Formation of Buoyant , Triglyceride-rich Lipoproteins, 933–942.

Yu, Q., Chen, D., König, R., Mariani, R., Unutmaz, D., & Landau, N. R. (2004). APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. *The Journal of Biological Chemistry*, *279*(51), 53379–86. doi:10.1074/jbc.M408802200

Zeegers, M. P., Tan, F. E., Dorant, E., & van Den Brandt, P. A. (2000). The impact of characteristics of cigarette smoking on urinary tract cancer risk: a meta-analysis of epidemiologic studies. *Cancer*, *89*(3), 630–9. doi:10.1002/1097-0142(20000801)89:3<630::AID-CNCR19>3.0.CO;2-Q

Zhang, J., & Webb, D. M. (2004). Rapid evolution of primate antiviral enzyme APOBEC3G. *Human Molecular Genetics*, *13*(16), 1785–91. doi:10.1093/hmg/ddh183